LoL-V2T: Large-Scale Esports Video Description Dataset

Tsunehiko Tanaka
Waseda University
tsunehiko@fuji.waseda.jp

Edgar Simo-Serra
Waseda University
ess@waseda.jp

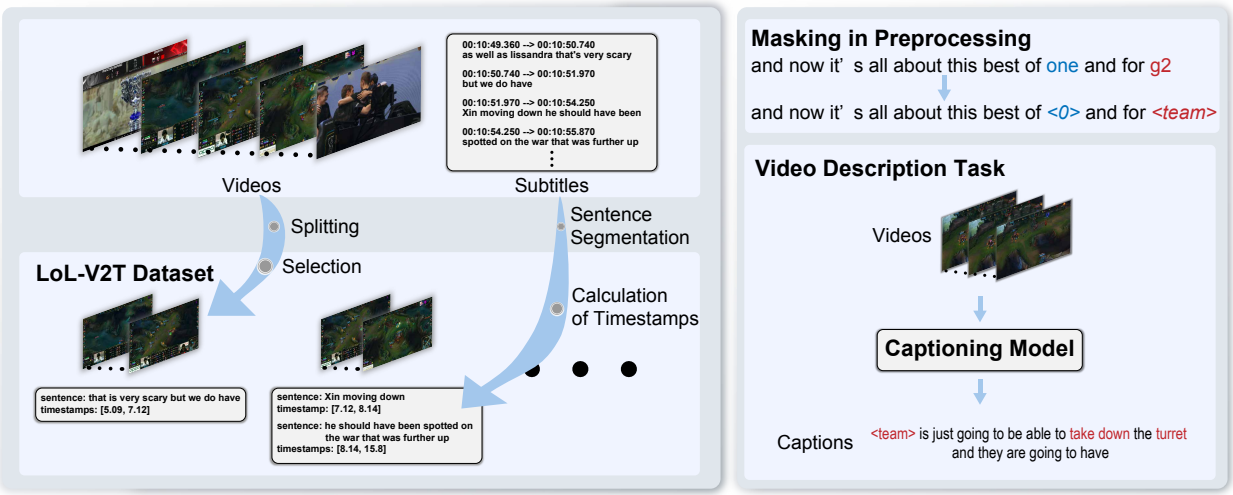


Figure 1: Overview of our video description approach for esports. Left: We created a new large-scale esports dataset consisting of gameplay clip with multiple captions. Right: We mask domain-specific words in captions to improve training and generalization abilities of our model.

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].

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.

[1] https://github.com/Tsunehiko/lol-v2t

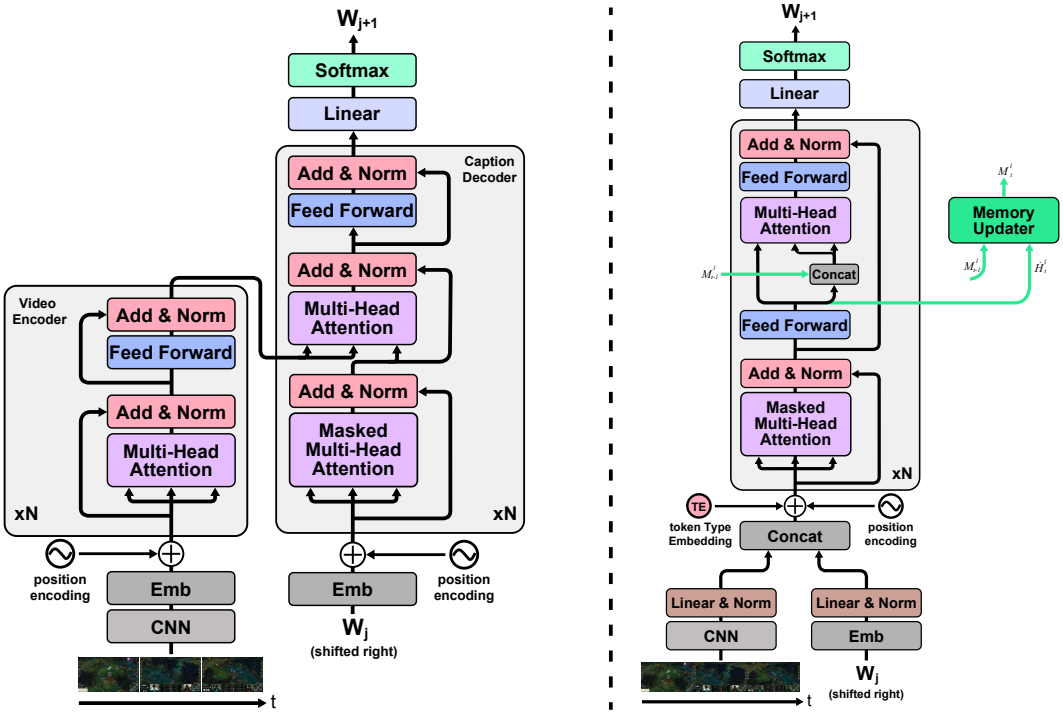


Figure 2: Overview of our captioning models. The left model is based on the vanilla transformer proposed by Zhou *et al*. [37], while the model on the right is based on the transformer with recurrent memory module [15].

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. *LoL-V2T* 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 *LoL-V2T* 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 *LoL-V2T* 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

Name	Domain	Clips	Captions	Duration	Multiple	Narration
Charades [24]	human	10k	16k	82h	-	-
MSR-VTT [31]	open	10k	200k	40h	-	-
YouCook2 [36]	cooking	14k	14k	176h	✓	-
ANet Captions [14]	open	100k	100k	849h	✓	-
TACOS-ML [22]	cooking	14k	53k	13h	✓	-
HowTo100M [18]	instruction	136M	136M	134,472h	-	✓
Getting Over It [16]	video game	2,274	2,274	1.8h	-	✓
LoL-V2T (Ours)	video game	9.7k	63k	76h	✓	✓

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.

compare key statistics in Table 1. Video description datasets can be divided into two types: one with one caption per clip, and one with multiple captions per clip. Furthermore, there are two types of captions: those generated by manual annotation, and those generated automatically from the narration. In this work, we propose a large dataset for esports where multiple captions correspond to a clip and are automatically generated from the narration. As shown in Table 1, *LoL-V2T* is the largest dataset in the video game domain and is comparable in size to datasets in other domains.

Video Description

Early work in video description [13, 3, 7] was based on template-based methods. The templates require a large number of linguistic rules manually set up, which are only effective in constrained environments. They also have limited applicability, and most research has focused on human actions. With the growth of deep learning, a method using an encoder-decoder framework for video description [27] was proposed, which overcomes the limitations of template-based methods. This method used CNN for encoder and RNN for decoder, demonstrating the high performance of CNN for video feature representation and RNN for sequential learning. Subsequently, mean pooling, which was used to aggregate the features of each frame in the encoder, was replaced by CRF [8], and the CNN used in the encoder was replaced by RNN [8, 26]. Following the success [1] of attention mechanisms in machine translation, several methods have been proposed: temporal attention [33], which focuses on the temporal direction, semantic attention [9], which focuses on tags of semantic concepts extracted from images, and methods using both [20]. Furthermore, Zhou *et al*. [37] applied the transformer architecture [25] to the Video Description Model. Self-Attention in the transformer can replace RNN, which is effective for modeling long-term dependencies in series data. However, transformer architectures are unable to model history information because they can operate only separated fixed-length segments. Lei *et al*. [15] solved this problem by MART which is an architecture with memory module like LSTM [11] and GRU [5] based on the transformer.

Video Description in Video Game

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

accuracy	precision	recall	F_1
0.963	0.928	0.285	0.437

Table 2: Performance of the model to detect whether a clip is related to gameplay.

on manual or ASR-generated subtitles.

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 gameplays 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. The content represented by a caption is often earlier temporally than the timestamps calculated by Section 3.3 because narrators talk about gameplays after watching them. In the next section, we propose a method for dealing with the first difficulty in *LoL-V2T*.

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]

[2] https://pyscenedetect.readthedocs.io/

[3] https://github.com/notAI-tech/deepsegment

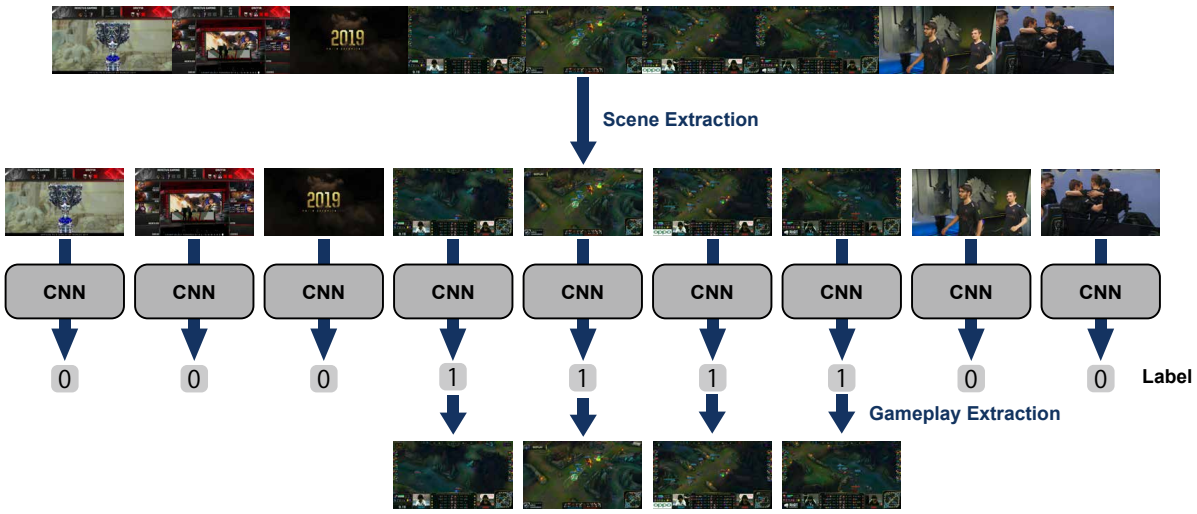


Figure 3: Overview of splitting and selection videos in *League of Legends (LoL)*. We split the *LoL* gameplay videos into clips for each scene because the average length of them is too long to be handled by the video description model. We remove clips not related to gameplay with ResNext-50(32x4d) [29] to minimize noise in training.

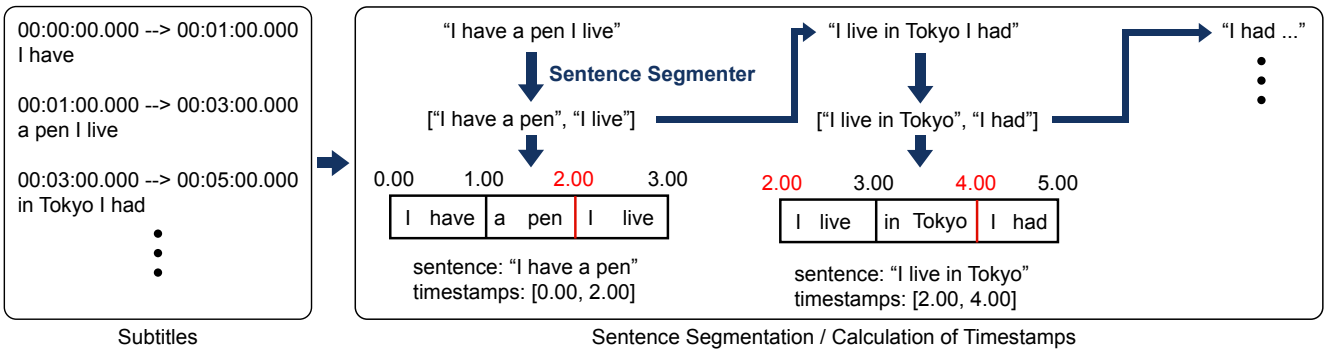


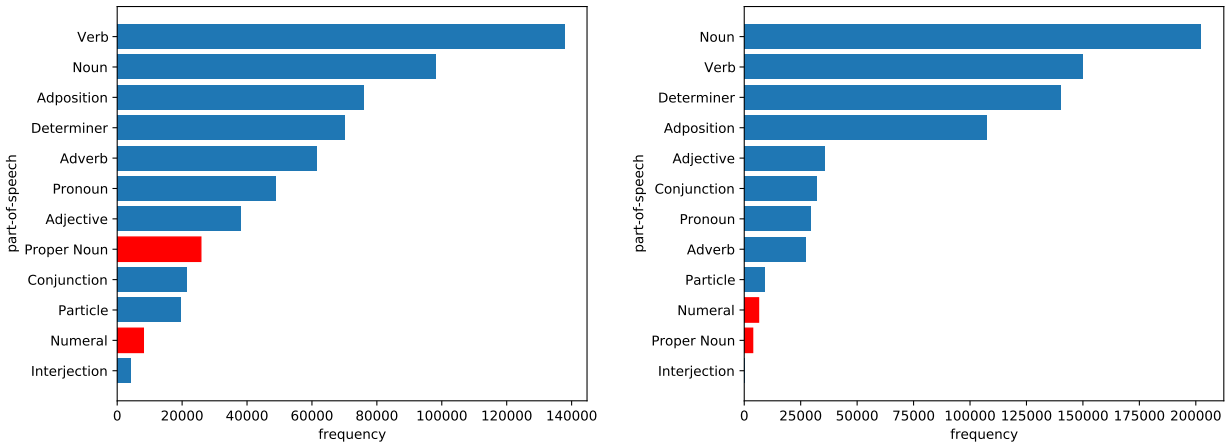
Figure 4: Sequence to create captions and temporal boundaries from subtitles. First, we concatenate several consecutive chunks and split them into complete sentences by a sentence segmenter. Next, the temporal boundaries are calculated according to the number of words in reconstructed sentences. In this 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.



(a) LoL-V2T

(b) ActivityNet Captions

Figure 5: Frequency of words included in captions for each part-of-speech. In *LoL-V2T*, there are more proper nouns and numerals than in ActivityNet Captions.

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

Origin 1
and now it's all about this best of one and for g2
Masked 1
and now it is all about this best of <0> and for <Team>

Origin 2
that's respect from the clutch gaming top laner and so now
they don't have the gangplank ultimate to drop on a Herald
fight but they do have more pressure bottom line
Masked 2
that is respect from the <Team> top laner and so now
they do not have the <Champion> ultimate to drop on a <Monster>
fight but they do have more pressure bottom line

Figure 6: An example of masking. *Origin* is the original caption and *Masked* is the caption after masking. For example, "one" and "g2" in *Origin 1* are converted to "<0>" and "<Team>", respectively. All numerals are converted to "<0>" regardless of their value.

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, out implementation was build upon the mmaction2 [6], we use the two different ResNet-50 [10] model pre-trained on

Group Name	Meaning	Examples
Team	names of teams competing in "Worlds"	fnatic, fanatic, Cloud9, C9, genji, Gen.G, g2
Player	names of players competing in "Worlds"	Baker, Matt, Svenskeren, Ceros, Diamondprox, dyNquedo
Champion	names of characters in *League of Legends*	Recon, Akali, Kaiser, Ryze, Ziya, Camille
Monster	names of monsters in *League of Legends*	Drake, Herald, Raptor, Krug

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.

Split	train	validation	test
clips	6,977	851	1,895
captions	44,042	5,223	13,412

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.

Method	BLEU@4↑	RougeL↑	METEOR↑	Repetition@4↓
VTransformer	1.53	12.76	8.58	37.33
MART	2.13	14.48	11.50	**8.76**
VTransformer+Masking	3.14	**16.57**	12.03	29.20
MART+Masking	**3.56**	15.39	**12.98**	9.74

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.

VTransformer: i think that 's a lot of damage that you can see that
MART: I think that this is a very good sign of a team that has been playing around
VTransformer+Masking: i think that is a very good start to the game for <Team> to be able to get
MART+Masking: i think that is kind of the best <Team> in the <League> and you can see how much of the
Ground-Truth: this is a <Team> that had a very good record against <Team> throughout the regular session

VTransformer: i think that 's a lot of damage that you can see that the <unk> <unk> is going
MART: I mean you can see that the gold lead is still in the mid lane for the side of the
VTransformer+Masking: i think that is a lot of the <unk> that is going to be a big <unk> for
MART+Masking: <Champion> is going to be a very big deal with so many of these fights and the fact that he
Ground-Truth: lane continue to fight for experience so as knows as the <Champion> he is never solar carrying this lane so instead he uses his advantage to help out the other side of the map

VTransformer: the <unk> of the <unk> <unk> <unk>
MART: the fight and then the end of the day and the fight
VTransformer+Masking: the <unk> of the <Team> that is the best <Team>
MART+Masking: he is got the ultimate available and he is got the kill
Ground-Truth: just kill <0> of the squished

VTransformer: <unk> <unk> is gon na be taken down but they 're gon na find the kill on the
MART: I think that this is a really good play for vitality to be able to do it
VTransformer+Masking: <Player> 's going to be able to get the kill but he is going to be able to
MART+Masking: but he is going to be able to get away from this <0> as well as the turret goes down
Ground-Truth: his old <Team> is actually going to die out of the time that was an oopsie other choice will die as well"

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

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 3

[2] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. 6, 7

[3] Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, et al. Video in sentences out. *arXiv preprint arXiv:1204.2742*, 2012. 3

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 7

[5] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014. 3, 6

[6] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 6

[7] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641, 2013. 3

[8] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 3

[9] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639, 2017. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 3, 6

[12] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 7

[13] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, 2002. 3

[14] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 2, 3, 4

[15] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal. MART: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2603–2614, Online, July 2020. Association for Computational Linguistics. 2, 3, 5, 7

[16] Chengxi Li, Sagar Gandhi, and Brent Harrison. End-to-end let's play commentary generation using multi-modal video representations. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*. Association for Computing Machinery, 2019. 3, 4

[17] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 6, 7

[18] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE international conference on computer vision*, pages 2630–2640, 2019. 2, 3

[19] newzoo. Newzoo global esports market report 2020 — light version. https://newzoo.com/insights/trend-reports/newzoo-global-esports-market-report-2020-light-version/, 2020. [Online: accessed 25-Feburuary-2021]. 1, 4

[20] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6504–6512, 2017. 3

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6, 7

[22] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In Xiaoyi Jiang, Joachim Hornegger, and Reinhard Koch, editors, *Pattern Recognition*, pages 184–195, Cham, 2014. Springer International Publishing. 2, 3

[23] Shukan Shah, Matthew Guzdial, and Mark O. Riedl. Automated let's play commentary. *CoRR*, abs/1909.02195, 2019. 3

[24] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 2, 3, 4

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia

Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3, 5

[26] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015. 3

[27] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014. 3

[28] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 6

[29] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 4, 5

[30] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 6, 7

[31] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2, 3, 4

[32] F. Yan, K. Mikolajczyk, and J. Kittler. Generating commentaries for tennis videos. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2658–2663, 2016. 1, 3

[33] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015. 3

[34] H. Yu, S. Cheng, B. Ni, M. Wang, J. Zhang, and X. Yang. Fine-grained video captioning for sports narrative. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6006–6015, 2018. 1, 3

[35] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007. 7

[36] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2, 3

[37] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 2, 3, 4, 5, 7