# Explainable Video Entailment with Grounded Visual Evidence

Junwen Chen and Yu Kong
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY, USA
{jc1088, yu.kong}@rit.edu

Table 1. Comparison results on VIOLIN dataset. "Visual" column indicates the visual features used in entailment judgment.

| Method | Visual | Text | Accuracy |
|---|---|---|---|
| VIOLIN [7] | Resnet [2] | BERT [1] | 67.60 |
| Ours | Resnet [2] | BERT [1] | 68.39 |
| VIOLIN [7] | Detection [4] | BERT [1] | 67.84 |
| Ours | Detection [4] | BERT [1] | 68.42 |
| HERO [6] | HERO | | 68.59 |
| Ours | HERO | | **69.16** |

In this appendix, we compare our method with a video+language representation learning method HERO [6]. HERO [6] aims at learning a large-scale video+language pretraining to solve many downstream tasks, such as video entailment. Specifically, it is firstly pretrained on the large-scale TVShow [5] and Howto100M [8] datasets by several pretraining tasks such as Masked Language Modeling. Then, it is finetuned on the video entailment task. This large-scale pretraining model outperforms VIOLIN [7] in the video entailment task.

In this appendix, we evaluate the proposed method using HERO as a backbone. Specifically, we replace the visual and textual feature extraction backbones by the HERO pretrained encoder. The results in Table 1 show that the proposed method using HERO as a backbone outperforms the original HERO in video entailment. This is because our method performs a fine-grained understanding of videos.

Following VIOLIN [7], we also evaluate our method using detection features as visual embedding. We run Faster R-CNN trained on Visual Genome [3] to detect object in each frame and use the regional features as frame representation. Our method using detection features outperforms the VIOLIN using detection features.

## References

[1] 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.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[3] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

[4] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[5] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018.

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

[7] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *CVPR*, 2020.

[8] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019.