Visually Guided Knowledge selection for Video Captioning

Ayush Gupta
Ashrya Agrawal
Poonam Goyal
Navneet Goyal
{f20180203,f20180210,poonam,goel}@pilani.bits-pilani.ac.in
ADAPT Lab, Birla Institute of Technology and Science
Pilani, India

ABSTRACT

Video captioning is a challenging task of modelling the objects, their temporal information and interaction in order to generate a textual description. Current models often fail to model these objects and their interactions correctly, due to lack of knowledge about them. In this paper, we propose approaches to provide this knowledge through knowledge bases like wordnet and conceptnet. We propose general encoder and decoder modules, which can be used on the top of any architecture to insert knowledge. Leveraging the advancements in attention architectures, we develop knowledge selection mechanism for the above modules. We demonstrate the efficacy of our model by extensive experiments on two benchmark datasets, MSVD and MSRVTT. The proposed model demonstrates better semantic consistency and makes significant improvement over the baseline. Our approach not only helps in object modelling, but also helps in further improving action prediction, as demonstrated in Figure 1.

ACM Reference Format:
Ayush Gupta, Ashrya Agrawal, Poonam Goyal, and Navneet Goyal. 2021. Visually Guided Knowledge selection for Video Captioning. In *Proceedings of* . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Video Captioning is a core task in vision-language research. An input video is used to automatically generate a natural language description for it. This task is challenging as it involves both, text modality and vision modality along with the time dimension. The most common architecture for Video captioning is the encoder-decoder architecture. The encoder module generates semantic representations of videos using frame-wise features, motion features, object level features etc. The decoder uses this semantic representation to generate a sequence of tokens as the natural language description of the input video.

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

A good video representation must have (i) Global context and (ii)Regional context. Global context across the spatial dimension can be captured using 2D CNNs. It is the encoding of the overall scene in a specific frame, including background and foreground in one single representation. Activities by subject and movement of objects cannot be captured by looking at one frame. Such activities can be encoded using 3D CNNs, which look along the temporal dimension in a video. These also form a part of the global context in the video representation.

Certain aspects of a video cannot be recognised by looking at the video as a whole. Specialised CNNs, in the form of object detectors, can be used to extract local object-level information to add to the video representation. These object detectors ignore the global information and just focus on a small area, called as ROIs, computing local features and adding them into the semantic video representations. For eg. consider a cropped image of a man. If we don't look at the full image, we don't know whether he is a cricketer/chef; but we can easily decide what he is based on the global information i.e. background of the kitchen or the cricket field. As a result, the local object-level information is poor and the resulting captions have poor diversity.

Structured knowledge graphs like ConceptNet and WordNet [6] are increasingly gaining popularity for NLG tasks. They provide a way to input real-world, structured information and rules in generated sentences.

Our proposed components are built upon SAAT[15]. SAAT explicitly predicts the actions to provide extra guidance apart from linguistic prior. We improve upon the object-level information by leveraging the real-world knowledge and rules from Knowledge graphs, guided by the context from global visual features of the video. We propose three different architectures to inject this additional information into the video representations.

Attention architectures have gained popularity in a wide variety of domains like Images, NLG, Reinforcement Learning. In this paper we use different variants of cross-attention to select external knowledge for text generation module. In our work, we focus on improving the regional context in a captioning model. We enhance the local information by injecting KB knowledge using global features in our VTKE module. The Knowledge base provides external knowledge by optionally being guided by the visual information.

In summary, the major contributions of our study are as follows:

- We propose different architectures to leverage external knowledge base(s) in vision-language tasks.
- We compare the effects of using different knowledge bases for the video captioning task

Baseline: a group of men are playing
Ours: a group of men are racing on a track
GT: a group of men compete in a track race

Baseline: a man is cutting a tomato
Ours: a man is cutting a piece of meat
GT: A man is cutting meat

Figure 1: Motivating examples for use of Knowledge Insertion in Video Captioning

- We compare the diversity in captions on using different object detectors

2 RELATED WORK

Video captioning: In earlier works on Video Captioning used template-based approaches[1, 10], where prominent. With the advancements in Attention[11] architectures, encoder-decoder based architectures[14, 15] took over the template based approaches. Video captioning has been of interest to a wide audience due to transferability to other vision-language tasks like Visual Question-Answering, Embodied Vision, Text based navigation, and so on.

Object detection: Object detection is a widely popular task in Computer Vision research, particularly because it is a sub-component of various architectures. While object detection research has been focused on predicting labels from a limited set of object categories [4], we are primarily interested in object detectors with large set of object categories like Yolo-9000 [9]. The progress on diverse object detection like Oscar, etc can be leveraged to further improve the results from our approach.

Attention and Transformers: Attention[11] has primarily been used to selectively attend to parts of a sequence to obtain probabilities corresponding to parts being attended. Transformers have been central to multiple breakthroughs in deep learning, because of their property to train parallely on GPUs. Recent breakthroughs in attention architectures are driven by (i) the work on Image-GPT by OpenAI, (ii) VilBert (iii) numerous other models of attention developed for Vision-Language research. In our study, we developed attention architectures, which can be further extended to build transformer architectures for selectively providing external knowledge to the base model.

Lexical Knowledge Bases: WordNet, ConceptNet[5], Dbpedia [3], NELL[7] are some of the commonly used knowledge bases. They are built mainly built using textual information. Thus, usage of knowledge base like VTKB, built using both visual and textual information, can further enhance the performance of proposed architectures on vision-language tasks. While WordNet and ConceptNet are manually constructed, NELL is automatically constructed from web. While these knowledge bases have been used for image captioning, they have not been used for video-language tasks to the best of our knowledge.

3 METHODOLOGY

3.1 Task Description

Given an input sequence of frames $\mathbf{F} = \{\mathbf{F_1}, ..., \mathbf{F_n}\}$, the task of video captioning aims to generate a sequence of tokens $\mathbf{S} = \{\mathbf{S_1}, ..., \mathbf{S_m}\}$ as the natural language description. We uniformly sample k frames and generate a sequence of 2D features using ResNet101, $\mathbf{f} = \{\mathbf{f_1}, ..., \mathbf{f_k}\}$. For a fixed c, we select a c^{th} frame, and apply object detector to get ROI features, bounding box coordinates and object labels $\mathbf{O} = \{\mathbf{O_1}, ..., \mathbf{O_L}\}$. Using the knowledge base KB, we obtain the related words of each object O_i and obtain the KB output as $\mathbf{R} = \{\{\mathbf{R_{1,1}}, ..., \mathbf{R_{1,KB_{max}}}\}, ..., \{\mathbf{R_{i,1}}, ..., \mathbf{R_{i,KB_{max}}}\}...\}$. A word embedding $\mathbf{E}$ is used to compute the representation $E[R_{i,j}]$ for each word $R_{i,j}$, which is used by the encoder and subsequently passed to the decoder.

Encoder's output is first used to generate SVO (subject, verb, object) tuples, which are subsequently used by decoder, along with the video's encoding to generate the caption. The architectures proposed by us can be classified into two types, depending on whether they are used in SVO generation or used directly for sentence generation. Architectures in section 3.2.1 and 3.2.2 belong to the former category, while 3.3.1 belongs to the latter.

We fetch KB_{max} related words for each detected object in the input video. These related words' embeddings for each of N words, result in a large size vector. This vast raw form of external knowledge is not directly usable by the model and a more efficient representation is required. To this end, we propose two techniques to encode raw knowledge embeddings into knowledge representations: (i) Context-Free Knowledge Targeting, and (ii)Visio-Textual Knowledge Embedding.

3.2 Encoder

Inserting External knowledge in the encoding phase helps in generating SVO triplets.

3.2.1 Context-Free Knowledge Targeting. This technique utilizes a 1D convolution over the stacked knowledge embeddings. Given the external knowledge vector R, we stack embeddings to form a large vector

$$R_{stack} = \{\{R_{1,1}; ...; R_{1,KB_{max}}\}....\{R_{N,1}; ....; R_{N,KB_{max}}\}\}$$

Running a 1D convolution across this vector gives us the external knowledge representation for each object

$$K = 1D\_Conv(R_{stack})$$

Setting the stride s and window size w appropriately can reduce the dimension of the external knowledge so that it can be utilised by the decoder.

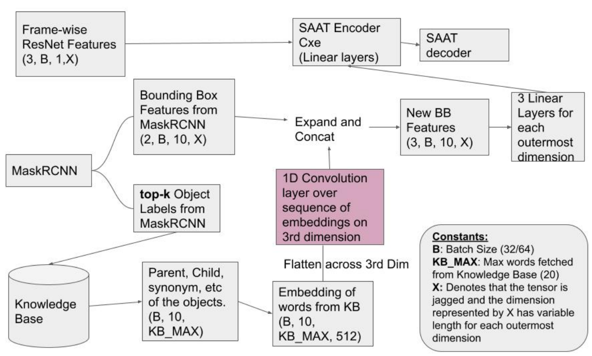


Figure 2: Context-Free Knowledge Targeting Architecture

3.2.2 KB Guided Visio-Textual Knowledge Embedding. This technique aims to generate a knowledge representation guided by the visual context in the frame. The c^{th} frame is chosen to provide the context. 2D CNN features of the chosen frames are used as a query to select from the available list of keys, which are the words R. The 2D CNN features have size F. The KB guided visio-textual knowledge embedding can be obtained by:

$$Q = W_q * q$$
$$K = W_k * k$$
$$a = softmax(Q^T K) \tag{1}$$

where q is the frame's 2D feature, k is a vector containing word embeddings and a is the final knowledge selection output. The matrices W_q and W_k are learnable parameters, which map the query and keys to a semantic $d-dimensional$ space. d is a hyperparameter which decides the size of the external knowledge representation in this semantic space.

The output a can be interpreted as the relevance score for each of the KB_{max} words based on the visual context. This can be used to compute the visio-textual knowledge embedding KB_{vt}

$$KB_{vt} = a * k^T \tag{2}$$

3.2.3 GCN Guided knowledge selection. Using the output from knowledge base, we construct star graph for each object, with O_i at centre and $R_{i,1}, ..., R_{i,KB_{max}}$ at the other ends of edges. Each of the $KB_{max} + 1$ nodes is represented by the embedding of word corresponding to that node. The graphs are passed through GCN, Relu, GCN and at last softmax to obtain a sequence of probabilities. These probabilities are further used to obtain the weighted external knowledge $KB_{GCN,i}$ from the graph of a given object. External knowledge vectors are stacked and passed to the the decoder.

3.3 Decoder

Inserting external knowledge in the decoder helps in selecting more diverse words in the final captions.

3.3.1 KB Guided Knowledge Selection. This technique utilizes the attention architecture to select knowledge from KB. Using frame-wise features as q, words fetched from Knowledge base as k and v, we compute $Q = W_q * q$, $K = W_k * k$, $V = W_v * v$

$$\alpha = softmax\left(\frac{Q^T K}{\sqrt{d}}\right)$$
$$KB_{vec} = \alpha * V \tag{3}$$

The KB_{vec} matrix represents the external knowledge selected using the visual features. We also leverage recent advancements in attention like the Multi-head attention[11] to further improve the knowledge selection. Multiple heads ensure that our architecture can focus on multiple frames to select knowledge. The default number of heads used for our study is 8, unless otherwise stated.

4 EXPERIMENTS

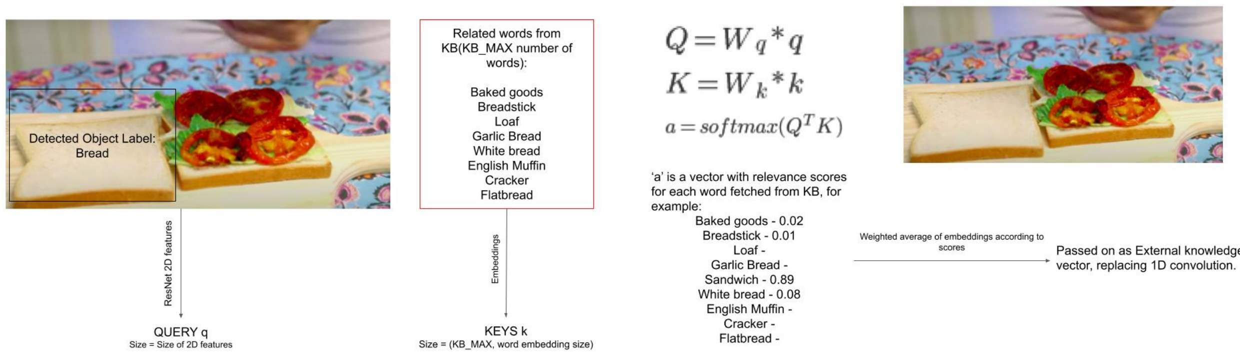
4.1 Datasets

MSVD: MSVD is a collection of short YouTube videos collected by Amazon Mechanical Turk (AMT) workers. The videos depict a single activity and are 10-15 seconds long. Each clip is annotated with 40 captions. Following the standard split, we use 1200 clips for training, 100 for validation and 670 for testing.

MSR-VTT: MSRVTT is a widely used benchmark for vision-language downstream tasks like video captioning. We use the initial version of MSRVTT which consists of 10K video clips categorised into 20 domains. Each video has 20 annotations performed using Amazon Mechanical Turk (AMT). We use the standard split [13] - 6513, 497 and 2990 clips for training, validation and testing respectively. MSR-VTT has higher diversity in the vocabulary and hence is better suited for our approach. Due to the larger size of the vocabulary, external knowledge is better incorporated in the model for this dataset. Further, the captions of MSRVTT are more diverse, which creates some more scope for external knowledge to have an effect on the generated captions.

4.2 Evaluation Metrics

We use the CIDEr [12] score to evaluate our model and optimize for hyperparameters. CIDEr focusses on consensus based evaluation, rating captions higher when they are similar to how other people describe the video. On the other hand, BLEU-n scores focus on n-grams to make captions similar to the ground-truth. We observe that if we optimize for BLEU-4 [8], the performance on other scores



(a) Extraction of Query and Keys **(b) Attention Computation**

Figure 3: Mechanism of KB Guided Visio-Textual Knowledge Embedding

deteriorates faster than when we optimize for CIDEr. For sentence generation, Cross Entropy loss is considered during training. We train using this loss function and select the best model we obtain using the the CIDEr score on the validation test.

4.3 Implementation Details

Data pre-processing, values of important relevant hyperparameters like batch size. Mention about usage of ResNet, 3D CNN, etc and other stuff done for preprocessing.

We use Resnet[2] 2D CNN features. A fixed number of equidistant frames are extracted from each video and a feature representation of each of the frames is mean pooled to get the final 2D visual features. For extracting temporal information, we use I3D as the 3D CNN and obtain feature representation of the videos. Apart from these global features, we experiment with various object detectors to obtain object labels and the ROI-bounding box features to obtain local information. All these correspond to the representation we get from a given input video. We also observe that λ is an important hyperparameter of the loss function, deciding the weightage of the SVO loss with the Sentence generation loss. We perform hyperparameter tuning on λ and observe optimal performance in the range 14-17. Videos are processed in mini-batches. We set the batch-size for MSVD as 8, and for MSRVTT as 20. We fix the number of heads in MultiHead Attention to 8 in most of our experiments. As evident from Table 1 and 2, word embeddings have a great effect on the scores. We performed experiments with learned/parametric embedding and word2vec.

4.4 Results

On the MSVD dataset, Visio-Textual knowledge embedding, when used with Multihead attention, wordnet and the yolo-9000 object detector, we obtain our best CIDEr score of 83.85, which is considerably higher than the Base SAAT score of 78.08 obtained after re-training the model using open-sourced code. This model was trained using Reinforcement Learning strategy of sequence critical sequence training.

On MSRVTT, Graph Convolution Network, when used with Multihead attention, wordnet and the yolo-9000 object detector, we obtain our best CIDEr score of 49.84, which is higher than the Baseline SAAT model's score of 49.21.

4.5 Ablation Study

We observe that Reinforcement Learning significantly boosted the CIDEr scores only in the case of MSVD dataset, but not for MSRVTT. Further, among encoders, Knowledge Guided Visio Textual Embedding performed the best. Among decoders, MultiHead Attention gave the best scores, but when no encoder was being used. This might be because injecting external knowledge at multiple places confuses the model.

We compare the effects of learned embeddings and pre-trained word2vec embeddings, and see that the former is significantly better. We also observe the effect of Graph Convolutional Networks in the encoder. Using a GCN in MSRVTT gives marginally better results. But in MSVD, the best performing models were obtained by Knowledge guided visio textual embedding in the encoder part, along with RL training.

5 CONCLUSION

In this paper, we propose knowledge insertion mechanism for video captioning models, which uses external knowledge bases to improve modelling of objects and their interaction. We also propose plug-in encoder and decoder modules to leverage the external knowledge. Additionally, we propose attention architectures to use visual information to select external knowledge more effectively. We demonstrate the efficacy on two benchmark datasets.

6 ACKNOWLEDGMENTS

We want to thank our advisors, Prof. Poonam Goyal and Prof. Navneet Goyal for their valuable guidance throughout the project. Their feedback and constant motivation was crucial for us. We also want to thank ADAPT Lab, BITS Pilani for the computing resources, server and GPUs.

Encoder	Decoder	Obj. Det.	KB	Embed.	CIDEr	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	Rogue_L	Spice
BASE SAAT					78.08	77.388	65.218	55.236	44.66	31.861	67.89	0.04764
GCN	MHA	Detectron	WordNet	Word2Vec	68.82	75.818	62.551	52.283	42.137	31.421	66.946	0.04697
VTKE	MHA	Detectron	WordNet	Word2Vec	74.01	76.417	62.945	52.496	42.068	31.359	67.291	0.0466
CFE	Att.	Detectron	Wordnet	Learned	77.4	76.165	62.664	52.408	42.421	31.315	67.48	0.04614
None	Att.	Detectron	Wordnet	Learned	78.54	77.355	64.171	53.296	42.779	31.293	67.229	0.04689
VTKE	MHA	Yolo	Wordnet	Learned	79.47	76.828	63.903	53.72	43.335	31.602	67.517	0.0482
CFE	MHA	Detectron	Wordnet	Learned	79.82	78.4	65.97	55.59	45.62	33.13	68.65	0.04765
None	MHA	Detectron	Wordnet	Learned	80.17	77.789	64.804	54.446	43.96	33.151	68.599	0.05019
VTKE	None	Detectron	Wordnet	Learned	80.86	77.401	64.203	53.885	43.584	33.172	68.582	0.05182
None	MHA	Detectron	Conceptnet	Learned	**81.65**	78.02	64.959	54.393	43.935	32.677	68.584	0.0513
VTKE	MHA	Detectron	Wordnet	Learned	**81.67**	78.682	66.601	56.505	46.281	32.839	68.774	0.05142
VTKE	MHA	Yolo	WordNet	Learned+RL	**83.85**	76.99	63.72	53.56	43.68	32.72	68.44	0.05164

Table 1: Our scores on the MSVD dataset

Encoder	Decoder	Obj. Det.	KB	Embed.	CIDEr	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	Rogue_L	Spice
BASE SAAT					49.21	80.096	65.892	52.285	40.349	28.189	60.853	0.06548
CFE	MHA	Yolo	Conceptnet	Learned	47.57	79.548	65.436	51.573	39.481	27.915	60.393	0.06503
None	MHA	Yolo	Conceptnet	Learned	47.75	80.119	65.777	51.932	39.896	28.341	60.478	0.06785
VTKE	MHA	Yolo	Conceptnet	Learned	48.88	79.385	65.019	51.535	40.021	28.369	60.515	0.06618
VTKE	MHA	Yolo	Wordnet	Learned	48.88	80.61	66.384	52.712	40.526	28.167	60.878	0.06614
None	Att.	Yolo	Conceptnet	Learned	49	79.758	65.621	52.284	40.401	28.159	60.727	0.06476
None	MHA	Yolo	Wordnet	Learned	**49.66**	79.677	65.684	52.283	40.43	28.073	60.93	0.06522
None	Att.	Yolo	Wordnet	Word2Vec	**49.68**	79.962	66.038	52.729	40.923	28.195	60.792	0.06501
GCN	MHA	Yolo	WordNet	Learned	**49.82**	79.964	66.391	53.099	41.135	28.221	61.243	0.06463
GCN	MHA	Yolo	Wordnet	Learned	**49.84**	79.96	66.39	53.1	41.13	28.22	61.24	0.06463

Table 2: Our scores on the MSRVTT dataset

REFERENCES

[1] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2013. YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-Shot Recognition. In *2013 IEEE International Conference on Computer Vision*. 2712–2719. https://doi.org/10.1109/ICCV.2013.337

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[3] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* 6, 2 (2015), 167–195. http://jens-lehmann.org/files/2015/swj_dbpedia.pdf

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs.CV]

[5] H. Liu and P. Singh. 2004. ConceptNet — A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal* 22, 4 (Oct. 2004), 211–226. https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d

[6] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (Nov. 1995), 39–41. https://doi.org/10.1145/219717.219748

[7] T. Mitchell, W. Cohen, E. Hruscha, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohammad, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2015. Never-Ending Learning. In *AAAI*. http://www.cs.cmu.edu/~wcohen/pubs.html : Never-Ending Learning in AAAI-2015.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[9] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6517–6525. https://doi.org/10.1109/CVPR.2017.690

[10] Jesse Thomason, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Raymond Mooney. 2014. Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 1218–1227. https://www.aclweb.org/anthology/C14-1115

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[12] Ramakrishna Vedantam, C. Zitnick, and Devi Parikh. 2014. CIDEr: Consensus-based Image Description Evaluation. (11 2014).

[13] J. Xu, T. Mei, T. Yao, and Y. Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5288–5296. https://doi.org/10.1109/CVPR.2016.571

[14] Wenqiao Zhang, Xin Eric Wang, Siliang Tang, Haizhou Shi, Haochen Shi, Jun Xiao, Yueting Zhuang, and William Yang Wang. 2020. Relational Graph Learning for Grounded Video Description Generation. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 3807–3828. https://doi.org/10.1145/3394171.3413746

[15] Qi Zheng, Chaoyue Wang, and Dacheng Tao. 2020. Syntax-Aware Action Targeting for Video Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

A SAMPLE RESULTS

Baseline: a man is cutting a fish
Ours: a man is cutting a piece of meat
GT: a chef carves some meat

Baseline: a woman is riding a horse
Ours: a woman is riding a motorcycle
GT: a man and woman are riding in the bike

Baseline: a man is putting some food
Ours: a man is putting butter on a tortilla
GT: a man is eating pizza

Baseline: a person is cutting a vegetable
Ours: a woman is slicing some vegetables
GT: a woman is chopping vegetables

Baseline: a man is driving a car
Ours: a man is lifting a car
GT: a man is lifting a car

Baseline: a woman is riding on a boat
Ours: a woman is riding a horse
GT: a man is riding a horse

Figure 4: Comparison of descriptions generated by our Model, Baseline and Ground truth, along the manually selected key frame of corresponding Video