MoP-VC: Mixture-of-PoS experts for video captioning via syntactic constraints (Supplymentary)

No Author Given

No Institute Given

A Experimental setup

A.1 Dataset and Evaluation Metrics

Experiments were carried out on two benchmark datasets: MSR-VTT [8] and MSVD [2]. The MSVD dataset comprises 1,970 videos, each annotated with approximately 40 captions and averaging around 10 seconds in length. Following the protocols of [9] and [8], we allocated 1,200 videos for training, 100 for validation, and 670 for testing. The MSR-VTT dataset contains 10,000 video clips, each paired with 20 captions and averaging approximately 15 seconds. Using the official split from [8], we employed 6,153 clips for training, 497 for validation, and 2,090 for testing. To assess our method against prior work, we adopted four standard evaluation metrics: BLEU@4 (B@4) [6], a precision-based measure; METEOR (M) [1], which computes sentence-level alignment; ROUGE (R) [5], quantifying the longest common subsequence between predicted and reference captions; and CIDEr (C) [7], a consensus-based metric.

A.2 Implementation Details

The maximum number of words in the generated caption was set to $T_{\max} = 48$. The model was trained using a batch size of 512 for a total of 50 epochs with the Adam optimizer [4]. The learning rate was initialized to 4e-3 with a warm-up proportion of 0.1, where the learning rate linearly increased during the warm-up period and then linearly decayed to zero. For each video, we uniformly sampled $N_f = 20$ frames. A 12-layer ViT-B/32 with a patch size of 32 [3] (abbreviated as $\text{CLIP}_{\text{ViT}-\text{B}/32}$) was used to extract features. The extracted video features were fed into a two-layer Transformer video encoder with a hidden layer dimension of $d_h = 512$ to obtain enhanced visual features. These were then passed to a two-layer Transformer decoder with the same hidden size to generate captions. $d = 256$. A dropout probability of 0.1 was applied across all attention and other layers. The Spacy toolkit was used to determine word PoS within the captions.

B Dimensionality of Linear Self-Attention

The performance of the linear self-attention directly impacts the effectiveness of the subsequent syntactic prior branch, making the dimensionality of this layer

critically important. We evaluated two configurations with dimensions set to 512 and 255, respectively. The experimental results are presented in Table 1.

Table 1. Comparison of performance with different embedding dimensions

dimension	B@4	M	R	C
256	0.4809 ± 0.0038	0.3110 ± 0.0014	0.6476 ± 0.0020	0.5822 ± 0.0020
512	0.4785 ± 0.0022	0.3103 ± 0.0012	0.6452 ± 0.0014	0.5768 ± 0.0044

Experimental results show that the model performs best with a dimensionality of 256, while performance degrades when the dimension is increased to 512. This decline may be attributed to a higher risk of overfitting introduced by the increased dimensionality at 512.

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 190–200 (2011)
3. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
5. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
7. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
8. 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 conference on computer vision and pattern recognition. pp. 5288–5296 (2016)
9. Yang, B., Cao, M., Zou, Y.: Concept-aware video captioning: Describing videos with effective prior information. IEEE Transactions on Image Processing (2023)