Multi-modal Representation Pre-training for Video Retrieval

Daya Guo
Sun Yat-sen University
guody5@mail2.sysu.edu.cn

ABSTRACT

This paper reports the 2nd place solution for QQ Browser 2021 AI Algorithm Competition Track1 in ACM CIKM 2021 AnalyticCup. As the number of short video in Internet grows rapidly, short video content understanding have become more crucial and attracted more attention from both academia and industry. Video content embedding that represents the semantics of video, plays an important role in scenes such as video de-duplication, relevance matching, ranking, diversity control, and etc. Therefore, QQ Browser hold the challenge for exploring video content embedding and provides millions of labeled videos in real business with tens of thousands of semantic video similarity annotations. The challenge defines the semantic video similarity as the cosine similarity between their content embedding, and evaluate the similarity by Spearman's rank correlation to the manual annotations. In this paper, we propose to pre-train cross-modal Transformer to learn multi-modal representation by four pre-training objectives, i.e. masked language modeling, inverse cloze task, SimCSE and video-text alignment. Experimental results show that our model achieve 2nd place in the challenge.

CCS CONCEPTS

• **Computing methodologies → Scene understanding**.

KEYWORDS

Multi-Modal Representation; Transformer; Pre-training.

ACM Reference Format:
Daya Guo. 2018. Multi-modal Representation Pre-training for Video Retrieval. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1122445.1122456

1 TASK DEFINITION

Given a short video represented in a sequence of frame features $H = \{h_0, h_1, ..., h_{n-1}\}$ with title $W = \{w_0, w_1, ..., w_{m-1}\}$ and speech $X = \{x_0, x_1, ..., x_{k-1}\}$, the task aims to encode the video into a semantic embedding $v \in R^N$, which can represents the semantic of short video and the dimension size N is less than 256. The challenge defines the cosine similarity of two embeddings as video semantic similarity between two short videos, and evaluate the similarity by Spearman's rank correlation to the manual annotations.

2 MODEL FRAMEWORK

Figure 1 gives an overview of our approach. As shown in the Figure, we regard the task as a retrieval task. We use a multi-modal encoder called VL-BERT [4] as the backbone to obtain multi-modal representation of two short video. Note that two VL-BERT in the Figure share parameters. The model takes text $T = \{W; X\}$ and video frame feature H as the input, and use multi-layer Transformer to encode short video. Finally, we apply mean pooling to representations from VL-BERT to get final vector.

3 PRE-TRAINING

In this section, we describe how to pre-train the model. To better encode short video, we propose to use four pre-training tasks to learn multi-modal representation.

3.1 Masked Language Modeling

We follow Devlin et al. [1] to apply masked language modeling (MLM) pre-training task. Specially, we sample randomly 15% of the tokens from the text. We replace them with a [MASK] token 80% of the time, with a random token 10% of the time, and leave them unchanged 10% of the time. The MLM objective is to predict original tokens of these sampled tokens, which has proven effective in previous works [4]. In particular, the model can leverage the video context if the text context is not sufficient to infer the masked word token, encouraging the model to align the natural language and video representations.

3.2 Inverse Cloze Task

Following Lee et al. [3], we also use Inverse Cloze Task (ICT) as our pre-trianing task, which has proven effective in retrieval tasks. Specially, we take a tile, speech or video as a query and other contexts as a document. The task requires retrieve the document in a batch by the given query.

3.3 SimCSE

SimCSE [2] is a simple strategy that provides positive examples for contrastive learning. We apply a dropout rate of 0.1 to Transformer encoders and the same input is fed to the encoder twice and one is used as the positive instance for contrastive learning.

3.4 Video-Text Alignment

We use the multi-modal representation that corresponds to the special token [CLS] to predict scores for the video-text alignment, which is similar to the BERT next sentence prediction task. We adopt the NCE loss to learn to discriminate against the positive from negative video-text pairs, which randomly sample 50% negative cases from dataset.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

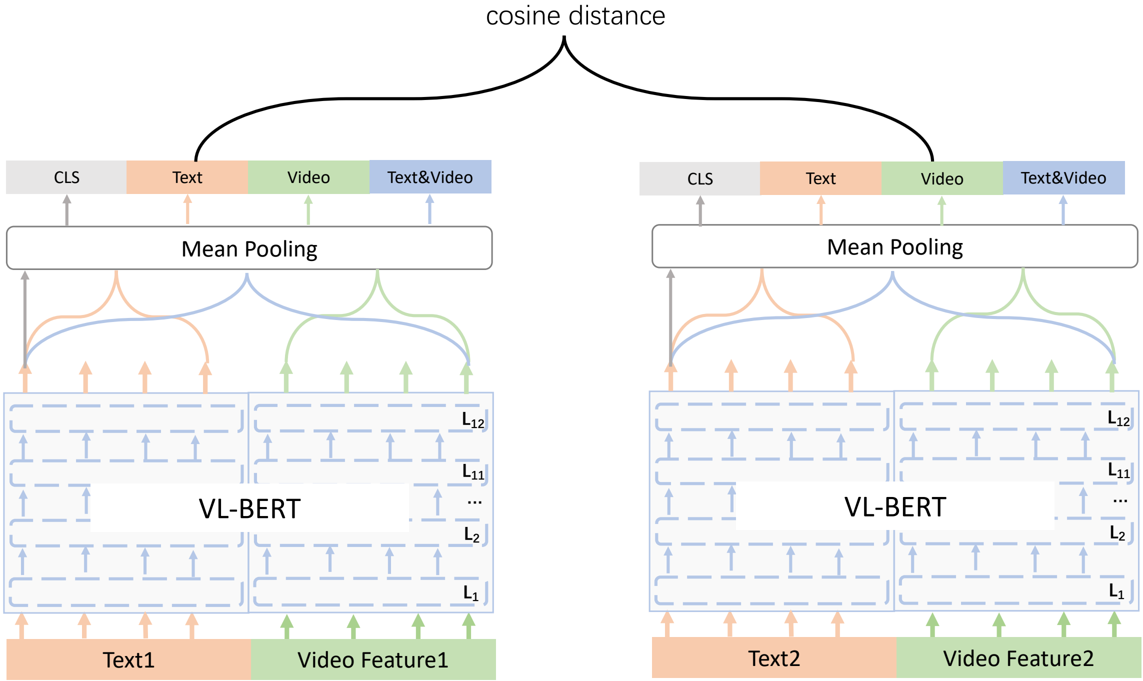

Figure 1: Overview of our proposed framework. We decouple the task of video ads content structuring into two subtasks. The first subtask is segment that generates proposals, which is shown in the left part. The second subtask is tagging that classifies each proposal, which is shown in the right part.

4 EXPERIMENT

We pre-train VL-BERT encoder to learn multi-modal representation from 1 million videos by proposed four pre-training tasks and fine-tune the model on 67,899 video pairs. Experiments show that our single model achieves 0.828 score and ensemble model ranks 2nd place in the challenge.

REFERENCES

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

[2] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv preprint arXiv:2104.08821* (2021).

[3] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300* (2019).

[4] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530* (2019).