Video Captioning

11-777 Final Report
Carnegie Mellon University
Spring 2016

Chaitanya Ahuja, Salvador Medina, Dheeraj Rajapogal

October 25, 2017

1 Introduction

The automatic generation of video descriptions or video captioning is a current problem of active research due to the vast amount of readily available video content. Digital video recording devices have become ubiquitous in modern society, and annotating each piece of video is impossible within our current capabilities. For this reason, an automatic video descriptor is desired to aid in different tasks that involve the analysis and description of videos. The search of information based on video content from large video repositories would be one of the most favored tasks. A video captioner might also be of aid to analyze surveillance systems in hospitals or crowded places such as airports, a system like this would alleviate the tedious task of watching multiple monitors for large periods of time. In another scenario, a video descriptor which could achieve human-like capabilities in real-time could be capable of aiding the visually-impaired population to describe their surroundings and give them the contextual information which they are deprived from. Although challenging, we believe that video captioning can be as accurate as Image Captioning methods with more investigation into models and supporting hardware architecture.

In this work, we propose a model which attempts to learn how to describe videos by learning from movie descriptions for the visually impaired. We use the two most commonly used video annotation datasets: the Montreal Video Annotation Dataset (M-VAD) and the Max-Planck-Institut für Informatik Movie Description Dataset (MPII). Although these datasets are relatively large, we see some obvious issues such as additional information, misalignment, just to mention a few, which will be later discussed in section 3, and it is to our knowledge that they have not been discussed in the literature so far. We believe that an additional filtering step might help us utilize the strength of deep multimodal

models. One aspect of our work addresses this by proposing a system that can help annotate videos to caption with a frame-level granularity.

Recent work in this area has mainly focused on solving the problem as a translation problem from video frames or video motion frames to text by means of recurrent neural network (RNN) models, more specific long short term memory (LSTM) networks. These models attempt to capture the temporal aspect of the videos through their memory gates and video frame or motion sequence is learned separately from the text sequence which later are fused within the model. The work proposed in this paper attempts to train the sequence of video frames and text jointly in one step by means of a relaxed version of grid-LSTM[5].

This report is organized as follows, in Section 2 we describe the most recent models for video and image captioning which inspire the proposed model. Section 3 describes how we address the challenge of creating a fine-grained video annotation dataset; Section 4 gives a full description of the proposed model, which is followed by Section 5 where we describe the experimental setup and the evaluation criteria for the models. In Section 7 we describe the results obtained from our experiments and discuss the strengths and shortcomings of various models. Finally, Section 8 describes how the proposed model can be improved and plan for the follow-up work in the summer towards achieving this goal.

2 Related Work

Most prominent approaches in video captioning tend to address the challenge by extending image captioning models which consider the video as a sequence of images, while some others attempt to describe the video through machine translation models by translating the video from the image or motion feature space to text.

Among the work that extends image captioning models into video captioning we can find [7], [22], [8], [1], and [14]. In their experiments, the authors consider the video as a sequence of still images to be described. The generated video descriptions are the result of processing pipelines that includes natural language processing (NLP) techniques such as: summarization, text mining for vocabulary expansion or text classification. For instance, Xu et al. [20] attempt to describe images by detecting the regions of interest by means of a 3D-CNN and generating the sentences by means of an LSTM as Karpathy and Fei-Fei did in the NeuralTalk [6].

On the other hand, there have been attempts to create a video captioning model through machine translation we can find [14], [23], [19], [18], [21], [20], and [13]. In this case the model attempts to translate the input from the video feature space into the natural language or text feature space through a latent variables. For this reason several experiments have incorporated RNNs as these are capable of learning latent spaces from sequences. This characteristic has shown to be of great use while generating complex and correct sequences of

text.

Recently, there has been an immense interest in video captioning using visual attention models, where the object or subject of interest are located through computer vision techniques and confine the captioning problem to only the detected region of interest. For instance, Sukhwani et al. [16] focus only on tennis matches and attempt to describe the match as humanly possible, by restricting the problem to only videos of people only playing tennis, this increased the readability and correctness of the game description. On the other hand, Rosenfeld et al. [12] were able to classify object-to-face actions with high scores to the task.

Venugopalan et al. [19] proposed the idea of implementing a sequence-to-sequence (Seq2Seq) model, which mainly consists of a CNN that takes as an input a sequence of video frames, then the newly acquired image embeddings are fed into an LSTM which generates the words that form the final video description. The main difference with previous models such as the one proposed by Sachdeva et al. [13], is that Seq2Seq is an end-to-end system.

Our work is conceptually similar to the work of Venugopalan et al. in [19] and in [18], as we tend to use a CNN for extracting image embeddings and add the temporality of of the video through an RNN. However, our work differs in the replacement of the dual layer LSTM based on Donahue et al. [4], with a Grid-LSTM as explained by Kalchbrenner et al. [5]. The reasoning behind this architectural decision is to reduce the information loss that occurs during the last mean pooling layer and the encoding of a 2D latent space by training the model with the full video caption text *au pair* with the full sequence of video frames.

3 Dataset Analysis

In this section we describe the used movie datasets along with their statistics and an analysis obtained from observing thousands of videoclips, which resulted in a series of observations about the problems that these videos convey while training models and some concealed challenges which might be of interest to the researching community.

3.1 Description and Statistics

Two commonly used datasets for the task are the MPII Movie Dataset [11] and the M-VAD Movie Dataset [17]. From the initial manual inspection, the M-VAD dataset, in general tends to have more details in the video compared to each video's corresponding caption whereas MPII dataset has relatively longer captions with many proper names. This makes it a challenging problem but we have to not only address the richness of models but also the ability of the training data to learn rich representations.

The M-VAD dataset is composed by short video clips from 92 different movies. The training data set has an average of 512.7 clips per movie with an average length of 5.7 seconds and and average of approximately 600 sentences of video descriptions per movie. The test set provided, only considers 10 out of the 92 movies, with an average of 495.1 clips per movie which have an average length of 6 seconds. The training and test do not overlap whereas the MPII dataset was collected from video snippets of 94 movies with about 64,000 captions.

Finally, among both datasets only three movies overlap, both instances of the movies were included in the newly created dataset.

3.2 Analysis

In a exploratory analysis of the new jointly created dataset we could determine that the videoclips would require to be reannotated. The main reason was that most of the videoclips would contain frames that were not relevant to its respective caption. In an attempt to reduce *frame noise* found we developed a publicly availabe annotation tool [1].

3.2.1 Frame Noise

After reannotating approximately 4000 videoclips we found that there are several sub-challenges that need to be overcome before developing complex models for video-captioning. Firstly, some videos are misaligned. By this we mean that the description contained in the caption does not match the same video clip, rather they might have been from the previous or the next clip.

The videos also contain fade-in and fade-out to black. This might represent a problem for the current sequential models as it might interpret it as a characteristic to be considered while learning the captions that contain it which could be a major source error. A possible solution to overcome this problem might be to insert a `<FADE>` token to the captions. This would allow the model to learn correctly the content within the video.

We also have to take into account the grammatical errors and misspellings in the captions. Couple of ways to address this is either by adding noise to the latent model or by re-annotating the captions for grammatical correctness.

Another obstacle that we need to overcome in the videoclip annotations is the use of proper names in captions, which can be confusing and misleading for learning the probabilites of the sequence. For instance, from the movie *Reservoir Dogs* the names of characters such as *Mr. Orange, Mr. White* and *Pink*, might lead to the incorrect sequence by adding *Mr.* to the colors that might appear in a text sequence.

[1] 1.`https://github.com/salmedina/MovieClipVideoAnnotator`

Another well known issue for video-captioning is the change of camera angles within the same scene while only one of the camera angles is being described. For example a person might be running in a track while the camera angle changes back and forth between the runner and different scenes from the audience watching the race, while the caption might only state: *Someone is running on a track.*

There are movies such as *Les Miserables* which contain long captions with sophisticated vocabulary including words from other language interspread with English for short videos which are a couple of seconds long. The extra vocabulary in really short videos might lead the model to learn long sentences with irrelevant adjective for short video clips as those which describe fast actions as jump or glance.

3.2.2 Concealed challenges

Working with DVS captions from movies also conveys complex challenges as those are aimed to describe a movie to a visually-impaired viewer. Therefore, the captions may ommit information contained in the video focusing only on the relevant aspects of the video or may require simple cognitive procedures for humans such as inference to comprehend the current scene.

To this extent, *coreference resolution* is one of the major challenge that we face in the task. Description writers tend to refer through pronouns to different persons or objects corresponding to people or things that happened in the previous scene in the movie and might be contained in another videoclip from the same movie. However, while training current deep learning models the videoclips are treated as a stand-alone piece of data for learning, without considering prior videos from the same movie to train the models.

Some descriptions are very subjective and they emit a judgement regarding the scene. "Darryll is checking himself out in the hall mirror, and it's obvious he likes what he sees.", "The room looks like it was decorated from a Sears catalog". A possible solution to this type of description require a classifier to infer whether the sentence should be considered as subjective or objective.

Another concern that we need to take care, comes from descriptions that require the viewer to complete the information given a visual cue as in: "It is the body of a long dead woman" while the video only shows the *head of the long dead woman*. For a human it is an easy task to complete the visual aspect of a *dead woman*, although current system might learn that the head is also the body.

A subset of videos contain description that go beyond the frames but that might be inferred from the sequence. For instance the caption is: *"The man is dressing for the party."*, while the video is about a man walks into a bathroom, closes the door, after a few moments opens the door and comes out properly dressed for a party. Current models require a more complex temporal semantic analysis of the video to properly describe this type of scenes. Until that time,

we presume that such captions should not be considered for training.

Finally, we have also encountered captions that are abstract enough and requires the viewer to complete the description through by inferring the content based on prior knowledge of the action occurring within the video. For instance, a videoclip that contains two persons interacting their captions only describe the action of one of the persons. This might also be considered as noise for those models that naively learn from global frame features and require more sophisticated models as the ones based on attention.

4 Mathematical Formulation

A video captioning model can be succinctly defined as the following

$$P(y_1, y_2, \ldots, y_K | v_1, v_2, \ldots, v_T) = \prod_{k=1}^{K} P(y_k | y_{k-1}, \ldots, y_1, v_1, \ldots v_T) \quad (1)$$

where y_ks are the words (as one-hot representation) in the caption in the increasing order of k and v_ts are the FC7 layers of VGG16 net of each frame of the video. All the methodologies implemented during our experiments are used to learn the probabilities of the target language model.

At the end of each time stamp of the LSTM, we predict a word using a softmax layer of length equal to the number of words in the vocabulary. Hence the appropriate loss function for this task is the categorical cross-entropy which is defined as:

$$\mathcal{L}_t(\hat{y_t}, y_t) = -y_t \log(\hat{y_t}) \quad (2)$$

where y_t is the ground truth and $\hat{y_t}$ is the estimated one-hot vector.

4.1 Sequence2Sequence

The Sequence to Sequence model was introduced in [18] which utilize the LSTMs to encode the video in an abstract feature space, which is then decoded by the language model. This is extremely similar to a translation model

If W^* are all the parameters of the model, the following MLE expression is used to estimate these weights.

$$W^* = \text{argmax}_W \sum_{k=1}^{K} \log P(y_{k+1} | y_k, y_{k-1}, \ldots, y_1 v_1, v_2, \ldots, v_T, W) \quad (3)$$

A potential issue posed by this model is that it encodes the complete video as one feature vector and hence loses the ability to use k features jointly with each frame, to predict $\hat{y_{k+1}}$.

4.2 NeuralTalk

A very naive baseline has been implemented which translates the video captioning problem to a sequential image captioning one. We sub-sample the videos and assign each frame with the same caption. This processed dataset is now used by the model proposed in [6] also popularly known as NerualTalk. While decoding, we we choose a frame at the center of the video, which is used for generating captions. We expect the results of this to act as the absolute baseline for the video captioning task.

4.3 Multimodal Language Model

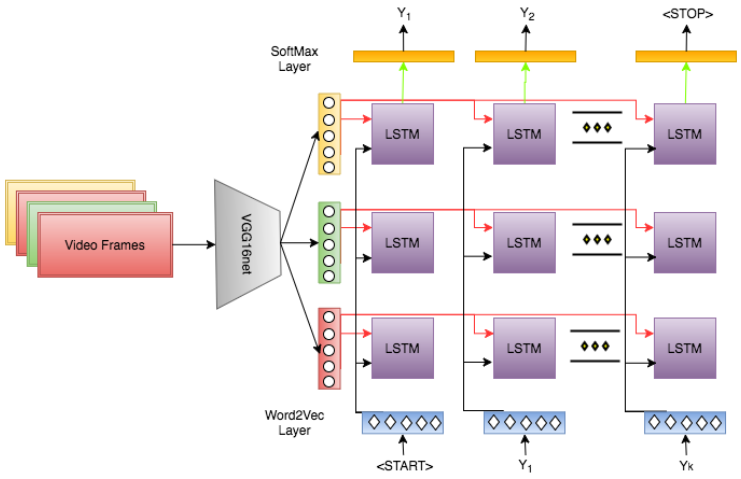

Figure 1: Flow Chart for the Multimodal Language Model

We propose a new model called Naive 2D grid, which is not based on the encoder decoder model like Sequence2Sequence. This model solves the issue of predicting the next word using the joint encoding of $\hat{y_{i-1}}$ and features of each frame to predict $\hat{y_i}$.

If W^* are all the parameters of the model, the following MLE expression is used to estimate these weights.

$$W^* = \text{argmax}_W \log P(y_{k+1}|y_k, v_1, v_2, \ldots, v_T, W) \tag{4}$$

While training, the captions are not split up into subsequent pairs of words and the first word of this pair is concatenated with the video features. Hence, the input to the first LSTM layer is

$$\{(f_{w2v}(y_k), v_1), (f_{w2v}(y_k), v_2), \ldots, (f_{w2v}(y_k), v_T)\}$$

where, $f_{w2v}(y_k)$ are word2vec [9] features of word y_k. The dimensions of the input features are reduced before feeding them as inputs to the LSTM layer. We train it using sequential classifier with the given input an the output as y_{k+1}.

While testing, the input is started with the <**START**> token to predict the first word. The predicted word is used as an input for the next decoding computation and this is repeated till the predicted word is the <**STOP**> token.

The flowchart for this model is described succinctly in Figure 1.

5 Experimental Setup

We use 53 movies which are a subset of the union of two movie description datasets MPII [11] and M-VAD [17]. These movies contain least number of videos compared to the other movies, hence providing diversity in terms of the number of movies for training and testing. There are a total of 16 thousand video clips being used for training the given models. We split the videos in the ratio of 6:2:2 for training, validation and testing respectively. It is important to note that the videos in the 3 different sets do not have any overlap of movies, Hence the model is movie independent as well.

5.1 Video Features

The videos were sub-sampled to 30 frames per video clip, in order to obtain the most from small actions and avoid the dominance of long video clips. After the sampled frames were extracted those were resized to 231X231 pixels, which were passed through a pre-trained 16 layer convolutional neural network (CNN) popularly known as VGG-Net [15]. The output from the last fully connected layer with a dimensionality of 4096x1 was taken as the feature vectors for the generative LSTM model. The CNN features have been proved to consistently perform across various vision tasks and hence the motivation to use this for the video representation as well.

5.2 Caption Features

Each caption is tokenized and aggreagated with a "<START>" and a "<STOP>" token at the beginning and end of the tokenized list respectively. The vocabulary is represented by word embedding obtained from Word2Vec [9]. The "<START>" and "<STOP>" tokens are calculated as the average of all the words in the vocabulary. The Word2Vec embeddings are used as input to the LSTM model. One of the advantages of using Word2Vec embeddings is the low-dimensionality of the calculated vectors while still containing the contextual information. This last characteristic from the word embeddings is the key to generate sensible sentences based on the structure of the captions used from our training set.

6 Baseline Error Analysis

Video Captioning is a challenging task due to the large dimensionality of the problem. As the task requires a generative model instead of a retrieval based method, the results need to be considered for a qualitative error analysis.

Videos can be described in many ways through natural language, the annotators can focus on many different aspects of the video itself leading them to use different sets of words and different levels of detail. Another prominent aspect to consider while using video captions the caption writer's bias while describing the video.

In this section, we analyze the errors of our baseline system for the video captioning task. This analysis also gave us some insights to the level of challenge in predicting captions for video compared to an image.

6.1 Complete mismatch

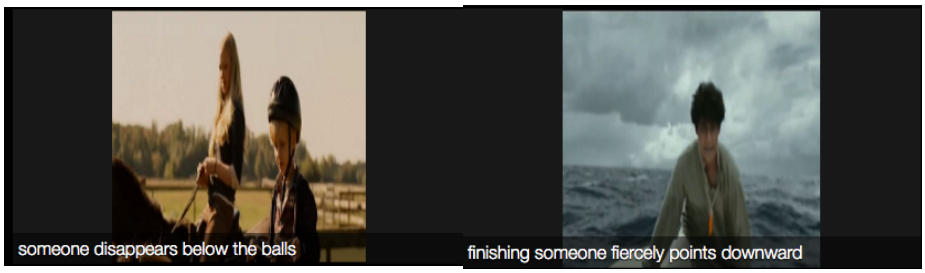

Figure 2: Examples of captions that are complete mismatch with the corresponding videos

A complete mismatch occurs when the video and the generated caption do not share a common idea or concept. A couple of examples are shown in Figure 2. There could be many reasons when the captions does not correspond to the video. We have found out that these errors require an in-depth low-level feature analysis. Our main suspicion is that the VGG features for such captions were not adequate enough to capture the caption for that video.

6.2 Captions with long sentences

It has also come to our attention that if the length of the caption is affected by the length of the video. In other words, if the video has a longer length, the caption generated is also longer. Our main concern with these behavior is that even if the video is long, this does not necessarily imply a larger caption as the video could describe a simple action in a longer period of time. As an example, in Figure 3, the descriptions tend to be long and incomplete. We attribute this

Figure 3: Examples of captions that have long captions

issue to the fact that longer videos naturally having longer captions since the DVS writers have more frames to write caption words. Our takeaway from this is that it might be a reasonable idea to approoximate the captions to certain number of words.

6.3 Captions with Emotion

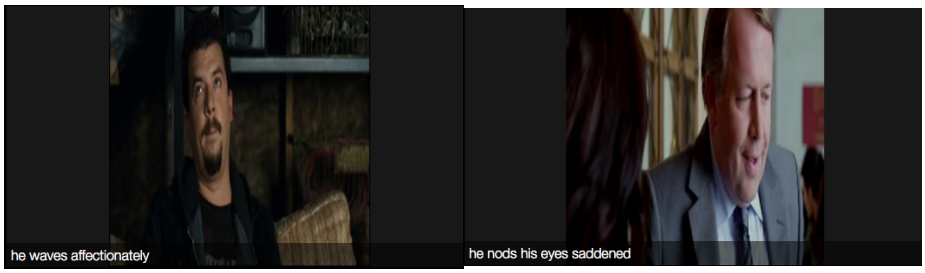

Figure 4: Examples of captions that have errors because of emotional words

A set of image captions that we inaccurate included images with emotional concepts as shown in figure 4. Firstly, not all videos that have an emotional component attached to them are annotated based on their emotional content. This is also related to length of the video where longer videos are annotated with more details and hence have associated emotions as opposed to shorter ones which focus on the main action in the video. We suspect the annotator's bias to be the primary reason for such an effect. When a caption is predicted with a particular emotion, it becomes increasingly challenging to verify whether the predicted emotion corresponds to the human emotion in the video. Since our model is aimed towards a general representation of videos rather than emotions in particular, we don't expect the model to perform well in terms of predicting the emotional aspect of the video. Emotion analysis might require a much deeper representation that takes face cues, change in body language, acoustic features such as tone change, etc. At this point of time, we decide not to focus on the emotional aspect of the videos and their corresponding captions.

6.4 Multiple action captions

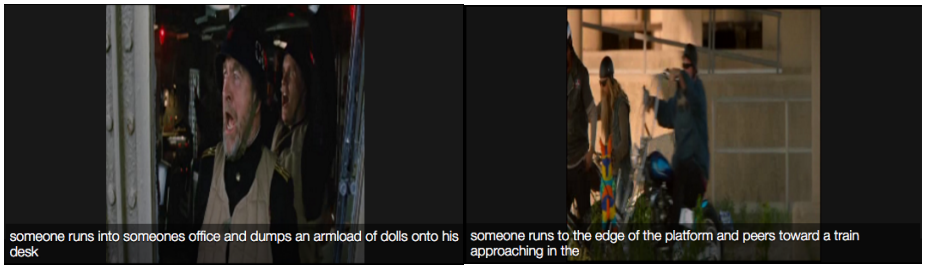

Figure 5: Examples of captions that have multiple actions

One of the most challenging aspects of video captioning is the task of describing multiple actions within the same caption as shown in Figure 5.

Even if in our work we train the model with the most relevant sentence from the captions, in most of the cases the captions portray multiple actions within the same sentence. This mainly occurs due to the nature of the video as it ha multiple events happening in the video clip. Generating this type of captions are more challenging and this is the main reason behind our hypothesis where training a model with single caption portraying one type of action might result in better video descriptions.

6.5 Partly right, partly wrong

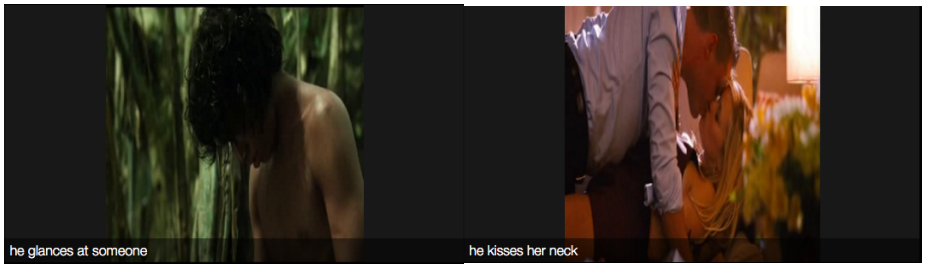

Figure 6: Examples of captions that are partly right/wrong

Some of the captions that were generated were partially correct. As you see in the first figure in figure 6, he glances at something. The video has no information about who or what he glances at. But the captions predict that *he glances at 'someone'*. Although the overall idea of the prediction is correct, we classify this caption as erroneous as its content is ambiguous. Another example of this error is shown in the second video shown in Figure 6, where the action *kiss* is present on the video however how the kiss is given is incorrect by saying *'kissing on the neck'*.

6.6 Errors in terms of granularity

Figure 7: Examples of captions that are too specific to any video clip

As we pointed out at the beginning of the section, a video can be described in multiple levels of granularity. As we observe in Figure 7, the first caption generated posses a large amount of detail while the second frame is described in a minimalist form with the sentence: 'he glances'.

It is debatable at which level of detail the captions should be generated for describing a video adequately, however we would like to point out that this is a potential error that might be caused by the original data with which the model is trained.

6.7 Incomplete information in captions

Figure 8: Examples of captions that have key missing information

If the caption generation model does not capture the key action in the video and manages to capture only minor details, then is is said that captions are incomplete. In the first caption of Figure 8, the action of 'kiss' is totally ignored, while in the second frame the 'chemistry lab' background is not included in the caption.

Perhaps stating that the second video shown in Figure 8 as incomplete information might be a little harsh, we would like to point out that this information is actually missing as it was originally found in the original captions, while some

others might not include it. This is an example of the issues found based on the caption writers' bias on how they describe each scene.

6.8 Gender mismatch

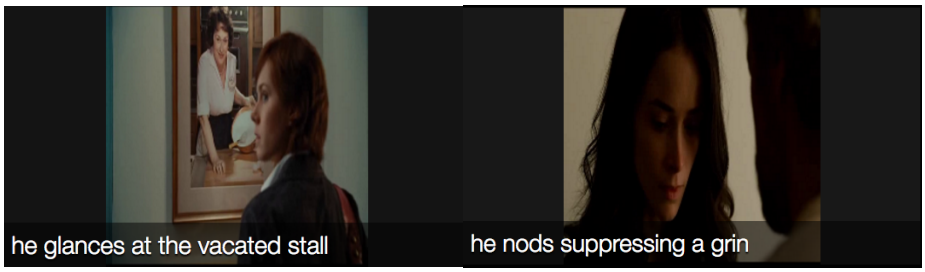

Figure 9: Examples of captions that got the wrong gender of the subject

Finally, another minor error found is the gender mismatch of the subject found in the video against the generated one in the caption as the ones found in figure 9.

7 Results and Discussion

Once our model was trained, we tested the performance based on the precision of the generated caption with respect to the original caption considering whether the generated caption contained the same verb or not. The test was using a subset of 16000 videoclips. Due to the diversity of the verbs, this result demonstrates that the model is indeed learning the actions according to the common patterns that could be found within videoclips of the same verb.

Our observations show that we often encounter not-so-meaningful patterns like *somemomma*, *someonesparkle* or *somejimmy* and cycled sentences as: *someone smiles someone smiles* or adding the prefix *someone* to proper names *someonedarius and someonehernandez watch the explosion grimly.* This result shows that datasets as M-VAD where proper names have been replaced by the word `SOMEONE` will overtrain the network with that term. As suggested by Correa et al. [3] our experiments might require to estimate a better number of hidden units alongside a different weight initialization method to avoid vanishing problems.

As shown in Table 1, it is seen that the quantitative scores are not as high as human annotated image datasets [2] and other baselines. Given the fact that an Image captioning based model (Neural Talk) performs much better than other state-of-the-art video captioning methods and our methods shows that we are working against a very strong baseline. As we discussed in the error analysis

Table 1: Quantitative Scores

	NeuralTalk	Seq2Seq Vid2Text	W2V CNN-LSTM (1 Layer)	W2V CNN-LSTM (2 Layer)
BLEU 1	0.563	0.050	0.018	0.018
BLEU 2	0.378	0.010	0.000	0.001
BLEU 3	0.216	0.002	0.000	0.000
BLEU 4	0.104	0.000	0.000	0.000
ROGUE	0.41	0.056	0.025	0.021
CIDER	0.361	0.022	0.002	0.004
METEOR	0.138	0.027	0.007	0.005

section 6, it is debatable as to whether we should use the same metrics for evaluation and make it a level-playing field for all such models.

One of our immediate next steps is to add a regularization term to the model as we suspect that the high dimensionality of the input and output might disrupt the learning phase of our model. The multimodal language model (W2V + LSTM) effectively trained a unigram model. Based on the qualitative analysis, we suspect that it learned the high frequency words as the most probable output in general. The <STOP> token that exists at the end of every sentence, made the model decode <STOP> more frequently than other words. After removing <STOP> token from the decoding vocabulary, we get words which cycle around. For example, one of the decoded sentences was "he up to he up to he up to ...". Firstly, the sentences are not able to learn a long term dependency, which is reasonable given the uni-gram training of the model. Secondly, words generated are high frequency words which indicate that the model learned the distribution of the vocabulary according to the captions.

In conclusion, introspecting the results made us understand our approach is in the right direction but still needs to address the challenges mentioned in the error analysis section 6 to overcome the common pit-falls that we faced in the language model.

8 Future Work

Currently, the major fallback of the approach is the lack of long term dependency in the grid based multimodal language model. We plan to include transfer of memory from each time-step to the next one and train it end to end. We believe that this could solve the problem of long-range dependencies and eventually obtain a better performing system in the end. It is to our knowledge, that a possible solution to the long-range dependency problem is to set a prior on the captions using a pre-trained language model that will impose constraints to the language with dependencies. Also, by using a beam-search over the caption-text-space could restrict the grammatical structure of the generated captions through the part-of-speech dependencies.

To overcome the vanishing problem with datasets that have a long-range dependency, Hinton et al. [10] suggest a *trick* (sic) where the weight matrix is the identity matrix and the bias are set to zero while considering ReLU as the activation function. This might be worth incorporating into our model since their results are up to par to the state of the art and reduce the complexity of the computation which is convenient due to the architecture of the multi-layered stack LSTM model we proposed.

References

[1] Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, et al. Video in sentences out. *arXiv preprint arXiv:1204.2742*, 2012.

[2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[3] Débora C Corrêa, Alexandre LM Levada, and José H Saito. Improving the learning speed in 2-layered lstm network by estimating the configuration of hidden units and optimizing weights initialization. In *Artificial Neural Networks-ICANN 2008*, pages 109–118. Springer, 2008.

[4] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[5] Nal Kalchbrenner, Ivo Danihelka, and Alex Graves. Grid long short-term memory. *arXiv preprint arXiv:1507.01526*, 2015.

[6] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[7] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, 2002.

[8] Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond J Mooney, Kate Saenko, and Sergio Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, volume 1, page 2, 2013.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[10] Geoffrey E. Hinton Quoc V. Le, Navdeep Jaitly. A simple way to initialize recurrent networks of rectified linear units. *arXiv:1504.00941*, 2015.

[11] Anna Rohrbach, Marcus Rohrbach, and Bernt Schiele. The long-short story of movie description. 2015.

[12] Amir Rosenfeld and Shimon Ullman. Face-space action recognition by face-object interactions. *arXiv preprint arXiv:1601.04293*, 2016.

[13] Shouvik Sachdeva and Ayush Mittal. Neural talk for videos. 2015.

[14] Rakshith Shetty and Jorma Laaksonen. Video captioning with recurrent networks based on frame-and video-level features and visual content classification. *arXiv preprint arXiv:1512.02949*, 2015.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[16] Mohak Sukhwani and CV Jawahar. Tennisvid2text: Fine-grained descriptions for domain specific videos. *arXiv preprint arXiv:1511.08522*, 2015.

[17] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015.

[18] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence–video to text. *arXiv preprint arXiv:1505.00487*, 2015.

[19] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

[20] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.

[21] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[22] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[23] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. *arXiv preprint arXiv:1510.07712*, 2015.