Hybrid Sequence Encoder Of Collaborative Experts For Video Retrieval

Kaixu Cui\textsuperscript{1,2}, Hui Liu\textsuperscript{1,2}, Chen Wang\textsuperscript{1,2}, Changliang Xu\textsuperscript{1,2}, Yudong Jiang\textsuperscript{1}\textsuperscript{1}XinHuaZhiYun Inc.
\textsuperscript{2}State Key Laboratory of Media Convergence Production Technology and Systems
{\{cuikaixu, liuhui, wangchen, xuchangliang\}@shuwen.com}
nebuladream@gmail.com(Work at in XinHuaZhiYun Inc.)

Abstract

In this report, we present our method in the CVPR 2020 Video Pentathlon challenge. In the challenge we use hybrid sequence encoder to extract information from the features of video and combine it with the Collaborative Experts. In the training phase, we use datasets fusion for pre-training. In the inference phase, we use the Query Ensemble improve performance of model.

1. Introduction

The goal of the CVPR 2020 Video Pentathlon challenge is to build a system for five video retrieval benchmarks (MSRVTT, DiDeMo, ActivityNet, MSVD, YouCook2), and the system only allowed to use provided multimodal features. In this report, we present our solution to the CVPR 2020 Video Pentathlon challenge. In the challenge, Firstly, we use a hybrid sequence encoder to modify the Collaborative Experts [1]. Secondly, in the training phase, we use datasets fusion to improve the generalization of the model. Finally, in the inference phase, we use query ensemble for MSRVTT and MSVD.

**Hybrid Sequence Encoder.** Only simple embedding is used in the Collaborative Experts, but the simple embedding ability is weak, not enough to represent complex video and text information, such as temporal information. We employ the Hybrid Sequence Encoder in combination with the Collaborative Experts to construct a common space. The hybrid sequence encoder not only use features extracted by different experts, but also extract more information on these features.

**Datasets Fusion.** In the training phase, we used datasets fusion. Firstly, we fused the five datasets for pre-training. Secondly, we fine-tuning on each dataset.

**Query Ensemble.** We found that each query of the MSRVTT and MSVD datasets corresponds to multiple captions, so we use query ensemble to obtain the similarity of the query and the video.

2. Method

2.1. Hybrid Sequence Encoder

In our solution, we use the Hybrid Sequence Encoder to extract information on the features. We tried two structures, the first one, we use encoder of Dual-encoder method [2] as the Hybrid Sequence Encoder, Figure 1(a) illustrates the first Hybrid Sequence Encoder. The second kind, we add GhostVLAD [3] and the Gate Convolution Network [4] to the first Hybrid Sequence Encoder, because it only focus on current input ignoring the whole distribution of the whole datasets and GhostVLAD is applied to embedded into the model so that the model can have a more general view among all training data. Figure 1(b) illustrates the second hybrid sequence encoder.

![The first hybrid sequence encoder.](image)

(a) The first hybrid sequence encoder.

![The second hybrid sequence encoder.](image)

(b) The second hybrid sequence encoder.

Figure 1: The overview of our hybrid sequence encoders.

In our models, text features encoder is the same as other experts features encoder and the text features are extracted using the Open-AI model. If the expert does not have segment features, we use its own simple embedding. We use Full-connection to make output dimension reach the common space dimension. The parameters of modules in
our model is shown in Table 1.

<table>
<thead>
<tr>
<th>Modules</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiGRU</td>
<td>512</td>
</tr>
<tr>
<td>Multi-kernel 1-d conv</td>
<td>Experts kernels size: 2,3,4,5 / Text kernels size: 2,3,4</td>
</tr>
<tr>
<td>GhostVLAD</td>
<td>Cluster: 32 or 16 / Ghost clusters: 1</td>
</tr>
</tbody>
</table>

2.2. Datasets Fusion

In the training phase, we combining different datasets for train model. According to our experiments, simply fusing the datasets is not useful because there are differences between different datasets. To solve this issue, divide the training stage into three steps and only use the fusion dataset for pre-training. The three steps of training are shown in Figure 2.

![Figure 2: The three steps of training. Firstly, fuse the five datasets for pre-training. Secondly, evaluate the pre-trained models on the five datasets, selected the best performing model of each dataset. Finally, Fine-tuning the best performing model on the corresponding dataset.](image)

2.3. Query Ensemble

We found that each query of the MSRVTT and MSVD datasets corresponds to multiple captions, so we use the Query Ensemble to obtain the similarity of the query and the video, the process is shown in Figure 3. Firstly, calculate the similarity between each caption of the query and the video. Then obtain the similarity by averaging the similarity of all captions of the query with the video. The average similarity is used as the similarity of each caption.

![Figure 3: The process of the Query Ensemble. According to our experiments, we found that the Query Ensemble can significantly improve the model's performance on MSRVTT and MSVD.](image)

3. Experiments

The bi-directional max-margin ranking loss used to learn the joint embedding space for our models. We compare our two Hybrid Sequence Encoder with CE[1]. The results that evaluated on the val sets provided by the organizer are shown in Table 2. “CE” denotes the Collaborative Experts model, “HSE1” denotes the first kind of hybrid sequence encoder, “HSE2” denotes the second kind of hybrid sequence encoder, “ALL” denotes train on the fusion dataset, “FT” denotes fine-tuning on each dataset.

![Table 2: The results on five datasets.](image)

4. Conclusion

From the results, we can see that our method has improved on the data set other than youcook2, but the result on the youcook2 data set is not as good as using CE directly, which might result from too much the parameters cause the model to overfit on the training set of youcook2.

References