Image Captioning Using Motion-CNN with Object Detection

Akshay Sreekumar Nair
Information Technology
National Institute of Technology Karnataka
Surathkal, India
a.191it103@nitk.edu.in

Kausthub Thekke Madathil
Information Technology
National Institute of Technology Karnataka
Surathkal, India
kausthubtm.191it125@nitk.edu.in

Sohanraj R
Information Technology
National Institute of Technology Karnataka
Surathkal, India
sohanrajr.191it149@nitk.edu.in

Dinesh Naik
Information Technology
National Institute of Technology Karnataka
Surathkal, India
din_nk@nitk.edu.in

***Abstract*—Image captioning is the process of generating a textual description of an image. Image captioning is a topic which involves computer vision and natural language processing and is very active research topic with researchers trying to improve the quality of the captions generated. The automation of image captioning has several important applications like assisting visually impaired people by depicting the visual content, index imaging of images on the internet. Given an image, the aim is to extract information and generate a valid caption for the image. The current method used includes image captioning with visual attention. Traditional image caption generation models just map the image features to the corresponding caption and is incapable of understand verb and ignore the motion features. Also, in this project we have considered object detection to extract only the necessary features and ignore the background features to improve the performance of the model. This project discusses the implementation of a image caption generating model with image features, motion features and object detection combined. The proposed model is trained on two datasets : MSCOCO and Flickr8k datatsets and the model is evaluated using BLEU-N metrics, METEOR and the results and analysis section shows that our proposed model improves quality of the captions generated.**

***Index Terms*—image captioning, object detection, deep learning, motion estimation, neural networks**

I. Introduction

Image captioning is the process of generating a textual description of an image. Image captioning is a topic which involves computer vision and natural language processing and is very active research topic with researchers trying to improve the quality of the captions generated. Image captioning is a widely used tool in areas like google images, social media and to generate description for visually impaired people. Previously, the captions for images where done manually and is a very tedious job and the need for automatically generating captions became the need of the hour.

Since, image captioning falls under the category of sequence to sequence model the encoder-decoder architecture is the most widely and commonly used architecture. The earliest automatic image captioning model used basic CNN model for extracting the image feature (encoder) and passes the feature vector to the RNN (decoder) to generate the corresponding captions. Later, to improve the quality of the captions generated they introduced attention or object detection to obtain specific important regions.

Even though the above model is very efficient in extracting the image features but it is capable of only mapping the image features with the corresponding image caption. The model incapable of understanding verbs and utilizing the motion features to generate the captions. In this project we try to introduce a motion-CNN model to extract and utilize the motion features along with the image features to generate a highly accurate image captions as compared to the traditional models. The motion-CNN extracts all the motion features even the background features but the unnecessary background features may reduce the performance of the model hence we utilize object detection to only extract the motion features from the important regions or the regions around object only ignoring the background motion features.

To summarize in this project we try to combine image features, motion features and object detection to generate highly accurate captions.

II. Related Work

In the paper [2] a cross-modal embedding method is learned for the images, topics, and captions. In the proposed framework, the topic, caption, and image are organized in a hierarchical structure, which is preserved in the embedding space by using the order-embedding method.
The paper [3] proposes a news image captioning method based on the attentional encoder-decoder model through

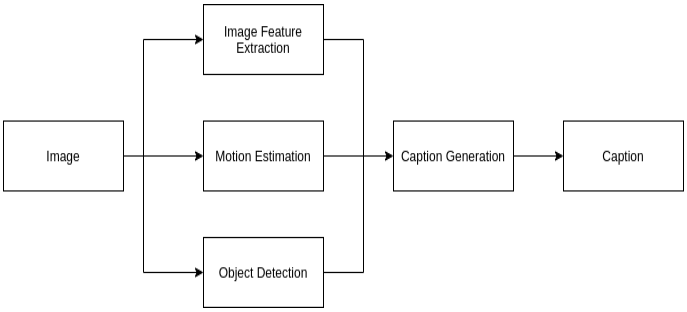

Fig. 1. Flowchart or components of the proposed method

summarizing the news text according to query image. The multi-modal attentional mechanism is proposed to compute the context vector. The proposed model is trained on the DailyMail news image captioning corpora which are created by collecting images, caption, news texts through parsing the html-formatted documents.

The paper [4] proposes a new framework which is a combination of generation and retrieval based image captioning. First a CNN-RNN framework combined with beam-search generates multiple captions for a target image. Then the best caption is selected on the basis of its lexical similarity with the reference captions of most similar images.

III. Problem Statement

To generate improved image captions using encoder-decoder model by leveraging motion features and object detection

- Most of the present image captioning methods are aimed at direct learning of relations between image features and corresponding image captions. However, this is not relevant for verbs that sometimes have no meaningful relation with image features. Hence, we must extract motion features from the image as well.
- Captions mainly describe contents of objects in the image. But when estimating, many background features are included than the object features. Also, the accuracy of the motion estimation is not always high. Hence, we must use object detection and use only motion features around the object region of an image.

IV. Objectives

The objectives of this project are :

- To generate image captions.
- To improve the quality of captions by considering motion features and the relation between motion features and verbs.
- To improve the quality of caption by considering only the motion features around the object region and ignoring background motion features using object detection.
- To try different datasets / models / parameter tuning and do a comparative study.

V. Methodology

Image caption generation is an example of sequence to sequence conversion problem to which the most commonly used solution is the encoder - decoder architecture. Even we have considered the same encode - decoder model. In the proposed model the encoder component involves extraction of the image features, motion features and object detection and passes the features extracted to the decoder component which generates the captions with the help of LSTMs assisted by attention architecture. Since, the decoder component in the proposed is quite similar to the traditional model we have not discussed it in detail, whereas we have given elaborate explanation for the encoder component which involves almost of novelty proposed.

A. Concept

Most of the present image captioning models have components like object detection, feature extraction of images and caption generation. These help improve performance of the model using specific sections of the image or the image overall. But, there is a catch. The feature extracted may not have meaningful relation with the image features. The model may confuse movement of legs in an image as running whereas the person might be swimming, in which legs assist in movement. Taking all such things into account a model [5] that inculcates the concept of motion estimation was proposed. The proposed model obtains motion feature from the background of the image. It can learn the relationship between the motion features obtained the verbs. The use of motion features from the image background have some drawbacks:

1) Background features don't hold much importance here because image captions in general describes the objects of the image. Extracting more background features may in turn hamper the caption generation task, which in turn might reduce the accuracy. Thus, important parts of the image are to be examined thoroughly.
2) Also, motion estimation might not be highly accurate always. Some motion features extracted might contain some errors. This error might propagate which can decrease accuracy of the model. Using all the motion features can further increase errors.
   This issue can be resolved by using the motion features wherever objects are detected in an image.These regions can be examined to generate the appropriate captions.

This paper proposes the use of image feature extraction, motion estimation, object detection and caption generation, in the model. Each of the components used gives some specific advantages.

- The image feature extraction component helps obtain important image information of objects present in the image.
- Object detection helps detect object specific regions of the image which in turn helps the motion feature in capturing required information

Authors	Methodology	Merits	Limitations	Additional Details
Iwamura, Kiyohiko, Jun Y. Louhi Kasahara, Alessandro Moro, Atsushi Yamashita, and Hajime Asama	Using Motion-CNN with Object Detection	eliminate unnecessary motion features	object detection sometimes generates bounding boxes that cover all image regions	evaluated using three datasets : MSR-VTT2016-Image, MSCOCO, and several copyright-free image
G. Hoxha, F. Melgani and J. Slaghenauffi	CNN-RNN framework combined with beam-search and lexical similarity	proposed framework results are slightly higher with respect to the simple encoder-decoder architecture	Some errors in image captions	Dataset used: RSICD dataset
N. Yu, X. Hu, B. Song, J. Yang and J. Zhang	CNN-based multi-label classifier; cross-modal embedding method	achieved competitive results with the state-of-the-art methods	Results inferior to some earlier models	MS-COCO and Flickr30K datasets used for validation
J. Chen and H. Zhuge	attentional encoder-decoder model	proposed method outperforms the generic image captioning and the text summarization method	Comparison of models only w.r.t. new image captioning	Experiments on the DailyMail test dataset

TABLE I
SUMMARY OF LITERATURE SURVEY

- Motion estimation enables the model to learn relation between image features and verbs
- Caption generation eventually helps generate appropriate captions

B. Architecture

The proposed model consists of various components as described in the Introduction section above namely image feature extraction, object detection, motion features and caption generation. The use of object detection assists motion features to extract information around the regions of objects in the image.

In the motion estimation component, motion-CNN is used. Since this model extracts motion features of foreground, background of the image, or all the motion features, the object detection component has been considered.

Figure 1 shows an overview of the model. All the components, Image Feature extraction, Object Detection, Motion Estimation and Caption Generation are represented in boxes. Image is initially passed to the image feature extraction component to extract features. Also, it is fed to Object Detection and Motion Estimation component, where it is passed to CNNestimate. It estimates the associated motion features and objects present in the image. Then, CNNextract helps extract the motion feature. These information are then passed to the Caption Generation component. Attention features are calculated using the Attention module. Attention features are then concatenated, fed into LSTMs. Appropriate caption according the input are generated.

C. Object Detection

Object detection is a computer vision and image processing problem which primarily deals with finding object regions from the input image. The identified regions are represented as bounding boxes which specifies the region's dimensions

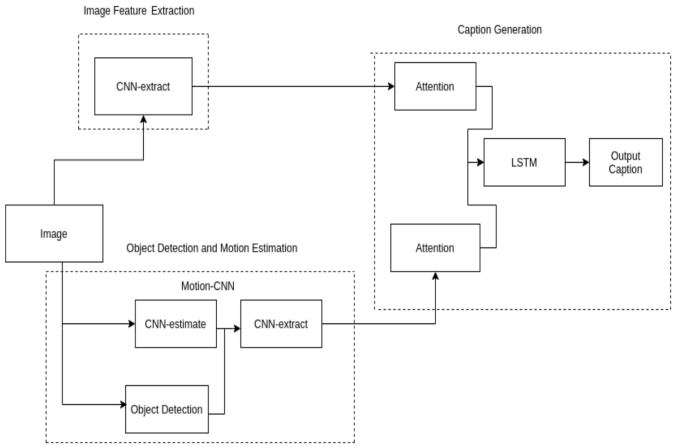

Fig. 2. Proposed Model

and its position. For the purpose of object detection we have used Faster R-CNN which is the most commonly used and sought out model. In short we pass an image to the Faster R-CNN model which identifies the objects and returns bounding boxes which represents the objects or the object regions. These bounding boxes are passed as input to the CNN-extract.

D. Motion CNN

1) Motion Encoding: Optical flow is a feature that encodes the apparent patterns of motion of objects in any visual scenarios, and is one of the most common and direct information of motion used in the process of action recognition. Tow grayscale images are often used to represent the optical flow. These matrices are used to encode quantized horizontal and vertical displacements. In recent times all the deep neural pre-trained models take a 3-channel RGB image as the input. So the two images are added with a

dummy third channel containing redundant values or all zeros. The base paper have used this form of input for their learning. They have also stated that directly predicting the optical flow of an image from the image vectors from all the pixels is a highly under-constrained problem, and they hypothesize that detangling the direction of flow and the strength may be an easier path . Therefore the optical flow is split into two components angle θ and mangnitude **A**.The prediction of flow is formulated pixel wise as a regression problem. Therefore the prediction of θ is inappropriate as the dimension of θ is circular(where 2Π is same as 0). so, θ is dived into two components horizontal and vertical angle which are represented by sin(θ) and cos(θ). We in the end conclude with a simple 3-channel encoding flow image **F**.

$$F_1 = sin(\theta) = v/A; F_2 = cos(\theta) = u/A; F_3 = A.$$

the motion encoding scheme that we have used has the following benefits: 1)The optical flow is divided into 3-channels namely vertical angle,horizontal angle and magnitude of motion. This makes it easy for a high dimension motion prediction problem to be more factored
2) Because sin(θ) and cos(θ) are non-circular and fall in the range [1,1], regression of angle is possible; 3) Encoding motion as a 3-channel picture makes its application efficient, easy, and suited for our framework, which is described next.

2) Im2Flow: The aim of our experiment in this section is to create a model that takes a static image as an input and give us a 3-channel optical flow map of motion features as output. Now let us consider X to be the domain consisting of all the static images and Y to be the set of corresponding flow maps.Now we need to create a model function such that G(X)-¿Y. During training we are given a set of static images as labeled pairs $x_i, y_i{}_{i=1}^{N}$ where x_i consist of static images and y_i the corresponding flow images. The base paper shows that they have devised a CNN(convolutional neural network) called as IM2Flow to obtain our model function G. they have adapted a U-net model architecture with a few modifications. Convolution-BatchNorm-ReLu is used by both the encoder and the decoder part. Dilated convolutions are used in the encoder part. Dilated convolutions expand their receptive field size exponentially while maintaining spatial resolution, allowing them to record long-range spatial relationships. Skip links link the encoder and decoder, allowing low-level information to be directly shuffled throughout the network, which is critical for our dense motion prediction issue. combination of two loses are being reduced in our Im2Flow network. Namely pixel error loss and motion content loss:

$$L = L_{pixel} + \lambda L_{content}^{\phi.j}$$

The agreement with the genuine flow is measured by pixel loss:

$$L_{pixel} = E_{p,q \in x_i, y_i{}_{i=1}^{N}} \left| |y_i - G(x_i)||_2 \right|$$

The motion content loss helps in retaining the very high level motion features.On the UCF-101 dataset , we fine-tune an 18-layer ResNet (pre-trained on ImageNet) for action classification utilising flow pictures as input to mimic realistic motion. The resulting network is ϕ . This ϕ is used to calculate $L-1$ and L_2 to get the final model function. The network is fine tuned to predict based on the predicting pixels that actually make some motion. The gradients of the first two channels are weighed against the motion magnitude.

E. Motion CNN with Object Detection Architecture

Its is mentioned in multiple sections before that the proposed model combines motion CNN with object detection. For object detection we have considered Im2Flow model which extracts or estimates motion of the entire image including the background details. The motion estimation is by calculating or findindg the optical flow in the image. Optical flow contains necessary details about the motion. The background motion features lead to a decrease in performance of the model. Hence, we perform object detection simultaneously to consider only the regions surrounding the objects and important regions. The rest of the part is set as 0 and thus ignoring the backgorund details and improving the model performance.

F. Caption Generation Component

In the caption generation component, attention features are calculated using Bahdanau Attention. Then, concatenated attention features are inputted to LSTMs and the corresponding caption is generated.

VI. Discussion and Conclusion

A. Dataset Description

We have used MS-COCO [Microsoft - Common Objects in Context] dataset which consists around 83,000 training images, each of which has at least 5 different caption annotations which adds up to a training dataset size of 415,000+. The validation set consist of over 41,000 images each of which has at least 5 different caption annotations which adds up to a validation dataset size of 200,000+. The other dataset we have considered is Flickr8k dataset which consists of 8,000 images which we have split into 6000 training, 1000 for validation and testing each. The dataset consists 5 different captions for each image which counts to 40460 captions. We have used this dataset mainly for performance comparison with MS-COCO dataset.

B. Experimental Procedures

The experiment makes use of two datasets namely - MS-COCO and Flickr8 as described in the Dataset Description section. Both these datasets consist of at least a set of 5 different captions for the image chosen.
This paper makes use of two models - the traditional model which uses InceptionV3 for encoding and feature extraction, and GRUs for decoding(or caption generation), and the

proposed model which makes use of motion estimation and object detection for image caption generation.
- equations for metrics
The model's performance was evaluated using the $BLEU_n$ metric [7]. BLEU-N metrics are calculated by the formula given below -

$$BLEU_n = min(1, e^{1-r/c}) * e^{(1/N)*\sum_{n=1}^{N} logp_n=1}$$

where r represents the reference sentence, c denotes the generated sentence, and p_n is the modified n-gram precision [1].
Other metrics used in this paper are METEOR metric [5] and ROUGE-L metric [6].

C. Metrics used for evaluation

BLEU - Bilingual Evaluation Understudy is used to evaluate how close the machine translation is to the human reference translation.
METEOR - Metric for Evaluation of Translation with Explicit ORdereing is another such metric used for evaluation. METEOR modifies the precision and recall computations.
ROUGE - Recall Oriented Understudy for Gisting Evaluation is mostly used for the summary evaluation. ROGUE is based on recall. ROGUE-L is taken into account for evaluation in this paper. It is based on the concept of longest common subsequence(LCS).

D. Results

Table 2 shows the results obtained by the proposed model and a comparison between the proposed model and the traditional model. Higher the values are the better the model is w.r.t MS-COCO dataset. The proposed model outperformed the traditional model as evident from values in the table. The use of motion estimation and object detection have helped achieve good results.

Model	BLEU 4	METEOR	ROUGE
Traditional Model + Attention	0.734	0.239	0.557
Proposed Model	0.752	0.267	0.558

TABLE II
RESULTS FOR MS-COCO DATASET

The same experiment as described above has been conducted for the Flicker8 dataset. The proposed model again outperformed the traditional model as expected from the results in the first experiment.

Model	BLEU 4	METEOR	ROUGE
Traditional Model + Attention	0.512	0.379	0.457
Proposed Model	0.552	0.367	0.498

TABLE III
RESULTS FOR FLICKR8K DATASET

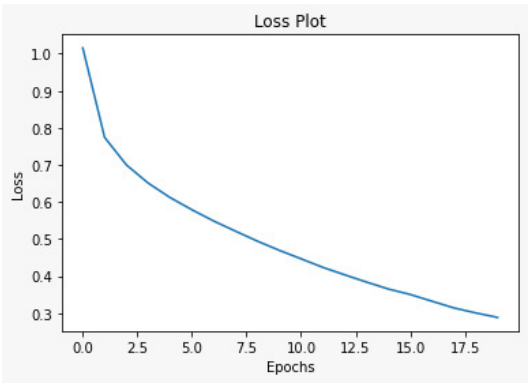

Fig. 3. Loss v/s Epoch

Figure 3 shows the Loss versus number of Epochs curve.

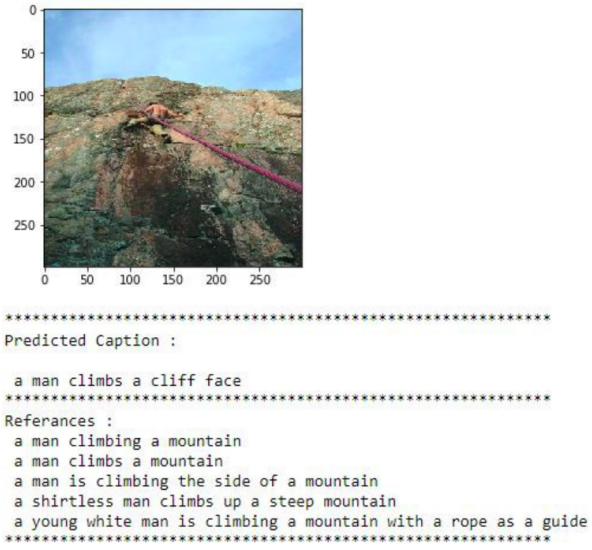

Fig. 4. Validation Image - 1

Figure Fig. 1 displays the input image and predicted caption for the input image. Some reference attributes used to predict the caption are also shown.

Fig. 5. Validation Image - 2

Figure Fig. 2 displays another input image and predicted caption for the input image. Some reference attributes used to predict the caption are also shown.

Fig. 6. Test Image

Caption generated for the test image is ”A guy on a surfboard in the ocean”

VII. Limitations

There were a few errors during predictions of captions on some images. Fig. shows a group of some children playing in a sand box, whereas the captions predict that a child in a stripe t-shirt is jumping in the air, which is a very far-fetched prediction from the original label. This might be due to some attention model issue where the model concentrates on one aspect of the image which weighs more into the result.

Fig. 7. Bad Test Case - 1

Fig. 8. Bad Test Case - 2

VIII. Conclusion

Our project experimentation was targeted to produce captions for images using basic image captioning and motion estimation using a pre-trained model Im2Flow. Image captioning is important as it helps in bringing out meaning to that images when the scenario is not known. These captions generated using NLP techniques helps in giving a better description of the image to visually impaired people. The motion estimation produces an image which determines direction of motion. This motion estimation helps in producing the captions more accurately. This combination of basic image features and motion estimation produces an accurate model for producing captions. Object detection is implemented into our model to

employ motion characteristics, as stated in this study. Object detection, on the other hand, can provide bounding boxes that encompass all picture areas, as discussed in the preceding discussion.Our future research will focus on how to better incorporate motion aspects.

REFERENCES

[1] Iwamura, Kiyohiko, Jun Y. Louhi Kasahara, Alessandro Moro, Atsushi Yamashita, and Hajime Asama. 2021. "Image Captioning Using Motion-CNN with Object Detection" Sensors 21, no. 4: 1270. https://doi.org/10.3390/s21041270

[2] G. Hoxha, F. Melgani and J. Slaghenauffi, "A New CNN-RNN Framework For Remote Sensing Image Captioning," 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), 2020, pp. 1-4, doi: 10.1109/M2GARSS47143.2020.9105191.

[3] N. Yu, X. Hu, B. Song, J. Yang and J. Zhang, "Topic-Oriented Image Captioning Based on Order-Embedding," in IEEE Transactions on Image Processing, vol. 28, no. 6, pp. 2743-2754, June 2019, doi: 10.1109/TIP.2018.2889922.

[4] J. Chen and H. Zhuge, "News Image Captioning Based on Text Summarization Using Image as Query," 2019 15th International Conference on Semantics, Knowledge and Grids (SKG), 2019, pp. 123-126, doi: 10.1109/SKG49510.2019.00029.

[5] Denkowski, M.; Lavie, A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In Proceedings of the EACL 2014 Workshop on Statistical Machine Translation, Baltimore, MD, USA, 26–27 June 2014; pp. 376–380.

[6] Lin, C.Y.; Cao, G.; Gao, J.; Nie, J.Y. An Information-theoretic Approach to Automatic Evaluation of Summaries. In Proceedings of theMain Conference on Human Language Technology Conference of the North American Chapter Association for Computational Linguistics, New York, NY, USA, 4–9 June 2006; pp. 463–470.

[7] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–

[8] Hui, T.W.; Tang, X.; Change, C. A Lightweight Optical flow CNN-revisiting Data Fidelity and Regularization. arXiv 2019, arXiv:1903.07414.

[9] Ansar Hani; Najiba Tagougui; Monji Kherallah. Image Caption Generation Using A Deep Architecture. Publisher : IEEE

[10] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1473–1482, 2015

[11] Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 375–383.

[12] Gao, R.; Xiong, B.; Grauman, K. Im2Flow: Motion Hallucination from Static Images for Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5937–5947.

[13] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

[14] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.B.; Li, F.F. Imagenet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. 2015, 115, 211–252. [CrossRef]

[15] Pennington, J.; Socher, R.; Manning, C.D. GloVe: Gloval Vectors for Word Representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.

[16] Xu, J.; Mei, T.; Yao, T.; Rui, Y. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5288–5296.