Video-and-Language (VidL) models and their cognitive relevance (Supplementary Material)

Anne Zonneveld[1] Albert Gatt[2] Iacer Calixto[3]
[1]Amsterdam Brain and Cognition Center, University of Amsterdam
[2]Department of Information and Computing Sciences, Utrecht University
[3]Department of Medical Informatics, Amsterdam UMC, University of Amsterdam

A. Pretraining tasks

Below a short list of common pretraining tasks.

Masked Language Modelling (MLM)[2] requires the model to predict masked words based on their surrounding words and their visually aligned video frames.

Masked Frame Modelling (MFM)[5] requires the model to predict masked out video frame features (as extracted with CNNs), given the text and remaining video frames.

Masked Visual-token Modelling (MVM) is similar to MFM, except that it uses 'tokenized' video frames instead of video frame features. Video frames are translated into discrete visual tokens, which can be used to reconstruct masked (regions of) video frames. The method is first used by [1] over the temporal dimension and later by [3] in both the temporal and spatial dimension.

Masked Object Classification (MOC) [9] is also similar to MFM but requires the model to predict masked out regional object features, instead of frame video features.

Masked Action Classification (MAC) [9] requires the model to predict masked out action features based on the remaining linguistic features and object features.

Video Subtitle Matching (VSM) [5] requires the model to predict whether a subtitle matches the input video, as well as to retrieve the relevant moment of localization, ensuring global and local temporal alignment.

Masked Modal Modelling (MMM) [6, 8] requires the model to predict all tokens from a completely masked out modality, based on the tokens from a other modality.

Frame Order Modelling (FOM)[5] requires the model to reconstruct the original timestamps of a set of randomly shuffled video frames, explicitly ensuring temporal alignment.

Sentence Order Modelling (SOM) [4] requires the model to reconstruct the original sentence order in a set of randomly selected and shuffled sentences.

Cross-Modal Matching (CMM) was introduced as 'the linguistic-visual alignment classification objective' [1], while [9] later called it cross-modal matching. By adding a linear layer followed by a sigmoid activation function on top of the output of the first token ([CLS]), a cross-modality score is achieved that indicates the relevance of the linguistic information and visual features. Alternatively, a similarity calculation module can be added to the network which calculates and optimzes the representational similarity between visual and textual information [7].

Language Reconstruction (LR) [6] requires the model to reconstruct words based on masked ground-truth text and video. LR is different from MLM in that LR focuses on next word prediction, i.e. the model only attends to previous word and video tokens when predicting the next word.

References

[1] Chen , Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[3] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end

video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 1

[4] Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. Understanding chinese video and language via contrastive multimodal pre-training. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2567–2576, 2021. 1

[5] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 1

[6] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 1

[7] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 1

[8] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996*, 2021. 1

[9] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 1