LoL-V2T: Large-Scale Esports Video Description Dataset

Tsunehiko Tanaka  
Waseda University  
tsunehiko@fuji.waseda.jp

Edgar Simo-Serra  
Waseda University  
ess@waseda.jp

Abstract

Esports is a fastest-growing new field with a largely online-presence, and is creating a demand for automatic domain-specific captioning tools. However, at the current time, there are few approaches that tackle the esports video description problem. In this work, we propose a large-scale dataset for esports video description, focusing on the popular game “League of Legends”. The dataset, which we call LoL-V2T, is the largest video description dataset in the video game domain, and includes 9,723 clips with 62,677 captions. This new dataset presents multiple new video captioning challenges such as large amounts of domain-specific vocabulary, subtle motions with large importance, and a temporal gap between most captions and the events that occurred. In order to tackle the issue of vocabulary, we propose a masking the domain-specific words and provide additional annotations for this. In our results, we show that the dataset poses a challenge to existing video captioning approaches, and the masking can significantly improve performance. Our dataset and code is publicly available.

1. Introduction

Esports are growing rapidly in popularity and have generated $950 million in revenue in 2019 with a year-on-year growth of +22.7% and is approaching the scale of existing sports leagues [19]. The main content in esports consists of tournament or gameplay videos, however, they are challenging for a new audience to understand given that they contain significant information (e.g., character hit points, skill cool time, and effects of using items). Captions are a useful tool for viewers to understand the status of a match. Recently, video description approaches have started to tackle the sports domain [32, 34], however, there are few approaches tackling the esports domain. We believe this is caused by a lack of large-scale datasets, and present the LoL-V2T dataset to address this issue.

1https://github.com/Tsunehiko/lol-v2t
Our dataset for video description in esports consists of narrated videos of the League of Legends world championships “Worlds”. We chose League of Legends as the sports title as it is one of the most popular esports games and thus it is easy to obtain significant amount of narrated videos. We split each video into different scenes, which are then filtered such that non-gameplay related scenes are removed. Afterwards, we extract complete sentence captions from the narrations, and compute the temporal boundaries corresponding to each of the captions. After this processing, we obtain a total of 9,723 clips with 62,677 captions, where each clip is associated to multiple captions. LoLV2T has three different challenges for training captioning models. First, the captions contain a significant amount of proper nouns specific to the esports title and numerals that describe the status of matches. Second, some important objects in the clips are difficult for captioning models to recognize because their size and subtle motions. Third, some pairs of a clip and captions are not necessarily aligned temporally, i.e., there can be significant lag between game actions and captions.

In this paper, we additionally tackle the difficulty corresponding to the large amounts of proper nouns and numerals. While they are important for describing gameplay, memorizing them is difficult for captioning models because of the large variety and low frequency of occurrence of each word. Furthermore, they change over time as the game is updated and renewed. To tackle this problem in the LoLV2T dataset, we rely on masking. In particular, our preprocessing approach consists in a masking scheme for proper nouns and numerals in the captions. First, we classify the proper nouns in the captions into several groups according to their meaning. We treat numerals as a group. Then, when a proper noun or a numeral appears in a caption, we replace the word with the name of the group to which it belongs. By using this method, we can increase the frequency of proper nouns and numerals while keeping the group names in the captions that are simple enough for a new audience to comprehend. Our key task is to generate multi-sentences for esports videos, and we use video paragraph captioning models that build upon the models of Vanilla Transformer [37] and MART [15]. An overview is shown in Figure 2. We show the challenges of LoLV2T and corroborate the effectiveness of our proposed approach through experiments.

2. Related Work

Video Description Dataset

Various datasets for video description have been proposed covering a wide range of domains such as cooking [36, 22], instructions [18], and human activities [24, 31, 14]. We summarize existing video description datasets and
Video Description Datasets

Several works [32, 34] on video description for existing sports videos have been proposed. Yan et al. [32] use a tennis video as input and generate captions using Structured SVM and LSTM. Yu et al. [34] used videos of NBA games as inputs and generated captions using LSTM and subnetworks suitable for basketball videos.

Esports footages contain much more information than existing sports videos, making video comprehension difficult. Some video description efforts target video games as complex as esports games and use “Let’s Play” videos as input. “Let’s Play” videos contain audio of players’ comments on gameplay, which can be converted to text by an Automatic Speech Recognition (ASR) system and used as captioning data. Shah et al. [23] generated a caption for each frame by training a simple CNN model that combines three conv layers with 75 minutes of Minecraft “Let’s Play” videos and 4,840 sentences. Li et al. [16] applied sequence-to-sequence networks with attention to this task, using a dataset of 110 minutes of “Let’s Play” videos by Getting Over It with Bennett Foddy and 2,274 sentences. Various inputs such as video, optical flow, and audio are compared.

In this paper, we propose LoL-V2T, a dataset using video game footages in esports that focuses on gameplay. LoL-V2T consists of 4,568 minutes of video and 62,677 sentences, which is much larger than existing datasets in the video game domain.

3. LoL-V2T Dataset

We create a new dataset for video description in esports, LoL-V2T. LoL-V2T consists of 9.7k clips of League of Legends playing video taken from YouTube and 63,000 captions. Each video is associated with multiple captions based

Table 1: Comparison of the existing video description datasets. Multiple indicates whether or not multiple captions are associated with a single clip. Narration indicates whether or not captions are generated from the narration contained in the video.
Table 2: Performance of the model to detect whether a clip is related to gameplay.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.963</td>
<td>0.928</td>
<td>0.285</td>
<td>0.437</td>
</tr>
</tbody>
</table>

3.1. Data Collection

We collect footages of 157 matches of the League of Legends world championships “Worlds” from YouTube. League of Legends is the most popular esports title, which is the most-watched game on Twitch and YouTube [19]. Besides, since Worlds has a large number of matches and commentators always provide commentary in them, it is easy to obtain narrations for the videos. While “Let’s Play” videos include narrations not related to gameplay, the quality of narrations in esports tournament footages is higher than that of “Let’s Play” videos because the purpose of commentators is not to enjoy viewers but to explain gameplays.

3.2. Splitting Videos and Selecting Clips

The average length of collected League of Legends videos is 44 minutes, which is too long to generate captions. The videos include scenes not related to gameplay, such as player seats and venue scenes. In order to reduce noise to the model training, we first split videos into clips by scene, and then remove the clips which are not gameplay. This process is shown in Figure 3.

We automatically split the video into clips of a length that the video description model can handle. For this splitting, we use PySceneDetect\(^2\) which is a tool that can detect scene changes in videos. The average length of the clips is 23.4 seconds.

To remove the clips not related to gameplay, we create a model to detect whether a clip is a gameplay clip. It takes a temporally centered RGB frame as input and outputs 0 or 1 to indicate whether it is relevant to gameplay or not, respectively. ResNext-50(32x4d) [29] is adopted as the model, and the clips segmented by the splitting tool are used as the dataset for training. As shown in Table 2, precision is much higher than recall in this model to reduce false-positive even the amount of data decrease and then preserve the quality of the dataset.

3.3. Generation of Captions and Temporal Boundaries

We produce full-text captions and temporal boundaries from the subtitles generated from narrations by YouTube. Subtitles are usually organized as a list of text chunks. Each chunk is not a complete sentence and associated with a specific time interval in the video. To help captioning models to understand the context, we re-segment the chunks into complete sentences with a sentence segmenter. For a sentence segmenter, we use DeepSegment\(^3\).

The relationship between chunk and temporal boundaries breaks down when the sentence is reconstructed. We compute the new temporal boundaries to which the reconstructed sentences correspond using the number of words in the sentences. The sequence of captions and temporal boundaries generation is shown in Figure 4.

3.4. Dataset Analysis

LoL-V2T is a labeled dataset with 4,568 minutes of video and 62,677 captions. As shown in Table 1, LoL-V2T is larger than the existing dataset [16] in the video game domain and contains video-text data as large as medium scale datasets frequently used in video description, such as Charades [24] and MSR-VTT [31]. The number of clips is 9,723, the average length of clips is 28.0 seconds, and the average length of intervals between temporal boundaries is 4.38 seconds. The mean number of words in a caption is 15.4.

In LoL-V2T, same as in ActivityNet Captions [14], multiple captions are associated with a single clip. This dataset can also be used for the task of inferring temporal boundaries (Dense Video Captioning) in the future. LoL-V2T has three features that make it more difficult to train captioning models compared to ActivityNet Captions. First, there are more proper nouns and numerals in LoL-V2T than in ActivityNet Captions as shown in Figure 5. They interfere with the training of captioning models. They also are important elements in describing the information of a game. However, too many of them can result in captions that become difficult for beginners to comprehend. Second, motions of important objects for gameplay in the clips are too subtle to be recognized by captioning models. The size of characters in League of Legends are smaller than that of people. This indicates that it is difficult for captioning models pre-trained on human activities to identify the clips in LoL-V2T. Finally, the clips and captions do not necessarily temporally match.

4. Proposed Method

In this section, we introduce a method for video description in esports. Our work builds upon the vanilla transformer for video description proposed by Zhou et al. [37]
Figure 3: Overview of splitting and selection videos in League of Legends (LoL). We split the LoL gameplay videos into clips for each scene because the average length of them is too long to be handled by the video description model. We remove clips not related to gameplay with ResNext-50(32x4d) [29] to minimize noise in training.

Figure 4: Sequence to create captions and temporal boundaries from subtitles. First, we concatenate several consecutive chunks and split them into complete sentences by a sentence segmenter. Next, the temporal boundaries are calculated according to the number of words in reconstructed sentences. In this example, four words appear between 00:01:00.000 and 00:03:00.000, half of which in the previous caption, so the last timestamp of the previous caption is 00:02:00.000.

and MART [15] based on the transformer with a memory module for modeling of history information. We modify them by masking proper nouns in preprocessing as we will next explain.

4.1. Model

**Vanilla Transformer.** This model is an application of the transformer [25] to the video description task, proposed by Zhou et al. [37], as shown in Figure 2-(left). It contains two components: a video encoder and a caption decoder. The video encoder is composed of a stack of two identical layers and each layer has a self-attention layer [25] and a position-wise fully feed-forward network. The caption decoder inserts a multi-head attention layer over the output of the video encoder stack in addition to the two layers in the video encoder. We also create masks with ground-truth temporal boundaries and apply them to the outputs of the video encoder to make the caption decoder focus on the event proposals in the clips. Although there is the proposal decoder that outputs event proposals and uses them to make the masks in [37], we remove it to simplify the task by focusing on the inference of captions.

**MART [15]** is a model based on the transformer [25] for the video paragraph captioning task, as shown in Figure 2-(right). The transformer model decodes each caption individually without using the context of the previously generated captions as it can operate only separated fixed-length segments. MART has two changes from the transformer to solve this problem. The first change is a unified encoder-decoder design, where the encoder and decoder are shared.
The second change is an external memory module similar to LSTM [11] and GRU [5] that enables the modeling of history information of clips and generated captions. With these two improvements, MART is able to use previous contextual information and generate a better paragraphs with higher coherence and less repetition.

4.2. Masking

As described in Section 3.4, the proportion of proper nouns and numerals to words in LoL-V2T is higher than in other datasets. Although proper nouns and numerals are important for explaining gameplay, they are difficult for captioning models to learn. To address this problem, we propose a masking method in preprocessing for captions. First, we classify the proper nouns in the captions into groups by meaning. Then, when a proper noun or numeral appears in a caption, we mask it with the name of the group to which it belonged. An example of masking is shown in Figure 6.

Since the masking increases the frequency of proper nouns and numerals, it makes them easier to recognize by the model. In addition, complex proper nouns with detailed meanings are replaced by simpler proper nouns, resulting in more comprehensible captions. We treat numerals as a group and deal with misspellings in ASR. The classification into groups is done manually by knowledgeable people in League of Legends. Examples of the created groups are shown in Table 3.

5. Experiments

In this section, we show the experimental results of video description for esports footages using LoL-V2T. We also demonstrate that our masking method improves generated captions. We measure the captioning performance with the automatic evaluation metrics: BLEU@4 [21], RougeL [17], METEOR [2], and Repetition@4 [30].

5.1. Implementation Details

We train the model on the LoL-V2T dataset with the split in the Table 4. To preprocess the videos, we down-sample each video every 0.26s and extract the TSN features [28] from these sampled frames. The TSN model extracts spatial features from RGB appearance and temporal features from optical flow and concatenates the two features. For TSN, our implementation was build upon the mmaction2 [6], we use the two different ResNet-50 [10] model pre-trained on...
<table>
<thead>
<tr>
<th>Group Name</th>
<th>Meaning</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team</td>
<td>names of teams competing in “Worlds”</td>
<td>fnatic, fanatic, Cloud9, C9, genji, Gen.G, g2</td>
</tr>
<tr>
<td>Player</td>
<td>names of players competing in “Worlds”</td>
<td>Baker, Matt, Svenskeren, Ceros, Diamondprox, dyNquedo</td>
</tr>
<tr>
<td>Champion</td>
<td>names of characters in League of Legends</td>
<td>Recon, Akali, Kaiser, Ryze, Ziya, Camille</td>
</tr>
<tr>
<td>Monster</td>
<td>names of monsters in League of Legends</td>
<td>Drake, Herald, Raptor, Krug</td>
</tr>
</tbody>
</table>

Table 3: Examples of created proper noun groups. Proper nouns related to competitions, such as Team and Player, and proper nouns related to gameplay, such as Champion and Monster, are separately grouped. Moreover, misnomers in ASR, such as “fnatic” and “fanatic”, are addressed in this group.

<table>
<thead>
<tr>
<th>Split</th>
<th>train</th>
<th>validation</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>clips</td>
<td>6,977</td>
<td>851</td>
<td>1,895</td>
</tr>
<tr>
<td>captions</td>
<td>44,042</td>
<td>5,223</td>
<td>13,412</td>
</tr>
</tbody>
</table>

Table 4: The splits of the LoL-V2T dataset.

Kinetics-400 [12] without fine-tuning. We also use tvl1 [35] for calculating optical flow. All captions are converted to lowercase. The words that occur less than 5 times in all captions are replaced with “<unk>” tags. All settings of our implementations for the vanilla transformer and MART are the same as in [37] and [15], respectively.

5.2. Evaluation Metrics

We measure the performance on the video description task with four automatic evaluation metrics: BLEU@4 [21], RougeL [17], METEOR [2], and Repetition@4 [30]. We use the standard evaluation implementation from the MSCOCO server [4].

5.3. Results and Analysis

We compare the case where our proposed masking is included in preprocessing (Masking) with the case where it is excluded (baseline) using Vanilla Transformer and MART trained on the LoL-V2T dataset. The vocabulary of the training data is 4,528 words after our proposed preprocessing and 5,078 words after the baseline preprocessing.

The results on the testing set in the LoL-V2T dataset are shown in Table 5. We can see that our proposed masking outperforms the baseline in BLEU@4, RougeL, and METEOR. It shows that by grouping proper nouns with a low frequency of occurrence using our proposed masking, the frequency of occurrence increases, which helps captioning models to recognize proper nouns. We can also display that MART outperforms Vanilla Transformer in Repetition@4, which show that the recurrent module in MART is effective in suppressing repetitive expressions.

Qualitative results are shown in Figure 7. The models trained on our proposed masking (VTransformer+Masking, MART+Masking) generate captions containing the names of the groups such as “<Player>” and “<Team>” and this tendency is particularly prominent in MART, so the captions with our method are more explicit than the ones with the baseline. The models also generate gameplay keywords such as “kill” and “fight”, and simple League of Legends terms such as “turret”, which indicates that the captions can explain the contents in more detail. On the other hand, the models trained with the baseline (VTransformer, MART) generate keywords such as “mid lane” and “damage”, but there are many “<unk>” occurrences. Besides, some of the captions in VTransformer are repetitions of “<unk>” and do not form sentences. It is shown that ours generates more comprehensible captions than the baseline.

6. Conclusion

In this paper, we introduced LoL-V2T, a new large-scale dataset for video description in esports with 9,723 clips and 62,677 captions, where each clip is associated with multiple captions. LoL-V2T has three difficulties for training captioning models. First, it has a lot of proper nouns for esports and numerals in captions. Second, the motions of objects in clips are subtle. Third, clips and captions do not necessarily have a direct temporal correspondence. We addressed the first difficulty initially, and proposed a preprocessing method consisting of masking proper nouns and numerals in captions. Our captioning models build upon [37] and [15]. Experimental results show that our proposed approach can improve the generated captions. In the future, addressing the remaining two difficulties should further improve performance. We hope that with the release of our LoL-V2T dataset, other researchers will be encouraged to advance the state of esports research.
<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU@4↑</th>
<th>RougeL↑</th>
<th>METEOR↑</th>
<th>Repetition@4↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>VTransformer</td>
<td>1.53</td>
<td>12.76</td>
<td>8.58</td>
<td>37.33</td>
</tr>
<tr>
<td>MART</td>
<td>2.13</td>
<td>14.48</td>
<td>11.50</td>
<td>8.76</td>
</tr>
<tr>
<td>VTransformer+Masking</td>
<td>3.14</td>
<td>16.57</td>
<td>12.03</td>
<td>29.20</td>
</tr>
<tr>
<td>MART+Masking</td>
<td>3.56</td>
<td>15.39</td>
<td>12.98</td>
<td>9.74</td>
</tr>
</tbody>
</table>

Table 5: Captioning results on testing set in LoL-V2T dataset. We evaluate the performance of our proposed masking using BLEU@4, RougeL, METEOR, and Repetition@4.

Figure 7: Qualitative results on the testing set in the LoL-V2T dataset. Colored texts highlight relevant content in clips. Captions generated by the model trained by the proposed method contain more words related to the screen in the sentence than the baseline. Some references are not complete sentences because they are generated by ASR.
References


