ENHANCING MULTIMODAL ALIGNMENT WITH MOMENTUM AUGMENTATION FOR DENSE VIDEO CAPTIONING

[1]*Yiwei Wei*[*], [2]*Shaozu Yuan*[*], [2]*Meng Chen*[†], [1]*Longbiao Wang*

[1]Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjing, China, [2]JD AI Research, Beijing, China

ABSTRACT

Dense video captioning aims to localize multiple events from an untrimmed video and generate corresponding captions for each event. Fusing different modalities(e.g. rgb, flow, audio) via transformer structure is a promising way to improve the caption performance. However, it is challenging for the cross-modal encoder to learn multimodal interactions due to their inherent disparities of distribution. In this paper, we propose a novel transformer structure with contrastive learning to align different modalities. Specifically, to avoid the limitation of small batch size and false contrastive targets, we design an event-aligned momentum augmentation strategy to apply contrast learning for dense video captioning. The experimental result shows that our proposals outperform all existing multimodal fusion methods for dense video captioning.

Index Terms— Dense video captioning, Multimodal fusion, Contrastive learning, Momentum augmentation

1. INTRODUCTION

Dense video captioning(DVC)[1, 2, 3] is a challenging task, where the captioner should first localize multiple events from an untrimmed video, then generate corresponding caption for each event. Formally, a series of approaches[4, 5, 6] model DVC as two sub-tasks, termed temporal localization and caption generation.

Given that videos typically contain information in multiple modalities (e.g. rgb, flow, and audio), it is critical to fully utilize these information. Some recent efforts[7, 8] seek to fuse different modalities via transformer structure, thereby promoting the captioning performance. However, it is insufficient to learn multimodal interactions with cross-attention modules due to their inherent different distributions in representation space as shown in Figure 1. Recent works[9, 10] have demonstrated that applying contrastive learning to align textual and visual information are beneficial for vision-language pretraining tasks. Thus, it seems reasonable to align the spatial distribution between different modalities with contrastive learning to handle this problem.

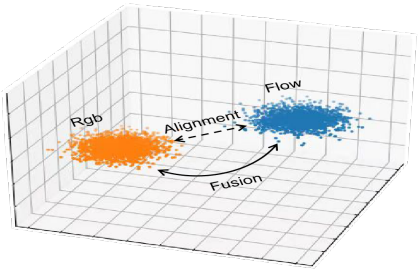

Fig. 1. The samples of two modalities marked in blue and red shows different distributions in representation space.

Nevertheless, less work tries to apply contrastive learning for DVC task, because there are still two challenges: 1) how to set the learning targets, and 2) how to enlarge the negative sampling space. For the first challenge, it is unreasonable to simply set the different events in a video as positive or negative samples, because there is no guarantee of whether the events in the same video has similar content or not. For the second, a large batch size can increase the negative sample space, thus facilitating contrastive learning. However, it is restricted by the GPU memory size and the computing cost of DVC task. These two challenges hinder the progress of multimodal alignment for dense video captioning.

In this paper, we design an event-aligned momentum augmentation(EAMA) strategy to apply contrastive learning for dense video captioning. To the best of our knowledge, this is the first work to apply contrastive learning to dense video captioning. Concretely, we first modify transformer to a bi-modal structure to accommodate different modalities for contrastive learning. It consists of three components: a unimodal encoder for unimodal encoding, a bi-modal encoder for cross-modal fusing, and a multimodal decoder for linguistic decoding. To construct the learning targets for contrastive learning, we set modalities extracted from the same event as positive pairs and exclude the events from the same video, since those events may disturb each other. Conversely, the cross-modal pairs of events from different videos will be treated as negative pairs. Moreover, we also designed a slow-update momentum queue to store adequate negative samples from other videos, thus enlarging the negative sampling space to facil-

[*]Equal contribution to this work
[†]Corresponding author.

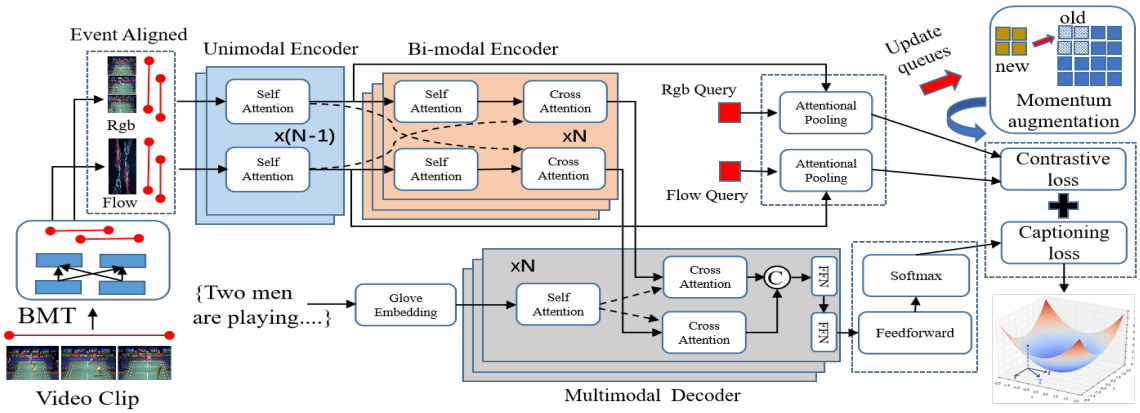

Fig. 2. The architecture of our model. It consists of two modules: the event localization module(BMT[5]) and the event captioning module. An event-aligned momentum augmentation strategy is designed to capture more negative samples for contrastive learning. The captioning objective is applied for encouraging the model to predict fluent captions, while the contrastive objective is exploited to align the unimodal representations before fusion. For brevity, we omit FFN layer in the encoder.

itate contrastive learning. Experiments conducted on ActivityNet Captions dataset demonstrate the effectiveness of our approach compared with existing multimodal fusion methods.

2. METHODOLOGY

In this section, we introduce each part of the proposed approach as shown in Figure 2. Note that we use rgb and flow modalities to illustrate our model.

2.1. Event Localization Module

Here, we employ the pre-trained event localization model[8] to obtain the event proposals $P = \{(start_j, end_j, score_j)\}_{j=1}^{K}$. Then, we map V and F to obtain event-aligned features based on the event proposals P, yielding a rgb sequence of each proposal $Z^V \in R^{L^j \times d^v}$ and a flow sequence of each proposal $Z^F \in R^{L^j \times d^f}$, where the rgb and flow features are defined as $V \in R^{L \times d^v}$ and $F \in R^{L \times d^f}$ respectively.

2.2. Captioning Module

Unimodal Encoder. We apply transformer[11] encoder to encode the event-aligned features Z^F and Z^V, for its powerful ability of representation learning. It contains a stack of N encoding layers, each of which is composed of two sublayers: multi-head self-attention and feed-forward neural network(FFN):

$$\begin{bmatrix} \nu^V = FFN(MultiHeadAttention(Z^V, Z^V)) \\ \nu^F = FFN(MultiHeadAttention(Z^F, Z^F)) \end{bmatrix}^N \quad (1)$$

Notably, we omit the residual connection and layer normalization in each layer for a concise explanation. The subsequent blocks follow the same principle.

Bi-modal Encoder. Bi-modal encoder are structurally similar to unimodal encoder, except that an additional multi-head cross attention is added in each layer to fuse unimodal outputs ν^V and ν^F, which is defined as follows:

$$\begin{bmatrix} \chi^V = MultiHeadAttention(\nu^V, \nu^V) \\ \chi^F = MultiHeadAttention(\nu^F, \nu^F) \\ \chi^{VF} = FFN(MultiHeadAttention(\chi^V, \nu^F)) \\ \chi^{FV} = FFN(MultiHeadAttention(\chi^F, \nu^V)) \end{bmatrix}^N \quad (2)$$

where $\chi^{VF} \in R^{L^j \times d^v}$ represents the flow-attended rgb features and $\chi^{FV} \in R^{L^j \times d^f}$ represents the rgb-attended flow features.

Multimodal Decoder. The multimodal decoder extends the bi-modal structure based on a N-layer transformer decoder, and adds a two-layer feed-forward network to integrate the bi-modal outputs χ^{VF} and χ^{FV}. Suppose the input for time step t is $W_{t-1} = (w_0, w_1, ..., w_{t-1}), w_i \in R^{d \times 1}$, the multimodal decoder is defined as:

$$\begin{bmatrix} \hat{\kappa}_t = MultiHeadAttention(w_{t-1}, W_{t-1}) \\ \kappa_t^{VF} = MultiHeadAttention(\hat{\kappa}_t, \chi^{VF}) \\ \kappa_t^{FV} = MultiHeadAttention(\hat{\kappa}_t, \chi^{FV}) \\ \kappa_t = FFN(FFN([\kappa_t^{VF}, \kappa_t^{FV}]) \end{bmatrix}^N \quad (3)$$

where κ_t is the output for the time step t.

We feed κ_t into a fully-connected layer with the softmax activation which generates the probability of the current word $p(w_t|W_{t-1})$ by mapping the caption features of size d into a dimension $|\Sigma|$ corresponding to the size of the vocabulary.

2.3. Event-aligned Momentum Augmentation

Event-aligned momentum augmentation aims to learn better unimodal representations before fusion by encouraging

matched event pairs to obtain higher similarity scores. We first employ an attentional pooling to aggregate the unimodal context feature and perform dimensionality reduction for each event in each mini-batch. Then, we combine the different context unimodal features as rgb-flow pairs for contrastive learning(CL). To overcome the size limitation of the mini-batch, inspired by [12], we maintain two queues to store the most recent M rgb-flow pairs sampled from other mini-batches. And the queues are dynamically updated by replacing the premier mini-batch with current mini-batch during training.

Different from [12], we also propose two improvements to adapt to dense video captioning task: (i) different events in the same video still have chance to be maintained in the queues, which may lead to false contrastive targets and negative effects for CL. To solve this problem, we mark each negative samples in the queues with a video id, those negative samples that have the same video id with current event will be masked when calculating the contrastive loss. (ii) Although the queues enlarge the negative sampling space, it is inevitable that those unimodal context features in the queues will be modified during training iteration. The frequently updated queues lead to unstable training and make the model hard to converge. To solve this issue, we design a slow-update momentum encoder composed of unimodal encoder and attentional pooling to stably produce unimodal context feature $\acute{q}^v$ and $\acute{q}^f$. Here, we apply momentum method[13] to optimize momentum encoder and restrain the updating speed:

$$\theta_m \leftarrow \beta\theta_m + (1-\beta)\theta_u \tag{4}$$

where θ_m denotes the parameters of the momentum encoder, θ_u denotes the parameters of the unimodal encoder and $\beta \in [0, 1)$ is a momentum coefficient.

2.4. Training Objective

Captioning Objective. For captioning generation, we directly optimize KL-divergence by loss function defined as:

$$\mathcal{L}_{Cap} = \sum_{i=1}^{T} KL(log(p(w_t|W_{t-1})), y_t^*) \tag{5}$$

where $y_t^* \in |\Sigma|$ is the current Ground-Truth caption.

Contrastive Objective. Suppose we have obtained the unimodal contexts and the momentum queues from the event-aligned momentum augmentation, we calculate the softmax-normalized rgb-to-flow and flow-to-rgb similarities as:

$$\rho_m^{v2f}(\hat{q}^v) = \frac{exp((\hat{q}^v \cdot \acute{q}_m^f)/\tau)}{\sum_{m=1}^{M} exp((\hat{q}^v \cdot \acute{q}_m^f)/\tau)} \tag{6}$$

$$\rho_m^{f2v}(\hat{q}^f) = \frac{exp(\hat{q}^f \cdot \acute{q}_m^v/\tau)}{\sum_{m=1}^{M} exp(\hat{q}^f \cdot \acute{q}_m^v/\tau)} \tag{7}$$

where $\hat{q}$ denotes the normalized unimodal contexts and $\acute{q}_m$ denotes the normalized momentum contexts in current queues. Let y_m^{v2f} and y_m^{f2v} denote the ground-truth one-hot similarity, where negative pairs have a probability of 0 and the positive pair has a probability of 1. The rgb-flow contrastive loss is defined as the cross-entropy between ρ and y:

$$\mathcal{L}_{Con}(\hat{q}^v, \hat{q}^f) = \frac{1}{2}[CE(y_m^{v2f}, \rho_m^{v2f}) + CE(y_m^{f2v}, \rho_m^{f2v})] \tag{8}$$

where CE represents the cross-entropy function. Finally, we jointly train the above two losses for dense video captioning.

3. EXPERIMENTS

3.1. Datasets and Evaluation

All our experiments were conducted on the ActivityNet Captions dataset[1], which contains $20k$ long untrimmed videos and follows a standard split including 10009 training videos, 4925 validation videos, and 5044 testing videos. Each video, on average, lasts 120 seconds and contains 3.65 temporally localized captions. To evaluate the captioning performance, we take the BLEU[14] and METEOR[15] as metrics.

3.2. Implementation Details

Following the settings of previous works[5, 7], we employ the I3D[16] network to obtain rgb and flow features and adopt VGGish model[17] to extract audio features. Note that the flow features are captured from the optical flow frames[18] that represent the motion between video frames. Besides, we use GloVe model to embed the captions. In transformer, the number of layers N is set as 2, and the number of attention heads is 8. In contrastive objective, the temperature τ is set as 0.07, and the dimension of learnable queries is 128. The momentum coefficient β is set as 0.995, and the size of the queue is set as 2048. During training, our model is trained for 40 epochs with a batch size of 32. ADAM is used as optimizer with a learning rate 10^5 and smoothing parameter γ as 0.7.

3.3. Main Results

Here we compare our model with following baselines: (1) **WSDEC**[19] applying cross-attention to fuse different modalities, (2) **MDVC**[7] and (3) **BMT**[8], both are transformer-based multimodal fusion models, (4) **DVCUSI**[20], which injects unsupervised semantic information into multi-modal fusion model. For fair comparison, we report the performances generated by the models trained with rgb and flow features. The main comparison results[1] are illustrated in Table 1. Here, we conduct the experiments with two different settings: captioning with ground truth events and captioning with predicted events. It can seen that our model surpasses all previous approaches in terms of all the metrics under

[1]Notice that we excluded audio information in the main experiments, for it is more difficult to align the audio and rgb modalities compared with flow and rgb modalities. For more analysis, please refer to section 3.5.

Unaligned: ["A woman is standing in a room talking to the camera", "the woman cleans the floor", "She then puts the mop in the bucket and puts it in the bucket"]

Aligned: ["A woman is seen speaking to the camera while holding a mop", "She puts the mop on the floor", "She then vacuums the floor with the mop"]

GT: ["A woman is sweeping the floor in a kitchen", "She grabs a green mop and pours it on the floor", "She then mops the floor with the mop"]

Fig. 3. The examples predicted by our model. The "unaligned" means removing contrastive learning.

Table 1. The results of the dense video captioning task in terms of BLEU4(B4), METEOR(M). † denotes the models trained with unsupervised semantic information.

Method	GT proposals			Predicted proposals		
	B3	B4	M	B3	B4	M
WSDEC[19]	3.04	1.46	7.23	1.85	0.90	4.93
MDVC[7]	-	1.73	9.87	-	-	-
BMT[8]	3.77	1.66	10.29	2.85	1.30	7.47
Ours	**4.30**	**1.86**	**10.52**	**3.01**	**1.43**	**7.58**
DVCUSI[20]†	4.32	1.85	10.55	3.68	1.81	8.26
Ours†	**4.43**	**1.94**	**10.55**	**3.75**	**1.87**	**8.33**

two different settings. Specially, our model also achieves superior performance against previous method[20] by feeding the unsupervised semantic information into the proposed model. Here, the unsupervised semantic information denotes a sequence of integers that captured from the video frames.

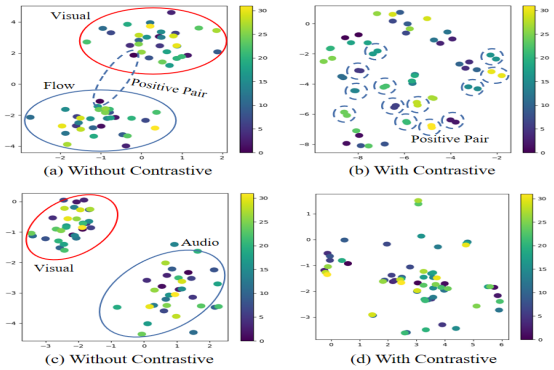

Fig. 4. Cluster visualization of unimodal contexts in a mini-batch. Dots of the same color represent the positive pairs.

3.4. Ablation Study

In Table 2, we conduct the experiments under different conditions. We first verify the performance by using different modalities. The results show that the rgb+flow yields better results compared to the rgb-only model or flow-only model. Besides, the model with contrastive learning(CL) achieves better performance, which shows the importance to align the multiple modalities. Furthermore, we also investigate the effectiveness of event-aligned momentum augmentation(EAMA) strategy. It shows EAMA improves the captioning performance further for it enlarges the number of negative pairs and benefits for feature alignment.

Table 2. The performances in terms of different conditions.

Conditions				Metrics		
Rgb	Flow	CL	EAMA	B3	B4	M
✓				3.66	1.53	10.26
	✓			3.24	1.38	9.85
✓	✓			4.19	1.72	10.51
✓	✓	✓		4.25	1.80	10.47
✓	✓	✓	✓	**4.30**	**1.86**	**10.52**

3.5. Case Study and Visualization

In the case study of Figure 3, we compare the dense captions predicted by our model with aligned feature and the unaligned feature. It shows that by directly fusing without alignment, the two different modalities failed to promote each other during generation. The video shows a woman is mopping the floor. The unaligned model generates "She then puts the mop in the bucket and puts it in the bucket," which is inconsistent with the rgb content. When we apply the aligned-fusion strategy, the generated captions are more coordinated and accurate.

We also conduct visualization experiment for the sampled cases in Figure 4. Here, we use the t-SNE algorithm to perform dimensionality reduction and obtain 2-dimensional vector for each case. Figure 4(a) is the case visualization without contrastive learning, while Figure 4(b) shows the visualization with contrastive learning. It can be seen that the distance between the positive pairs get closer with contrastive learning, which demonstrates that the CL can help the model align the multimodal features. Moreover, Figure 4(c,d) reveals it is more difficult to align the audio and rgb modalities compared with flow and visual modalities. We conjecture this is due to the weak correlation between the rgb content and the audio content. Therefore, it is more beneficial and effective to align rgb and flow content with contrastive learning.

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel approach with event-aligned momentum augmentation to produce cross-modal alignment between different modalities. This was the first work to apply contrastive learning to DVC. The experimental results demonstrated that our model outperforms all compared methods. In future work, we will employ the distillation model to overcome the mismatch between rgb-flow pairs.

5. REFERENCES

[1] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles, "Dense-captioning events in videos," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 706–715.

[2] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8739–8748.

[3] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo, "End-to-end dense video captioning with parallel decoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6847–6857.

[4] Dali Yang and Chun Yuan, "Hierarchical context encoding for events captioning in videos," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1288–1292.

[5] Vladimir Iashin and Esa Rahtu, "A better use of audio-visual cues: Dense video captioning with bi-modal transformer," in *The 31st British Machine Vision Virtual Conference*. British Machine Vision Association, BMVA, 2020.

[6] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu, "Bidirectional attentive fusion with context gating for dense video captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7190–7198.

[7] Vladimir Iashin and Esa Rahtu, "Multi-modal dense video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 958–959.

[8] Maitreya Suin and AN Rajagopalan, "An efficient framework for dense video captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 12039–12046.

[9] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[10] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu, "Coca: Contrastive captioners are image-text foundation models," *arXiv preprint arXiv:2205.01917*, 2022.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[13] Vijay V Phansalkar and P Shanti Sastry, "Analysis of the back-propagation algorithm with momentum," *IEEE Transactions on Neural Networks*, vol. 5, no. 3, pp. 505–506, 1994.

[14] 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*, 2002, pp. 311–318.

[15] Michael Denkowski and Alon Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 376–380.

[16] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[17] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[18] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8934–8943.

[19] Tanzila Rahman, Bicheng Xu, and Leonid Sigal, "Watch, listen and tell: Multi-modal weakly supervised dense event captioning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8908–8917.

[20] Valter Estevam, Rayson Laroca, Helio Pedrini, and David Menotti, "Dense video captioning using unsupervised semantic information," *arXiv preprint arXiv:2112.08455*, 2021.