Natural Language Processing - Final report

Multimodal Language Models : Generating Image Descriptions

Dimitris Alivanistos
11163879

Selene Baez Santamaria
10985417

Thanos Efthymiou
11426381

Sindi Shkodrani
11128348

Charalampos Tamvakis
11158581

Abstract— **Describing images on a visual context is still an unsolved problem in the field of Artificial Intelligence and Machine learning. More specifically, the areas of Natural Language Processing and Computer Vision must come together to create a system that could perform the aforementioned task. In this paper we replicate the work of a recent Image Caption Generator that uses of Deep Recurrent Neural Networks to generate captions from images in a state-of-the-art end-to-end manner.**

I. Introduction

The goal of this paper is to replicate, to the best of our abilities, the work presented by Google for Image Caption Generation[6]. In their paper, the authors create an end-to-end system that processes images and generates sentences from their visual features. This is a data driven technique which illustrates the power and limitations of the latest Deep Learning research wave. Our work is heavily inspired by the same approach. In the next paragraphs we expand on the problem itself and present related and recent work on the subject of Image Caption Generation. We also show our approaches and assumptions on the problem and how we tackle specific difficulties. Lastly, we show the results of our experiments and discuss the results, quantitatively and qualitatively.

A. Image Caption Generation

Humans are able to accurately describe images with Natural Language from an early age. Interestingly, the descriptions are not restrained to recognizing objects in a scene, but inferring context from other visual clues as well. This level of abstraction has yet to be achieved by an artificial system.

From a scientific perspective, the problem of Image Caption Generation is interesting in that it lies in the intersection between two mayor AI fields: Natural Language Processing and Computer Vision. As such, techniques from both fields must be merged in order to create one system that is able to extract and express the semantics of a visual scene.

B. Research Question

As any newly developed area of AI, Image Caption Generation poses a general question: Can we make machines achieve human-like accuracy in this task? Therefore, before tuning for parameters and going deeper into optimization algorithms, what is needed is a proof of concept.

The more specific question of this paper is: **Can data driven techniques provide sufficient results in the context of Image Caption Generation?**

Recent work has investigated the capabilities and constraints on answering this question. We recreate research done by Google, aiming to validate their results, but do so on a smaller scale and with time and computational limitations, thus exploring the boundaries of deep neural networks architectures in this task.

C. Relevance

If successful, achieving human-like accuracy on Image Caption Generation could be beneficial in many fields. For example, creating intelligent agents that help visually impaired people or allowing robotic agents to make better assumptions on their environment through visual input, are relevant applications of this research.

A secondary goal is to find underlying causal relations of visual context and natural language that can answer on how even humans associate these two. The areas of Psychology as well as Linguistics would greatly benefit from these insights.

II. The problem

The problem in this task lies, as mentioned above, in the intersection of two domains: Natural Language Processing (NLP) and Computer Vision. While the second domain has seen remarkable progress in the latest years through the use of Convolutional Neural Networks (CNN) [5], NLP still struggles to achieve reasonable performance in applications such as Image Captioning. Due to the inherent domain complexity and problems that arise in NLP, increases in performance are not as significant as in vision tasks. We should mention that while there are many annotated datasets (MSCOCO,flicrk8,flicrk30), in the field of NLP there is still lack of data which even the authors of Show and Tell [6] struggled with and showed inconsistent results from dataset to dataset. Additionally, training language models in such large datasets shows computationally poor performance and takes too much time to train, especially when using large corpus's.

A. Related Work

Creating Multimodal Language Models is a challenging task because it requires semantic understanding of different inputs, and correct expression of such understanding.

Previous work focused on interpreting images, and using language templates to describe them. Lebret et al. achieved reasonable performance on phrased-based captions using a simple bilinear model. However, though their proposed model is simple, it heavily relies on hand-crafted features for phrase syntax [2]. Similarly, Heuer et al. explicitly translate object nouns found in images into captions [1]. Yet, their approach produces captions where objects (nouns) are the center of the descriptions, while we can imagine scenes where actions (verbs) are more important than the objects in it (e.g. A person is drowning vs A person is in the sea).

Show and tell is unique in that it approaches the problem from a machine translation perspective. In that sense, the task is to translate from visual language to Natural language, specifically to English. The key to their success relies on data driven techniques and end-to-end training.

The first part of the end-to-end training is the use of a CNN to extract meaningful image features. In the paper, they use the GoogleNet architecture, also known as Inception Model, to extract image features [5]. The second part is the use of Recurrent Neural Networks (RNN), specifically Long Short Term Memory (LSTM) networks to produce captions related to the image features. Other details about their approach will be mentioned through the rest of this paper, since they serve as inspiration for our approach.

Lastly, we should also mention that the above end-to-end training has inspired others, leading to progress in the field of Image Captioning using attention [7].

B. Assumptions and Observations

Just as in Show and Tell, we assume that a certain level of differentiation is needed between visual and language inputs. This also helps in ”keeping things simple”. We neither need large or multiple captions to explain pictures, and also we do not need to describe every image feature in a sentence to capture the context of the images. Furthermore we assume that increasing the complexity of the caption does not necessarily mean that we described the image better. Some images are better explained in a simple straightforward way.

Differences from Show and Tell arise in two manners. First, our visual features come from a pretrained VGG network [4] instead of the GoogleNet used in their paper. However, we believe this is not a drawback, as VGG-16 has showed similar performance with GoogleNet when trained on ImageNet. Therefore the features from both CNNs are bound to be similar.

Secondly, due to time and computational constraints, we trained a small version of the original Show and Tell model. Using LSTMs for Language Modeling requires extensive training for achieving reasonable performance. We trained on less iterations and with a subset of the original dataset.

Finally, a most basic constraint in the task is the quality of the captions and the vocabulary. While the datasets are enriched with multiple captions per image, the enrichment is not entirely successful. If unique image features are associated with multiple words in the corpus, we predict this would create noise, thus reducing the strength of the visual context in the model.

C. Our method

Our approach is to extract the learned features from the last convolutional layers of the VGG architecture, and use an LSTM Recurrent Neural Network that takes as input visual embeddings of these features, as well as word embeddings from a corpus created by the annotated captions.

To tackle computational constraints, we approach the problem with using subsets of the main dataset.

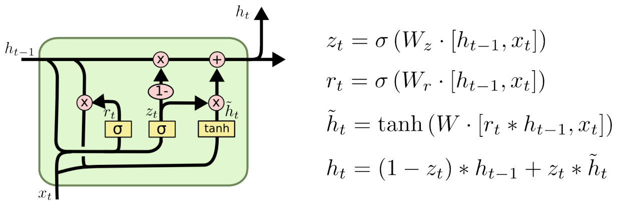
III. Approach

A. Model

We focus on the second part of the task: expressing the semantic knowledge in a meaningful and grammatically correct manner. To do so, we use processed image features instead of raw images. We assume the features contain the semantic knowledge we need to convey, thus allowing us to explore in more detail the behavior of NLP techniques.

We follow the main idea of Show and Tell creating an LSTM network in order to obtain output word sequences of various lengths. An LSTM memory block differs from a traditional RNN as it makes use of a cell which is controlled by the input, output and forget gates. These gates get activated by a sigmoid function and are integrated by multiplication. The output is fed to a Softmax to get the word prediction. Figure 1 shows a representation of an LSTM memory block.

While basic RNNs can be used , LSTMs are much fitter for the task since they avoid problems, such as vanishing and exploding gradients, which RNNs suffer from. One the one hand vanishing gradient appears when the parameters of our model go close to 0 , while on the other hand exploding gradient appears when the parameters reach values close to infinity.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Fig. 1: An LSTM memory block

B. Inference

The inference of our model is described in figure 2. The LSTM architecture we use is similar to the one in Show and tell. We learn word embeddings using a fully connected embedding layer that is similar to Word2Vec. As for the visual embeddings, we also pass them through a fully connected layer and embed into a lower dimensional space. We add dropout layers before and after the LSTM layer, in order to avoid overfitting. Lastly we use a fully connected Softmax layer to output words probabilities.

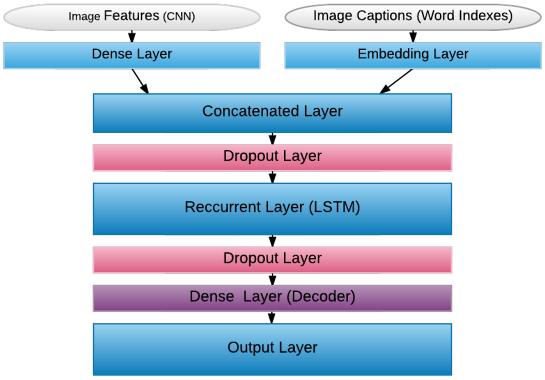


Fig. 2: Deep Network Architecture

C. *Learning*

As it is commonly practice in Deep Learning, we coupled an Softmax output with a Cross Entropy loss. The latter is defined by:

$$\mathcal{L} = \sum_{i}^{N} \log p_i, \qquad p_i \text{ the output of the softmax layer}$$

As a consequence from our divergence from Show and Tell's architectural design (we take extracted features as opposed to concatenating a CNN and LSTM), a major difference is that we cannot backpropagate our error back to the CNN model, as the authors did. This would be useful if we wanted to fine tune features according to their captions. However, it could also introduce severe problems if the captions are not evaluated correctly.

D. *Visual input*

The image features were obtained from a pretrained VGG-16 network , trained on Imagenet [4]. They are the feature extractions of the fifth layer activations of the VGG-16 network.

Attached to each image are five descriptive captions. Each caption consists of a single sentence. We processed this in order to create pairs of a single image and a single caption. All samples are seen by the network as independent from each other, thus enriching our datasets.

E. *Vocabulary*

Words from the training captions are counted and stored. With Zipf's law in mind, we assume that low frequency words in our sample are also rare in the English language. Therefore, words with less than n appearances in the total training set are discarded in order to control over an extremely large vocabulary size.

Because our images are associated with at least five captions, we require at least four frequencies for a word to be added in the vocabulary. Our reasoning is that if a word does not at least appear in more than five captions of an image, then it is not essential for properly describing its context and can introduce noise to the image features.

After counting and selecting the words to be included, we sort them by decreasing count, and store them for latter use.

F. *Caption input*

Since the goal of this project is not to create optimal word embeddings, we encode words in the vocabulary in a basic way. For every instance, the words are first vectorized with their corresponding indices in the vocabulary.

Word embedding is applied to the image captions using a fully connected embedding layer right before feeding the captions to the network. Implementing one-hot encoding proved difficult due to memory incapacity.

1) Start/Stop tags: In some of the experiments we introduce two special start/stop tags in the vocabulary. We used these special codes in order to represent the start and end of a caption. This also serves as a mechanism that signals the network to stop generating words.

IV. Empirical evaluation

A. *Data*

The training set consists of 143,161 images, creating 716,076 samples in total. The test set consists of 20,000 images, creating 100,061 samples. A sample image and its caption is shown in Figure 3

Fig. 3: A man who is jumping in the air while playing tennis.

B. *Training*

We create to setups for the model. the first is a vanilla LSTM, with basic training and optimization parameters obtained from literature. The second is an optimized LSTM for this task. The setups are summarized in the following table:

TABLE I: Settings for 2 trained models

Model	vanilla-LSTM	optimized-LSTM
Iterations	50000	50000
Embedding Size	512	1024
Sequence Length	16	32
Optimizer	sgd	adam
Batch size	32	64
Learning rate	1e-4	1e-4
Special Tokens	False	False
Grad Clipping	[-5,5]	[-5,5]
Exploding Gradient	False	True
Bleu-4	0.0112585396374	0.0203055154838

C. Evaluation

1) Quantitative: We evaluate our model using BLEU metrics. According to Papineni et al [3], it provides a quick, inexpensive and language independent method of evaluating machine translation. Though largely criticized, this metric is still used in most research around Image Caption Generation, so it provides a good mechanism for us to compare to other work.

The BLEU metric scores a translation on a scale of 0 to 1, but is frequently displayed as a percentage value. Our captions are given a score between 0 and 1, where 1 means that the caption we generated was identical to our reference. A score less than 0.15 means that the system is not performing optimally, while a score greater than 0.5 is a very good score.

In our approach we use the BLEU-1 and BLEU-4. These metrics calculate ngram precision for unigrams and 4-grams respectively. We decided to include 4-gram precision because unigram precision methods may lead to incorrect translations with high-precision. For example, situations in which a word of reference repeats several times obtain very high precision. Finally, we report scores comparing the generated caption only to the first target caption for each image from the dataset.

2) Qualitative: In order to have a more direct insight of the model performance, we also evaluate subjectively, by comparing generated captions with target captions.

We consider two aspects for evaluation. First, the network must generate meaningful sentences that account for natural language. Second, the descriptions generated need to be as close as possible to the annotated captions. We compare some target captions to the generated captions.

D. Results

1) Vanilla LSTM: In this setup we have a basic training and optimization techniques, as well as a smaller embedding feature space and shorter output captions. The model performance is shown in Figures 4 and 5 for Cross entropy and L2 regularized loss, and on Figure 6 for BLEU-1 and BLEU-4 scores.

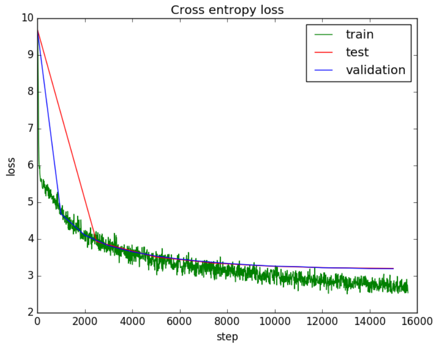


Fig. 4: Loss for vanilla LSTM.

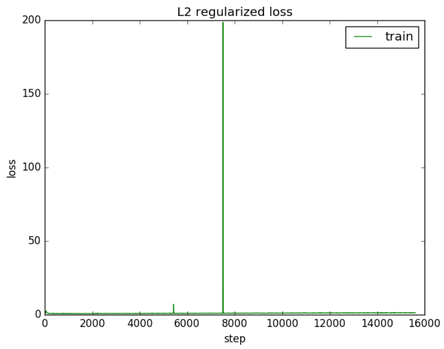


Fig. 5: L2 norm regularization for vanilla LSTM.

Figure 5 clearly shows a problem with exploding gradients. This means that the weights in the network suddenly grow too fast, as seen at iteration 7500. In fact, at iteration 16k the gradients explode, after which the network cannot recover and returns null.

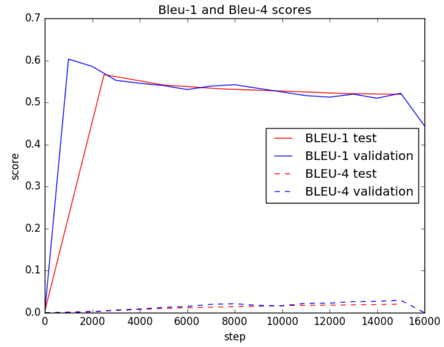


Fig. 6: Accuracy for vanilla LSTM.

Figure 6 illustrates the great difference between BLEU-1 and BLEU-4 scores, with the latter only visible in the lower part of the graph.

2) Optimized LSTM: We decided to try to improve the performance by tuning training parameters parameters, introducing more elaborate optimization algorithms, and increasing the architecture complexity. The new model performance is shown in Figures 7 and 8 Cross entropy and L2 regularized loss, and on Figure 9 for BLEU-1 and BLEU-4 scores.

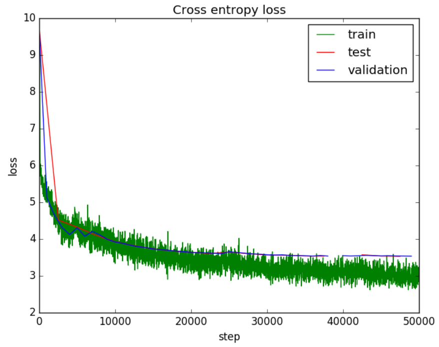


Fig. 7: Loss for optimized LSTM.

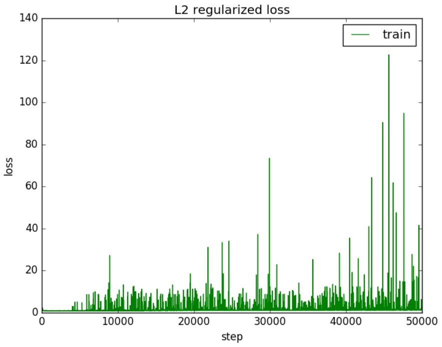


Fig. 8: L2 norm regularization for optimized LSTM.

From Figure 8 we gather that the optimization aided in controlling for exploding gradients, since the model correctly concluded after 50k iterations.

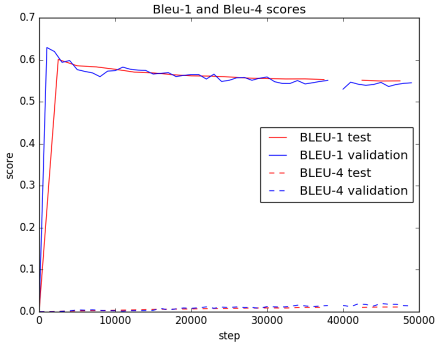
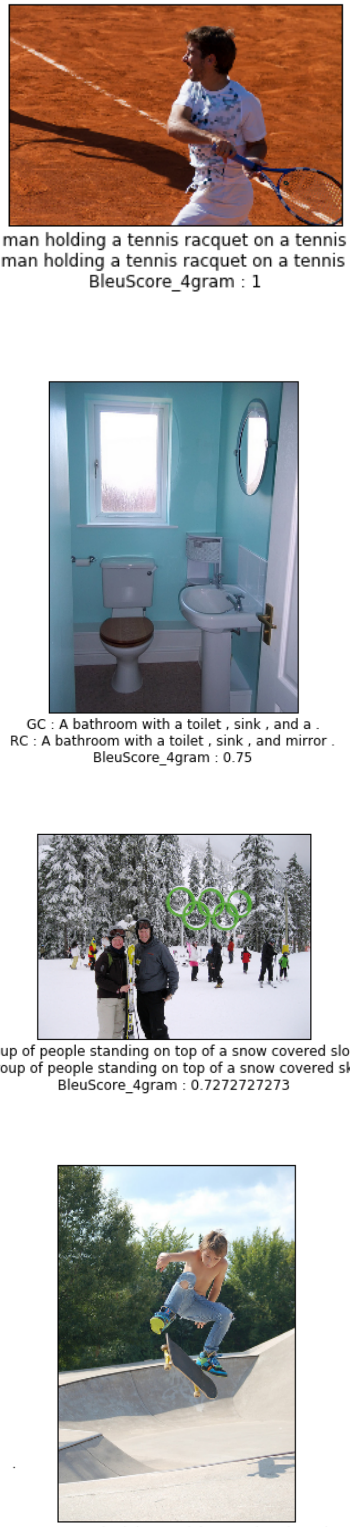


Fig. 9: Accuracy for optimized LSTM.

Figure 9 shows that, regardless of our optimization, the BLEU scores maintain their behaviour.

3) Qualitative Evaluation: To further explore the results, we visualize some images and their target/generated captions. Figure 9 shows results that achieved a high BLEU-4 score, while Figure 11 shows misleading captions.



GC : A man holding a tennis racquet on a tennis court .
RC : A man holding a tennis racquet on a tennis court .
BleuScore_4gram : 1

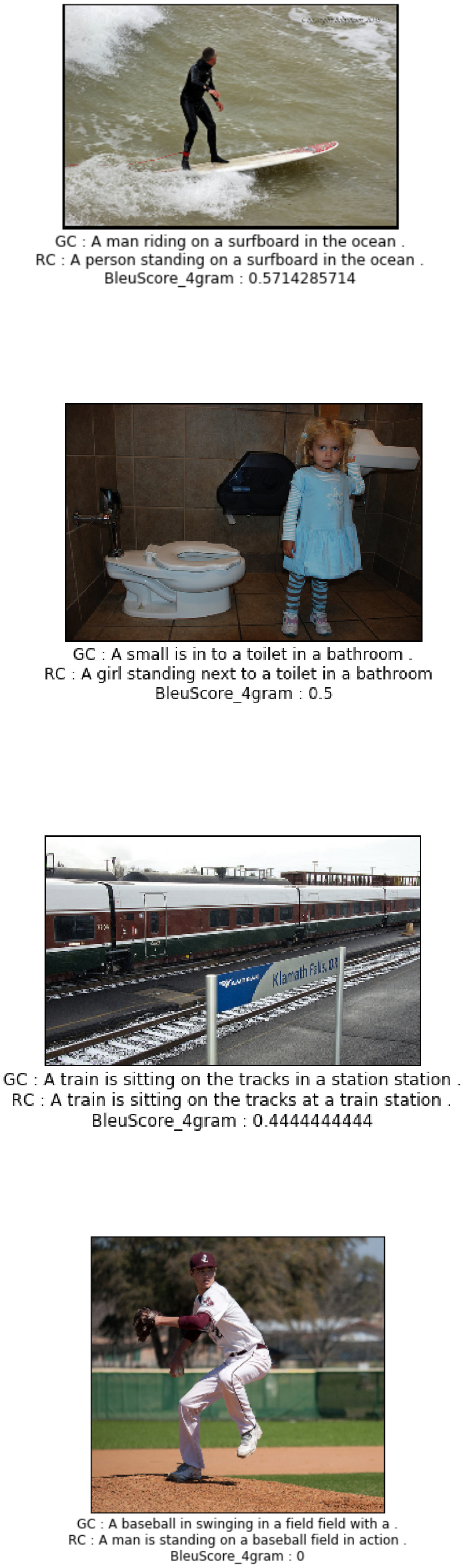
GC : A bathroom with a toilet , sink , and a .
RC : A bathroom with a toilet , sink , and mirror .
BleuScore_4gram : 0.75

GC : A group of people standing on top of a snow covered slope slope .
RC : A group of people standing on top of a snow covered ski slope .
BleuScore_4gram : 0.7272727273

GC : A man is doing a trick on a skateboard .
RC : A child is doing a trick on a skateboard .
BleuScore_4gram : 0.7142857143

Fig. 10: Good Results

The top sample image in Figure 10 is a clear example of identical generated and real captions, while the other three images have generated captions very close to the original one.



GC : A man riding on a surfboard in the ocean .
RC : A person standing on a surfboard in the ocean .
BleuScore_4gram : 0.5714285714

GC : A small is in to a toilet in a bathroom .
RC : A girl standing next to a toilet in a bathroom
BleuScore_4gram : 0.5

GC : A train is sitting on the tracks in a station station .
RC : A train is sitting on the tracks at a train station .
BleuScore_4gram : 0.4444444444

GC : A baseball in swinging in a field field with a .
RC : A man is standing on a baseball field in action .
BleuScore_4gram : 0

Fig. 11: Bad Results

Figure 11 shows four sample images with poor performance. We divide this set in two:

On the one hand, we have generated captions that are not grammatically correct (top right and bottom images). We assume that these cases are due to the limited variety of consistent image features and captions. While the network can generate meaningful sentences, it cannot compare them with the image features accurately because these features only exist empirically and have not been really extracted from the CNN.

On the other hand, we have generated captions that correctly express the scene but are scored poorly (top left image). We believe this reflects on the drawbacks of BLEU metrics, and motivates for finding better, more specialized comparison metrics for Image Caption generation.

E. Analysis

At first glance, both vanilla and optimized models have similar Cross Entropy and BLEU performance. However, a crucial difference comes in when dealing with exploding gradients since only the optimized model is able to do so.

Furthermore, our accuracy graphs support the idea that BLEU-4 is a stricter and probably more correct metric. The authors in Show and Tell also touched on this topic, and widely criticize these metrics.

As for the qualitative evaluation, a deeper look into the generated captions revealed a trend to learn captions beginning with ”A man”. We believe this is due to overfitting and that different sampling techniques, or a richer dataset could help in avoiding the problem.

V. Discussion and Conclusions

A. Discussion

Coming back to the original research question posed, we have provided with a proof of concept that data driven techniques can tackle the Image Caption Generation problem. Our model actually learned how to construct proper phrases to some extent, taking into account the start and end of the sentence. However, the methodology requires more tuning and more data to be considered successful. We firmly believe that provided the previous, our solution may generate better results.

Regarding related work, in this paper we tried to replicate to the best of our abilities and resources the Show and Tell paper. However in the paper, major technical details are not described in detail, which made it quite challenging to succeed in replicating. They have achieved the second best accuracy on the MSCOCO dataset, which by all means is a high challenge to tackle. As further work, it would be really interesting to add features from other papers on the field, such as visual attention [7].

B. Conclusion

Data driven techniques depend on the quantity and quality of data. Though we provide a proof of concept that the task in hand can be achieved through Neural Networks, our limitations prevent us from creating a fully functional system.

The biggest challenge was certainly running the experiments with massive amounts of data while having limited time and computational resources. Not only is the amount of data crucial to the learning, but also different learning rates and other parameters had to be experimented with in order to get the network to generate sentences in natural language.

Additionally, we did not have the images included in the dataset, only their features and captions. It was relatively easy to compare to annotated captions, but time consuming to validate the results to the actual images. Linked to the previous, and as mentioned before, is the fact that we could not backpropagate to the VGG network error. We had to work with the features as given, being unable to improve upon noisy feature output.

C. Future Work

For future work, it will be interesting to see how we can improve the accuracy of the model with more data. Furthermore, we would like to transfer the implementation to a different framework like Tensorflow, which is faster, allows for easier control of model parameters and can help with the visualization.

Another thing we were not able to implement in our approach, but that was present in the original Show and Tell approach, is the Sampling and the Beam Search on the captions generated. Beam Search uses the best n sentences at every time step. On the paper they report that a greedy search vs Beam Search of size 20 decreased their BLEU scores by 2 points on average. We would like to investigate how this would affect our results in the future.

VI. Team responsibilities

Out team dynamics consisted on working together during the initial phases of the project, in order to have the concepts clear in our heads and plan on the implementation. The Dimitris and Selene took care of the enrichment of the dataset as well of the preprocessing of samples to fed the model with. Sindi and Charalampos continued to implement and tune the LSTM. Thanos then took care of the evaluation. In the end, we all dicussed the results and wrote and revised the report.

References

[1] H. Heuer, C. Monz, and A. W. Smeulders. Generating captions without looking beyond objects. *arXiv preprint arXiv:1610.03708*, 2016.

[2] R. Lebret, P. O. Pinheiro, and R. Collobert. Phrase-based image captioning. *arXiv preprint arXiv:1502.03671*, 2015.

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015.