MSc Dissertation
Generating Stories from Images

David Wilmot (s1569885)

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh

2016

Abstract

The ReelLives project generates stories from personal social media image collections. This project extends it by: Supporting generation of stories for the existing system for the MS-COCO Image dataset. Creating new image features based on R-CNN (Region - Conventional Neural Networks), and Semantic features based on Word embedding averaging, with the capability of performing a Word embedding based search. Implements a new MILP (Mixed Integer Linear Programming) based selector for generating Triptychs (3 image stories). Evaluation is performed against random subsets of MS-COCO, and sampled queries. The project finds that the new features substantially outperform the previous LSA text only based implementation with regard to topic cohesiveness, and moderately with regard to telling a story. Correct ordering of images is found to be a weakness of the approach.

Acknowledgements

I would like to thank Prof. Mirella Lapata and Dr Matthew Aylett for their continuing support throughout the MSc project. Dr Carina Silberer for her assistance and scripts for extracting Fast R-CNN features from the MS-COCO dataset. Elaine Farrow for assistance with running the ReelLives system, and developing additional utilities to support the evaluation. My Sister Emma Wilmot for help with proof reading. A couple of students briefly assisted with reviewing trials - Andreea Pascu and Adam McCarthy, Adam also suggested the Tabu search for the selector.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Stories though Images

Telling stories though images has always been an important part of art and culture (Steiner, 2004; Gottschall, 2012). From modern day through classical civilisations such as the *Romans* (Brilliant, 1984) to the earliest days of modern human culture in the *Ice Age* (Cook, 2012). *The Garden of Earthly Delights* (figure 1.1) tells the story through a Triptych (3 image panels) of the fall of man from the Garden of Eden in the left panel to Hell in the right. In religious works these kind of pictorial representations of stories whether on canvas or silk, or in frescoes, illuminations, sculpture or stained glass have a long history. As has it been in artistic works such as Hogarth's *Rakes Progress*, Picasso's *Guernica*, or contemporary Grayson Perry's ceramics and tapestries.

Figure 1.1: *The Garden of Earthly Delights*, by Hieronymus Bosch (https://en.wikipedia.org/wiki/The_Garden_of_Earthly_Delights)

Outside art and religion, pictorial narratives have played an important role in current affairs; such as political cartoons (a classic historical example Figure

1.2); comic strips that feature daily in newspapers, magazines, and websites; and illustrations in Children's book; and in moving form comedies from Charlie Chaplin to Mr Bean.

1.2 ReelLives

There are also historical examples of more personal journeys such as the commemorative depiction of Pony Express to California in the 19th Century (Figure 1.3). It is to this more personal experience that the ReelLives project (http://reellives.net/) is dedicated. ReelLives (Aylett et al., 2015; Farrow et al., 2015) is a collaborative research project between several research groups. The aim is to build personal documentaries ("Reels" or stories) out of the collections of personal image collections on social media using methods developed from machine learning, NLP, and computer graphics.

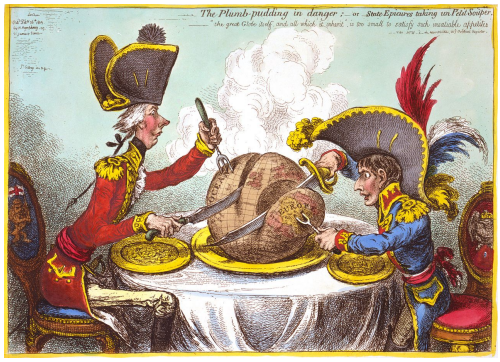


Figure 1.2: *The Plumb Pudding in Danger*, by James Gillroy. (https://en.wikipedia.org/wiki/File:Caricature_gillray_plumpudding.jpg)

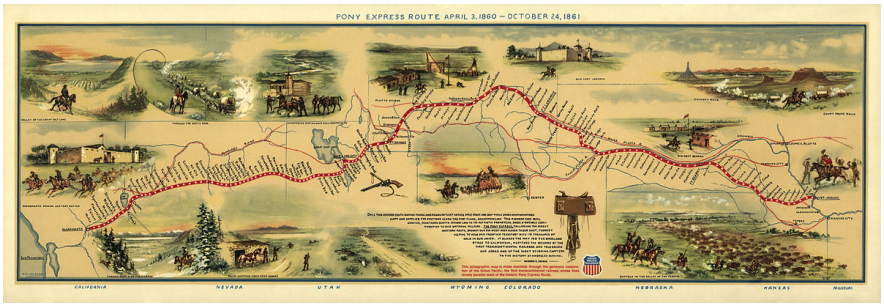


Figure 1.3: The Pony Express (https://en.wikipedia.org/wiki/Pony_Express)

Enormous quantities of photographic data is being captured everyday, for Facebook alone as of 2014 there were over 350 million images uploaded per day (Ericsson, 2014). With other sites such as Instagram, Flickr, Google Photos, Twitter, Snapchat, Weixin (WeChat) and Weibo the total volume is far higher, and many photos never leave the phones or cameras they are captured on. Some

work is already being done by these social media companies with this intent: Google Photos (https://photos.google.com/) will automatically create photo albums based on location and time, e.g. *weekend in Barcelona*, and Facebook (https://www.facebook.com/) is creating friendship movies (e.g. *celebrating 5 years of friendship*) with a chronological set of photos two people have appeared in together. There is clearly a lot of interest in this area, and wide ranging opportunities to apply NLP and machine learning techniques to create interesting and relevant stories from images.

1.3 MSc Project

The aim of this project is to generate Triptychs from a large collection of images - MS-COCO (Lin et al., 2014) and related annotations. This project extends the work of *Aylett et al* (2015) in the following ways:

1. Represents the MS-COCO dataset in a format compatible with ReelLives supporting the existing features.
2. A new set of features for selection based on R-CNN (Region - Convolutional Neural Network) representations of the images, and recent advances in Word and Document embeddings.
3. New Triptych generator using MILP (Mixed Integer Linear Programming) methods.

It makes use of a widely used imageset MS-COCO , extends the linguistic analysis to take advantage of many of the recent advancements in word and document embeddings, , and creates

Chapter 2

Background

2.1 Stories and Images

What makes a good story? [1] Referring back to the examples in the introduction: Bosch's (figure 1.1) painting tells a visual story through an emotional arc enlightenment to doom, or the reverse such as Hans Christian Anderson's *The Ugly Duckling* journey to a Swan. In a recent paper (Reagan et al., 2016)[2] using sentiment analysis techniques the authors have analysed project Gutenberg's archive and suggest that written stories at least follow 6 emotional arcs (classic books searchable online http://hedonometer.org/books/v1/): "Rags to riches" (rise), "Tragedy" (fall), "Man in a hole" (fall rise), "Icarus" (rise fall), "Cinderella" (rise fall rise), "Oedipus" (fall rise fall). While this positive/negative sentiment is an oversimplification it is important part of narrative forms.

In visual forms stories often make use of word play, visual metaphor and allegory (such as the cartoon Figure 1.2). In a more documentary form (figure 1.3) it can tell the story of a journey: It could be a day out at the beach - drive, swim, make sand castles, shelter from the rain, or of someones weddings - arriving at the Church, the service, reception, throwing the bouquet. There needs to be an element of commonality running through the images: in time, space, characters, scene, action but there also needs to be progression and changes, it's in balancing the two that makes it engaging.

[1] A nice introductory video (Stanton, 2012)
[2] See also this humorous Youtube clip (Vonnegut, 2010)

There are a variety of forms that could be used for evaluating visual story telling: A slideshow sequence of many images, movies that are slide shows for static images, or the form that has been focused on for evaluation in ReelLives a Triptych (Aylett et al., 2015; Farrow et al., 2015). A Triptych is form with a long history in Art; As per Figure 1.1 *The Garden of Earthly Delights* it is a narrative told through 3 images with a central, left-hand, and right-hand panel. It simplifies a story to its a basic constituents start, middle, end: It is a form standard to Theatre (3 act plays), traditional poetic forms such as the *Petrarchan Sonnet* where the form of *quatrains* are called *proposition*, *resolution*, and *volta* (turn). Being a simple visual form, it is also easy for people to understand quickly and thus to evaluate.

Our aim is to establish how well recently developed techniques in distributed semantics and computer vision can be used to create engaging pictorial Triptych stories by adjusting the degrees of similarity and dissimilarity (or convergence and divergence) between elements such as objects, composition, linguistic descriptions, and sentiment.

2.2 The ReelLives System

2.2.1 Details

The ReelLives system (Aylett et al., 2015; Farrow et al., 2015) creates personal photographic stories out of image collections from Social Media, adaptors exist to extract users personal images for Twitter, Facebook, and Instagram into an XML feature file. The linguistic data is extracted as is from the social media source, for example Facebook captions on the photo, or the Tweet the image was posted with - there is no structured formatting or tagging. In practice this can mean this can be lots of photos have no linguistic content, or meaningless generated tags like *DSC1001, DSC1002*.

ReelLives selects images for a Triptych using a Viterbi search (Forney, 1973) . The Viterbi algorithm scores each candidate image based on its semantic and temporal similarity with subsequent position's (*transition probabilities*). It also scores based on criteria that have been applied to each position (*emission probabilities*): These

include sentiment (a trend can be created so a Triptych flows from negative, to neutral, to positive for example), a chosen Theme, Named Entity, Location or Keyword search. This allows a wide degree of flexibility when creating Triptychs.

Text is parsed using POS (Part of Speech) tagging from TweetNLP (Gimpel et al., 2011; Owoputi et al., 2013) a POS tagger targeted at social media which is often very different from traditional written language - abbreviations, slang, emoticons, often incomplete sentences, etc. LSA (Latent Semantic Analysis) (Landauer & Dumais, 1997) is applied to the text from the users personal collections, with Cosine similarity to score the similarity of the caption or one image to another. The LSA implementation is *Gensim* (Řehůřek & Sojka, 2010).

The sentiment of images' captions is classified into Positive, Neutral, and Negative using *Sentistrength* (Thelwall, 2013), designed specifically for social media content. Desired sentiment can be targeted at specific positions to create Triptychs that have an emotional narrative to them.

Other features are Themes (a hierarchical classification), Named Entities, and Locations are extracted with the proprietary cloud service *Alchemy* (IBM, 2016a). These features are unchanged by this project, more details are in the Themes and Named Entities (Section 3.4.2 p. 24) section.

ReelLives was evaluated by assessing random, best, and a sentiment narrative (running from negative to positive) on 26 public Instgram accounts photos queried on the term "life" (Aylett et al., 2015). They found that participants were sensitive (they could identify) the different types of narrative. The selection narratives are assessed as better than random, but the level of perceived story is low. Sentiment narratives also scored significantly high on object cohesiveness.

2.2.2 Extensions

Whilst capable of producing meaningful Triptych there are a number of opportunities for improvement. Firstly, the current system doesn't use any features from the image relying solely on the text. This has a few problems: Social media images often don't have any or very short messages leaving the selector not much to use. Often they are also very similar - *Walking on the beach 1/2/3.* This can lead to the selector choosing 3 identical images because they have close text but that

does not make an interesting story. Being able to identify objects, composition, textures, colours and other image attributes and use these in selection greatly enhances the potential of the system.

Secondly, LSA has been surpassed in performance by new word and document based text representations (see Section 2.6.1, p. 10), so there is opportunity to evaluate if these methods can improve on LSA in this context.

Thirdly, the Viterbi algorithm (Forney, 1973) uses a beam search to simplify the complexity of selection. This has the potential problem of discarding solutions early that ultimately be optimal, and there are alternatives that can find a more globally optimum solution.

Fourthly, the algorithm treats the text as a single block whereas there are different constituent components such as the subject, objects, and verbs of sentences that can be extracted and queried as separate features (Angeli et al., 2015).

Fifthly, ReelLives is at the moment limited to participants personal collections with concerns over the privacy and their highly subjective nature. Using a large widely used image collection provides an opportunity to extend the work in other domains, as well as develop and more robustly evaluate new techniques without these concerns.

2.3 Visual Stories

There is a growing body of other interest in automatic generation of pictorial stories in various forms. *Zhu et al* (Zhu et al., 2007) created text-to-picture that makes collages that describe sentences from news articles and Children's stories. Their work uses the notion of *centrality*; that is key clauses and words in the sentences should be more central in the final image. Subsequently (A. B. Goldberg et al., 2009) used a Triptych layout (they called it ABC) with conditional random fields as a method for selecting images based on keywords. Both of these approaches rely, as ReelLives does, entirely on the text and not features of the image.

There have been several very recently published paper, since this project was started, aimed at telling visual stories. SIND (Sequential images Narrative Dataset)

(Huang et al., 2016) is a new dataset aimed directly at the problem of visual stories. The dataset consists of 81K images with 21k sequences that have been crowd sourced via Amazon Mechanical Turk. Each of the images has 3 sets of captions: Descriptions of images in isolation (DII), Descriptions of images in sequence (DIS), and Stories for images in sequence (SIS). The rationale is that narrative language is different from descriptive adds layers of depth that can be useful in building not only story telling but also relevant to other areas such as natural question answering systems.

Based on SIND, *Sort story* (Agrawal et al., 2016) uses an LSTM (Hochreiter & Schmidhuber, 1997), Skip-Thoughts (Kiros et al., 2015) (Section 2.6.1, p. 10), CNN for visual representation (Section 2.6.2, p. 15), and what they call an NPE (Neural Pairwise Embedding) a learnt embedding for sequencing information. They train the network to predict the ordeering in the SIND Gold standard stories; evaluation found the different representations are complementary, and strong results can be achieved via an ensemble voting system that combines all features.

Similar work has been done using an S-RNN (Skipping Recurrent Neural Networks) (Sigurdsson et al., 2016; Liu et al., 2016) . Rather than predicting every step in a sequence the RNN is trained to skip over closely related photos in storylines, the S-RNN learns a series of subsets of the storyline. The idea behind skipping is that in a sequential story, images close to each often are very similar. For example in a wedding album pictures of the church, then photos of the couple, the reception, etc will occur next to each other; so there are often little differences between neighbouring pictures and descriptions that the RNN can use to learn the narrative structure. Skipping reduces this repetition problem, and allows learning of the broader narrative.

SIND was created (and related papers published) during the implementation of this project, and so too late for consideration. Nevertheless the problem all these papers are tackling is one of learning the order of events over sequences using multi-modal (image and Linguistic) features is highly relevant to later discussion.

2.4 Related Work

2.4.1 Captioning

A related area relevant to the task is in the area of caption generation from images, since it requires learning a relationship between the image and text that is also relevant to story telling. Approaches (Feng & Lapata, 2013) have been taken that use *SIFT* (Lowe, 1999) based features as a visual bag of words and use LDA (Latent Dirichlet Allocation) to model a distribution over learnt topics. Other recent work has concentrated on deep learning neural network based approaches: A joint model that combines a DCNN (Deep Convolutional Neural Network) with an RNN (Recurrent Neural Network, specifically LSTM decoder (Vinyals et al., 2015). The DCNN encodes a representation of the image as a vector which is then decoded by the LSTM based on the image vector and recurrent state from previous words generated as a sentence. Importantly the model is trained as one in a supervised fashion maximising the probability of the generated image description. Others have employed a BRNN (Bi-Directional RNN) (Karpathy & Fei-Fei, 2015) to train an alignment between regions of the images and the captions that describe them that can then be used to generate descriptions. *Xu et al* (2015) extend the CNN encoder - RNN decoder model to include an attention mechanism that learns a distribution over the words being generated. All these models are focused on generating descriptive text. Storytelling is quite different in that it is about connecting objects and events that flow from one to the another. Nevertheless there is much to learn from this work. Core to all these are DCNN representations for image features covered further in Section 2.6.2.

2.4.2 Question Answering

Ren et al (2015) try a variety of the techniques discussed in caption generation such as CNNs and RNNs to create an image question and answer system that can answer questions such as “where is the cat sitting?”, and “what colour is the hat?”, *Ma et al* (2015) solely use CNNs. The various network architectures are used to create forms of Multimodal embeddings representing both the text and images and can be used to infer answers to questions, although the overall accuracy of the systems is still far from ideal making naive mistakes in a lot of

cases. As per the caption generation this work provides a way of linking text and image representations. Before discussing how these various techniques can inform new features for creating stories, datasets will be considered.

2.5 Imagesets

One of main issues that needs to be resolved with ReelLives is the reliance on limited amounts of personal data. In this project we extend the narrative analysis to some recently developed techniques for object detection, captioning and question answering type tasks that have become subject of recent research interest. The most widely known is imageNet (Deng et al., 2009; Russakovsky et al., 2015) having over 14 million images (http://image-net.org/explore) with glosses and semantic sysnets from Wordnet (Miller, 1995). While excellent for training computer vision detection systems it's not suitable for this task because it is not story based and does not have descriptive captions, but rather definition based glosses and categorisation.

MS-COCO (Lin et al., 2014; Chen et al., 2015) - http://mscoco.org/ - whilst not offering narrative annotations has been created for a wide variety of uses including image classification and object detection. Each image 123k+ has 5 descriptive captions and the locations of objects marked, see Figure 2.1 for example. There are 80 objects types including kites, pizza, giraffes, and people so covering a wide enough subjects and providing both text and visual features to be able to create interesting stories.

2.6 Embeddings

2.6.1 Semantic Composition

The more complex integrated captioning, storytelling, and other models discussed are attempting to integrate multiple modalities - semantic representation of text and image features - into a single model. Rather than integrate the models there is also work to represent these separately via vector representations using similar

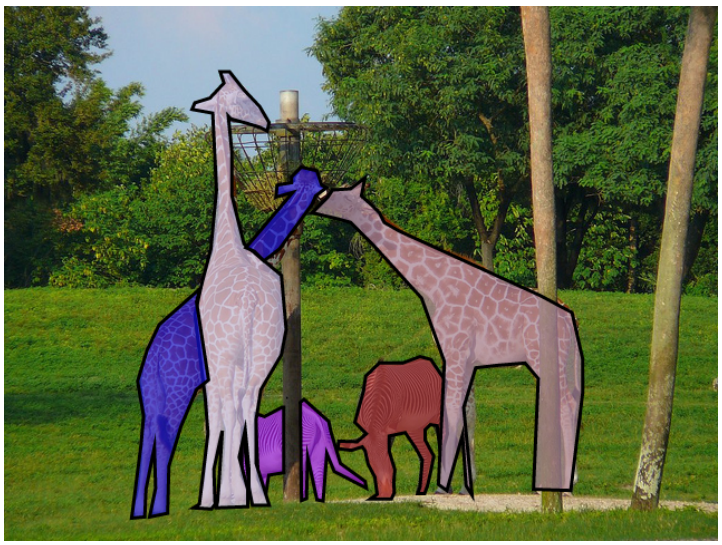
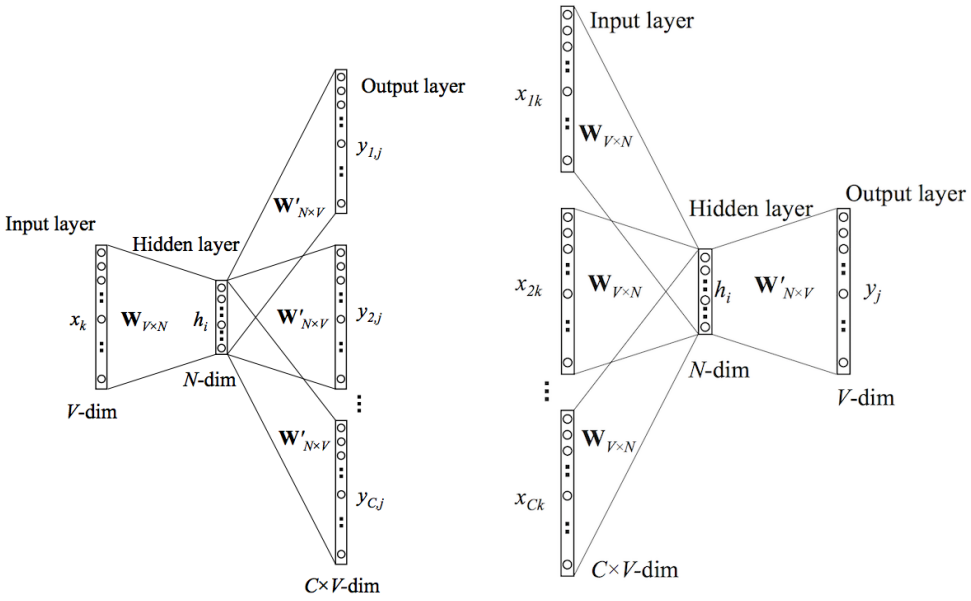


Figure 2.1: Example of the 5 annotated captions that are provided for each image in MS-COCO:
three giraffes standing next to two zebra on a lush green field.
some giraffes and zebras in an exhibit at zoo
a group of giraffes and zebras feeding and grazing in a grassy field.
three giraffe and two zebras are grazing in the grass together.
three giraffes and two zebras are feeding in a zoo.

techniques. It is these representations that will be looked at more detail in this section.

LSA (Landauer & Dumais, 1997) is a form vector representation that factorises matrix cooccurence counts using SVD (Single Value Decomposition). By taking the top n *eigenvectors* it is possible to get a compressed vector representation that encodes information from the cooccurences. This kind of model is a distributed representation because the meaning of the text is distributed across the single vector representation rather than individual elements representing particular units, such as words. The resulting vector representations have some nice properties to encapsulate meaning: Semantically similar documents are close to each other in space using distance measures such as *Cosine* or *Euclidean* distance, likewise dissimilar documents are further away. There can also be useful representations from composing together vectors via means such as addition or multiplication (Mitchell & Lapata, 2010). Being able to represent the similarity



(a) Skip-gram model, reproduced from https://commons.wikimedia.org/wiki/File:Cbow.png

(b) CBOWmodel, reproduced from https://commons.wikimedia.org/wiki/File:Skip-gram.png

Figure 2.2: CBOW and Skipgram, both from Moucrowap under the license https://creativecommons.org/licenses/by-sa/4.0/deed.en

of text using a simple distance has attractive qualities for story telling in being able to select related pieces of text that have a certain level commonality (or not) with another.

There are other distributed semantic models such as LDA (Latent Dirichlet Allocation) (Blei et al., 2003) that are also widely used, but most recent interest has shifted to word embedding based approaches. Word2Vec (Mikolov, Chen, et al., 2013) encodes the meaning in words via training neural networks to predict missing words. There are two models Skip-Gram and CBOW (Continuous Bag of Words), both shown in figures 2.2a and 2.2b. In the case of CBOW the surrounding words over a window are presented and the network learns to predict the missing word. In the case of Skip-Gram a single word is presented and the network learns to predict the surrounding words in a window. In both cases by learning to predict the missing words encodes a representation of the meaning of the word in a vector in the neural network which can then be reused in other contexts. Implicitly via prediction the Word2Vec model is performing a form of factorisation similar to the decomposition performed by LSA (Levy &

Goldberg, 2014), however the models perform significantly better than LSA type count based methods on a wide variety of tasks (Baroni et al., 2014). Words that similar, for example *frog* and *toad*, as they are used in similar contexts will be represented with similar vectors that are close to one another space, so will synonyms. They also support attractive composition properties via simple vector addition, and have been found to represent analogies to some extent (Mikolov, Chen, et al., 2013).

Glove (Pennington et al., 2014) is an alternative word embeddings model that rather than predict missing words has a neural network learning objective that predicts the ratio of concurrence counts - how likely are other words to appear given a word. This allows it capture global information that Word2Vec does not. Glove performs slightly better in the original paper than Word2Vec but the results are very close.

One of they key benefits of word embeddings models is that they can be trained on large corpora of text and can be reused, and fine tuned in other contexts. There are though several limitations: They do not model the syntax of the sentence. Glove (as well as other distributed representations such as LSA and LDA) doesn't take account of order, for example "Dog chasing the Swan" will learn the same representation as "Swan chasing the Dog", though they have different meanings. This bag-of-words approach has a potential limitation with story telling in that while the model may be able to connect text with things in common, without understanding entailment it there could be limiting factor in being able to reliably order and connect sentences that follow on from each as opposed to just being closely related. Word embeddings also embed all the different senses of a word into one representation such as *crane* - as in stretching the neck, the bird, or construction machinery. Usually this is seen as a problem, but lots of visual forms of storytelling such an political cartoons (example in Figure 1.2) and comics use puns and visual metaphors that could benefit in some cases from linking different senses of a word.

Simple compositional techniques such as vector averaging can get surprisingly close to some far more sophisticated compositional models (Socher et al., 2013). However there are other approaches that extend the prediction based Word2Vec models to be able to encode paragraphs or whole Documents. One such approach is Doc2Vec (Le & Mikolov, 2014) that has two models analogous to those

of Word2Vec PV-DM (Distributed Memory) and PV-DBOW (Distributed Bag of Words). In each case the text to be encoded is represented by an Id label - this could be a sentence, paragraph, document, or represent a classification label such as a topic or sentiment. In PV-DBOW the paragraph Id the equivalent of the word in Skip-Gram; a neural network is trained to predict sampled words from the paragraph (these can be random vectors or word vectors trained simultaneously). Thus the paragraph vector learns a representation of words that occur in paragraph. PV-DM is very similar to the CBOW (word embedding) model except as well as have windows of neighbouring word there is also a paragraph vector which is used to predict the missing word; the paragraph vector should learn the context of the paragraph that is missing from the neighbouring words. On Wikipedia topic classification (Dai et al., 2015) Paragraphs Vectors has 93% accuracy compared to 84.9% for averaged word embeddings, both outperformed LDA.

Skip-Thoughts (Kiros et al., 2015) is an alternative model that rather than sampling from words in a sentence uses and encoder-decoder model to generate predictions of the neighbour, and then has a loss function that measures the error against the real neighbours in order to train the model. Other models have tried to model structure better than the bag of words alternatives such as recursive autoencoders (Socher et al., 2013) for sentiment, and more recently Tree-LSTM (Tai et al., 2015). Tree-LSTM is an extension of LSTM that supports tree structures as well as sequences; this allows it with pre-parsing (either constituency or dependency) to take account of the syntax or word order implications for meaning, which allowed improved results on semantic similarity benchmark.

This project uses word vector averaging supporting either Word2Vec or Glove embeddings, and a second alternative text similarity feature using Paragraph Vectors. Word embedding averaging still performs relatively well and are computationally quick to calculate (useful for dynamic querying). Paragraph Vectors while slightly behind the best results on semantic measures has strong software support via *Gensim* and *DeepLearning4J* (section 3.6 p. 36), and should be expected to improve substantially on LSA.

2.6.2 Image Features

In representing the image for narrative story telling the main requirement is being able to understand the scene - who and what objects are present, where they are in relation to each other, what they are doing, the composition of the photo (such as the level of the sky line), and details of the location such as whether it is indoors or out, the lighting, etc. Scene recognition is a well developed area in computer vision (see for example (Xiao et al., 2010)). The problem is while scene recognisers do a good job of categorising the type of scene - beach, forest, indoor living space, etc - they don't contain the other details of the composition or objects contained in the image, making them unsuitable when these are the main elements the system needs to be able to differentiate between.

There are other general feature models used to represent images such as SIFT (Lowe, 1999). As well as other models for picking out the contours of objects (Arbelaez et al., 2011). While these would be useful for a general impression of the image they would not be able to tell which objects are in the image. More recent attention for both identifying objects and and their locations has focused on using DCNN (Deep Convolutional Neural Networks). These models have in recent years achieved results that have far surpassed alternative models (Simonyan & Zisserman, 2014; Krizhevsky et al., 2012; Girshick et al., 2014a). Later improved models have been developed such as Fast R-CNN (Girshick, 2015a, 2015b), and Faster R-CNN (Region - Convolutional Neural Network), both following on from the earlier R-CNN (Girshick et al., 2014b) and VGG16 (Simonyan & Zisserman, 2014) as a method for predicting objects and their locations in images. The usefulness from a story telling point of view is in that as these models are able to predict objects locations with a probability they can be used as features to represent scenes, for example a photo with a Zebra in the foreground would have some similarity (but also difference) with a scene with a few zebras in the background. By controlling the degree of similarity between features the intent is control how much consecutive photos in the story have in common in terms of the objects they contain and their positions.

Figure 2.3 (reproduced from (Girshick, 2015a)) illustrates the Fast R-CNN pipeline. The network takes a set of object proposals - simply the categories of object that may occur in the image to be predicted. Fast R-CNN applies convolutions and

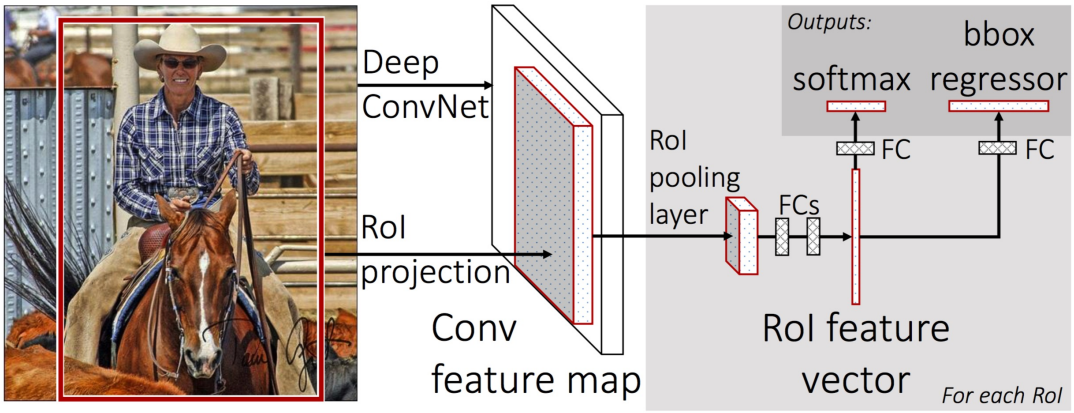


Figure 2.3: Fast R-CNN Architecture.

max pooling over the whole image to create a feature map. An external algorithm - selective search (Uijlings et al., 2013) - to the neural network projects ROI (Regions of Interest) for each of the object proposals. These are overlapping boxes for each of the object classes overlayed onto regions of the image. This is then pooled in an ROI pooling layer, an overlapping variant of max pooling. The ROI pooling layer is fed though various FC (Fully Connected) layers. The final FC layer is fed to two outputs: One is a Softmax output giving the probability of each object type occurring. The other is a regressor layer that gives the coordinates of the object locations. The training loss is the sum of both. It is implemented on Caffe (Jia et al., 2014).

There have been several later variants, Faster R-CNN (S. Ren et al., 2015) which integrates region proposal into the network (rather than an external algorithm), and YOLO (You Only Look Once) (Redmon et al., 2015), newly published, that dispenses with the region proposals entirely. Both changes are geared towards improving speed for better realtime object detection, and so are not as relevant on a collection of static images. As far as performance goes on the MS-COCO 2015 detection challenge (`http://mscoco.org/dataset/#detections-challenge2015`), all of the prize winners - MSRA (He et al., 2015), FAIR (Zagoruyko et al., 2016) and ION (Bell et al., 2015) are enhanced variants of Faster/Fast R-CNN. Fast R-CNN features are used to represent the image in this project.

2.6.3 Multimodal Features

As well as the integrated system already discussed for question and answering generation there has been interesting work to combine image and semantic features into MMDM (Multi-Modal Distributional Semantics). With stacked bimodal autoencoders (Silberer & Lapata, 2014) denoising autoencoders are used learn representation of the semantic and image data and the concatenated together. A second level then users a semi-supervised layer to predict original object labels. Embeddings can also be trained separately and just concatenated (Kiela & Bottou, 2014). This simple technique can be effective, the visual information is shown to enhance performance. Another integrated approach (Lazaridou et al., 2015) uses a Skip-gram model to train word embeddings but use a training objective that additionally measures the Cosine similarity of the image with those of the predicted words (that have corresponding images), so *pizza* might be compared with *dough*, *plate*, or *cheese*.

All of these approaches have shown improvement over unimodal embeddings as they are able to take advantage visual similarities between objects as well as their use in language that have semantic relevance such as between *pliers* and *tongs*, or *eagle* and *owl*, or *horse* and *zebra*. The disadvantage is as they are integrated it is difficult to control the influence of the different modalities; this project therefore implements them as separate features so that weights and thresholds can be used to independently control the influence of each feature.

2.7 Narrative Arrangement

The narrative selection involves a mixture of features that convey topic cohesiveness (similarity between elements), and narrative development (order or progression over the images). Similarity between Semantic text or image features, or selecting based on Theme or Named Entity can be used to create a cohesive topic running though the generated story, for example about a Dog. The narrative features control the progression so for example the first image must be negative sentiment, and the last positive, or the first contains water and the last a Frisbee. By inverting the features influence image and Semantic similarity can also be used to create narrative flow by causing there to be differences in the image

(objects contained or composition), or text (different actions or subjects).

A Viterbi search (Forney, 1973) is used in the existing ReelLives implementation for selecting images and is widely used elsewhere for NLP and machine learning applications. In the arrangement of images a Viterbi stages proceeds forwards from the first image considering possible alternatives; a probability of each is estimated based on the emission probabilities, the probability that the images is best in the position, and the transition probability of it being the right continuation of previous images. A emission probability for example could be specifying the image must be positive sentiment, or contain the word *skateboard*, the transition probability is based on LSA similarity and relative time.

The potential problem with Viterbi is that is that the probability is maximised step by step at a local level, which can lead to ultimately better global solutions being discarded. An alternative for finding an optimum global solution is MILP (Mixed Integer Linear Programming). MILP has been used in a variety of contexts for NLP tasks such as building coreference chains (Finkel & Manning, 2008), and document summarising (Gillick & Favre, 2009; Woodsend & Lapata, 2012), and Image Centriction generation (Kuznetsova et al., 2012). In MILP programming an objective equation is defined that the solver tries to optimise; this can be defined as weights of features for positions or across transitions between images. This equation is subject to constraint equations that can be used to enforce rules of transitivity or any other relational constraints between attributes. The Mixed part of MILP is that solutions are required to be Integer, a picture must either be present at each position or not. MILP has the advantage of being able to potentially find globally optimum solutions and is also highly adaptable to be able to tune relative features to be able balance image and linguistic similarity. The disadvantage is that considering all possible combinations is eventually an NP hard problem, so optimum global optimum solutions are not possible beyond a certain size (discussed further in 3.5.4, p. 29).

2.8 Additional Features

All of the other existing features are reimplemented in MS-COCO with only changes made where necessary to adapt the features - sentiment analysis, POS

tagging - where the existing implementation is tailored for social media uses and is not appropriate for the MS-COCO annotations.

Relation extraction (Fader et al., 2011; Schmitz et al., 2012; Angeli et al., 2015) (or Open Information Extraction) parses and analyses sentences producing Subject, Relation (usually Verbs), and Object triples for each sentence. For example "A small plane flying through a cloudy blue sky" can be extracted as *subject(small, plane)*, *relation(flying, through)*, and *object(cloudy, blue, sky)*. Producing triples that splits sentences into role is useful as it allows more flexibility in creating stories by specifying of the particular role that a word should play in sentence such as *plane* being the subject of the sentence not the object. It provides another query option beyond keyword matches.

Chapter 3

Methodology and Implementation

3.1 Stories

A MILP selector will be used to generate Triptychs, using 4 main types of controls: Prefiltering the selection based on theme, keyword, IE triples or sentiment. Weights to control the desired level of similarity and dissimilarity between features. This is intended to allow Triptychs of forms that have similar text (or are about the same thing), and different images composition and elements, for example a mid-distance shot of a dog with a Frisbee shot in long, medium, and close distance from different angles. Or the converse, similar looking images with different text - a dog chasing a Frisbee, drinking water, sleeping - applying a sequence of action. To augment this upper and lower thresholds can be placed on the degree of similarity. There are also general constraints such as the soft queries using word embeddings (section 3.5.2) that can be weighted and control degrees of similarity to a users search query. These controls make the implementation highly adaptable in being able to choose desired degrees of similarities when generating Triptychs.

3.2 Evaluation

The recently developed SIND (Huang et al., 2016) has a set of Gold standard stories that can be evaluated using measures such as BLEU (Papineni et al., 2002). The ability of automated validation has advantages, however huge amount

of labour and cost are required to put together such a dataset. The focus of this study is on evaluating how semantic, visual, sentiment and other features can perform without learning from a predefined narrative. The evaluation therefore shall directly survey participants following on from previous studies (Aylett et al., 2015) using the same question to provide a clear basis for comparison with the earlier work.

There are a number of areas of story telling that are of interest: A story should have a cohesive theme to it, the ordering should be correct, it should be interesting as a story, and have emotional impact (exact survey questions Section 4.3). The first aspect should be the simplest as in both the reviewed linguistics and computer vision literature performance of the proposed methods on semantic similarity, and in predicting object location have performed well. The second is an interesting test as there are no features that have ordering information, a skier may get a ski lift up, stand at the top, and then ski down. Both visually and semantically these can be related but can the system reliably get the correct order from semantic and visual similarity? An interesting story is the most difficult as the selector could easily select 3 nearly identical photos which would be boring, or in contrast select photos that are apparently unrelated; achieving a balance of continuity of entities flowing through the triptych is important. Sentiment is another difficult element. Previous trials with ReelLives has found from social media sources it can be difficult to create an emotional narrative as a high proportion of posts are positive. There may be a similar difficulty with MS-COCO in that the descriptive information used for Sentiment is mainly neutral.

The evaluation is via Amazon Mechanical Turk (`https://www.mturk.com/mturk/welcome` (AMT)). It is convenient for getting feedback quickly it does however have some potential issues: Generally it is not controlled for population or cultural factors. In previous studies (Aylett et al., 2015; Farrow et al., 2015) participants were evaluating their own images made into Triptychs, whereas now they are evaluating generic Triptychs generated from a stock collection so the emotional response may well be quite different just because the participants are more disconnected from the images.

A limitation of not having an automated evaluation set is that there are a large number of permutations in building the features (e.g. word embedding size, training iterations), and in parameter configurations, and only a small portion of them

Table 3.1: Unique counts for captions and identifiable objects in MS-COCO

Stat	Count
Tokens	6967142
Unigrams	38017
Bigrams	488160
Trigrams	1568864
Themes	922
IE Triples	1131639

can be evaluated. To this end there is a prescribing trial phase where parameters are tuned on relevant tasks to establish the configurations that are worth evaluating (section 4.1). Best practice from earlier work is used where possible as it's extremely difficult to assess small variations in features contributions to the overall generated Triptychs.

3.3 MS-COCO

The MS-COCO dataset has *123,287* images on a wide variety of subjects. Each image has 5 separate captions, each a sentence long. Table 3.1 has counts for the caption text, and Figure 3.1 is a tag cloud of the most common words in the corpus. As the tag cloud illustrates the most common words are related to people (*man, woman, person*), common actions (*standing, sitting,holding*, with fewer but still relevantly frequent nouns (*field, plate,train*), and adjectives mainly relating to colour (*red, white*), position (*front, next, top*), or counts (*many, few, two*). This is relevant as the language is quite different from that typically used on social media and so it would be expected to affect all of the text related features such as semantic similarity, themes, and named entities.

There are 80 types of objects with locations identified in the dataset (see http://mscoco.org/explore/ to explore). These don't just cover the main subjects (*people, dog, elephant*) but also items being used or worn (*bike, surfboard, tie*), and peripheral items that make up the scene (*potted plants, cutlery, benches*). Taken as a whole then the objects and their locations are descriptive of a scene,

Figure 3.1: Tag cloud of the most common words in captions with the stop words removed.

and often cover objects not mentioned in the captions.

A number of utilities have been built as part of the implementation to support the editing and management of feature files required in order to conduct experiments: Merging files, splitting them according id, filtering with the available Criteria (Section 3.5.1, p. 27), copying features from one file to another, and creating random subsets of the data. Help text is available for all the commands via the Spring Shell interface.

3.4 Features

This section discusses the features that have been extracted to be able to create interesting stories from the MS-COCO dataset. It starts with a brief overview of features that replicate the existing features in the system - tokenizing, POS tagging, themes, named entity recognition, and sentiment analysis - and then goes onto discuss in more detail newly implemented features - relation extraction, word embedding based text similarity, paragraph vectors, and Fast RCNN based image similarity.

3.4.1 Tokenizing and POS Tagging

ReelLives does tokenizing and POS using TweetNLP (Gimpel et al., 2011; Owoputi et al., 2013); a library specifically tailored for the challenges of social media in dealing with short messages, abbreviations, slang, emoticons, etc. The MS-COCO captions are more conventional descriptive language. For this the widely used Stanford Core NLP library (Manning et al., 2014) tagger (Toutanova et al., 2003) is more suitable. The tags are based on the Penn Treebank set.

3.4.2 Themes and Named Entity Recognition

Alchemy API (Turian, 2013) (http://www.alchemyapi.com/) is a proprietary suite of NLP tools offered by IBM via a cloud based REST web service [1]. The existing ReelLives system implements the themes and named entity recognition features using the service; these features are replicated with the MS-COCO dataset using the same services.

The theming (or taxonomy) service returns the 3 closest related themes to the provided Image Centrictions with a confidence score (between 0.0 and 1.0) from a predetermined set of more than 1000 (in a hierarchical structure) (IBM, 2016b). The taxonomy is geared towards industrial and business applications. Examples include *Music / Musical Instruments / Guitars*, *Automotive and vehicles / motorcycles*, and *Home and garden / appliances / microwaves*.

The named entity recognition service (IBM, 2016b) identifies a wide variety of entities such as People, Organizations, Locations, Hobbies, Movies, Songs, etc. The named entities from the MS-COCO message text are a list of the type, e.g. *Person*, and value *Clinton*. In keeping with the existing structure location related entities such as cities, landmarks, towns, geographical features are split into a separate features from the others that can be filtered on separately when creating stories.

[1] Details on the algorithms used are not published because of commercial secrecy.

3.4.3 Sentiment Analysis

The existing ReelLives system uses the Sentisense library (Thelwall, 2013). Sentisense is targeted at social media applications with support for identifying emoticons and other Twitter relevant elements. Instead the Core Stanford NLP libraries implementation (Socher et al., 2013) is used. The model uses a recursive neural network over parse trees of sentences to identify the positivity to negativity of sentences, and achieves a class leading 85%+ accuracy on the Stanford Sentiment Treebank. The API reports a scale of 0-4 from completely negative to most positive. In order to keep compatibility with the existing XML feature format these have been mapped to negative/neutral/positive tags [2].

When there is more than one sentence the score for each sentence is calculated and averaged to get the overall label. Because of the averaging the ratios gradually change as the number of sentences is increased from 1 through to 5: The ratios *negative:neutral:positive* for 1 sentence per image are *40:42:18*, and for 5 sentences per image *33:60:7*. Generally though the sentiment across the board is quite neutral as the text is descriptive and doesn't carry match sentiment. There are also significant number of sentences that appear to be neutral but are tagged differently, for example "a person skiing in an open area of snow" is tagged negative, while "a man riding a board over the top of the wave" is tagged positive. Implications are discussed later in the evaluation of the sentiment based stories.

3.4.4 Relation / Event Extraction

There are various good software relation extraction tools such as Reverb (Fader et al., 2011), Ollie (Schmitz et al., 2012), but Open IE (Angeli et al., 2015) is used as it outperforms earlier implementations. It works via splitting sentences into entailment clauses, maximally shortening each one, and then deleting less probable inter-clause dependencies to produce the most likely triples for a sentence. Multiple triples can be produced for a single sentence so *object(cloudy, blue, sky).* and *object(sky)* might both be produced; each is stored in the feature file. The extracted features turn up the *maximum entailments per clause* to 1000 the recommended limit, and switches on the coreference resolution that attempts

[2]Though a useful future enhancement would be to change it to a score so it can be optimised via a desired sentiment weight.

to replace pronouns such as *she* or *that* with the person or object being referred to. See Figure 6.1 for how the features are represented in XML.

3.4.5 Word Embeddings Composition

There are two implementations of linguistic similarity, vector averaging and Doc2Vec. Vector averaging supports either Word2Vec (Mikolov, Chen, et al., 2013) or Glove embeddings (Pennington et al., 2014). A few different embeddings have been tried during pre-screening trials but the main embeddings used in the experiments are Glove *300* dimension embeddings trained on the Common Crawl [3] and so it is suited to a wide variety of tasks [4][5]).

A second linguistic similarity measure has been built using Doc2Vec DBOW (Le & Mikolov, 2014) trained on the whole corpus with a dimensionality of 300 (for consistency with word averaging), a window size of *8*, negative sampling of *10* (Y. Goldberg & Levy, 2014), and over *10* training epochs.

3.4.6 image and Object Prediction Features

R-CNNs (Girshick et al., 2014b; Girshick, 2015a) predicts both the object likelihood and their locations. The layers of weights in the model thus will learn a form of representation of the scene. This should mean similar outputs at each layer for images with similar objects and composition, whereas those with differing objects, or the same objects in different numbers or positions the image would be expected to be further away using Cosine similarity; thus the feature layer should be a good proxy representation for the image. It is a form of visual embedding that can be used optimise the similarity/dissimilarity of the scene in the image across the story. MS-COCO features are reused from *NeuralTalk* (Karpathy & Fei-Fei, 2015)[6]. These features represent the last fully connected layer before

[3]Common Crawl (http://commoncrawl.org/ trained on 840 billion tokens crawled from the web. The context is highly general rather than News or Wikipedia based corpora where the context would be more specialised towards events or definitions.

[4]The Glove embeddings (http://nlp.stanford.edu/data/glove.840B.300d.zip)

[5]Other vectors trained directly on the Corpus, or Word2Vec pretrained alternatives such as the Google News 300 vectors (https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing) could also be used as is via a command line parameter.

[6]They are available from https://github.com/karpathy/neuraltalk

object and location predictions, the layer size is *4098*. Sparse feature vectors are used unchanged rather than applying dimensionality reduction as this would result in loss of information, and the system performs computationally well with the full vector. Additionally a separate feature has been experimented with that uses the Fast-RCNN (Girshick, 2015a) predictions directly; they are probabilities of each object type - the 80 MS-COCO types appearing anywhere in the image [7].

3.5 Selection

3.5.1 Criteria

A selector has been built to create Triptychs using a MILP (Mixed Integer Linear Programming) solver GLPK. A separate selector has been created rather than changing the existing one as it would not be feasible to complete the refactoring necessary to allow the existing selector to work with the different MILP approach in the project timescales. The current ReelLives system (Aylett et al., 2015; Farrow et al., 2015) allows the user to be able to generate Triptychs by querying on keywords, themes, entities, and sentiments via a web, command line, and Java Swing GUI. The new selector provides similar capabilities for generating Triptych extensions for new features via a command line interface.

- **Ids -** images must match a list of ids provided.
- **Word tokens -** The message text of the image must contain the word token(s) provided. Constraints can be specified to allow basic Boolean matching in a comma separated list, .e.g. "grass,and,giraffe,or,elephant,not,lake".
- **Sentiment -** images must match the sentiment(s) specified Positive/Neutral/Negative.
- **Theme -** Based on Alchemy API identified themes. For example "Sport/Skiing/Nordic Skiing", the filter will match any part of the theme hierarchy. Each theme also has a confidence so the query can specify the theme must be above a confidence threshold.

[7] Potentially the location predictions could be used as a separate feature to identify images with known objects in particular locations.

- **Entity / Location** - Queries based on the Alchemy API named entity recognition either for entities or locations, e.g. "City/London", "Restaurant/Dino's", or "Person/John".
- **Open IE Triples -** Consist of subject, relations (usually verbs), or objects. There can be multiple triples per image and each part can be filtered on or together, for example "–subject young man –relation flying –object red kite".
- **Date / Time -** images can be selected that are before/after the specified time or within the given range.

All of features are implemented as filters using the Java 8 predicate API with streams. Each of the above filters can be applied to the Triptych as a whole or to images in individual positions. This allows narrative flow across the Triptych to be specified so the sentiment can start out *negative* and finish *positive*, or start with an image on a subject say a *Dog*, finish with a *kite* and the selector will find the optimum images to connect them. With all of the above criteria the images are filtered before the MILP Solver (Section 3.5.4, p. 29) which limits the number of images the solver needs to optimise reducing computational complexity. The next section describes a softer natural language alternative that can be used.

3.5.2 Natural Language Queries

Simple vector addition and averaging with Cosine Similarity has been shown to work well in representing the meaning and enabling comparisons between longer sentences and documents (Mikolov, Sutskever, et al., 2013). It can be surprisingly close in performance to more sophisticated autoencoder models (Blacoe & Lapata, 2012). The selector supports a natural language query [8] created via combining vector averaging of words of the vocabulary either of Word2Vec (Mikolov, Chen, et al., 2013) or Glove embeddings (Pennington et al., 2014) [9] in a file provided as a parameter; or alternatively via Paragraph Vector similarity (DBOW) (Le & Mikolov, 2014), with word vectors trained simultaneously with the Paragraph Vectors. By comparing the similarity of the query vector to those of the image

[8] A useful further enhancement for the system would be a similar image feature based soft search to generate a Triptych related to an initial image.

[9] The natural language query allows stopwords to be included or excluded in the query vector based on a parameter.

text a soft query that should be able to take advantage of the semantic space properties of related terms being close to each other in space. This isn't possible with the hard filtering criteria described in the last Section [10]. The obvious limitation is that it is a bag-of-words model that doesn't take into account the compositional semantics of a sentence, and all different word senses will also be represented the same way.

3.5.3 Text, image, and Prediction Similarity

Similarity between the text, overall image and object predictions is based on the Cosine Similarity of the respective feature vectors for each image. The similarity (figure 3.1) is the dot product of the vectors normalised by the unit length of each multiplied together. It is then normalised so it is a value between 0.0 (completely different) and 1.0 (the same).

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3.1}$$

Cosine Similarity has been widely used for comparing the similarity of word compositions of vector representations of words using methods such as LSA and LDA (Mitchell & Lapata, 2010), and as effective means of comparing the similarity of word embeddings (Mikolov, Sutskever, et al., 2013; Pennington et al., 2014). It has been used to measure similarities in multimodal embeddings (Kiela & Bottou, 2014). The implementation uses Cosine similarity as the measure between the image text, image features, and object predictions.

3.5.4 MILP Model

The principle aim of MILP is to maximise an objective equation (a global optimum) subject to equations specifying constraints (Chandru & Rao, 2010). This is relevant as the goal of the selector is to generate Triptychs that tell interesting

[10] There are various ways it could be approximated by using thesaurus such as Wordnet (Miller, 1995) to lookup synonyms, closely related words, and score them based on the distance but the implementation is far more cumbersome than the approach used.

stories. Part of that is being able to select images that have degrees of similarity and dissimilarity; photos that are nearly the same aren't going to be interesting, likewise if they are entirely different. Whether it is the text, objects, or general scene of the image some things need to be shared for it to be a narrative - *a Dog walking on a beach, a Dog swimming, a Dog sleeping on the sand.* MILP allows an objective function that put weights of different features for the desired degree of similarity/dissimilarity and thresholds on the maximum or minimum degrees and find an optimum over a set of images.

The key difference between ILP and the mixed form is that in MILP some variables have to be discrete. This is a requirement for this selection task as either an image needs to be at a particular position at a Triptych or it not, it cannot be partially present. The algorithm used for solving this type of problem is branch-and-bound (Clausen, 1999). The problem is initially solved as a real numbered one (in the case of the GLPK (GNU, 2016) library a *revised simplex* algorithm). Branch-and-bound then branches the solution space into a tree structure, and computes lower bounds on the candidate solutions; the whole process is a top down recursive search to find the optimum solution that meets the constraints. The real numbered solutions are gradually relaxed until they satisfy the Integer constraints. Additionally cuts (Cornuéjols, 2007) are used to prune and tighten the solution space during the search.

In earlier work (Finkel & Manning, 2008) transitivity constraints are enforced over triples in coreferences so that if a path from i from j, and j to k then there is from i to k. This enforces a single chain that runs through all the coreferences according to the objective score. This approach would be more appropriate for longer stories where n images may be part of a story, but it is NP hard. Instead constraints are specified directly on the triples which simplifies the computational costs.

$$
\begin{aligned}
n \in N & \quad \text{Number of images} \\
S = \text{Start images} & \quad \text{images in each Triptych position} \\
M = \text{Middle images} & \\
F = \text{Finish images} & \\
X = (i \in S, j \in M, k \in F : i \neq j \wedge j \neq k) & \quad \text{Valid Triptychs, binary variable} \\
E = (n_1 ... n_z, n_1 ... n_z : n_x \neq n_y) & \quad \text{Non recursive edges between images} \\
a(p,q) \in E & \quad \text{paragraph Doc2Vec similarities} \\
b(p,q) \in E & \quad \text{object prediction similarities} \\
c(p,q) \in E & \quad \text{word text similarities} \\
d(p,q) \in E & \quad \text{image similarities} \\
Y = \{a, b, c, d\} & \quad \text{Set of features} \\
g \in N & \quad \text{Similarity between query and text} \\
\alpha = \{\alpha_a, \alpha_b, \alpha_c, \alpha_d\} & \quad \text{Weights} \\
\beta = \{\beta_a, \beta_b, \beta_c, \beta_d\} & \quad \text{Threshold Caps}
\end{aligned}
\tag{3.2}
$$

3.2 defines the sets and parameters required in the model. As images can be fixed to individual positions in the Triptych three sets are defined S, M, F (Start, Middle, and Finish) corresponding to the images allowed to be at each position. If there are no criteria specified for each position then each set will be same. The sets are made up of the images after the hard filtering described previously has been applied. X is a set variables of all valid triples over the start, middle, and finish set. It is either on (1) when a triple is selected or (0) when it is not. It is the key variable in the model as the solver is attempting to determine the optimum n triples to maximise the objective equation. E is the set of edges from all the images to another. b, c and d represent the feature Cosine similarities between all the images. α parameters are respective weights representing the strength of each of the features, β parameters are cap thresholds that can be placed on the total overall similarity.

$$maximise = \sum^{x \in X} \sum^{y \in Y} y(i,j) \cdot x(i,j,k) \cdot \alpha_y + y(j,k) \cdot x(i,j,k) \cdot \alpha_y$$
$$+ \sum^{x \in X} g(i) \cdot x(i,j,k) \cdot \rho + g(j) \cdot x(i,j,k) \cdot \rho + g(k) \cdot x(i,j,k) \cdot \rho, \text{ if } g \neq \{\} \tag{3.3}$$

Equation 3.3 is the objective equation for the model. Each respective feature similarity between the start and middle, and middle and finish is multiplied with the value of the x tuple, and the respective feature weight. If x is 0, the triple is not active, then the particular triple has no effect on the objective function value. The weights (α, β, γ) influence the degree of similarity/dissimilarity the objective function will optimise for: A positive weight produces similarity across the resulting Triptych, whereas negative weights produces dissimilarity. A weight of 0 will switch off the feature. Higher values both positive and negative will make the feature more influential. In combination the features can be used to create combinations of stories: Text that is similar but with dissimilar object predictions, or images that looking similar in but have dissimilar captions. The additional feature g is for the soft natural language query. It specifies the similarity between the query and each image. It has a tunable weight ρ so that the relative importance of the query can be changed: Make the weight higher so the resulting Triptych is closer to the query, or lower which may mean the generated Triptych is less related to the query but will satisfy other criteria stronger image similarity. For example maybe the query included *Dog* and there is no Dog in an image but could still be part of the narrative because it contains the same object or looks similar to the other pictures in the Triptych. All of the discussed weights are provided as parameters from the command line interface.

$$\text{Number of Triptychs} = \sum^{x \in X} x(i,j,k) \tag{3.4}$$

3.4 is a subject to constraint that ensures the *number of Triptychs* found as a solution is the correct one; so if 10 is specified then only 10 $x(i,j,k)$ variables can be 1 so only the the 10 highest scoring Triptychs by the objective equation will be selected. If there are not n valid Triptychs that meet the the solution and error is reported to the user.

$$\forall x(i,j,k) \in X, \forall y \in Y y(i,j) \cdot x(i,j,k) + y(j,k) \cdot x(i,j,k) <= \beta_y, \text{ if } \beta_y > 0 \tag{3.5}$$

$$\sum^{x \in X} \sum^{y \in Y} y(i,j) \cdot x(i,j,k) \cdot \alpha_y + y(j,k) \cdot x(i,j,k) \cdot \alpha_y <= 1 - \eta, \text{ if } \beta_y > 0 \tag{3.6}$$

3.5 defines threshold caps on the similarity for the feature across the Triptych, 3.6 is the same for dissimilarity. Normally the objective equation will maximise either similarity or dissimilarity. The cap allows a maximum average similarity to be placed on a feature stopping the generated Triptychs being too similar/dissimilar along a dimension. The dissimilarity is defined differently because the *for all* constraint performs faster but is not possible with dissimilarity as the values would need to be away greater than 0.0, and this would not be satisfied for triples that are not active. Both however place a computational performance penalty on the solver; more so the further the cap is away from the optimum because of the way the branch-and-bound algorithm relaxes the optimum real value solution.

3.5.5 Tabu Search

Together the feature weights, threshold caps, and filters provide a wide variety of possibilities in generating Triptychs. The MILP model will find an optimum solution given the constraints. However the number of constraints is $(n)^3$, and with 123k+ images it becomes computationally unfeasible [11]. While some prefiltering will reduce the size of the set of images there still should be a way to generate Triptychs for cases where a direct solution is not feasible.

To resolve this issue, a simplified form of Tabu search has been implemented (Glover, 1989, 1990). Tabu Search is an iterative method than considers a subspace of the problem, keeps a memory of the solutions encountered, and moves to a new space in the neighbourhood looking for a better solution. This process continues iteratively until a stopping condition is met. The pseudocode is in Algorithm 1. The algorithm is applied when the problem is too big to solve directly based on number of images under consideration (this is a configurable). If the

[11] The problem will as the size of the pool increases become NP-Hard.

Tabu search is used then a random sample is taken from images allowable in each position, and solved. If the solutions are better than those already found the list of best solutions (*short term memory*) then the best solutions are updated (with the best n kept after each iteration). At the start of each iteration the best n solutions found so far are added as candidates; this allows MILP solver to look for improved solutions in the proximity of the existing best solutions. The sample sizes from each position are weighted according to how many images can occur there. This is so the sampling across the Triptych is aligned with the total number of possibilities. For example if the generation query specifies that the final image must be a *Giraffe* with a *positive* sentiment and the others positions are unrestricted then the number of candidates images will be far lower in the final position so the sampling should reflect this.

There are more complicated models where *medium-term memory* (storing rules for promising areas to explore) and *long-term memory* (areas in the space that are unexplored) is kept to reduce computation on unpromising areas of the solution space, and the probability of getting stuck in local minima. Since a previously discounted image could again become relevant as the other images in the best solutions list change these more complex implementation would not offer much benefit which is why simple random sample approach has been used. [12]

With a large collection of images and random resampling there needs to be a mechanism for stopping in a reasonable time with a near optimum solution. The following parameters all defined in the command line interface can be used to control when the search stops:

- **Max Time -** Elapsed wall clock time in minutes.
- **Max Iterations -** Max number of iterations.
- **Iterations without Improvement -** The maximum number of iterations without improvement. A threshold parameter can also be used so small improvements aren't counted so iterations aren't wasted when there would be no discernible difference to the user.

If for example 5 iterations without improvement are specified and the random sample size is 100 then it equates to the search stopping after 500 images have

[12] If the system was extended to generate longer sequences than Triptychs then chains of full or partial solutions would be stored.

Algorithm 1 Psudocode for Tabu Search

Require: $Start \neq IsEmpty$
Require: $Middle \neq IsEmpty$
Require: $Finish \neq IsEmpty$
Require: $nBest \geq 1$
  $BestSolutions \Leftarrow Empty$
  if Simple($Start, Middle, Finish$) **then**
    $BestSolutions \Leftarrow$ Solve(S, M, F)
  else
    $BestScore \Leftarrow 0$
    while not StoppingCondition($BestScore$) **do**
      $S, M, F \Leftarrow$ Sample($Start, Middle, Finish$)
      $S, M, F \Leftarrow$ Add($BestSolutions$)
      $Solutions \Leftarrow$ Solve(S, M, F)
      $Score \Leftarrow$ Score($Solutions$)
      if $Score > BestScore$ **and** all $Solutions$ **not** in $BestSolutions$ **then**
        $BestScore \Leftarrow Score$
        $BestSolutions \Leftarrow$ AddAndKeepBestN($BestSolutions, Solutions$)
      end if
    end while
  end if

been tried in each position without improvement.

3.6 Technologies

The primary programming language used is Java JDK Version 8 (Oracle, 2016). Java was chosen primarily for compatibility with the chosen libraries Stanford Core NLP (natively implemented in Java), ease of use accessing Alchemy services in Java, for having robust high-performance libraries for XML and JSON mapping, support for Word2Vec and Glove embeddings and required matrix operations. It is also relatively simple to develop for (compared to say C++), has good IDE and build tool support, and has strong support for concurrency via built in features such as parallel streams, and executor services. Brief details of libraries and build tools used:

- **Spring Shell 1.2 (Pivotal, 2016) -** The project requires a large amount of manipulation of XML files in feature building, and applying different criteria in selection. This suits a command line interface (CLI). Spring shell makes it easy to build CLIs via annotations with support for type checking, mandatory and default values, and code completion. It is also built on the Spring Framework, a dependency injection framework that makes wiring and configuration of the application straightforward.
- **Gradle 2.13 (Gradle, 2016) -** Is a flexible build tool written in the Groovy programming language that provides dependency management, test integration, and support for automated packaging and deployment.
- **Spock 1.1 (Niederwieser, 2016) -** A BDD (Behaviour Driven Development) style unit testing framework.
- **Stanford Core NLP 3.6.0 (Manning et al., 2014) -** A powerful NLP library used for the features: Tokenizing, POS tagging, Sentiment Analysis, and Open IE extraction.
- **Alchemy API (IBM, 2016a) -** Accessed via REST web services. Provides themes (taxonomy), and entity extraction. Features are the same as the existing ReelLives system.
- **Jackson 2.7.4 (Jackson, 2016) -** JSON to XML mapping from MS-

COCO annotations to ReelLives XML feature format. Implementation uses annotations to map to/from Java classes, making import/export easy, and comfortably handles the full data set of 130K images with all features.

- **DeepLearning4J / ND4J 0.5 (Skymind, 2016a, 2016b) -** DeepLearning4J is a machine learning library written for JVM (Java Virtual Machine) languages that aims to be highly scalable across clusters (building on Hadoop / Apache Spark). It has built in support for Word2Vec and Glove embeddings. DeepLearning4J is built on ND4J a JVM based tensor library that intends to be the the JVM equivalent of Numpy.
- **GNU Linear Programming Kit (GNU, 2016) -** Is an Open Source ILP (Integer Linear Programming) and MILP (Mixed Integer Linear Programming) solver that uses a version of a branch-and-bound algorithm (Clausen, 1999) together with Gomory cuts (Cornuéjols, 2007) to solve MILP problems. A lib-GLPK a bridge library has been used to call the solver from Java, and the constraints have been specified in MathProg, a mathematical expression language.

The source code for the system is available on a GIT repository at https://gitlab.com/david.wilmot/reelout-mscoco-feature-extractor/tree/MSc-Submission [13]

[13] Alchemy API based components will fail because the API key has been revoked. A new API key will need to obtained from http://www.alchemyapi.com/api/calling-the-api to continue using these services.

Chapter 4

Evaluation

4.1 Prescreening Trials

4.1.1 Types of Trials

The combinations of features and parameters as described in the previous chapters are too numerous and extensive to be able to evaluate. Therefore a prescreening phase has been conducted on a variety of tasks to discover setups worth evaluating. Generally the subjective nature of this pre-evaluation is a problem but this is discussed further in the following discussion. Ideally experiments would be conducted with 1-5 sentences per caption to assess the effect of increasing amounts of linguistic content, but due to constraints on the combinations that could be tried 1 sentence and 5 sentences per image have been used. The following types of trials have been conducted:

- **Best Stories vs ReelLives system -** Comparing the earlier version of ReelLives with MS-COCO against the new system with a variety of different parameter configurations on random samples of 1000 and 5000 images [1].
- **Sentiment Stories vs ReelLives system -** Similar but adding a sentiment narrative to the generated Triptychs.
- **Theme / Entity / Search Based Stories -** Generating Triptychs on particular themes, Open IE triple based searches, and named entities. In the

[1]The ReelLives system will not run on the full dataset as LSA is too computationally expensive, which is why smaller random samples have been used for comparison.

Table 4.1: A series of parameter trials that have been performed.

Trials	**Params**
Equal Weighting	-1.0,0.0,1.0
Caps	0.95, 0.9, 0.8, 0.675, 0.6
Finer	(-) 0.75, 0.5, 0.25, 0.33, 0.20, 0.10
Finer with Caps	Best caps combined with best weights

trials different criteria are sampled from high, medium, and low frequency buckets to ensure the trials covers adequately the variety in the dataset.

- **Connections -** Choose a random start and end subject to test how features are able to link together subjects by finding common elements either linguistic or in image features.

A wide variety of parameter settings have been used in pre-training trials. In summary a series of progressions have been tried with the weighted features word embedding based linguistic similarity, paragraph vector based similarity, image features, and object prediction:

The feature exploration progressed (Table 4.1) from trying each weighted feature individually, and then in combination of equal weights. Following on, a positive predominant weight combined with smaller negative weights, for example the text of the Triptych should be similar and the image dissimilar, and then combining these smaller thresholds with caps.

4.1.2 Prescreening Findings

There are a number of findings from the prescreening trials that inform the later experiments. There are problems with the reproduced Themes and Named Entity features provided by Alchemy API. With Themes a significant portion of images seem to be misclassified, possibly because of the industrial nature of the taxonomy, and there are a substantial portion where there are no themes of more than 50% confidence. With named entities there are few that are interesting because descriptions are of the form “A man is hitting a Tennis ball” rather than “Novak Djokovic is playing Tennis in the Rod Laver Arena, Melbourne” where it may

pick up some interesting named entities that can be used for queries. Therefore neither feature is used in the main experiments, rather the Open IE Triples are used because they are more directly related and representative of the captions. Sentiment as well is generally quite neutral in descriptions but this is tested for comparisons to previous ReelLives experiments.

As far as the linguistic features go, generally when maximised similarity vector averaging and paragraph vectors produced similar results. Overall the Cosine similarity between the paragraph vectors was higher, had a narrower range, and was more difficult to balance with other features [2]. To keep the variations manageable for evaluation only word averaging with pre-trained vectors is used.

The image features Cosine similarity in general has a lower upper bound a higher variance than the semantic based features. This means that the image features need to have smaller weights or otherwise they become too dominant in the generated Triptychs. When unrestricted by search such as in random samples trials image similarity pulls the selection to certain types of images. The resulting Triptychs often feature zebras, skiing scenes, or trains; all of these have strong visual similarities - stripe patterns, predominantly white, and a strong diagonal in the image. Object prediction features seem to produce quite a strong overlap with image features, which isn't surprising as the image features must incorporate this information. For the participant evaluated experiments only the image features are used in configurations; more work is needed to identify how object prediction (and possibly location) can be used in combination to improve generated Triptychs.

Often the samples generated from the trials were good but where there was a clear ordering events should take such as baseball (pitch, hit, catch) the ordering is often incorrect. Other findings are when the soft query related search is used then the weight needs to set considerably higher than other features (a factor of 3 is used in the experiments), otherwise the other features can lead to the generated Triptych not being closely related to the search. In general the weight of features is highly sensitive, and small changes can either produce Triptychs that are boring because all the images are similar, or unrelated. Caps on maximum

[2]The paragraph vectors were trained solely on the corpus, with word vectors trained at the same time. A further variant worth evaluating for future studies would be to reuse the existing word embeddings for training.

similarity also don't work well if other features weights are negative (dissimilar); the generated Triptychs are pulled too easily to unrelated items even with small negative weights.

4.2 Configurations

From the trials the following configurations are used for evaluation with the new system. All of the example figures are the best Triptychs taken from the 5000 image samples generated for the experiments (Section 4.4).

- **Semantic similarity -** To provide a direct comparison between word embeddings and the LSA in the existing features, maximises semantic similarity. Example Figure 4.1.
- **Image similarity -** To purely assess how image features on their own perform in generating Triptychs, maximises image similarity. Example Figure 4.2.
- **Image Cap -** images as close as possible below a given threshold. Example Figure 4.3.
- **Word centric -** Similar word features (semantics), with dissimilar image features. This creates Triptychs that are about the same thing but have different views in terms of composition, distance and background. Example Figure 4.4.
- **Image Centric -** Similar image features, with dissimilar word features. This creates Triptychs that have similar objects and similar composition, but the differing text draws the Triptych towards different actions, e.g. a dog sleeping, a dog walking, a dog drinking. Example Figure 4.5.

4.3 Questions

To provide for a strong basis of comparison with earlier published papers the evaluation questions are the same as previous ReelLives papers (Aylett et al., 2015):

Figure 4.1: Example of a Word semantic similarity maximising Triptych.

Figure 4.2: Example of an Image feature maximising Triptych.

Figure 4.3: Example of an Image feature capped similarity Triptych.

Figure 4.4: Example of a Word Centric Triptych.

- Q1: To what extent do these pictures share a common topic?
- Q2: Do the pictures seem to be in the wrong order?
- Q3: How much does this sequence of pictures tell a story?

Figure 4.5: Example of an Image Centric Triptych.

- Q4: What emotional feeling is conveyed by this picture sequence?

The first 3 questions are Likert scales from *not at all* to *very much* with 3 being the midpoint and neutral. The 4th question is a 2D emotional compass (figure 4.6, from (Russell, 1980)) that allows participants to express emotional sentiments about the Triptychs by selecting any point on the grid. For each question the user only sees a composite image Triptych and not captions so it is only the images as stories that are evaluated.

To assess variations from different groups in the population participants are also asked for their gender and age range in the categories under 20, 20-29, 30-39, 40-49, 50-59, and 60+.

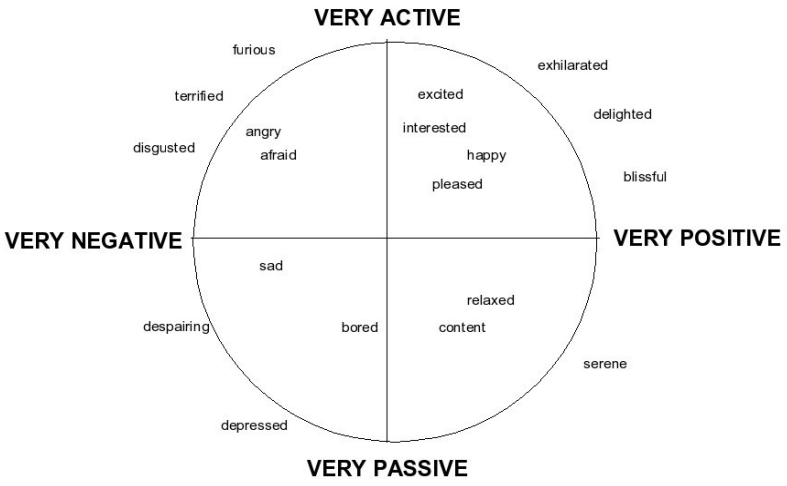


Figure 4.6: 2D Sentiment selection, axes from positive to negative, active to passive.

4.4 Experiment: Comparison with ReelLives system

4.4.1 Setup

The first experiment directly compares ReelLives Triptychs for Best and Sentiment based on stories against the new features and MILP selector on the MS-COCO dataset. The ReelLives system has a limitation in that LSA does not computationally perform well on the whole dataset. So instead random samples of 5000 images were sampled from the whole dataset and the best Triptychs for each of the following configurations generated [3]:

- **ReelLives -** Maximises the semantic similarity using LSA across the image. Example Figure 4.7.
- **Random -** Randomly selects 3 images. Example Figure 4.8.
- **Image Similarity -** *image weight* = 1.0 , maximise image similarity.
- **Image Cap -** *image weight* = 1.0, *max image similarity* = 0.675
- **Word Similarity -** *word weight* = 1.0 , maximise word embedding semantic similarity.
- **Word Centric -** *word weight* = 1.0, *image weight* = −0.2,
- **Image Centric -** *word weight* = −0.6, *image weight* = 1.0,

Figure 4.7: Example of a ReelLives configuration Triptych.

The Tabu search is run for a maximum of 45 minutes or 10 iterations without improvement with a minimum improvement threshold of 0.01. In practice in most cases the search stops without improvement before the time expires. The first

[3]This experiment doesn't use queries. Each example presented to the user is the best Triptych given the Semantic, Image, and Sentiment configuration provided.

Figure 4.8: Example of a random Triptych.

survey consists of 10 Triptychs for each of the above 7 types with a randomised order [4]. The second survey is the same but includes a sentiment narrative; the first image has a constraint to be negative, the middle neutral, and last positive. This should create a more positive sentiment over the Triptych.

4.4.2 Hypotheses

From the earlier background and discussions come the following hypotheses:

- **H1:** Both ReelLives systems' implementation of all new variations considerably outperform the randomly generated stories on Q1 and Q3. They will be relatively weaker on ordering Q2 than they are on Q3 and Q1 but still outperform random.
- **H2:** Word Embeddings based semantic similarity will produce more cohesive stories than the existing LSA implementation according to Q1, and better but with a smaller difference on Q3.
- **H3:** Image features on their own will perform as well as linguistic features on Q1, Q2, and Q3.
- **H4:** Capping the image will produce Triptychs that score lower overall on Q1 (common topic), but produce a more interesting stories (Q3).
- **H5:** Image and Word centric Triptychs will be less cohesive (Q1), but produce stronger stories (Q3).
- **H6:** Image Centric stories will be more cohesive (Q1) than Word Centric.

[4] To keep the survey to a reasonable size to be completed, a lot more samples Triptychs have been generated, from 40 different samples of 500, and are available in the project folder.

- **H7:** Word Centric stories will produce more interesting stories (Q3) than Image Centric.
- **H8:** Sentiment stories will perform better on Q3 Story than neutral.
- **H9:** The sentiment stories will show stronger sentiment along the positive axis than the best stories.

H1 is a baseline comparison of ReelLives against the new dataset. H2 and H3 are from the performance discussed in related work in the literature. H4, H5, and H6 are all based on the notion that maximising similarity will produce Images that are too similar to be judged as stories, and either capping or having a dissimilar element will create more variation of action, objects, and composition and thus be more of a story. H6 and H7 come from observations in the trials that Image Centric Triptychs are more similar in the objects they contain than Word Centric (and so expected to have greater topic cohesion), yet the Word Centric are more likely to contain different objects but with a common verb such as *play*, *ride* or subject such as *party* linking them and hence will be judged more highly as stories. H8 and H9 are from earlier findings (Aylett et al., 2015).

4.4.3 Results

4.4.4 New Features Results

The data was analysed in SPSS using a MANOVA (Multivariate analysis of variance) with the different configuration types as in subject factors, with Greenhouse-Geisser correction applied. For the best Triptychs there are 31 complete survey responses, and for the Sentiment stories 33 included in the analysis. The approach adopted to thre results is two tier, first checking the significance of overall multivariate tests, and if significant performing a T-test pairwise comparisons Bonferroni correction. Figure 4.9 shows a histogram of questions Q1 to Q3 for each configuration with a 95% confidence interval, Figure 4.10 is the same for the Sentiment story (in Appendices are Figures 7.1 and 7.2 that show Standard Deviation error bars).

The overall significance for Q1 Topic is high showing significant results on all four multivariate tests - *Pillai's trace*, Wilks' lambda, Hotelling's trace, and Roy's

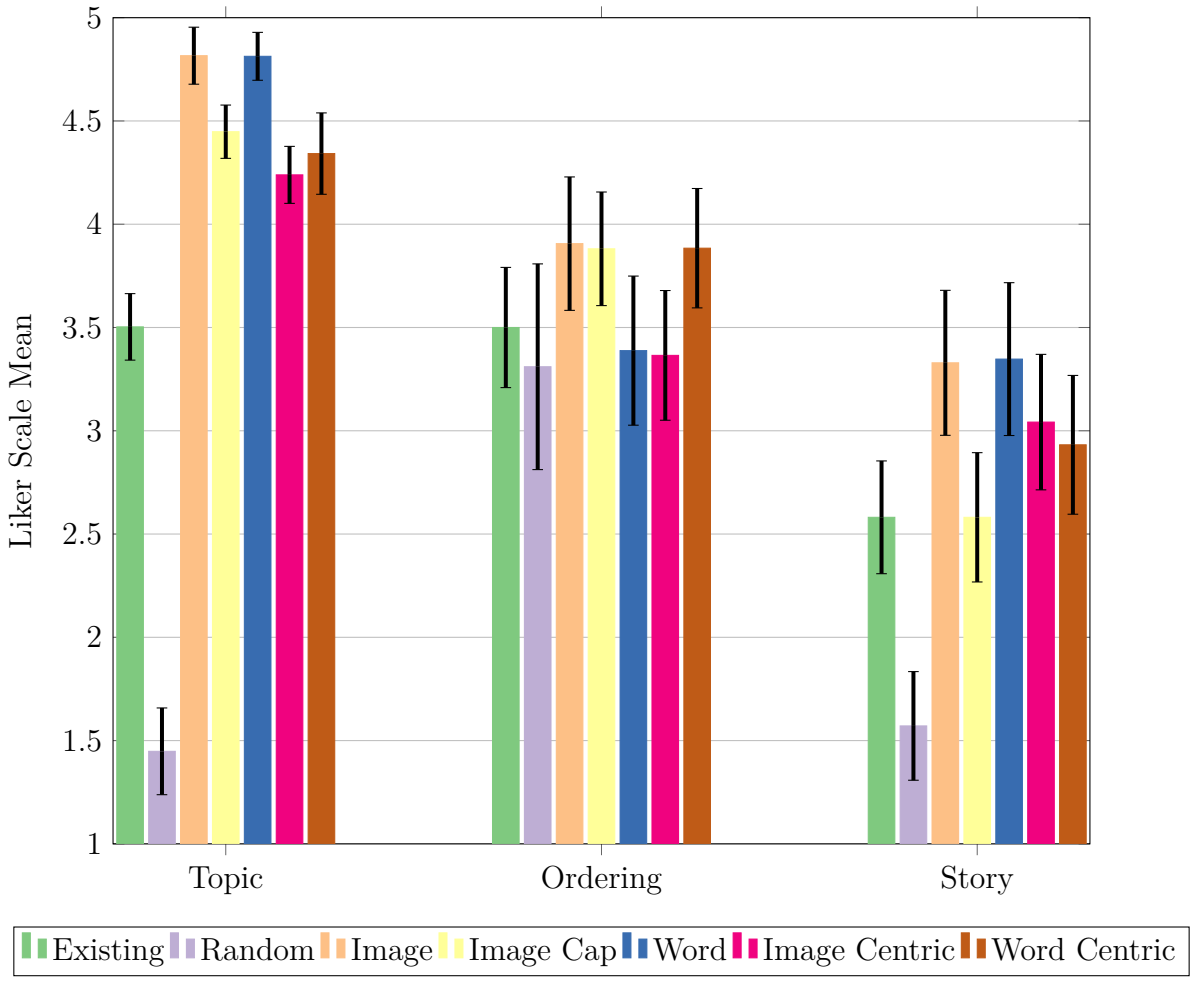


Figure 4.9: Shows averages and 95% confidence interval error bars for the Best Triptychs Q1-Q3 (Topic, Ordering, Story). With Q1 and Q3 higher is better, for Q2 lower is.

largest root for both best and sentiment stories, for the Best stories *Pillai's trace* -0.932, $F = 105.252$, $p < 0.0001$ and Sentiment 0.939, $F = 68.948$, $p < 0.0001$. For Q1 (topic coherence) random as expected does badly on both the Best and the Sentiment stories scoring significantly far lower than the ReelLives system in pairwise comparison T-tests (using *Bonferonni* interval adjustment) - 2.055 mean difference for best and 2.208 for sentiment stories both with $p < 0.0001$, and the best of new configurations image 3.368 with best and sentiment 3.091 $p < 0.0001$. For Q3 (telling a story) the results are still significant but considerably weaker than Q1. Overall the scores are substantially lower which isn't a surprise as telling an interesting story is far harder than collating images on the same topic. Compared to random for the best triptychs the ReelLives system is 1.010 better,

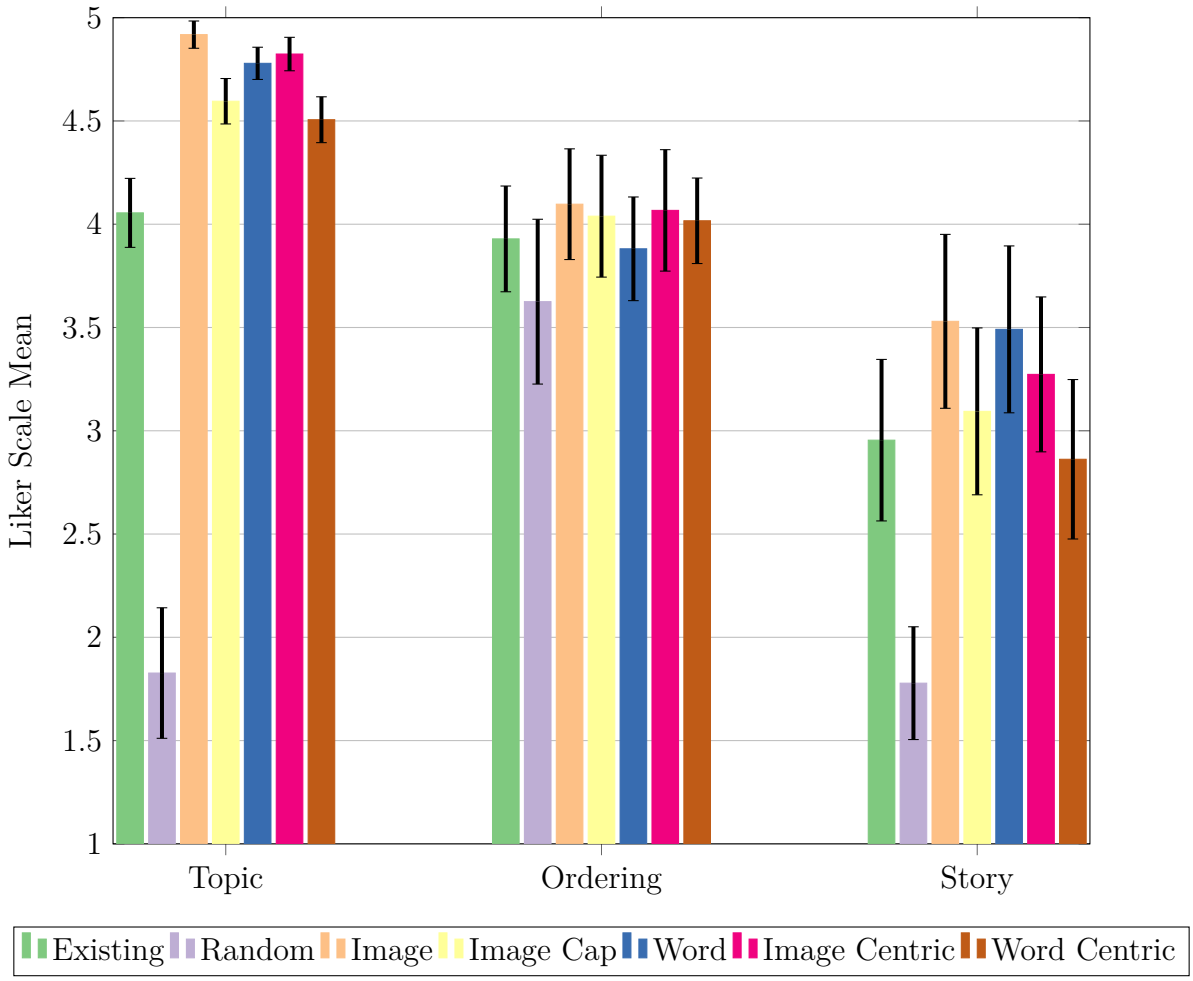


Figure 4.10: Shows averages and 95% confidence interval error bars for the Sentiment Triptychs Q1-Q3 (Topic, Ordering, Story). With Q1 and Q3 higher is better, for Q2 lower is.

for sentiment 1.176, for the new configuration means are at least 1.326 high for the best stories, for the sentiment stories 1.084, all $p < 0.0001$. Overall the data supports H1, the ReelLives system and new variants considerably outperforms random selections. Though the new configurations do not score much higher than the midpoint of the scale of 3 so it is not convincing overall that participants judged that the Triptychs tell a story. Q2 is a different matter, H1 is that correct ordering is much weaker than other measures but it was expected to be better than random. However for both the best and sentiment stories the ReelLives system and new variants do no better than random with a 95% confidence, though the differences with image are close to this threshold. So H1 is supported for Q1 topic cohesiveness and Q3 story, but not for Q2 ordering.

H2 tests the word embedding based similarity directly against LSA. The word embeddings significantly outperforms LSA on Q1 (topic cohesiveness) - 1.310 mean pairwise difference for best and 0.724 for sentiment, less so on Q3 (story) - 0.767 pairwise mean difference with best and 0.318, all with significance $p < 0.0001$. H2 is supported, the word embeddings do represent topics better and there is a smaller but significant improvement to the story.

H3 is a comparison between the word embedding based similarity and the image features. For the best stories the results are nearly identical with 0.03 difference for Q1 and 0.019 for supporting the hypothesis that the image features on their own represent the a topic as well as the word embedding based features. For the sentiment stories there is a bigger difference; image is better than Word on both Q1 0.139 ($p < 0.01$), and Q3 0.318 mean difference (only $p < 0.16$). Overall the results show that image features are able to generate Triptychs at least as well as word embedding features.

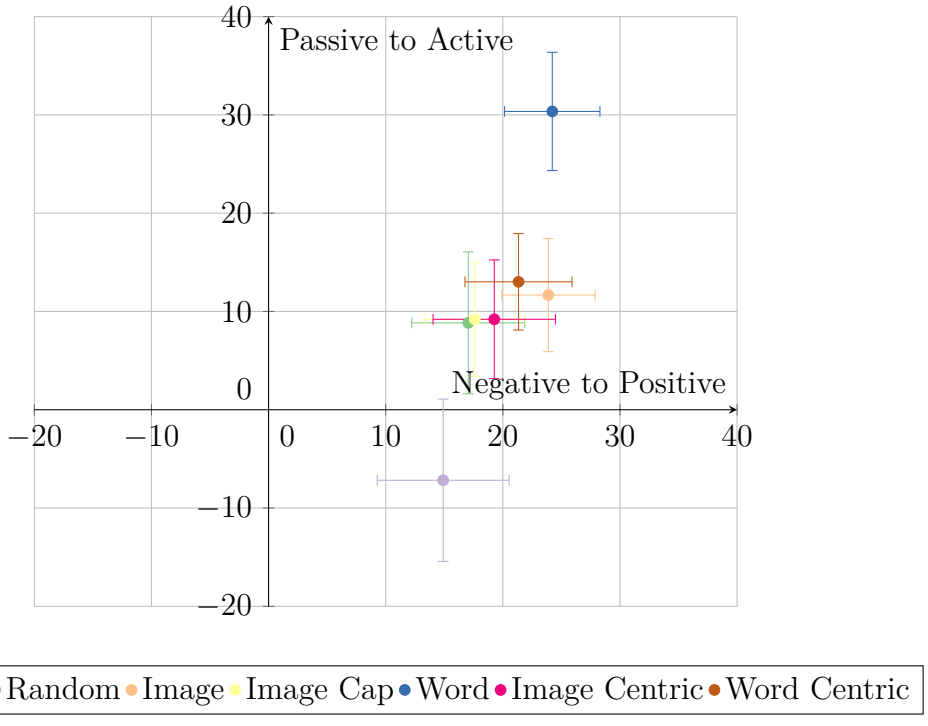


Figure 4.11: Sentiment Plot for Q4 Best Triptychs showing 95% confidence interval error bars.

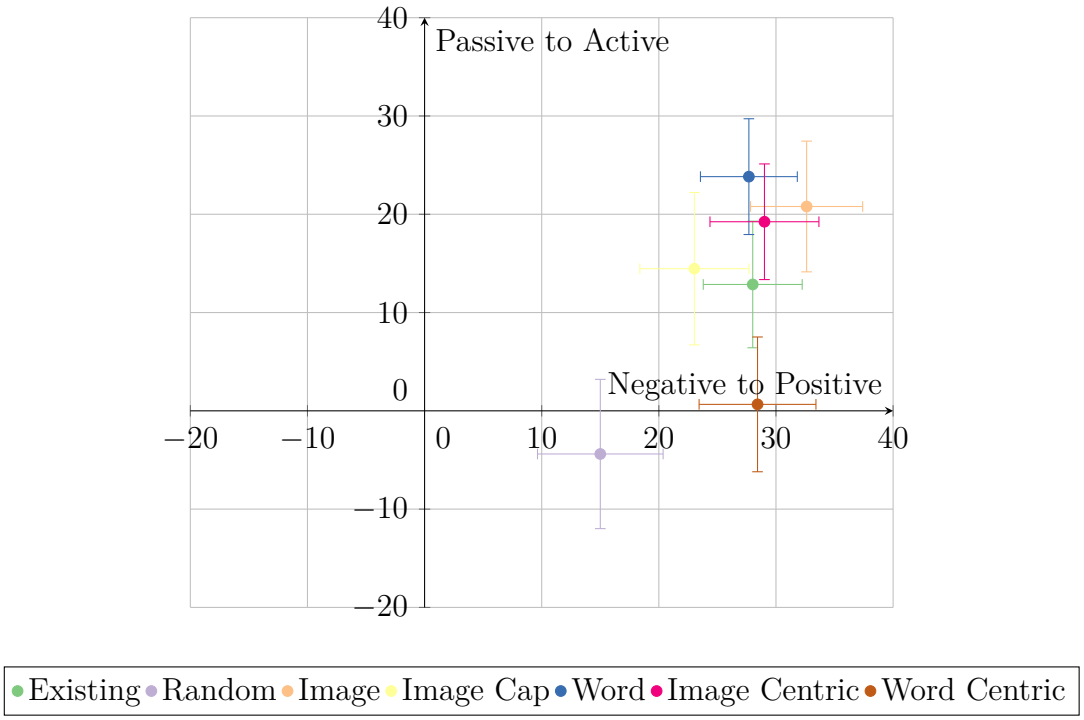


Figure 4.12: Sentiment Plot for Q4 Sentiment Triptychs showing 95% confidence interval error bars.

4.4.5 Dissimilarity Results

H4-H7 relate to the other variations Image Cap, Image Centric and Word Centric. These all have some type of limit or divergence to try and create more interesting narratives (by having some elements that are similar, and some that are different across the Triptych). In all cases for the Best stories the hypotheses are not supported: For H4 (Image Cap) and H5 (Image Centric and Word Centric) the topic cohesiveness is indeed significantly lower than either Word or Image (Best means: Image Cap 4.448, Image Centric 4.239, Word Centric 4.342, Word 4.814, Image 4.816 - all significant at 95% confidence). For Sentiment stories the differences are not significant when compared to Word based stories, but are with Images. However the resulting Triptychs do not create a more interesting narrative: Image Cap performs significantly worse than either Word or Image (For Best: Image Cap 2.897, image 3.329 Word 3.348, for Sentiment: Image Cap 3.094, image 3.529 Word 3.273). For Image Centric and Word Centric they are either worse or not significantly better than Word and Image (Best: Image Centric 3.042, Word Centric 2.939 vs image 3.329, Word 3.348, Sentiment: Image Centric 3.491, Word

Centric 2.861 vs Image 3.529, Word 3.273) using a 95% confidence threshold, so the hypotheses are rejected. H6 (Image Centric having more topic cohesion than Word Centric) is significantly better only for the Sentiment based stories, for the Best overall it is not. For H7 Word Centric stories do not produce better stories than Image Centric for the Best results, or the Sentiment results, and although not significant at 95% confidence it is more likely the reverse is true. So overall using dissimilarity (divergence), or threshold cap to try and create a more interesting story isn't supported for either H4, H5, or H7. Only H6 which says Image Centric produces more cohesive Topics than Word Centric is partially supported. In summary the dissimilar elements have lowered topic cohesiveness but haven't improved the story.

4.4.6 Sentiment Results

H8 is that the Sentiment Stories will be. For Q3 analysing both the Best and Sentiment stories *Pillai's trace* is 0.932, and *Wilk's Lamdba* 0.58 at $F22.292$. For all the configuration types apart from random there are slight increases in the ranges of between 0.1 and 0.25 but they are not statistically significant at 95% confidence. With the larger sample size they may well be but if there is any effect it would be small. One of the reasons for this maybe that proportionally there are far fewer positive image captions so the selector has far few images to choose from hence the connections that can be made with other images are fewer, and so there is not the improvement that is expected. Q2 ordering is slightly lower for Sentiment but the difference is much smaller and not significant.

H9 is that creating a negative to positive will increase the positivity. Figures 4.11 shows the view of the sentiment compass (figure 4.6) for the Best Triptychs; the user is asked to select in Q4 with the same axes. The scale for the whole compass runs from -100 to 100 and is compressed for the charts. For H8 the chart shows most of the configurations (with two exceptions) are clustered around 20 positive sentiment, and 10 active (closest to where *pleased* is in the compass) and clearly not on the 0/0 axis, and so not neutral as per the hypothesis. However the fact they are all close, insignificantly different at 95% confidence suggests it is facet of the data rather than individual parameters The two significant outliers are random which is much more passive (compared with Image Cap 16.381 with

95% confidence according to pairwise T-Tests). Likewise Word is significantly stronger on the active dimension (17.339 mean difference with Word Centric at 95% confidence) in a close position to *happy* in this axis. Figure 4.12 is the same view for Sentiment Triptychs. The figure shows that generally the results look to be shifted towards positive but none of the differences are significant so it's not possible to support the hypothesis H8. This time Word Centric is an outlier but less positive than the others. With both this and Word in the Best Triptychs there wouldn't seem to be a causal reason why one would be shifted active and the other active when the weights in the MILP are independent of the sentiment which is a pre-filter.

The sample size was not big enough and the variation in participants not wide enough for there to be statistically significant results between Gender and Age Groups. [5].

4.4.7 Primary Results Discussion

There are a number of major points to come out of this experiment. Word embeddings based feature represent Q1 topic coherence better than LSA; this is consistent with other factors literature discussed in the background. The image based convolution features also perform as well in this regard as the word embedding features showing that they are able to represent a scene well. This is particularly important as to some extent having a set of well defined captions alongside images is rare; a high proportion of photos on social media have no captions at all, or are of mixed quality and so do not have the descriptive detail of the MS-COCO descriptions.

There do seem to be two major issues: One is on their own while both image features and word embeddings are good at finding cohesive topics they perform considerably worse in produce good story Triptychs. The wide Standard Deviations for Q3 (Figures 7.1 and 7.2) though also indicate there are broad opinions about what makes an interesting story. Other attempts to produce more interesting stories either by capping similarity, or introducing a dissimilarity produce interesting results that have quite a different look but are not judged to have

[5]Further research with larger sample sizes and stratified samples to get a representative balance of age and gender would be needed.

produced better stories. Part of this is down to the sensitivity of the weights; for each of these alternative configurations there are one or two examples in the experiment that don't look coherent because the images are too dissimilar, but if the weights are slightly lower as in some of the pretrials then they do not look different from either the Image or Word configurations. These distributed matrix based similarities with Cosine distance are coarse controls; they allow the choosing of how much similarity but not where the difference should be. So while the dissimilarities produce differences they don't necessarily follow on from each other in the way a narrative should while also lowering topic cohesiveness, hence the worse outcomes. An example of this is a Triptych that was generated for Image Centric that has a Green, Red, and Yellow train with nearly identical angle and composition, while it is a cohesive topic as a collection of trains, it doesn't tell a story.

The random Triptychs provide a lower bound on performance that the configurations should be expected to be better than. However a weakness of the evaluation is there is no Gold Standard upper bound. Hand selected best stories would serve as a target or upper bound that the system should expect to reach. If hand selected stories only scored 3.5 on average on the Q3 Story then the results would already be impressive. It's likely hand picked stories would do better, but without a specifically evaluating them it cannot be said for certain.

4.4.8 Further Observations

Retrospectively maximising semantic and image similarity together is another configuration that should have been in the survey. Examples were generated but the images were judged too similar to each in the pre evaluation trials, and would make the survey unmanageable-ably large. Given maximising semantic and image features on their own scored highest overall it would have made a useful point of comparison.

Another strong cause of not creating stories is the the weakness of ordering. The previous (Aylett et al., 2015; Farrow et al., 2015) studies were collected from peoples personal social media archive which is much more limited in range of subjects, and also made use of time information in selection, both makes correct ordering easier to achieve. Without these while many of the examples do look in

the correct order when there is clear order to events such as in baseball - pitch, hit, catch - then while the features can correctly identify hose closely related they are, there is not the ordering information available to consistently get them correct.

Sentiment isn't a big factor as most of the configurations were within a very narrowly clustered around the same area on the chart. This is another area though where a dataset based on someones own personal images is likely to evoke a different response from those generated from a stock collection. Useful further work though would be to try different variants such as *positive, neutral, negative* or *positive, neutral, negative* to test if it significantly effects the sentiment.

4.5 Experiment: Soft Search

4.5.1 Setup

The previous experiment tested the ability to select the Best or Sentiment from a random sample, but one of the main functions of the system is to be able generate Triptychs on particular topics. This experiment test variants on the the new system configurations seen earlier, but uses a soft search word embedding based query. A soft search query is based on the Cosine similarity between averaged word embedding vectors. The existing configurations have to be adjusted (negative weights increased) to take account of the fact that there is a new weight introduced for the query weights which creates text similarity across the image. The soft *sentence weight* is also set to a comparatively high value of 3. This is as a result of the pre-evaluation trials where it was found that smaller weights would sometimes generate Triptychs that were not closely related to the query. With the divergent Image Centric and Word Centric stories is was also necessary to make sure positive weights (either image or semantic) are higher than the negative as otherwise the Triptychs would split; a completely unrelated image in the middle with tow images closely related to the search on either side.

The search queries have been randomly selected from the Open IE Triples. These search queries (such as *man riding a surfboard*) are used in the soft query, but also used as a hard filter to preselect images that match the triples that are then

randomly selected from to generate the Triptych - Open IE Random. This is a strong baseline comparison as the Open IE Triples will tightly select images that belong to a single subject, making it a much tougher test to beat than purely random selection [6]. The distribution of triples follows a typical Zipf distribution (Newman, 2005); so it order that the search queries are representativeness each was selected randomly using equal probability sampling from high (50+ occurrences), medium (10 - 50 occurrences), and low (3 - 10 occurrences) subset [7]. The dataset used for the experiment has 5 sentence captions per image rather than 1 as in the previous experiment to given more linguistic context so that the Open IE search has a wider distribution of triple counts. Selecting from 3 different frequency subsets means different degrees of freedom are placed over thr Open IE Random search. In theory the soft search should not be effected in the same way as it matches using Cosine distance against distributed representations and so doesn't rely on exact wording. The queries for the experiment are in table 4.2 [8][9].

Table 4.2: Queries and frequency sets used in the experiment. Randomly sampled from High, Medium, and Low frequency Open IE Triple sets.

Query	**Frequency**
zebras are grazing	High
man hit tennis ball	High
dog laying on bed	High
plate sitting on table	High
man riding skateboard	High
dog holding frisbee	High
man swinging baseball bat	High
man flying kite	High
pizza is topped	High
Continued on next page	

[6] In trials there were also a number of Hard selection variants generated that use the Open IE Triples to preselect and then the MILP selection to choose the best of them. These are not tested to keep the the number of configurations to a feasible size for creating surveys.

[7] Most triples only occur once but these cannot be used for the obvious reason that 3 images are needed to generate a Triptych.

[8] As with the previous experiment some were skipped other randomly selected queries were skipped over because the images have now been removed online.

[9] All the pronouns used are male. While *man* is more common than *woman* in the dataset this is just a coincidence of the random selection process.

Table 4.2 – Queries for experiment

Topic Id	Top 10 words
person riding brown horse	Medium
cat sits on table	Medium
path is in forest	Medium
umbrellas is in rain	Medium
man is in red jacket	Medium
Living room filled with furniture	Medium
pizza sitting on plate	Medium
two people sitting at table	Medium
fire hydrant is located	Medium
male is in brown shirt	Low
couple standing next to tree	Low
this is cat laying	Low
man doing stunt	Low
dog sitting on ground	Low
city buses are parked	Low
court is with rackets	Low
stove top with tea kettle	Low
man walking next building	Low

The following are the configuration weights used:

- **Open IE Random -** Pre-filters all image matching the Open IE Triple, and randomly selects 3 of these.
- **image Similarity -** $sentence\ weight = 3.0$,$image\ weight = 1.0$, maximise image similarity.
- **Image Cap -** $sentence\ weight = 3.0$, $max\ image\ similarity\ = 0.675$
- **Search Similarity -** $sentence\ weight = 1.0$, maximise similarity with the query. [10]

[10]This is effectively the same as the word similarity from the previous experiment only now instead of maximising similarities between the text in the Triptych it maximises similarity between the query and each candidate image text.

- **Word Centric -** *sentence weight* = 3.0, *word weight* = 1.0, *image weight* = −0.4,
- **Image Centric -** *sentence weight* = 3.0, *word weight* = −0.8, *image weight* = 1.0

9 Triptychs of each of the above type were created for 3 different surveys with the high, medium, and low frequently occurring Triples make 45 separate Triptychs in total. These were published on AMT as 3 separate surveys completed by different participants. Splitting of the survey was necessary to make it a manageable size of participants could complete each one in under an hour.

The Tabu search is run for a maximum of 60 minutes, other MILP selection parameters are unchanged. The increase in time allowed is because the search is sampling from a much bigger dataset. Most searches however finished using early stopping before the time limit is reached.

4.5.2 Hypotheses

This experiment has the following hypotheses:

- **H10:** The new configurations are able to outperform the Open IE random selection in Q2, and Q3.
- **H11:** Participants are able to differentiate between Triptychs generated with different configurations.
- **H12:** Sentiment will be not have any distinctive variations across the configurations.

H10 is on the basis that Open IE is restricted to exact matches, whereas the soft embedding based search can more flexibly find closely related images that may not have the same caption wording - having different composition, objects, or other details. This should produce better stories. This is not necessarily the case for Q1 topic cohesiveness as exact match on subject, verb, and object should generate a Triptych strongly about the topic. H11 is asserting that the configurations produce discern-ably different outcomes. H12 is following on the previous experiment and is rechecking in the context of specific searches that Sentiment is not an important factor.

4.5.3 Results

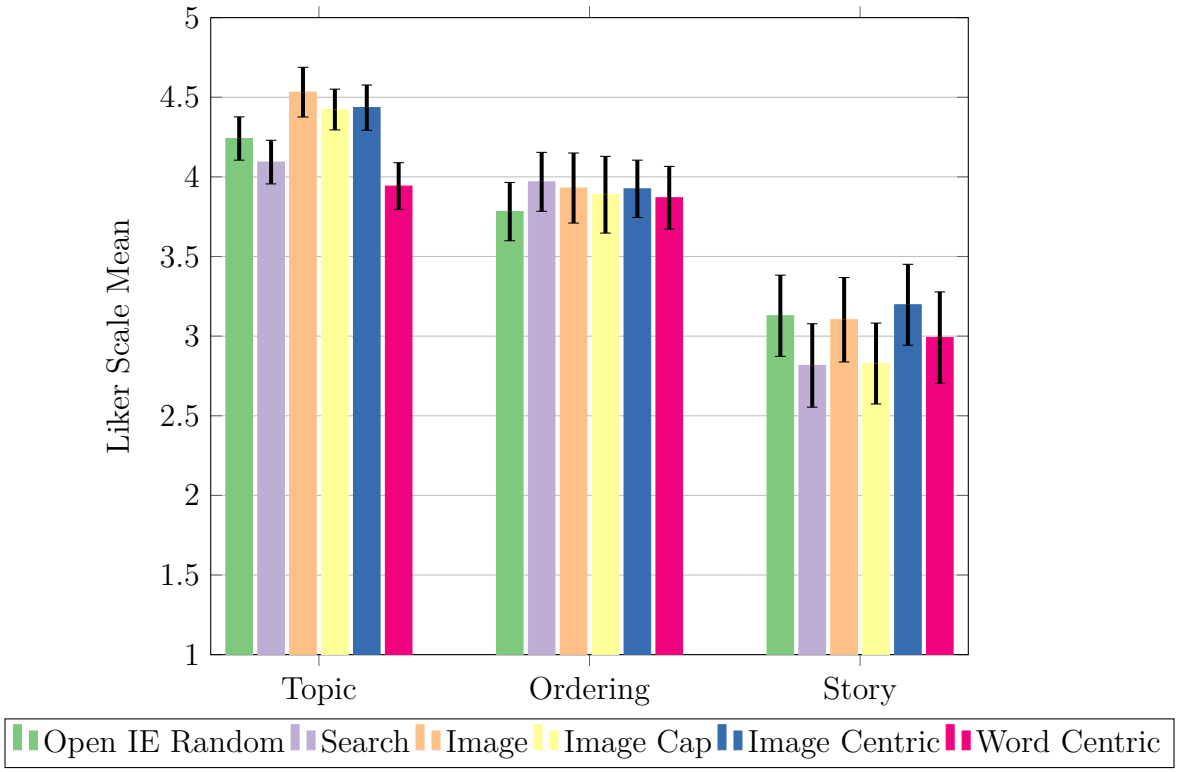


Figure 4.13: Shows averages and 95% confidence interval error bars for the Soft Search Triptychs.

There are 25 complete results used from the High frequency survey, 23 from Mid frequency, and 23 from Low frequency. These 71 results have been combined into Figures 4.13 and 4.14 showing the averages and 95% confidence interval, and sentiment in the same format as before . For Q1 attributes *Pillai's trace* is 0.611, and *Wilk's Lamdba* 0.389, at $F = 20.773$, $p < 0.0001$ indicating significance. Once again pairwise T-tests (with Bonferroni correction) is used to assess the significance of differences. The differences in mean between Open IE and the others for Q1 Topic Cohesiveness are (positive is Open IE random is better than and negative is worse) - Search 0.148, image -0.291, Image Cap -0.182, Image Centric -0.194, Word Centric 0.299. All are significant at a 95% confidence interval. The differences are small but all 3 of the image focused configurations do slightly better, and the Word based one slightly worse. So in regard to H10, as far as topic cohesiveness goes, it is not supported as a whole. Yet it does show that the word embedding based search is competitive in being able identify cohesive topics as the Open IE parser. Although the difference is small the results

also demonstrate that the image features are able to add a small amount to topic cohesiveness.

Q2 ordering is significantly worse than the earlier experiment, but this shalt not be dwelt on further as the problems have already been discussed. For Q3 (Story) the only significant difference with Open IE Random is Search which performs 0.337 worse with 95% confidence (*Pillai's trace* is 0.333, and *Wilk's Lamdba* 0.667, at $F = 20.773$, $p < 0.0001$). H10 overall is not supported as the best of the new configurations is no better than the Open IE Random.

For H11 there are significant differences for Q1 between those configurations that maximise image similarity - Image, Image Centric, and Image Cap, and for Q3 image and Image Centric over the purely semantic based search, and word centric (which has a dissimilarity negative weight for image features). In all cases these differences are small on average maximum of 0.338 and min of 0.112 on the pairwise comparison tests. Overall it supports the hypothesis with regard to image features being able to differentiate Triptychs from each other.

Figure 4.14 the sentiment is once again fairly neutral (the total scale is -100 to 100). For H12 all but one are close being slightly active in terms of sentiments. There are some tiny significant differences in the context on the scale between a couple but very small in the context of the scale. The one outlier Word Centric that is more active with a mean difference of 8.349 (with Image Cap) and 16.847 (Open IE Random) at 95% confidence. However this is still only a difference at the upper end of 8.4%, so overall these configurations are not important in altering general sentiment.

Relevant to all of these Questions is the extremely wide Standard Deviations across the board (See Figures 7.5 and 7.6 in the Appendices). This indicates generally the highly subjective nature of participants judgements on topics, ordering, story, and sentiment.

4.5.4 Further Discussion

While the word embedding based search combined image features is not significantly better than producing a result than the Open IE based selection, it is still at least good. This still represents a reasonably strong result; the Open IE

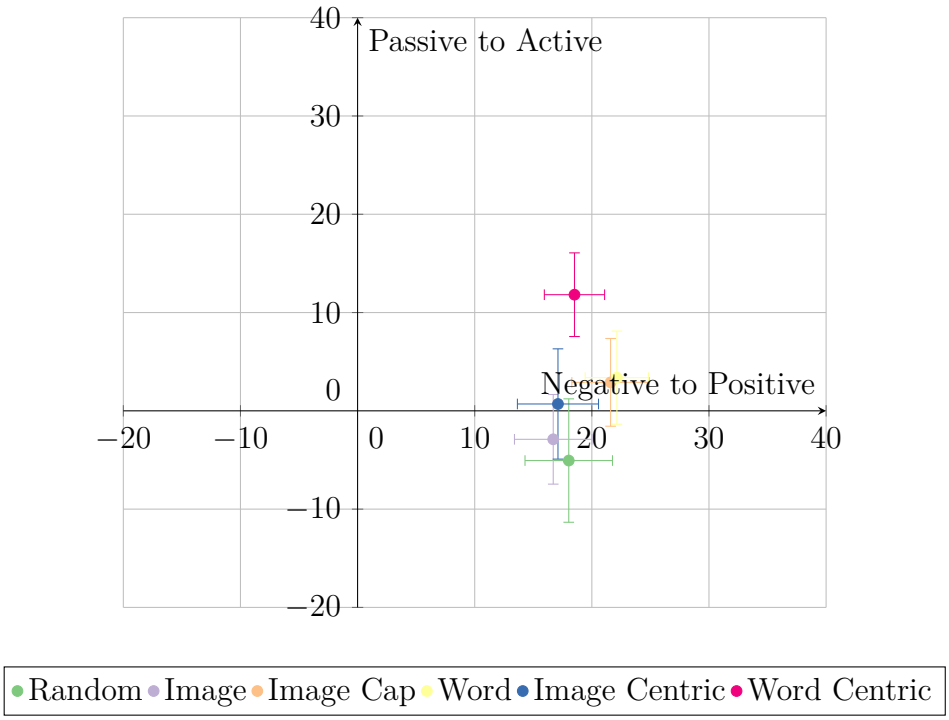


Figure 4.14: Sentiment Plot for Search Triptychs showing 95% confidence interval error bars.

Random selection is based on a sophisticated triple extractor is working on 5 separate sentence captions for each image. This is more advantageous to this approach than is likely to be encounter in another context where captions shorter such as in social media uses where exacts matches are likely to cause problems. It also demonstrates again that the word embeddings similarity are competitive be being able to identify related topics either based on image or semantic features.

4.6 Experiments Summary

There are several important results from the performed experiments. The R-CNN Image features perform strongly in representing topic cohesiveness, as do the word embeddings, both outperform the ReelLives LSA implementation. For producing stories the new features are also better but the margin is much smaller, and overall strength of the stories is significantly weaker than topic cohesiveness. Ordering is generally weak and performs no better than random. Configurations that attempted to create more successful stories using dissimilar element do no better but predominantly worse than those that maximise the similarities for

features. Sentiment broadly isn't much of a factor being fairly similar across the dataset. In search related queries the Soft search is able to match, but not outperform, the Hard filtering search.

4.7 Limitations and Further Work

4.7.1 Evaluation

There are though some potential limitations and discussion points from using MS-COCO vis-a-vis the current dataset: The existing data is used to create stories from users own images that is aimed at being relevant to them, a stock library such as MS-COCO stories will not be personal in that way. Some context is also lost such as time information which is not relevant in the same way as the images aren't taken in a sequence related to a single event. It also doesn't have the context of the thread in which the image is posted, who liked, who replied to it and other interaction data that social media would have, nor does it have precise locations (though the ReelLives system does not use latitude/longtitude locations).

The ReelLives system has been evaluated against MS-COCO, but it would also be beneficial to evaluate the reverse and rerun the new features and selector on the Instagram experiment used in the original evaluation.

4.7.2 Scale

The evaluation has demonstrated that the image features can represent a scene well on their own. There is the problem though in that they have specifically trained to identify and predict 80 objects. The question is whether this will scale when outside the context of the dataset as a general method where there many more objects of interest, and with the highly labour intensive supervised approach need to build the object position masks. There is also related interesting work of studying if other features such as object or location predictions can be used in addition as separate features to fine tune and improve overall performance.

The limitation in needing to rely on a relatively small number of participants with

surveys severely curtails the number of combinations that can be evaluated. It also leads to the possibility of potentially the best combination of parameters not being evaluated because they are screened out prematurely, and there not being the capacity (or budget) to try some comparisons at all such as Word Vector Averaging with Paragraph Vectors. SIND (Huang et al., 2016) is a step in being able to automate some of the evaluation using many of the techniques that have been standard across machine learning in trying combinations, parameter tuning, and validation (via means such cross validation). There is still the problem that extending these datasets is time consuming and expensive. There is enough public information under suitable licenses to be able to collect related images that are already sorted into collections with time information for ordering but it is story (or narrative text) that goes alongside it that is the difficulty.

4.7.3 Selection, Narrative and Ordering

While the MILP selector is a promising alternative in order to be evaluated fully it would need to be evaluated directly against a Viterbi, Greedy, and other selectors. As touched on in the background it is not clear whether attempting to find a global optimum (even when falling back to the Tabu search) is beneficial over a more point to point selector. A countervailing point to this approach is newly published research (Huang et al., 2016) finding a Greedy search outperformed Viterbi suggesting a local rather than global approach may be preferable; a direct evaluation of selection methods is needed in further work.

Ordering has been shown the a major limitation in the system; often it can be reasonable but more so my chance than design. While it's possible for the system to assess commonality it is not for the ordering of events. The SIND dataset (and successors) would seem to be needed in order to train some notion of ordering of events, and this can be approached by training over sequences (using something like the Skipping RNNs touched on in the background or improved versions thereof). The problem with SIND as a Gold Standard is that it faces a similar problem to machine translation; give 20 different people a collection of photos from a holiday or a wedding say and ask them to pick a few out and write a story to go with it, and you are likely to end up 20 different selection of images, with different ordering, and written annotations. This is why machine transla-

tion evaluation techniques such METEOR (Lavie & Denkowski, 2009) and BLEU (Papineni et al., 2002) were used in the discussed SIND papers to score across multiple alternative stories, but having to provide multiple alternative stories adds substantially the cost and time of building the dataset. As well as ordering geographical information (Geotagging) would be a productive addition to a dataset such as SIND, as particularly with personal stories progression over space (e.g. travel log) can be important, as could for example social media information - who liked or was interested in specific parts of the source collection.

The approach of trying to use dissimilarity to create stories as per the Image Centric and Word Centric configurations produces results that look different from the others, but have not achieved the aim of being judged as better stories. There needs to some level on commonality and progression to create interesting narratives in pictures; however the mechanism employed in this project has not achieved it. Creating stories needs to good representations of ordering, but also something more around common themes, entities, sentiment, and progression. Using distributed vector features (whether linguistic based or image) in isolation can be seen to have a weakness in that while it's possible to control degrees of similarity it's a crude way to control progression in a narrative. There is a clear opportunity in trying to use Deep Recurrent Neural Network with multi-modal (or grounded) representations to train networks over datasets such as SIND to jointly learn representations of different elements of visual stories, with the possibility of pre-training visual, linguistic, sentiment or other elements of much larger single media corpora. It though remains to seen how much progress can be made with this approach.

In addition there are still a wealth of other work in exploring other forms of generated outputs for visual stories in forms such as in longer sequences of images beyond the Triptych, movies, timelines, and collages.

Chapter 5

Conclusion

This dissertation has found that newer distributed semantic methods such as word embeddings can outperform LSA as far as representing topic cohesion, as can R-CNN features to represent scenes, identify common elements and composition. Importantly, purely image based features are likely to be more applicable to Social Media where captions are often short or absent. It has also demonstrated that a MILP selector is a viable alternative to Viterbi in finding solutions that optimise globally rather than locally. Sentiment has not been found to be important largely because of the descriptive nature of the captions in MS-COCO.

There are as has been discussed limitations with the approach used with regard to ordering and creating a story as opposed to closely related images, though the evaluation found a great deal of subjectivity and variation in surveys participant judgement about what constitutes one. More fine grained control is needed than is offered by Cosine similarity over Semantic or Image features to create stronger narratives, including learning some form of ordering over multi-modal representations. Deep Learning based sequence modelling approaches may be fruitful (Section 4.7) in being able to create interesting stories, but this kind of work does still seem to contain a gap. The approach stands a good chance reproducing the type of pictorial stories seen in the *Pony Express*, or a family story. However as in the introduction there is a broader type of pictorial story such as political cartoons or art which relies more on creativity, metaphor, and plot twists that is a far harder challenges. While these have been explored in other contexts (such as metaphor (Shutova, 2010)) they remain far more distant problems to tackle.

Chapter 6

Appendix A: Feature Format

```
<rlunit id="111032" src="mscoco" time="2013-11-14T21:26:18">
  <tag>DT</tag>
  <tag>NN</tag>
  <tag>IN</tag>
  <tag>NN</tag>
  <tag>VBG</tag>
  <tag>IN</tag>
  <tag>NN</tag>
  <tag>IN</tag>
  <tag>DT</tag>
  <tag>NN</tag>
  <tag>JJ</tag>
  <tag>TO</tag>
  <tag>DT</tag>
  <tag>NN</tag>
  <tag>IN</tag>
  <tag>NN</tag>
  <tag>.</tag>
  <token>A</token>
  <token>bottle</token>
  <token>of</token>
  <token>wine</token>
  <token>sitting</token>
  <token>on</token>
  <token>top</token>
  <token>of</token>
  <token>a</token>
  <token>table</token>
  <token>next</token>
  <token>to</token>
  <token>a</token>
  <token>glass</token>
  <token>of</token>
  <token>wine</token>
  <token>.</token>
  <image>http://farm3.staticflickr.com/2067/2191335794_5997bd1236_z.jpg</image>
  <sentiment>
    <value>negative</value>
  </sentiment>
  <theme confidence="0.992971">
    <value>food and drink</value>
    <value>beverages</value>
    <value>alcoholic beverages</value>
    <value>wine</value>
  </theme>
  <theme confidence="0.0330866">
    <value>technology and computing</value>
    <value>consumer electronics</value>
    <value>home video and dvd</value>
  </theme>
  <theme confidence="0.0281593">
    <value>food and drink</value>
    <value>beverages</value>
    <value>non alcoholic beverages</value>
    <value>bottled water</value>
  </theme>
  <relationtriple>
    <subject>bottle</subject>
    <relation>sitting</relation>
    <relation>next</relation>
    <object>glass</object>
  </relationtriple>
  <relationtriple>
    <subject>bottle</subject>
    <relation>sitting</relation>
    <relation>next</relation>
    <object>glass</object>
    <object>wine</object>
  </relationtriple>
  <relationtriple>
    <subject>bottle</subject>
    <relation>sitting</relation>
    <relation>on</relation>
    <object>table</object>
  </relationtriple>
  <messagetext>A bottle of wine sitting on top of a table next to a glass of wine.</messagetext>
</rlunit>
```

Figure 6.1: An example of a single images features.

Chapter 7

Appendix B: Additional Experiment Charts

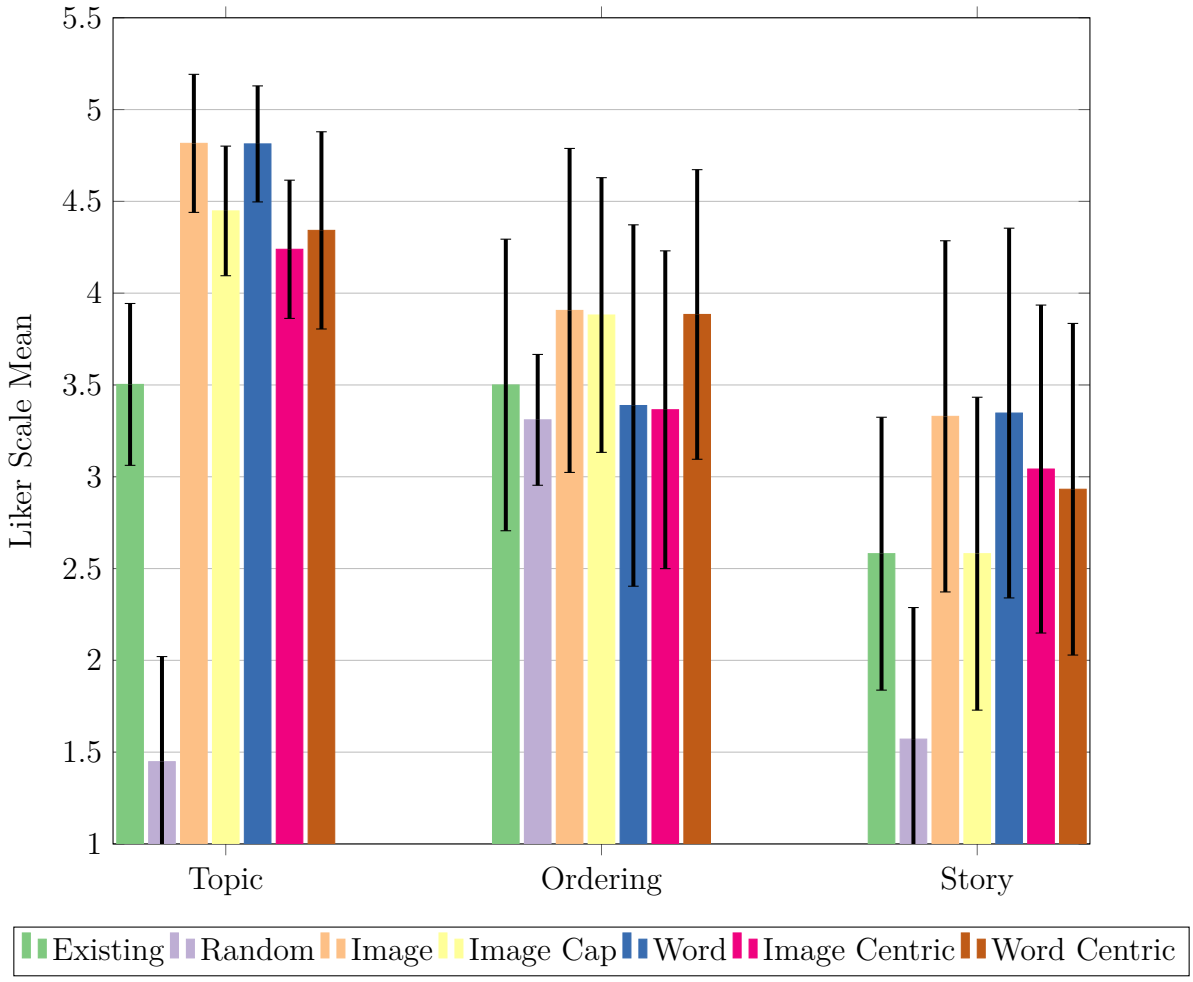


Figure 7.1: Shows averages and Standard Deviation error bars for the Best Triptychs Q1-Q3 (Topic, Ordering, Story). With Q1 and Q3 higher is better, for Q2 lower is.

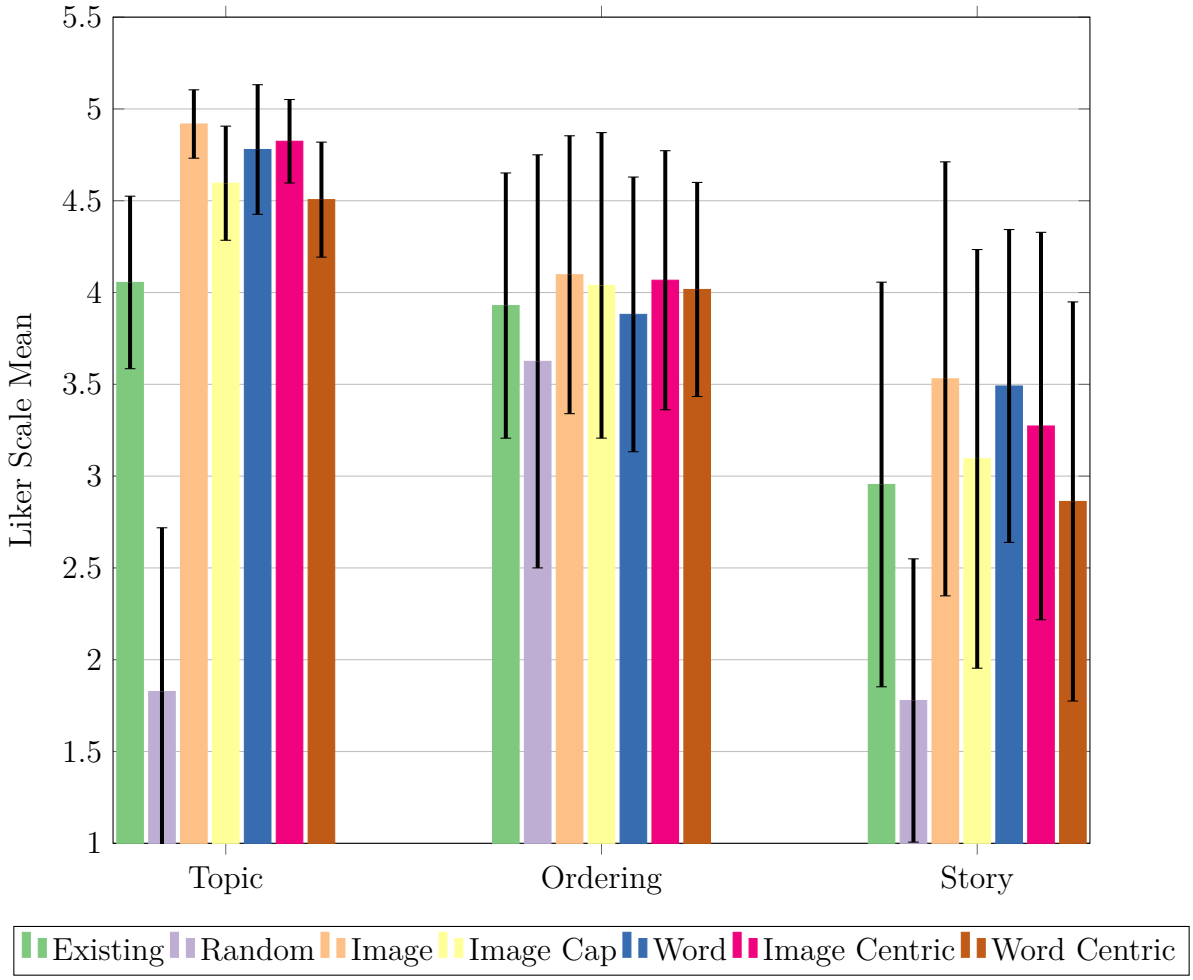


Figure 7.2: Shows averages and Standard Deviation error bars for the Sentiment Triptychs Q1-Q3 (Topic, Ordering, Story). With Q1 and Q3 higher is better, for Q2 lower is.

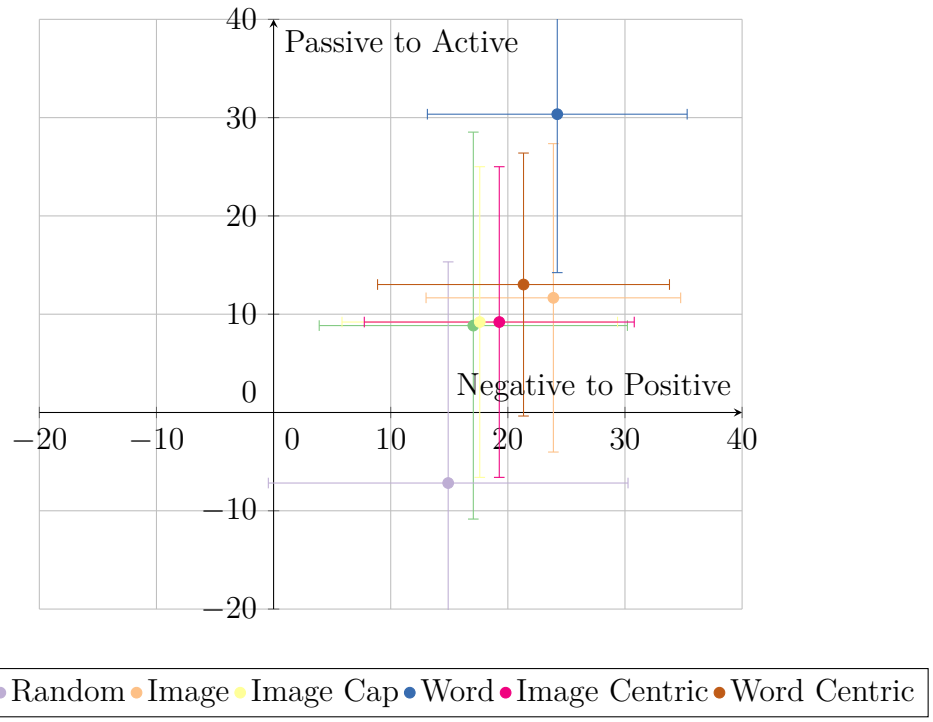


Figure 7.3: Sentiment Plot for Q4 Best Triptychs showing Standard Deviation error bars.

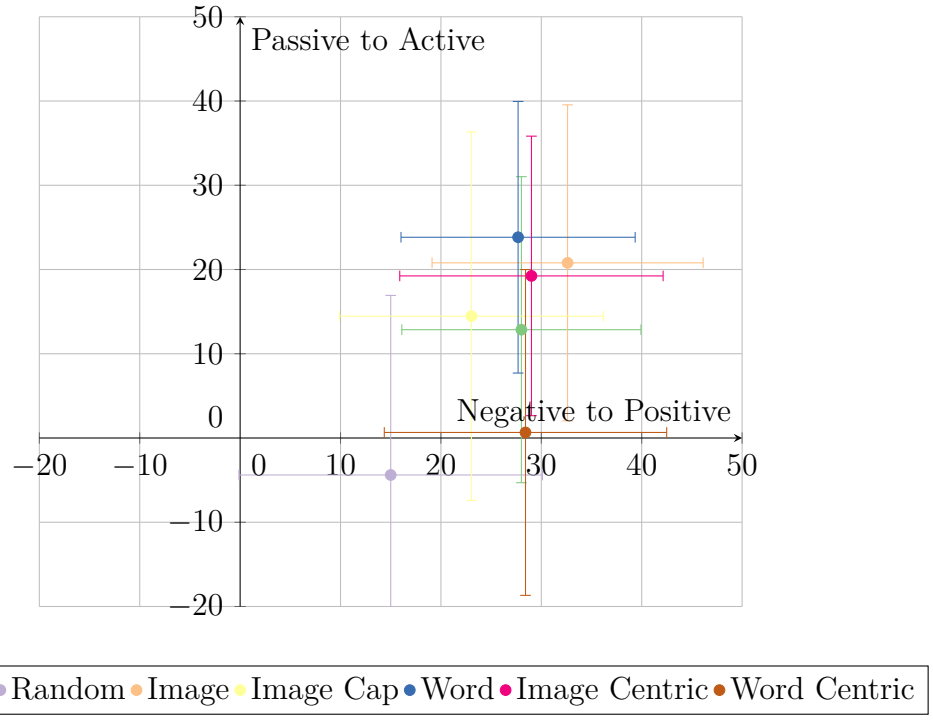


Figure 7.4: Sentiment Plot for Q4 Sentiment Triptychs showing Standard Deviation error bars.

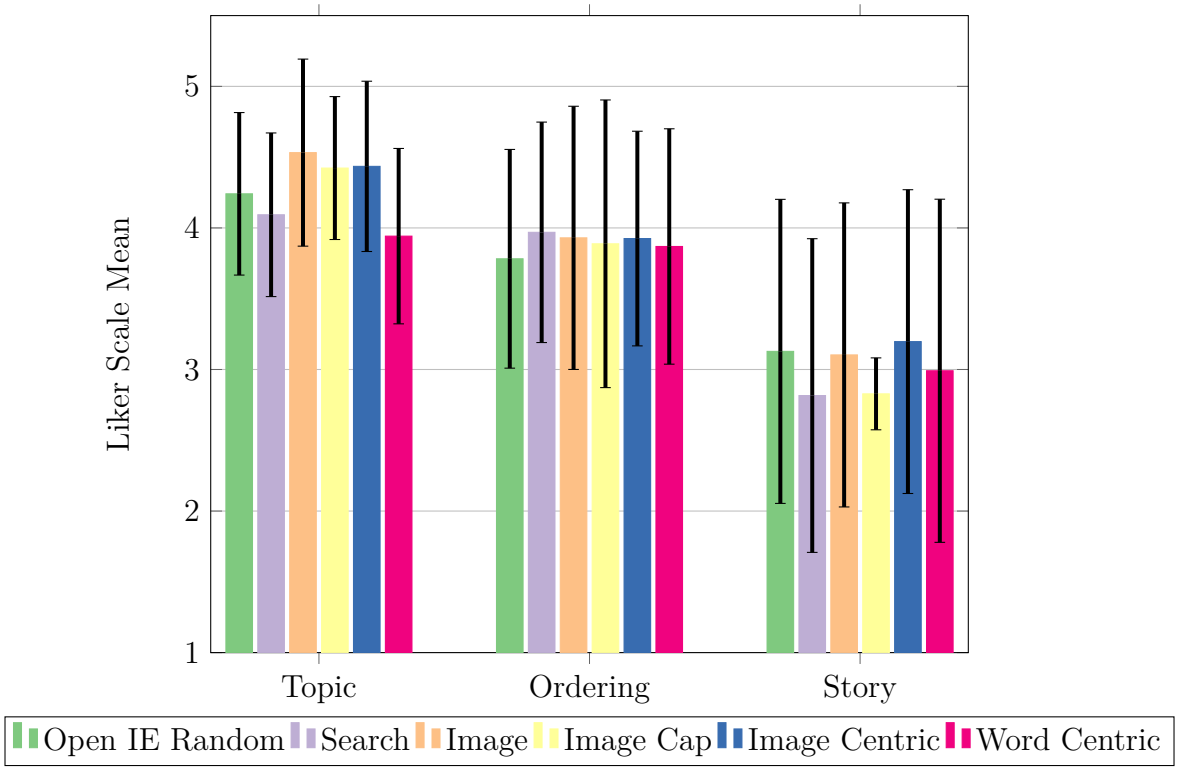


Figure 7.5: Shows averages and Standard Deviation error bars for the Soft Search Triptychs.

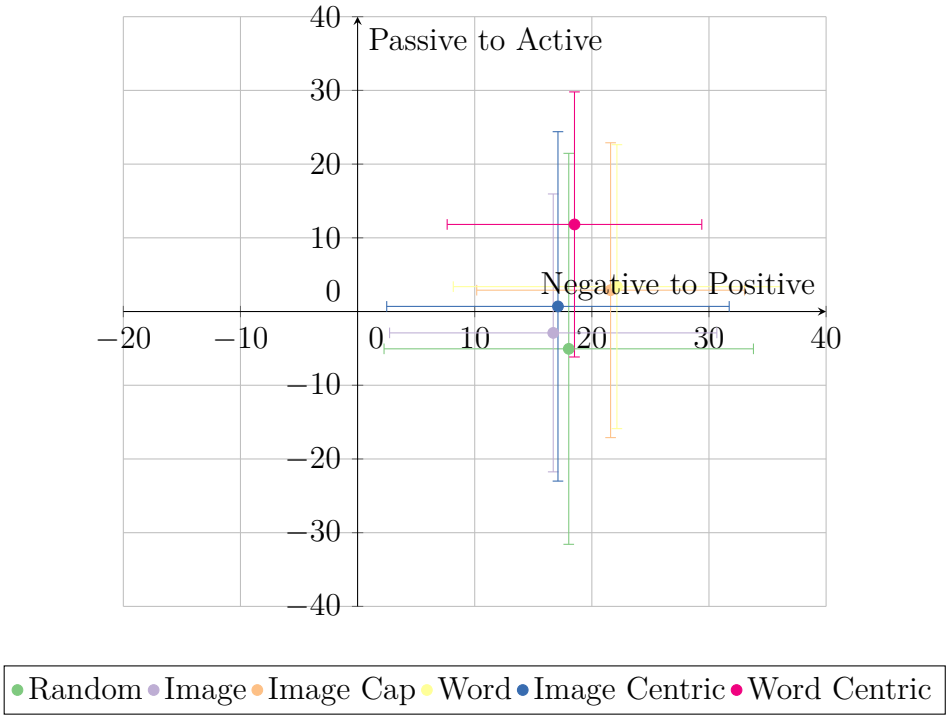


Figure 7.6: Sentiment Plot for Search Triptychs showing Standard Deviation error bars.

References

Agrawal, H., Chandrasekaran, A., Batra, D., Parikh, D., & Bansal, M. (2016). Sort story: Sorting jumbled images and captions into stories. *CoRR*, *abs/1606.07493*. Retrieved from http://arxiv.org/abs/1606.07493

Angeli, G., Premkumar, M. J., & Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing of the asian federation of natural language processing, acl* (pp. 26–31).

Arbelaez, P., Maire, M., Fowlkes, C., & Malik, J. (2011). Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *33*(5), 898–916.

Aylett, M. P., Farrow, E., Pschetz, L., & Dickinson, T. (2015). Generating narratives from personal digital data: Triptychs. In *Proceedings of the 33rd annual acm conference extended abstracts on human factors in computing systems* (pp. 1875–1880).

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Acl (1)* (pp. 238–247).

Bell, S., Zitnick, C. L., Bala, K., & Girshick, R. B. (2015). Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. *CoRR*, *abs/1512.04143*. Retrieved from http://arxiv.org/abs/1512.04143

Blacoe, W., & Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 546–556).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003, March). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022. Retrieved from http://dl.acm.org/citation.cfm?id=944919.944937

Brilliant, R. (1984). *Visual narratives: Storytelling in etruscan and roman art*. Cornell Univ Pr.

Chandru, V., & Rao, M. R. (2010). Algorithms and theory of computation handbook. In M. J. Atallah & M. Blanton (Eds.), (pp. 30–30). Chapman & Hall/CRC. Retrieved from http://dl.acm.org/citation.cfm?id=1882757.1882787

Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *CoRR*, *abs/1504.00325*. Retrieved from http://arxiv.org/abs/1504.00325

Clausen, J. (1999). Branch and bound algorithms-principles and examples. *Department of Computer Science, University of Copenhagen*, 1–30.

Cook, J. (2012). *Ice age art: The arrival of the modern mind*. British Museum.

Cornuéjols, G. (2007). Revival of the gomory cuts in the 1990s. *Annals of Operations Research*, *149*(1), 63–66.

Dai, A. M., Olah, C., & Le, Q. V. (2015). Document embedding with paragraph vectors. *CoRR*, *abs/1507.07998*. Retrieved from http://arxiv.org/abs/1507.07998

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Cvpr09*.

Ericsson, F., XL Axiata. (2014). *Measuring and improving network performance: An analysis of network and application research, testing and optimization.* Retrieved 2016-03-29, from http://scontent-lga3-1.xx.fbcdn.net/hphotos-xpa1/t39.2365-6/12057133_958179554220316_1236052925_n.pdf

Fader, A., Soderland, S., & Etzioni, O. (2011, July 27-31). Identifying relations for open information extraction. In *Proceedings of the conference of empirical methods in natural language processing (EMNLP '11)*. Edinburgh, Scotland, UK.

Farrow, E., Dickinson, T., & Aylett, M. P. (2015). Generating narratives from personal digital data: Using sentiment, themes, and named entities to construct stories. In *Human-computer interaction–interact 2015* (pp. 473–477).

Springer.

Feng, Y., & Lapata, M. (2013). Automatic caption generation for news images. *IEEE transactions on pattern analysis and machine intelligence*, *35*(4), 797–812.

Finkel, J. R., & Manning, C. D. (2008). Enforcing transitivity in coreference resolution. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers* (pp. 45–48).

Forney, G. D. (1973, March). The viterbi algorithm. *Proceedings of the IEEE*, *61*(3), 268-278. doi: 10.1109/PROC.1973.9030

Gillick, D., & Favre, B. (2009). A scalable global model for summarization. In *Proceedings of the workshop on integer linear programming for natural language processing* (pp. 10–18).

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., ... Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: short papers-volume 2* (pp. 42–47).

Girshick, R. (2015a). Fast r-cnn. In *International conference on computer vision (ICCV)*.

Girshick, R. (2015b). *Training r-cnns of various velocities: Slow, fast, and faster.* Retrieved 2016-07-25, from http://mp7.watson.ibm.com/ICCV2015/slides/iccv15_tutorial_training_rbg.pdf

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014a). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 580–587).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014b). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer vision and pattern recognition*.

Glover, F. (1989). Tabu search-part i. *ORSA Journal on computing*, *1*(3), 190–206.

Glover, F. (1990). Tabu search part ii. *ORSA Journal on computing*, *2*(1), 4–32.

GNU. (2016). *Gnu linear programming kit, 4.60.0.* Retrieved 2016-07-20, from http://www.gnu.org/software/glpk/glpk.html

Goldberg, A. B., Rosin, J., Zhu, X., & Dyer, C. R. (2009). Toward text-to-picture synthesis. In *Nips 2009 mini-symposia on assistive machine learning for people with disabilities.*

Goldberg, Y., & Levy, O. (2014). word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Gottschall, J. (2012). *The storytelling animal: How stories make us human.* Houghton Mifflin Harcourt.

Gradle. (2016). *Gradle.* Retrieved 2016-07-19, from https://gradle.org/

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, *abs/1512.03385*. Retrieved from http://arxiv.org/abs/1512.03385

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Huang, T., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., ... Mitchell, M. (2016). Visual storytelling. *CoRR*, *abs/1604.03968*. Retrieved from http://arxiv.org/abs/1604.03968

IBM. (2016a). *Alchemy api.* Retrieved 2016-07-19, from http://www.alchemyapi.com/api

IBM. (2016b). *Alchemy api taxonomy.* Retrieved 2016-07-25, from http://www.ibm.com/watson/developercloud/doc/alchemylanguage/download/Taxonomy-Classifier-IAB++.pdf

Jackson. (2016). *Jackson.* Retrieved 2016-07-19, from https://github.com/FasterXML/jackson

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the acm international conference on multimedia* (pp. 675–678).

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3128–3137).

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Emnlp* (pp. 36–45).

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., &

Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems* (pp. 3294–3302).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., & Choi, Y. (2012). Collective generation of natural image descriptions. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1* (pp. 359–368).

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Lavie, A., & Denkowski, M. J. (2009). The meteor metric for automatic evaluation of machine translation. *Machine translation*, *23*(2-3), 105–115.

Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. In *NAACL HLT 2015, the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies, denver, colorado, usa, may 31 - june 5, 2015* (pp. 153–163). Retrieved from http://aclweb.org/anthology/N/N15/N15-1016.pdf

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Icml* (Vol. 14, pp. 1188–1196).

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* (pp. 2177–2185).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision (eccv)*. ZÃijrich. Retrieved from /se3/wp-content/uploads/2014/09/coco_eccv.pdf,http://mscoco.org

Liu, Y., Fu, J., Mei, T., & Chen, C. W. (2016). Storytelling of photo stream with bidirectional multi-thread recurrent neural network. *CoRR*, *abs/1606.00625*. Retrieved from http://arxiv.org/abs/1606.00625

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. the proceedings of the seventh ieee international conference on* (Vol. 2, pp. 1150–1157).

Ma, L., Lu, Z., & Li, H. (2015). Learning to answer questions from image using convolutional neural network. *CoRR, abs/1506.00333*. Retrieved from http://arxiv.org/abs/1506.00333

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for computational linguistics (acl) system demonstrations* (pp. 55–60). Retrieved from http://www.aclweb.org/anthology/P/P14/P14-5010

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR, abs/1301.3781*. Retrieved from http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Miller, G. A. (1995). Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM, 38*, 39–41.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*.

Newman, M. E. J. (2005). Power laws, pareto distributions and zipfs law. *Contemporary Physics*.

Niederwieser, P. (2016). *Spock*. Retrieved 2016-07-19, from http://spockframework.github.io/spock/docs/1.1-rc-1/index.html

Oracle. (2016). *Java version 8*. Retrieved 2016-07-19, from http://www.oracle.com/technetwork/java/index.html

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters..

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical methods in natural language processing (emnlp)* (pp. 1532–1543). Retrieved from http://www.aclweb.org/anthology/D14-1162

Pivotal. (2016). *Spring shell*. Retrieved 2016-07-19, from http://projects

.spring.io/spring-shell/

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *CoRR, abs/1606.07772*. Retrieved from http://arxiv.org/abs/1606.07772

Redmon, J., Divvala, S. K., Girshick, R. B., & Farhadi, A. (2015). You only look once: Unified, real-time object detection. *CoRR, abs/1506.02640*. Retrieved from http://arxiv.org/abs/1506.02640

Řehůřek, R., & Sojka, P. (2010, May 22). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA. (http://is.muni.cz/publication/884893/en)

Ren, M., Kiros, R., & Zemel, R. S. (2015). Image question answering: A visual semantic embedding model and a new dataset. *CoRR, abs/1505.02074*. Retrieved from http://arxiv.org/abs/1505.02074

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems (NIPS)*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211-252. doi: 10.1007/s11263-015-0816-y

Russell, J. A. (1980). circumplex model of affect. *Journal of personality and soc. psych.*, *A39*(6), 1161.

Schmitz, M., Bart, R., Soderland, S., Etzioni, O., et al. (2012). Open language learning for information extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 523–534).

Shutova, E. (2010). Models of metaphor in nlp. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 688–697).

Sigurdsson, G. A., Chen, X., & Gupta, A. (2016). Learning visual storylines with skipping recurrent neural networks. *CoRR, abs/1604.04279*. Retrieved from http://arxiv.org/abs/1604.04279

Silberer, C., & Lapata, M. (2014, June). Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp.

721–732). Baltimore, Maryland: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P14-1068

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR, abs/1409.1556*. Retrieved from http://arxiv.org/abs/1409.1556

Skymind. (2016a). *Deeplearning4j.* Retrieved 2016-07-20, from http://deeplearning4j.org/

Skymind. (2016b). *Nd4j.* Retrieved 2016-07-20, from http://nd4j.org/

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (emnlp)* (Vol. 1631, p. 1642).

Stanton, A. (2012). *The clues to a great story.* Retrieved from https://www.youtube.com/watch?v=KxDwieKpawg

Steiner, W. (2004). Pictorial narrativity. *Narrative across media: The languages of storytelling*, 14577.

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. *CoRR, abs/1503.00075*. Retrieved from http://arxiv.org/abs/1503.00075

Thelwall, M. (2013). Heart and soul: Sentiment strength detection in the social web with sentistrength. *Proceedings of the CyberEmotions*, 1–14.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology-volume 1* (pp. 173–180).

Turian, J. (2013). *Using alchemyapi for enterprise-grade text analysis* (Tech. Rep.). Technical report, AlchemyAPI (August 2013).

Uijlings, J. R., van de Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. *International journal of computer vision*, *104*(2), 154–171.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3156–3164).

Vonnegut, K. (2010). *The shape of stories.* Retrieved from https://www.youtube

.com/watch?v=oP3c1h8v2ZQ

Woodsend, K., & Lapata, M. (2012). Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 233–243).

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (cvpr), 2010 ieee conference on* (pp. 3485–3492).

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., ... Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd international conference on machine learning, ICML 2015, lille, france, 6-11 july 2015* (pp. 2048–2057). Retrieved from http://jmlr.org/proceedings/papers/v37/xuc15.html

Zagoruyko, S., Lerer, A., Lin, T., Pinheiro, P. H. O., Gross, S., Chintala, S., & Dollár, P. (2016). A multipath network for object detection. *CoRR*, *abs/1604.02135*. Retrieved from http://arxiv.org/abs/1604.02135

Zhu, X., Goldberg, A. B., Eldawy, M., Dyer, C. R., & Strock, B. (2007). A text-to-picture synthesis system for augmenting communication. In *Aaai* (Vol. 7, pp. 1590–1595).