VLCAP: VISION-LANGUAGE WITH CONTRASTIVE LEARNING FOR COHERENT VIDEO PARAGRAPH CAPTIONING

Kashu Yamazaki, Sang Truong, Khoa Vo, Michael Kidd, Chase Rainwater, Khoa Luu, Ngan Le

University of Arkansas, Fayetteville, AR 72701 USA

ABSTRACT

In this paper, we leverage the human perceiving process, that involves vision and language interaction, to generate a coherent paragraph description of untrimmed videos. We propose vision-language (VL) features consisting of two modalities, i.e., (i) vision modality to capture global visual content of the entire scene and (ii) language modality to extract scene elements description of both human and non-human objects (e.g. animals, vehicles, etc), visual and non-visual elements (e.g. relations, activities, etc). Furthermore, we propose to train our proposed VLCap under a contrastive learning VL loss. The experiments and ablation studies on ActivityNet Captions and YouCookII datasets show that our VLCap outperforms existing SOTA methods on both accuracy and diversity metrics. Source code: https://github.com/UARK-AICV/VLCAP

Index Terms— Contrastive Learning, Video Captioning, Vision, Language

1. INTRODUCTION

Video paragraph captioning (VPC) aims to generate a paragraph description of untrimmed videos with several temporal event locations in a coherent storytelling. VPC can be considered as a simplified version of dense video captioning by eliminating the requirements for generating event proposals. VPC takes a video with its corresponding event proposals as the input and returns a coherent paragraph as the output. A typical VPC contains two components corresponding to (i) feature extraction to encode each event into a feature and (ii) caption generation to decode features into a list of sentences. An essential requirement of VPC is maintaining the intra-event coherence between words within a sentence describing an event and inter-event coherence between sentences within a paragraph describing an entire video.

Zhou, et al. [4] first leveraged the success of Transformer [5] to dress VPC task, known as Vanilla Transformer VPC. In their approach, intra-event coherence is decoded by a Transformer but there is no mechanism to model the inter-event coherence i.e., each event is decoded individually. Later, [6] tackled this limitation and proposed MFT by utilizing LSTM [7]. In MFT, the last hidden state of the current sentence is used as an initial hidden state for the next sentence. However, the coherence between sentences in MFT is ill-favored, facing the gradient vanishing problem [8] and unable to model long-term dependencies [9]. Being inspired by the recent transformer language model, Transformer-XL [1], which is able to resolve context fragmentation for language modeling, [2] proposed MART. While Transformer-XL directly uses hidden states from previous segments, MART is designed as a unified encoder-decoder to prevent overfitting and reduce memory usage.

Clearly, to understand and describe a video, we not only observe the entire scene but also pay attention to both element scenes such as human and non-human objects (e.g., vehicles, animals, tools, etc.), visual and non-visual elements (e.g., actions, relations, etc). Furthermore, vision and language are two primary capabilities of humans language influences basic perceptual processing [10]. However, most of the existing VPC
approaches [4, 11, 1, 12, 13] decode caption description by applying a backbone, e.g. C3D [14], I3D [15], 2Stream [16, 17], or Slowfast [18] to extract global visual information of the entire scene. By doing that, they ignore the interaction between the entire scene and relevant elements as well as disregard the fact that language and perception are two central cognitive systems.

In this paper, we propose a multi-modal VL representation consisting of the global visual feature of the entire scene and linguistics relevant scene elements. While maximum likelihood estimation (MLE) is the most widely used loss function for supervised learning VPC, it does not guarantee that the learnt latent features represent the groundtruth captions. In this paper, we leverage contrastive learning [19, 20] and propose VL Loss, which consists of two terms corresponding to captioning loss ($L_{\text{cap.}}$) and a contrastive contextual loss ($L_{vl}$). The network comparison between our proposed VLCap with other existing VPC networks is shown in Fig. 1.

## 2. PROPOSED METHOD

Our proposed VLCap is designed as a unified encoder-decoder architecture and contains two main modules, i.e., VLCap Encoder and VLCap Decoder. Both modules are trained in an end-to-end framework by our proposed VL loss function. The entire architecture of VLCap is shown in Fig. 2.

In this section, we first introduce all notations and VPC problem formulation as follows: Given an untrimmed video $V = \{v_i\}_{i=1}^{|V|}$, where $|V|$ is the number of frames, and a list of its important events $E = \{e_i = (e_i^s, e_i^e)\}_{i=1}^{|E|}$, where $|E|$ is the number of events within a video and event $e_i$ is defined by a pair of beginning and ending timestamps $(e_i^s, e_i^e)$. Our objective is to generate a coherent paragraph $P = \{s_i\}_{i=1}^{|P|}$ that describes the whole video $V$. In this setup, a sentence $s_i$ aims to describe its corresponding event $e_i$. We use notation $e = (e^s, e^e)$ to denote an event and it is presented by a sequence of frames $V_e = \{v_i\}_{e^s \leq i \leq e^e}$.

### 2.1. VLCap Encoder

This module aims to extract VL feature $f^{VL}_i$ given an event $e_i$, presented by a sequence of frames $V_e$. Follow the standard setup [4, 11, 1, 12, 13], we divide $V_e$ into $L$ snippets, $\{S_i\}_{i=1}^L$, each snippet $S_i$ consists of $\delta$ consecutive frames, where $L = \lceil \frac{|V_e|}{\delta} \rceil$ and $|V_e|$ is the number of frames in $V_e$. VLCap Encoder processes a snippet $S_i$ to extract $f^{VL}_i$. As a result, VLCap Encoder processes the event $e$ to extract feature $f^{VL} = \{f^{VL}_i\}_{i=1}^L$ as shown in the Fig.2 (left). The VLCap Encoder contains three modalities as follows:

**i. Vision Modality** This modality aims to extract visual content by applying a C3D network [14] into snippet $S_i$. The output feature map of C3D network $\phi$ is processed by average pooling to reduce the entire spatial dimension followed by channel multilayer perceptron (MLP). As a result, each snippet $S_i$ is represented by a feature $f^V_i$.

**ii. Language Modality** This modality aims to extract element-level linguistic details of each snippet $S_i$. We leverage the success of recent works [24, 25] which have proved the effectiveness of feature representation learned via Contrastive Language-Image Pre-training (CLIP) [26]. Given a snippet $S_i$, the linguistic feature $f^L_i$ is extracted by the following steps: (i) - Word embedding: We construct a vocabulary $W = \{w_1, \ldots, w_N\}$ using the groundtruth captions from training dataset. Each word $w_i \in W$ is encoded by a Transformer network [5] into a text feature $f^w_i$. We then project feature...
Table 2. Performance comparison of VLCap with other SOTA models on ActivityNet Captions ae-val. † denotes results by us.

| Methods            | Year       | Input                  | B@4↑ | M↑ | C↑ | R↑ | Div@2↑ | R@4↓ |
|--------------------|------------|------------------------|------|----|----|----|--------|------|
| Vanilla Transformer [4] | CVPR2018   | Res200 + Flow          | 9.75 | 15.64 | 22.16 | 28.90† | 77.40 | 7.79   |
| AdvIn [11]        | CVPR2019   | C3D + Object           | 10.04 | 16.60 | 20.97 | –    | –      | 5.76   |
| GVD [12]          | CVPR2019   | Res200 + Flow + Object | 11.04 | 15.71 | 21.95 | –    | –      | 8.76   |
| Transformer-XL [1] | ACL2019    | Res200 + Flow          | 10.39 | 15.09 | 21.67 | 30.18† | 75.96† | 8.54   |
| Transformer-XLRG [2] | ACL2020    | Res200 + Flow          | 10.17 | 14.77 | 20.40 | –    | –      | 8.85   |
| MART [2]          | ACL2020    | Res200 + Flow          | 10.33 | 15.68 | 23.42 | 30.32† | 75.71† | 5.18   |
| PDVC [13]         | ICCV2021   | C3D + Flow             | 11.80 | 15.93 | 27.27 | –    | –      | –      |
| **VLCap (ours)**  | –          | C3D + Language         | **14.00** | **17.78** | **32.58** | **36.37** | **78.01** | **4.42** |

Table 3. Performance comparison of VLCap with other SOTA models on ActivityNet Captions ae-test. † denotes results by us.

| Methods            | Year       | Input                  | B@4↑ | M↑ | C↑ | R↑ | Div@2↑ | R@4↓ |
|--------------------|------------|------------------------|------|----|----|----|--------|------|
| Vanilla Transformer [4] | CVPR2018   | Res200 + Flow          | 9.31 | 15.54 | 21.33 | 28.98† | 77.29† | 7.45   |
| Transformer-XL [1] | ACL2019    | Res200 + Flow          | 10.25 | 14.91 | 21.71 | 30.25† | 76.17† | 8.79   |
| Transformer-XLRG [2] | ACL2020    | Res200 + Flow          | 10.07 | 14.58 | 20.34 | –    | –      | 9.37   |
| MART [2]          | ACL2020    | Res200 + Flow          | 9.78  | 15.57 | 22.16 | 30.85† | 75.69† | 5.44   |
| MART w/ COOT [23] | NIPS2020   | COOT                   | 10.85 | 15.99 | 28.19 | –    | –      | 6.64   |
| **VLCap (ours)**  | –          | C3D + Language         | **13.38** | **17.48** | **31.29** | **35.99** | **78.29** | **4.18** |

Table 4. Performance comparison of VLCap with other SOTA models on YouCookII validation set.

| Methods            | Year       | Input                  | B@4↑ | M↑ | C↑ | R↑ | Div@2↑ | R@4↓ |
|--------------------|------------|------------------------|------|----|----|----|--------|------|
| Vanilla Transformer [4] | CVPR2018   | Res200 + Flow          | 4.38 | 11.55 | 38.00 | –    | –      | –     |
| GPaS [28]         | IEEE-TM2020| Res200                 | 1.64 | 12.20 | 41.44 | 27.98 | –      | –     |
| MART [2]          | ACL2020    | Res200 + Flow          | 8.00  | 15.90 | 35.74 | –    | –      | **4.39** |
| MART w/ COOT [23] | NIPS 2020  | COOT                   | 9.44  | **18.17** | 46.06 | –    | –      | 6.30   |
| **VLCap (ours)**  | –          | C3D + Language         | **9.56** | **17.95** | **49.41** | **38.17** | **67.97** | **5.16** |
2.3. VL Loss

Maximum likelihood estimation (MLE) loss, which is trained to increase the likelihood between predicted captions and groundtruth, is the most common in VPC. However, it is unable to address the question of how well the learnt latent features represent the groundtruth captions. In this paper, we proposed Visual-Linguistic (VL) Loss, which tackles the aforementioned concerns while maintaining the likelihood between predicted caption and groundtruth. Particularly, we leverage the recent advantages of constructive learning to propose $\mathcal{L}_{vl}$ to pull all snippets of the same event and push snippets of different events. Let consider a set of $N$ events $\{e_i\}_{i=1}^N$, each event $e_i$ consists of $L$ snippets $\{S_i\}_{i=1}^L$. Each event $e_i$ has its corresponding groundtruth caption $c$, which is then presented as $f^T$ by the pretrained Text Transformer from CLIP [26]. Apply our proposed VLCap network into $e_i$, we obtain the event embeddings $F_i$ which is then processed as a vector $f_i = \text{mean}(F_i)$. $\mathcal{L}_{vl}$ is computed as follows:

$$\mathcal{L}_{vl} = -\sum_{i,j=1}^{N} \mathbb{1}_{i \neq j} \log e^{\rho(f_i \cdot f_j^T)} + \log (1 - e^{\rho(f_i \cdot f_j^T)})$$

where $\rho$ is a learnable temperature parameter, which is initialized to $\log(1/0.07)$, to prevent scaling of the dot product and reduce training instability.

Our VL loss $\mathcal{L}_{VL}$ consists of two terms corresponding to caption-caption loss ($\mathcal{L}_{MLE}$) and a vision-language loss ($\mathcal{L}_{vl}$) as follows:

$$\mathcal{L}_{VL} = \mathcal{L}_{MLE} + \mathcal{L}_{vl}$$

3. EXPERIMENTS

3.1. Datasets, Metrics and Implementation Details

We benchmark our VLCap on two popular VPC datasets, YouCookII [22] and ActivityNet Captions [21]. Information of these datasets are summarized in Table 1. We follow the previous works [2] to split the original validation set into two subsets: ae-val with 2,460 videos for validation and ae-test with 2,457 videos for test.

We benchmark VLCap on four standard accuracy metrics, i.e., BLEU@4 (B@4) [30], METEOR (M) [31], CIDEr (C) [32], ROUGE (R) [33] and two diversity metrics i.e., 2-gram diversity (Div@2) [34] and 4-gram repetition (R@4) [6].

Adam optimizer was used to train our VLCap with an initial learning rate of 1e-4, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $L_2$ weight decay of 0.01, and learning rate warm up over the first 5 epochs. During the training, we use the label smoothing with a value of 0.1 and $\lambda = 0.1$.

3.2. Performance and Comparison

Tables 2, 3 report the performance comparison on ActivityNet Captions corresponding to ae-val and ae-test sets whereas Table 4 shows the performance comparison on YouCookII validation set. In each table, we highlight the best and the second-best with **bold** and underline. On YouCookII, VLCap obtains the best performance on B@4, C and R metrics whereas it gains compatible on other metrics. On ActivityNet Captions, VLCap obtains the best performance with large gaps on both accuracy metrics and diversity metrics compared to the second-best score. Take ActivityNet Captions as an example, corresponding to ae-val and ae-test sets, our VLCap gains (2.2%/1.18%/5.31%/6.05%) and (2.53%/1.49%/3.10%/5.14%) higher on BLEU@4/METEOR/CIDEr/ROUGE metrics while improves (0.61%) and (1.0%) on Div@2 as well as reduces (0.31%) and (1.26%) on R@4 compare to the second-best achievement.

To evaluate the effectiveness of our proposed VL feature as well as VL loss, we conduct ablation studies as shown in Table 5. The capability of the proposed VL loss ($\mathcal{L}_{VL}$) is shown in comparisons between Exp.#1 v.s #2 and Exp.#3 v.s #4 where we compare between VL loss and MLE loss. The advantage of the proposed VL feature is shown in comparisons between Exp.#1 v.s #3 and Exp.#2 v.s #4 where we compare between vision feature (i.e. C3D) and VL feature. Both of our proposed VL feature and VL loss contribute in improving the accuracy and diversity metrics.

4. CONCLUSION

In this work, we present a novel VLCap network for video paragraph captioning. Our VLCap network is trained in an end-to-end framework with a two-fold contribution: (i) VL feature, which extracts the global visual features of the entire scene and local linguistics feature of scene elements; and (ii) VL loss, which is trained by a constrastive learning mechanism. In VLCap network, the intra-event coherence is learnt by a Transformer whereas the inter-event coherence is modeled by GRU-like memory. Comprehensive experiments and ablation studies on ActivityNet Captions and YouCookII datasets demonstrate the effectiveness of our VLCap, which outperforms the existing SOTA approaches on both accuracy (BLEU@4, METEOR, CIDEr, ROUGE) and diversity (Div@2, R@4) metrics.

---

Table 5: Ablation study on the contribution of the proposed VL feature and VL loss $\mathcal{L}_{VL}$ on ActivityNet Captions dataset.

| Exp. | Vision | Lang | $\mathcal{L}_{MLE}$ | $\mathcal{L}_{VL}$ | ae-test  | ae-test  | ae-val  | ae-val  |
|------|--------|------|---------------------|---------------------|---------|---------|---------|---------|
|      |        |      |                     |                     | B@4↑    | M↑      | C↑      | R↑      |
| #1   | ✓      | ×     | ✓                   | ✓                   | 11.10   | 13.72   | 27.88   | 31.75   |
| #2   | ✓      | ×     | ✓                   | ✓                   | 11.17   | 16.27   | 30.22   | 31.72   |
| #3   | ✓      | ✓     | ×                   | ✓                   | 13.56   | 17.42   | 30.10   | 35.78   |
| #4   | ✓      | ✓     | ✓                   | ✓                   | 13.38   | 17.48   | 31.29   | 35.99   |
|      |        |      |                     |                     | 78.29   | 4.18    | 78.01   | 4.42    |

---

3659
5. REFERENCES

[1] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov, “Transformer-XL: Attentive language models beyond a fixed-length context,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 2978–2988, Association for Computational Linguistics.

[2] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal, “MART: Memory-augmented recurrent transformer for coherent video paragraph captioning,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 2020, pp. 2603–2614, Association for Computational Linguistics.

[3] K. Vo, H. Joo, K. Yamazaki, S. Truong, K. Kitani, M. Tran, and N. Le, “AEI: Actors-Environment Interaction with Adaptive Attention for Temporal Action Proposals Generation,” BMVC, 2021.

[4] Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong, “End-to-end dense video captioning with masked transformer,” in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp. 8739–8748, IEEE Computer Society.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, Eds., 2017, pp. 5998–6008.

[6] Yilei Xiong, Bo Dai, and Dahua Lin, “Move forward and tell: A progressive generator of video descriptions,” in Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, Eds. 2018, vol. 11215 of Lecture Notes in Computer Science, pp. 489–505, Springer.

[7] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] Razvan Pascanu, Tomáš Mikolov, and Yoshua Bengio, “On the difficulty of training recurrent neural networks,” in Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013, 2013, vol. 28 of JMLR Workshop and Conference Proceedings, pp. 3101–3108, JMLR.org.

[9] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber, et al., “Gradient flow in recurrent nets: the difficulty of learning long-term dependencies,” 2001.

[10] G. Lupyan, R. Abdel Rahman, L. Boroditsky, and A. Clark, “Effects of Language on Visual Perception,” Trends Cogn Sci, vol. 24, no. 11, pp. 930–944, 2020.

[11] Jae Sung Park, Marcus Rohrbach, Trevor Darrell, and Anna Rohrbach, “Adversarial inference for multi-sentence video description,” in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp. 6598–6608, Computer Vision Foundation / IEEE.

[12] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, and Marcus Rohrbach, “Grounded video description,” in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp. 6578–6587, Computer Vision Foundation / IEEE.

[13] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo, “End-to-end dense video captioning with parallel decoding,” in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pp. 6827–6837, IEEE.

[14] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, “3d convolutional neural networks for human action recognition,” in Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel, Johannes Fürnkranz and Thorsten Joachims, Eds. 2010, pp. 495–502, Omnipress.

[15] João Carreira and Andrew Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 4724–4733, IEEE Computer Society.

[16] Karen Simonyan and Andrew Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, Eds., 2014, pp. 568–576.

[17] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, “Convolutional two-stream network fusion for video action recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 1933–1941, IEEE Computer Society.

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, “Slowfast networks for video recognition,” in 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pp. 6201–6210, IEEE.

[19] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karon Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio, “Learning deep representations by mutual information maximization,” in 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, 2019, OpenReview.net.

[20] Yonglong Tian, Dilip Krishnan, and Phillip Isola, “Contrastive multiview coding,” in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds. 2020, vol. 12356 of Lecture Notes in Computer Science, pp. 776–794, Springer.

[21] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles, “Dense-captioning events in videos,” in IEEE International Conference on Computer Vision, ICCV 2017.
[22] Luowei Zhou, Chenliang Xu, and Jason J. Corso, “Towards automatic learning of procedures from web instructional videos,” in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. Sheila A. McIlraith and Kilian Q. Weinberger, Eds. 2018, pp. 7590–7598, AAAI Press.

[23] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox, “COOT: cooperative hierarchical transformer for video-text representation learning,” in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, Eds., 2020.

[24] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski, “Styleclip: Text-driven manipulation of stylegan imagery,” in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, 2021, pp. 2065–2074, IEEE.

[25] B. Yang and Y. Zou, “CLIP Meets Video Captioners: Attribute-Aware Representation Learning Promotes Accurate Captioning,” arXiv e-prints, 2021.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, “Learning transferable visual models from natural language supervision,” in Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, Marina Meila and Tong Zhang, Eds. 2021, vol. 139 of Proceedings of Machine Learning Research, pp. 8748–8763, PMLR.

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. 2021, OpenReview.net.

[28] Z. Zhang, D. Xu, W. Ouyang, and L. Zhou, “Dense Video Captioning Using Graph-Based Sentence Summarization,” IEEE Trans. Multimed., vol. 23, pp. 1799–1810, 2020.

[29] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholi, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu, “UNITER: universal image-text representation learning,” in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds. 2020, vol. 12375 of Lecture Notes in Computer Science, pp. 104–120, Springer.

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318, Association for Computational Linguistics.