Neural Caption Generation for News Images

Vishwash Batra, Yulan He, George Vogiatzis
School of Engineering and Applied Science, Aston University
{batrav1, y.he9, g.vogiatzis}@aston.ac.uk

Abstract

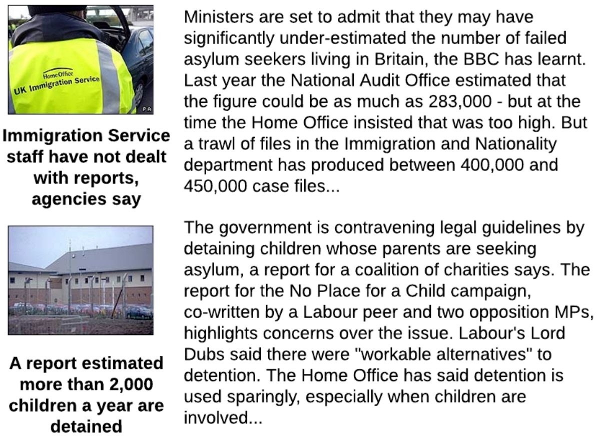
Automatic caption generation of images has gained significant interest. It gives rise to a lot of interesting image-related applications. For example, it could help in image/video retrieval and management of vast amount of multimedia data available on the Internet. It could also help in development of tools that can aid visually impaired individuals in accessing multimedia content. In this paper, we particularly focus on news images and propose a methodology for automatically generating captions for news paper articles consisting of a text paragraph and an image. We propose several deep neural network architectures built upon Recurrent Neural Networks. Results on a BBC News dataset show that our proposed approach outperforms a traditional method based on Latent Dirichlet Allocation using both automatic evaluation based on BLEU scores and human evaluation.

Keywords: Recurrent Neural Networks, Image caption generation, Deep learning, Order Embedding.

1. Introduction

There is rich information available on the Internet. Many on-line news sites like CNN, Yahoo, BBC etc. publish images with their stories and even provide photo feeds related to current events. These news sites are a good resource for multimedia files containing information in the form of videos, images and natural language texts. Presence of this vast amount of multimedia data has provided strong impetus to develop machine-learning based applications that jointly model data from different modalities. For example, Ngiam et al. (2011) develops a speech recognition system where they jointly model audio and visual modality. They focus on learning representations of audio data which are coupled with the videos of the lips. Another such application is image caption or visual description generation, which aims to generate text descriptions for an image, often times capturing all the different objects depicted and their spatial relationships.

News image caption generation, however, is different from the typical image captioning task. The input to news image caption generation is both a news article and its accompanying image, as opposed to the traditional image captioning task where the input is only an image. Hence, rather than enumerating objects in a given image and describing their properties or relationships to each other as in the traditional image captioning task, the output of news image caption generation is informative text not only describing the key semantics conveyed in the given image, but also summarising the content of its relating news article (Berg et al., 2004). An example is shown in Figure 1. It can be seen that the captions of news images provide more information than what have been depicted in images only. For example, a reasonable caption for the second image would be "A building". But its actual caption conveys much more information and it is evident that the text content of news articles would also need to be considered when generating good captions for news images.



Immigration Service staff have not dealt with reports, agencies say

Ministers are set to admit that they may have significantly under-estimated the number of failed asylum seekers living in Britain, the BBC has learnt. Last year the National Audit Office estimated that the figure could be as much as 283,000 - but at the time the Home Office insisted that was too high. But a trawl of files in the Immigration and Nationality department has produced between 400,000 and 450,000 case files...

A report estimated more than 2,000 children a year are detained

The government is contravening legal guidelines by detaining children whose parents are seeking asylum, a report for a coalition of charities says. The report for the No Place for a Child campaign, co-written by a Labour peer and two opposition MPs, highlights concerns over the issue. Labour's Lord Dubs said there were "workable alternatives" to detention. The Home Office has said detention is used sparingly, especially when children are involved...

Figure 1: BBC News Corpus shows sample news articles containing text, image and caption in the bold.

News caption generation tools can assist journalists in creating descriptions for the images associated with their articles or in finding images that appropriately illustrate their text. It also helps in increased accessibility of web for visually impaired individuals (blind or people with partially impaired vision) users who cannot access the content of many sites in the same way sighted users can (Ferres et al., 2006). A wide variety of techniques exist for caption generation ranging from semantic space learning (Karpathy and Fei-Fei, 2017), where both supervised and unsupervised methods exist to learn associations between features extracted from image and words, to latent variable models (Feng and Lapata, 2013). There are models inspired by information retrieval and instantiations of noisy-channel model (Lavrenko et al., 2004). Semantic space learning models learn parameters to map an image to a caption, whereas latent variable models are probabilistic in nature. Recently, there has been a surge of interests in neural caption generation methods due to ground-breaking results produced by deep learning. Mainly, they all have a fundamental architecture in common which is inspired by encoder-decoder architecture from neural machine translation. (You et al., 2016) (Karpathy et al., 2014) (Chen and Zitnick, 2015)

In the encoder-decoder models, caption generation is seen as a translation problem where image is translated to a natural language. Convolutional Neural Networks (CNNs) are typically used as an image encoder, whereas Recurrent Neural Networks (RNNs) are used for decoding sentences, because of their sequence modeling capability. Although there are other variants proposed, for example, with attention mechanisms included, the encoder-decoder architecture is at the heart of these methodologies (Xu et al., 2015).

Existing work to news image captioning generation is scarce. An early approach tackled the problem with a two-stage process, content selection and surface realization. The first stage consists of an image annotation model, where a given image is tagged with a set of keywords based on topics learned from both news article texts and images using a variant of Latent Dirichlet Allocation (LDA) (Blei, 2004). The second stage uses extractive and abstractive summarisation techniques in forming a sentence from these set of keywords. Word-based models are highly specific in nature and may results in ambiguous results. There is need of sentential integration with the images, as a sentence describes an image without any ambiguity.

In this paper, we propose a sequence-to-sequence deep learning model to address the news image caption generation problem. Specifically, we first encode each sentence of a given news article using an order-embedding vector and extract semantic features from the accompanying image using a pre-trained CNN Network, which are further projected to same semantic space, such that both text and image vectors reside in a common semantic space (Vendrov et al., 2015). We then feed the sentence vectors together with the image vector to a Long Short-Term Memory (LSTM) network (Sak et al., 2014) to generate a vector representation of the image caption. Finally, we use the generated vector to retrieve the most similar sentence from the original news article based on cosine similarity measurement as the caption of the given image. We also explore a number of variants of our proposed architecture and compare them with the previous work on the news image captioning task.

Our experimental results on the BBC News Corpus show that our proposed strategies outperform traditional methods according to automatic evaluation metrics like BLEU scores (Papineni et al., 2002) and are comparable in terms of Meteor Scores (Lavie and Agarwal, 2007). Since automatic evaluation metrics are currently limited by their capability to measure the quality of caption generation models, a human evaluation experiment has also been conducted, where users were shown the news articles from our test dataset.

Our evaluation results show that captions generated by our proposed approach were more favoured than captions generated by an existing model based on LDA. In what follows, we first discuss related work and then describe our proposed methodology, followed by experiments and results, and finally conclude the paper and outline future research directions.

2. Related Work

Our work is related to two lines of research, image captioning and encoder-decoder architecture.

2.1. Image Captioning

The most fundamental problem that connects computer vision and natural language processing is of automatic caption generation of an image. For a long time, there has been significant work in image classification, object detection and image annotation, but a relatively little focus on generating sentential descriptions. So, some of the obvious solutions consist of using the results of these methods, that annotates an image with a set of keywords. These keywords are fed to another stage, that arranges these keywords in the form of a sentence. All of these methods fall under two-stage architecture methods.

Two stages are content selection and surface realization. The former stage, content selection consists of an image annotation model that analyses the content of an image and identifies "what to say" of the image. The latter stage, surface realization consists of a language model, that analyses the keywords and identifies "how to say" of the image.

Image Annotation Methods. Much work within computer vision has focused on image annotation, a task which is very much related but distinct from image description generation. The goal in image annotation is to label an image with keywords relating to its content. Image annotation methods can be classified as *supervised* and *unsupervised*. Supervised image annotation is similar to image classification, as the keywords (or categories) are fixed and pre-defined at training time. The fixed set of categories are identified usually in the form of classes of vocabulary words. Machine learning algorithms are applied to learn a one-to-one correspondence between an image and these categories. The core notion behind is to learn a mapping between visual feature vectors and semantics of the image. A detailed review of supervised methods for image annotation can be found in F Tsai and Hung (2008). Unsupervised image annotation methods do not have a fixed set of pre-defined classes. Instead, algorithms attempt to learn the connections between visual features and words and automatically cluster them into classes of words, which will finally denote the semantics of the image. Typical solutions to this involve introducing latent variables such as LDA models (Pan et al., 2004). Standard latent semantic analysis (LSA) and it's probabilistic variant (PLSA) have been applied to this task (Pan et al., 2004). Barnard et al. (2003) provide a more sophisticated model, they estimate the joint-distribution of words and regional image features while treating annotation as a problem of statistical inference in a graphical model. The final output is clusters of words, which appropriately describe the content of the image.

Surface Realization. The output of the previous stage is a set of keywords that appropriately describe the content of the image. The aim of this stage is to go from keywords to a sentence. Two methods are generally popular for this approach, *extractive methods* and *abstractive methods*. The main idea behind extractive methods is to retrieve most relevant sentences from a document database given the keywords identified in image annotation. Various metrics could be used to calculate the relevance of a sentence with a set of keywords, for example, word-overlap based sentence selection score, vector-space based sentence selection score or topic-based sentence selection score. Jones (1998) provided a comprehensive overview on sentence-selection algorithms. Although extractive methods help in coming up with grammatically correct sentences and require relatively less linguistic analysis, there are few caveats to consider. Sometimes, there is no single sentence in the document database that best describes the image and is one big limitation for such methods. Abstractive methods try to compose a sentence from the words based on language models. Examples of language models are probabilistic generative models or neural-language based models.

In Farhadi et al. (2010), images are parsed into $< object, action, scene >$ triplets. A more complex graph of detections beyond triplets is used by Kulkarni et al. (2013). State-of-the-art object recognition and language generation techniques are used in their model Babytalk. Feng and Lapata (2013) provide a news article caption generator. They use an LDA-based model for image annotation and use wide variety of surface realization techniques.

All of these two-stage architecture methods have some serious limitations. As mentioned before, a list of keywords is often ambiguous. A set of keywords "blue, sky, car, green" could depict "a blue sky and a green car" or "a blue car and a green sky". Therefore, the models should be designed such that there is strong correlation between phrases and images or sentences and images that are semantically relevant. In other words, a direct leap is taken from image to sentence and vice versa. Moreover, these approaches are heavily hand-designed and rigid when it comes to text generation. So, their applicability becomes limited and they cannot be generalized for new domains (Hofmann, 2001).

In recent years, some deep learning approaches like neural networks are used to co-embed images and sentences in the same vector space also called semantic space (Socher et al.,). Karpathy et al. (2014) co-embed image-crops and subsentences into semantic space. But even such approaches cannot solve the problem of limited applicability. These cannot describe unseen compositions of features even though individual features might have been observed in the training data.

Recently, an encoder-decoder architecture inspired from machine translation has been applied to image captioning and has achieved state-of-the-art performance. In the next subsection, we describe this architecture.

2.2. Encoder-Decoder Architecture

In neural machine translation, an encoder is used to read a sentence in the source language and transforms it into a rich fixed-length embedding vector representation. This embedding vector is in turn fed to a decoder that generates the sentence in the target language. This encoder-decoder architecture has been adopted in image captioning and a class of methods called "neural image captioning" methods have been developed. The idea here is to view captioning as a translation problem, where image is a source language and caption is a target language. Typically a CNN is used as an encoder for the image and RNN as a decoder.

Over the last few years, it has been convincingly shown the CNNs can produce rich representation of the input image by embedding it to a fixed-length vector, such that this representation can be used by a variety of vision tasks (Sermanet et al., 2013). Therefore, it is natural to use CNN as image encoder,

In specific, CNN is first pre-trained for classification. This network is subsequently used as an off-the-shelf feature extractor, where the last hidden layer of the network is used as a feature vector. This hidden representation is fed to the decoder to generate descriptions for the image. Vinyals et al. (2014) proposed a model with a similar architecture. Karpathy and Fei-Fei (2017) developed a deep neural network that infers the latent alignment between segments of sentences and region of image they describe. They use CNN for encoder and a bi-directional RNN over sentences as decoder.

Rather than compressing an entire image into a static representation, attention mechanisms have been introduced which allow salient features to dynamically come to forefront as needed. Using representations from top layer of a CNN that distill information in an image down to the most salient objects is one effective solution. But it has a potential drawback of losing information present in the lower layers which could be useful for generating richer and more descriptive captions. You et al. (2016) propose a soft and hard attention mechanism for image captioning tasks. They use a CNN to encode the images and a RNN with attention mechanism to generate a description. By visualizing attention weights, they switch what the model is looking at while generating a word. You et al. (2016) propose a CNN with an attention mechanism that weights the image features and RNN to generate captions to describe weighted image features.

3. Methodology

Our problem is formulated as follows: given a news image I, and its associated article D, create a sentence description S that best describes the image given D. The training data thus consists of document-image-caption tuples like the ones shown in Figure 1. During testing, we are given a document and an associated image for which we need to generate a caption.

In this section, we propose a novel deep Neural Network (NN) architecture to automatic caption generation of news images. Figure 2 provides a block diagram of the model architecture. We first convert sentences in a news article into a sequence of vectors using a pre-trained order-embedding model (Vendrov et al., 2015). We then encode the accompanying image into an image embedding using the pre-trained Oxford VGGNet (Simonyan and Zisserman, 2014) as an off-the-shelf feature extractor. The VGG Features are further projected to the same order-embedding space. Both sentence and image vectors are represented in a 1,024 dimensional semantic space. The sentence vector sequence is then fed to a LSTM network, which is a specific type of Recurrent Neural Network (RNN). The output of the the network is fed into another LSTM cell which also takes the image vector as an additional input. The final output is considered as a representation which captures the semantics conveyed in both text and image. The cross entropy between the output vector and caption order-embedding vector is used as an objective function to train the LSTM pa-

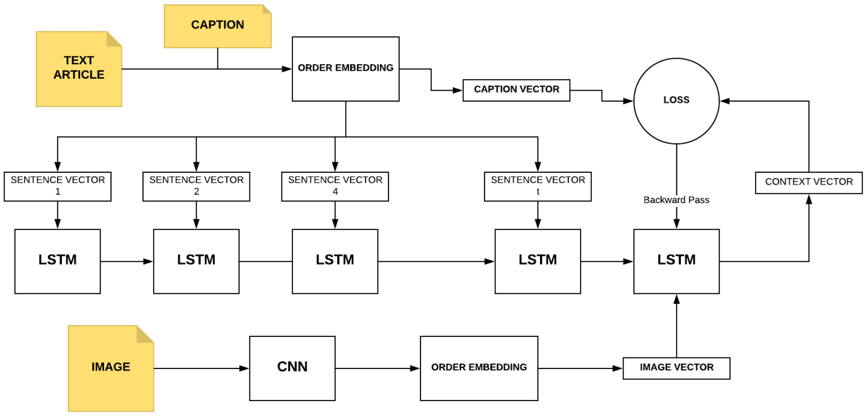


Figure 2: Our proposed deep Neural Network (NN) architecture for news image caption generation.

rameters.

3.1. Text and Image Representation

For encoding sentences, we use a pre-trained order-embedding model (Vendrov et al., 2015) to encode sentences using distributed representations. Order-embeddings exploit the partial order structure of the visual-semantic hierarchy by learning a mapping between sentences and semantic vector space. This projects each sentence into a 1024-dimensional embedding space.

For encoding images, we first use a pre-trained Convolutional Neural Network (CNN), which is an important class of learnable representations applicable, among others, to numerous computer vision problems. Deep CNNs, in particular, are composed of several layers of processing, each involving linear as well as non-linear operators. We use pre-trained Oxford VGGNet as an off-the-shelf feature extractor. The whole network consists of 22 layers. We use the fc7 features, that is the output of the penultimate fully-connected layer, as a representation for the image. The VGG features are projected to same order-embedding space, where sentence vectors reside. As such, both image and sentence vectors reside in a common semantic space which enables direct comparison between them.

3.2. LSTM Training

RNNs surely do a great job at modelling sequences. Unfortunately, the shortcoming of such networks is that they are unable to carry forward information when the length of the chain grows beyond a measure. This is called vanishing gradient effect. To solve this problem, a forgetting mechanism has been proposed in LSTM. LSTMs have many variations. One cell consists of three gates i.e. input, output and forget. Gates typically use sigmoid activation, while input and cell state is often transformed with the hyperbolic tangent function, $\tanh$.

At timestep t, an LSTM has two inputs, x_t the input vector at that timestep and h_{t-1}, the hidden state vector of previous timestep. All the W are weight matrices and b are biases, which are learnable model parameters. In the forward pass, this is how updates are done in the input gate i_t, forget gate f_t, the output gate o_t, the input transform cin_t is taken and the state c_t and h_t is updated in this manner.

$$i_t = g\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right)$$

$$f_t = g\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right)$$

$$o_t = g\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right)$$

$$cin_t = \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_{cin}\right)$$

$$c_t = f_t c_{t-1} + i_t cin_t$$

$$h_t = o_t.\tanh\left(c_t\right)$$

In encoder-decoder based models, information is encoded to a context vector which is then fed to the decoder.

At training time, in the forward pass, both sentence vectors and an image vector are fed to a LSTM network to obtain a context vector, as shown in Figure 2. It is assumed that the context vector summarises the information conveyed in both textual and visual formats. We use the cross-entropy between the output of the LSTM network and the order-embedding vector of the image caption as the loss function to backpropagate and update model weights. We set the learning rate to 0.6, momentum to 0.9 and train the model with 30 epochs using stochastic gradient descent.

During testing, given a news article and its accompanied image, we retrieve the most relevant sentence from the article based on the cosine similarity measurement between the output vector from the LSTM and the order-embedding vector of each sentence.

3.3. Variant Architecture

There are multiple ways, in which sequential information can be propagated through an LSTM network. Another variant of the proposed architecture is to feed the image vector at each timestep of the LSTM such that the input to each LSTM cell is a concatenation of a sentence vector and the image vector. Figure 3 shows a variant of our proposed architecture which is called the Deep NN Dual Architecture.

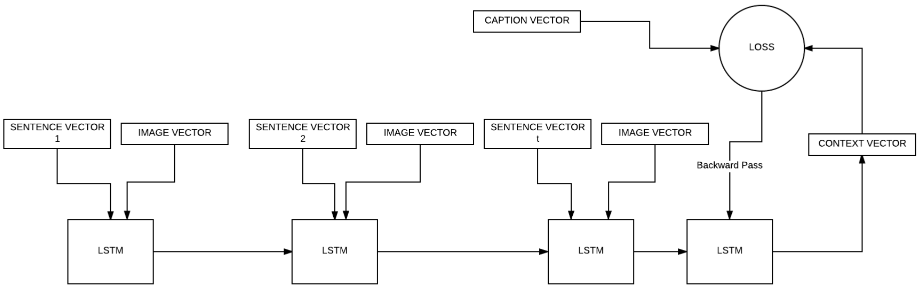


Figure 3: A Deep NN Dual architecture for news image caption generation.

4. Experiments

4.1. Experimental Setup

Dataset. We use the BBC News dataset collected in Feng and Lapata (2013), which contains 3,361 news documents in total. The dataset covers a wide range of topics. Each news article consists of a text article, an image which are normally 200 pixels wide and 150 pixels high, and a caption of the image which has an average length of 20.5 words. On an average each news article contain 421.5 words. The caption vocabulary is 6,180 words and the document vocabulary is 26,795 words. The vocabulary shared between captions and documents is 5,921 words. Some example news articles with their accompanying images and image captures are shown in Figure 1. The original dataset was split into a training set consisting of 3,115 news articles, and a test set consisting of 237 remaining news articles.

Baselines. We compare our proposed model with the following baselines:

- **LDA-based (KL)**. We reproduced the results from Feng and Lapata (2013). For content selection, we first synthesized textual and visual dictionaries where a textual dictionary was created by assigning a unique token id to each word present in any of the articles and visual dictionary was made by clustering SIFT descriptors into 2,000 different visual words. We then trained a LDA model with 1,000 topics on the BBC news dataset containing both text and images. For surface realization, we only used extractive summarisation. It has been shown in Feng and Lapata (2013) that retrieving sentences based on the Kullback-Leibler (KL) divergence between the topic distribution of a sentence and the topic distribution of a news article with its accompanying image gives the best results in terms of human evaluation.

- **Nearest Neighbour**. We also implemented a Nearest Neighbour approach in the order-embedding space. Since both sentences and images are projected to the same semantic space, we can simply choose the sentence which is nearest to a given image as its caption. We use cosine similarity measurement to calculate the similarity score between a sentence vector and an image vector.

- **Deep NN (text input only)**. We explore a variant of our proposed architecture where the input is only text from news articles. This is similar to news headline generation based on text input only except that what we generated here are image captions.

- **Deep NN (dual)**. This is the variant of the architecture shown in Figure 3 where the input to an LSTM cell at each timestep is the concatenation of a sentence vector and the image vector.

Evaluation Metrics. We compare the generated image captions with the actual captions using both BLEU and Meteor scores. The BLEU scores are typically used to evaluate machine translation models. They are calculated based on number of n-gram matches. The Meteor score overcomes the limitation of BLEU by also taking synonyms into consideration. Apart from objective evaluation using BLEU and Meteor, we have also invited human participants to evaluate the generated results by various models. For human evaluation, we have invited 16 human evaluators to choose between the caption generated by the baseline models and our approach for each pair of news article and image presented to them. If human evaluators found none of the captions generated can describe the image well, they can choose the option "none".

4.2. Experimental Results

Method	**BLEU**	**Meteor**
LDA-based (KL)	0.3002	**0.0706**
Nearest Neighbour	0.3237	0.0672
Deep NN (text only)	0.3315	0.0642
Deep NN (dual)	0.3303	0.0609
Deep NN	**0.3427**	0.0677

Table 1: News image caption generation results.

The objective evaluation results are shown in Table 2. It can be observed that the simple Nearest Neighbour approach already outperforms the LDA-based method in terms of the BLEU score. Deep NN with text input only improves Nearest Neighbour slightly on BLEU. Deep NN (dual) performs almost the same as Deep NN (dual). This shows that feeding an image vector at each time step somehow diffuse the semantic information captured in images. Our model (deep NN), where the image vector was only fed in the last timestep in the LSTM network, gives the best overall BLEU score of 0.3427, which outperforms the LDA approach by 4%.

In terms of Meteor scores, both Deep NN and Nearest Neighbour give similar results and they slightly outperform other variants of the deep NN model. Deep NN also performs on par with LDA since the difference of their Meteor scores is only 0.003.

For human evaluation, 38.3 percent of times, the caption generated by our approach was selected as the most appropriate image description by the users, whereas only 28.8 percent of times, the caption generated by the LDA-based model was preferred. We also notice that a staggering 32.91 percent of times, no caption was picked by the users, which could be due to the limited capability of extractive summarisation techniques. Figure 4 shows qualitative study of generated captions.

When only using text content of news articles as the input to our NN architecture, the original model reduces to one-sentence summarisation based purely on text content. As expected, without taking into account the image information, the model has a difficulty in producing appropriate description of a given image. As such, the results are worse than the full approach taking both text and image as input.

Method	**Human Evaluation**
LDA-based (KL)	28.8%
Deep NN	**38.3%**

Table 2: Human Evaluation results.

5. Error Analysis

In this section, we present more results from the experiments conducted. Figure 4 shows three cases of results. The first case, shows the case, where majority of users picked "Deep NN" caption as a right caption for the given article. In this case, Deep NN methodology is clearly able to identify the subject "Chris Langham" in the picture. It is also able to capture background knowledge of the article. The second case, is where the majority of users picked "LDA" caption as a right caption for the given article. In this case, LDA methodology is able to identify the subject. However, the third case shows, where majority of users picked "No" caption as a suitable description for the given article. This is an example case, where both "LDA" as well as "Deep NN" methodologies have failed to capture the content of the articles. It is quite a challenging case. The gold standard caption is "Parts of Charlie and Chocolate factory were also filmed there.", which is not clearly evident from the image.

38.3% of times, "Deep NN" caption has been picked as a right choice by the users. 32.9% of times, "No" caption has been picked as a suitable choice. 28.8% of times "LDA" caption has been picked as a right choice by majority of users.

6. Conclusion

In this paper, we have proposed a novel deep NN-based architecture for the task of automatic caption generation for news images. The experimental evaluation on the BBC News corpus show that proposed methodology gives a better BLEU score than baseline models and performs similarly compared to the LDA approach on Meteor scores. Nevertheless, we notice that the captions generated by our approach were favoured over the captions generated by the LDA based model most of time by human evaluators. In future, this model can be extended to a full-fledged encoder-decoder architecture, where the context vector from the LSTM cell used in our model can be passed to another LSTM cell, which acts as a decoder for word sequence generation.

7. References

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I. (2003). Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, March.

Berg, T. L., Berg, A. C., Edwards, J., Maire, M., White, R., Teh, Y.-W., Learned-Miller, E., and Forsyth, D. A. (2004). Names and faces in the news. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–848–II–854 Vol.2, June.

Blei, D. M. (2004). *Probabilistic Models of Text and Images*. Ph.D. thesis, Berkeley, CA, USA. AAI3183785.

Chen, X. and Zitnick, C. L. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2422–2431, June.

F Tsai, C. and Hung, C. (2008). Automatically annotating images with keywords: A review of image annotation systems. 1:55–68, 01.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In Kostas Daniilidis, et al., editors, *Computer Vision – ECCV 2010*, pages 15–29, Berlin, Heidelberg. Springer Berlin Heidelberg.

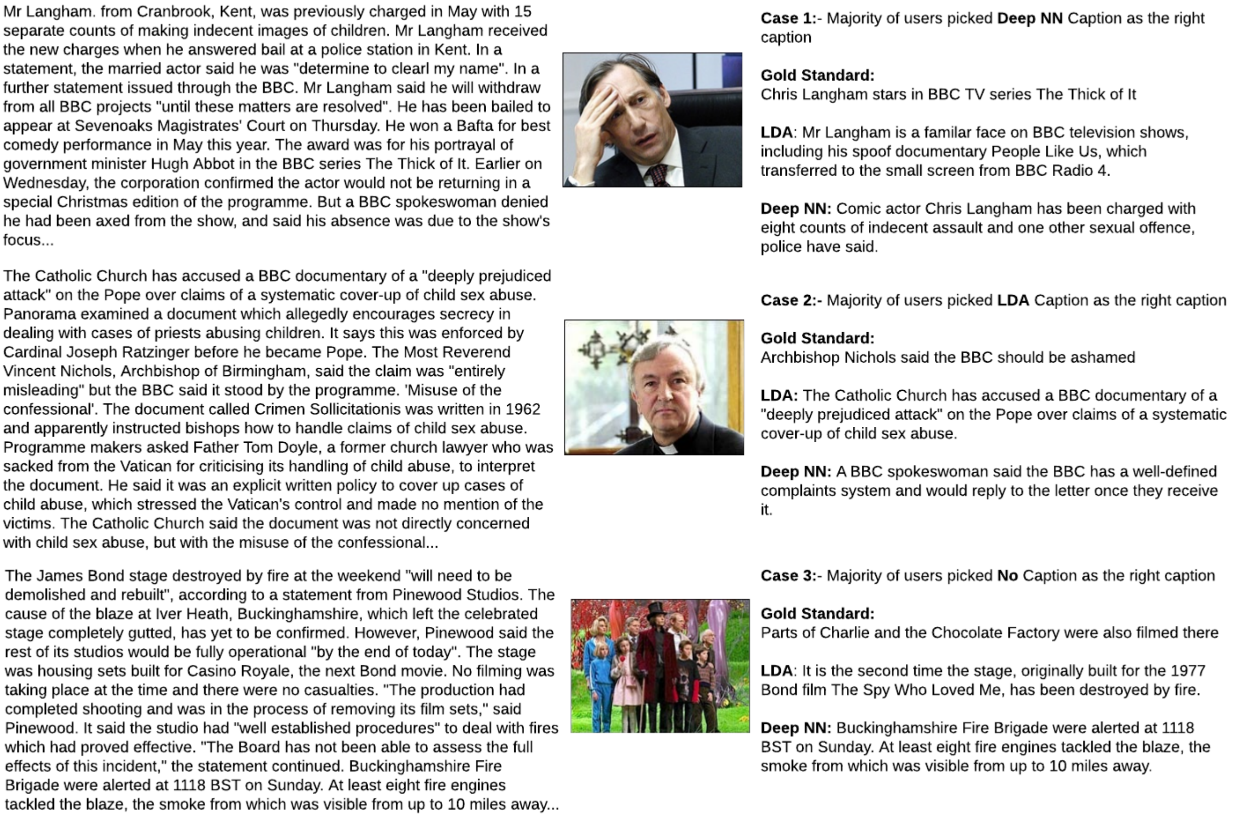
Feng, Y. and Lapata, M. (2013). Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812, April.

Ferres, L., Parush, A., Roberts, S., and Lindgaard, G. (2006). Helping people with visual impairments gain access to graphical information through natural language: The igraph system. In *Proceedings of the 10th International Conference on Computers Helping People with Special Needs*, ICCHP'06, pages 1122–1130, Berlin, Heidelberg. Springer-Verlag.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1/2):177–196, January.

Jones, K. S. (1998). Automatic summarising: Factors and directions. In *Advances in Automatic Text Summarization*, pages 1–12. MIT Press.

Karpathy, A. and Fei-Fei, L. (2017). Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, April.



Mr Langham. from Cranbrook, Kent, was previously charged in May with 15 separate counts of making indecent images of children. Mr Langham received the new charges when he answered bail at a police station in Kent. In a statement, the married actor said he was "determine to clearl my name". In a further statement issued through the BBC. Mr Langham said he will withdraw from all BBC projects "until these matters are resolved". He has been bailed to appear at Sevenoaks Magistrates' Court on Thursday. He won a Bafta for best comedy performance in May this year. The award was for his portrayal of government minister Hugh Abbot in the BBC series The Thick of It. Earlier on Wednesday, the corporation confirmed the actor would not be returning in a special Christmas edition of the programme. But a BBC spokeswoman denied he had been axed from the show, and said his absence was due to the show's focus...

Case 1:- Majority of users picked **Deep NN** Caption as the right caption

Gold Standard:
Chris Langham stars in BBC TV series The Thick of It

LDA: Mr Langham is a familar face on BBC television shows, including his spoof documentary People Like Us, which transferred to the small screen from BBC Radio 4.

Deep NN: Comic actor Chris Langham has been charged with eight counts of indecent assault and one other sexual offence, police have said.

The Catholic Church has accused a BBC documentary of a "deeply prejudiced attack" on the Pope over claims of a systematic cover-up of child sex abuse. Panorama examined a document which allegedly encourages secrecy in dealing with cases of priests abusing children. It says this was enforced by Cardinal Joseph Ratzinger before he became Pope. The Most Reverend Vincent Nichols, Archbishop of Birmingham, said the claim was "entirely misleading" but the BBC said it stood by the programme. 'Misuse of the confessional'. The document called Crimen Sollicitationis was written in 1962 and apparently instructed bishops how to handle claims of child sex abuse. Programme makers asked Father Tom Doyle, a former church lawyer who was sacked from the Vatican for criticising its handling of child abuse, to interpret the document. He said it was an explicit written policy to cover up cases of child abuse, which stressed the Vatican's control and made no mention of the victims. The Catholic Church said the document was not directly concerned with child sex abuse, but with the misuse of the confessional...

Case 2:- Majority of users picked **LDA** Caption as the right caption

Gold Standard:
Archbishop Nichols said the BBC should be ashamed

LDA: The Catholic Church has accused a BBC documentary of a "deeply prejudiced attack" on the Pope over claims of a systematic cover-up of child sex abuse.

Deep NN: A BBC spokeswoman said the BBC has a well-defined complaints system and would reply to the letter once they receive it.

The James Bond stage destroyed by fire at the weekend "will need to be demolished and rebuilt", according to a statement from Pinewood Studios. The cause of the blaze at Iver Heath, Buckinghamshire, which left the celebrated stage completely gutted, has yet to be confirmed. However, Pinewood said the rest of its studios would be fully operational "by the end of today". The stage was housing sets built for Casino Royale, the next Bond movie. No filming was taking place at the time and there were no casualties. "The production had completed shooting and was in the process of removing its film sets," said Pinewood. It said the studio had "well established procedures" to deal with fires which had proved effective. "The Board has not been able to assess the full effects of this incident," the statement continued. Buckinghamshire Fire Brigade were alerted at 1118 BST on Sunday. At least eight fire engines tackled the blaze, the smoke from which was visible from up to 10 miles away...

Case 3:- Majority of users picked **No** Caption as the right caption

Gold Standard:
Parts of Charlie and the Chocolate Factory were also filmed there

LDA: It is the second time the stage, originally built for the 1977 Bond film The Spy Who Loved Me, has been destroyed by fire.

Deep NN: Buckinghamshire Fire Brigade were alerted at 1118 BST on Sunday. At least eight fire engines tackled the blaze, the smoke from which was visible from up to 10 miles away.

Figure 4: Error Analysis

Karpathy, A., Joulin, A., and Fei-Fei, L. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 1889–1897, Cambridge, MA, USA. MIT Press.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, Dec.

Lavie, A. and Agarwal, A. (2007). Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lavrenko, V., Manmatha, R., and Jeon, J. (2004). A model for learning the semantics of pictures. In S. Thrun, et al., editors, *Advances in Neural Information Processing Systems 16*, pages 553–560. MIT Press.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 689–696, USA. Omnipress.

Pan, J.-Y., Yang, H., and Faloutsos, C. (2004). Mmss: multi-modal story-oriented video summarization. In *Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on*, pages 491–494, Nov.

Papineni, K., Roukos, S., Ward, T., and 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*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sak, H., Senior, A. W., and Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *CoRR*, abs/1402.1128.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (). Recursive deep models for semantic compositionality over a sentiment treebank.

Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. (2015). Order-embeddings of images and language. *CoRR*, abs/1511.06361.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach et al., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul. PMLR.

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. *CoRR*, abs/1603.03925.