Order-Prompted Tag Sequence Generation for Video Tagging
Supplementary Materials

Table 1. Statistics for the CREATE-210k dataset. “IS.” is the abbreviation of intersection.

<table>
<thead>
<tr>
<th>Data</th>
<th>Videos</th>
<th>Unique tags</th>
<th>Union</th>
<th>IS.</th>
<th>Common</th>
<th>Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>210k</td>
<td>17573</td>
<td>18464</td>
<td>2795</td>
<td>705</td>
<td>2090</td>
</tr>
<tr>
<td>Test</td>
<td>5k</td>
<td>3686</td>
<td>2795</td>
<td>2795</td>
<td>705</td>
<td>2090</td>
</tr>
</tbody>
</table>

Table 2. Statistics for the Pexel dataset. “IS.” is the abbreviation of intersection.

<table>
<thead>
<tr>
<th>Data</th>
<th>Images</th>
<th>Unique tags</th>
<th>Union</th>
<th>IS.</th>
<th>Common</th>
<th>Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>162k</td>
<td>27752</td>
<td>28094</td>
<td>5669</td>
<td>1627</td>
<td>4042</td>
</tr>
<tr>
<td>Test</td>
<td>5k</td>
<td>6003</td>
<td>28094</td>
<td>5669</td>
<td>1627</td>
<td>4042</td>
</tr>
</tbody>
</table>

Figure 1. Tag distributions of the CREATE-210k dataset (a) and the Pexel dataset (b). The horizontal axis represents the index of the tag, and the vertical axis is the square root of the tag frequency for better visualization.

1. Benchmarks Details

1.1. CREATE-tagging Benchmark

The core part of CREATE-tagging is the CREATE-210k dataset [8]. As shown in Table 1, CREATE-210k contains 162k videos with 17573 unique tags for training and 5k videos with 3686 unique tags for test. A total of 18464 unique tags appear in CREATE-210k, of which 2795 tags appear in both the training and test data and are further divided into 705 common tags and 2090 rare tags to evaluate the performance of different models. The distributions of the 2795 tags in the training and test data are shown in Figure 1 (a), it can be seen that the tag distribution of the test data (right side) is not exactly consistent with the tag distribution of the training data (left side), i.e., some rare tags in the training data appear frequently in the test data while some common tags appear infrequently in the test data.

1.2. Pexel-tagging Benchmark

Pexel-tagging is built on the newly collected and challenging Pexel dataset. According to the statistics in Table 2, Pexel consists of 162k images with 27752 unique tags for training and 5k images with 6003 unique tags for test. There are 28094 unique tags appearing in Pexel, of which 5669 tags appear in both the training and test data and are further divided into 1627 common tags and 4042 rare tags to evaluate the performance of different models. The distributions of the 5669 tags in the training and test data are shown in Figure 1 (b), compared with CREATE-210k, the tag distribution of the test data is more consistent with that of the training data.

2. Comparisons of Tagging Datasets and Multi-Label Classification Datasets

We list the details of video/image tagging datasets and video/image multi-label classification datasets in Table 3 to compare their differences more intuitively, and we observe two important distinctions: (1) Video/image tagging datasets are oriented towards open scenarios. In contrast, multi-label classification datasets are restricted to specific scenarios, i.e., video multi-label classification datasets THUMOS14 and ActivityNet1.3 focus on actions, and im-
Figure 2. Architectures of different models: (a) Our OP-TSG, (b) Model F in Table 7 of the main paper, which uses prompt labels for handling meaningless prompts. (c) Model G in Table 7 of the main paper, which uses prompts for classification rather than generation.

3. Implementation details

3.1. Training schedules of OP-TSG

The training schedules of OP-TSG on CREATE-tagging and Pexel-Tagging are as follows: (1) For CREATE-tagging, we pre-train the model for 20 epochs with a batch size of 1024 on 16 NVIDIA V100 GPUs, followed by 30 epochs of fine-tuning with a batch size of 512 on 8 NVIDIA V100 GPUs. (2) For Pexel-tagging, we train the model for 20 epochs with a batch size of 512 on 8 NVIDIA V100 GPUs. Other methods used for comparison in Table 1 and Table 2 of the main paper also adhere to the same training schedules.

3.2. Comparisons of Model Architectures

The architectures of our OP-TSG, model F and model G in Table 7 of the main paper are shown in Figure 2 (a), Figure 2 (b) and Figure 2 (c), respectively. OP-TSG adopts the pre-defined [PAD] tags to assign to the meaningless order prompts. Model F trains a binary classification head that takes the order prompts as input, and predicts label 1 for the prompts aligned with meaningful tags and label 0 for the prompts aligned with [PAD] tags. Then prompts with label 1 are fed into the tag decoder and the target sequence is the concatenation of the aligned tags of the prompts. Model G directly attaches a multi-classification head on the order prompts, and trains the multi-classification head to predict the aligned tags of the prompts.

4. Inference Time Measurement

We evaluate the inference time of different models on Pexel-tagging using a single NVIDIA V100 GPU with a test batch size of 8, and the results are presented in Table 4. The classification model ASY runs the fastest but has the worst performance. Compared with the generation model open-book, our method improves the F1 score by 2.4% on all tags and expands the inference time to 1.7 times, the reason for the increase in inference time is that our method needs to generate longer tag sequences containing [PAD] tags.

5. More Visualizations

5.1. Tag inference results on CREATE-tagging

In Figure 3, we show examples of tag inference results of different methods on CREATE-tagging benchmark. We can observe that: (1) Our method is better at generating a more comprehensive tag set. As shown in Figure 3 (a), the classification method misses the tags “appetizers” and “cold noodles” and the generation model misses the tags “appetizers” and “summer noodles”, while our method provides a complete set of tags. (2) Our method is able to infer the wrong and meaningful tag that is outside the annotations but present in the training data and consistent with the video content, e.g., the tag “western dessert” in Figure 3 (b). (3) Our method may also produce some wrong and meaningless tags, such as the tags “amateur billiard”, “top skills” and “waiters” in Figure 3 (c). (4) Our method can generate novel meaningful tag that is beyond the annotations and not present in the training data, such as the tag “Tianmen Mountain” in Figure 3 (d).

5.2. Tag inference results on Pexel-tagging

We also present examples of tag inference results of different methods on Pexel-tagging benchmark in Figure 4,
Figure 3. Examples of tag inference results from multiple methods on CREATE-tagging. “Cls.” and “Gen.” indicates the classification method Asy and generation method Open-Book, respectively. The tags in black, green, red, and purple are common tags, rare tags, incorrect tags, and novel tags, respectively.

Figure 4. Examples of tag inference results from multiple methods on Pexel-tagging. “Cls.” and “Gen.” indicates the classification method Asy and generation method Open-Book, respectively. The tags in black, green, red, and purple are common tags, rare tags, incorrect tags, and novel tags, respectively.
and we make the following observations: (1) Our method is able to generate an accuracy and complete tag set, e.g., in 4 (a), the classification method provides the wrong tags “leaves” and “season” and the generation model misses the tag “brown leaves”, while our method generates all tags accurately. In addition, when our method misses tags, other methods will tend to miss more tags as well. As shown in Figure 4 (b), the classification method, the generation method and our method miss three, two and one tags, respectively. (2) Our method has the ability to infer wrong and meaningful tag that is outside the annotations but present in the training data and consistent with the image content, such as the tag “drumsticks” in Figure 3 (c). (3) Our method may also make the mistake of generating some wrong and meaningless tags, such as the tags “eye level shot” and “aesthetic” in Figure 3 (d).

References


