CVPR 2020 Video Pentathlon Challenge:  
Multi-modal Transformer for Video Retrieval

Valentin Gabeur\textsuperscript{1,2}  
Chen Sun\textsuperscript{2}  
Karteek Alahari\textsuperscript{1}  
Cordelia Schmid\textsuperscript{2}  
\textsuperscript{1}Inria\textsuperscript{*} \textsuperscript{2}Google Research

Abstract

The task of retrieving relevant videos with natural language queries plays a critical role in effectively indexing large-scale video data. In this report, we present a framework based on a multi-modal transformer architecture, which jointly encodes the different modalities in video, and allows them to attend to each other. The transformer architecture is also leveraged to encode and model the temporal information. This novel framework allowed us to achieve the top result of the CVPR 2020 video pentathlon challenge. More details are available at http://thoth.inrialpes.fr/research/MMT

1. The Video Pentathlon Challenge

In this report, we present the method that we implemented for the CVPR 2020 video pentathlon challenge. This challenge tackles the task of caption-to-video retrieval. Given a query in the form of a caption, the goal is to retrieve the videos best described by it. The challenge considers 5 datasets: ActivityNet, DiDeMo, MSRVTT, MSVD and YouCook2. For each dataset, our task is to provide, for each caption query of the test set, a ranking of all the test video candidates such that the video associated with the caption query is ranked as high as possible.

2. Methodology

Our overall method is greatly inspired by MEE \cite{6} and CE \cite{5}. It relies on learning a function $s$ to compute the similarity between text and video, as shown in Fig. 1. All videos in a dataset are then ranked according to their similarities with the query caption.

Video representation. We begin our video encoding with pre-trained video feature extractors called “video experts” \cite{5, 6}. They were provided by the challenge organiz-
Figure 1. Our cross-modal framework for similarity estimation. We use our Multi-modal Transformer (MMT, right) to encode video, and BERT (left) for text.

We use a batch size of 64 pairs except for YouCook2 where we use a batch size of 300 pairs. We set the epoch size at 200 optimization steps for YouCook2, 500 optimization steps for ActivityNet and DiDeMo, 1000 optimization steps for MSRVTT and 2000 optimization steps for MSVD.

To compute our caption representation, we use the “BERT-base-cased” checkpoint of the BERT model that we finetune. For our multi-modal transformer video encoder, we use a hidden size of 512 and an intermediate size of 3072. Training a model on ActivityNet takes about 3 hours on a single Nvidia V100 16GB GPU.

3.2. Hyperparameter search

For each dataset, we have run a hyperparameter search using the Tune implementation of the Asynchronous Successive Halving algorithm over 3 brackets with a maximum number of 30 epochs, a grace period of 3 epochs and a reduction factor of 2. Parameters we considered for the searches: learning rate, learning rate decay rate, dropout rate for BERT and MMT, loss margin, weight decay, number of experts to use (random combination), proportion of each dataset in the pre-training dataset mix, after how many epochs do we stop pre-training on the dataset mix and start finetuning on the target dataset only, number of layers and number of attention heads for MMT.

3.3. Ensembling

We stop the hyperparameter search once 20 experiments have terminated, and select the 5 best performers to re-run them on 8 randomly selected validation splits. For each of the 8 validation splits, we select the 2 best performing models, therefore keeping 16 models for ensembling. Each model provides a ranking of the candidate videos for each query. We ensemble the results following a penalty approach: if a video is ranked at the first position, it gets 0 penalty, a video ranked at the second position gets penalty of 1, and so on. The candidate videos are finally ranked according to the sum of penalty points across the models.

References