

A Descriptive Basketball Highlight Dataset for Automatic Commentary Generation

Anonymous Authors

1 SUPPLEMENTARY OVERVIEW

In this document, we offer supplementary material of our BH-Commentary dataset. In specific, we go into additional implementation details in section 2, and we provide more details related to our dataset in section 3. Additional corresponding discussions are presented in section 4.

2 IMPLEMENTATION DETAILS

Our model is implemented based on PyTorch [2]. Each module of the model is jointly trained in an end-to-end manner. The visual encoder is pre-trained on Kinetics-600 and is initialized with corresponding weights, the multi-modal encoder is transformer-based and is initialized randomly, with 12 layers and 768 dimension for hidden state, and the prompt embedding is initialized randomly with the number same to 1% of multi-modal embeddings and the dimension same to hidden state. The vocabulary size are 30,522 and 21,128 for English and Chinese version, respectively. Both the English and Chinese commentaries are truncated to a maximum of 50 words. And for Chinese commentary, instead of using the raw Chinese characters, we use the segmented Chinese words the processed via the open-source tool¹ for words segmentation. Our model is optimized using Adam optimizer [1], with an initial learning rate of 3e-5 and use a learning rate warm-up during the early 10% training steps followed by linear decay. In addition, all hyperparameters are tuned on the validation sets, and this process is consistent for both English and Chinese commentary generation training.

3 DATASET COLLECTION

BH-Commentary dataset is a descriptive sports highlight commentary generation dataset. Due to the professional characteristics of the sports videos, here especially for basketball highlights videos, the commentary label must be accomplished by professional commentators possessing extensive basketball knowledge and experience to ensure the the high quality of the dataset in terms of accurately describing actions and presenting content in a compelling manner.

3.1 Chinese Version Collection

We display the IFlytek interface² for Chinese commentary collection in Figure 1. As the basketball highlights videos we collected already contain professional commentaries based on the video content, we initially used the IFlytek interface to obtain the coarse Chinese commentary label. However, as shown in the case in Figure 1, the recognized text may contain errors that cannot be adopted directly. Therefore, additional manual proofreading work is necessary. In order to ensure the quality of the Chinese commentary label, we implement a rigorous verification process. Native Chinese-speaking workers are required to watch the video content to verify

¹<https://github.com/fxsjy/jieba>

²<https://fanyi.iflyrec.com/video-translate>

that the collected descriptions match. Additionally, each transcribed Chinese commentary label must undergo review and approval by another independent worker. After the initial round of review, all labels are assigned scores, and those scoring below 90 undergo further process. This process ensures the accuracy of the collected Chinese commentary labels.

3.2 English Version Collection

Since the collected commentary labels from professional commentators are in Chinese, the English commentary labels are obtained based on the verified Chinese commentary labels. Given the high accuracy of the verified Chinese commentary labels, we utilize a combination of machine translation and manual proofreading to obtain the English commentary labels. Moreover, to mitigate annotation bias towards a specific translation system, we employ three advanced Chinese→English translation systems (Google, Microsoft, and ChatGPT). After receiving the preliminary English commentary through machine translation, we apply the same verification mechanism as described above to ensure the accuracy of the English commentary labels.

In all, we hire 15 workers for dataset building, and the entire dataset collection process last for about 3 months.



Figure 1: The interface for collecting the commentary.

Dataset	verb per sent	noun per sent	adj per sent	adv per sent
MSR-VTT [3]	1.84	3.20	0.60	0.15
BH-Commentary	1.42	6.31	1.30	0.55

Table 1: Comparison of the average number of verbs, nouns, adjectives and adverbs per sentence of the our dataset and MSR-VTT dataset.

3.3 Dataset Statistics

Our BH-Commentary dataset, dedicated to basketball highlights commentary generation, stands out from traditional video captioning datasets that typically feature shorter video descriptions. Figure

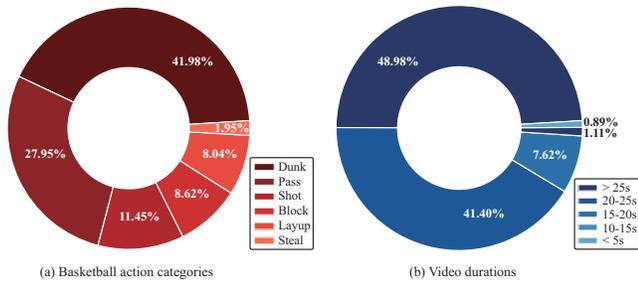


Figure 2: The distribution of basketball action categories and video duration.

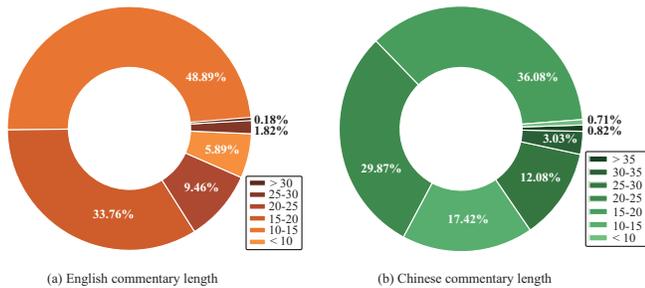


Figure 3: The commentary length distribution of BH-Commentary dataset in English and Chinese version.

2 illustrates the distribution of basketball action categories and the distribution of video duration contained in our dataset. Figure 3 illustrates the distribution of commentary length of our dataset on the two version. For basketball highlights videos, our BH-Commentary dataset presents longer descriptive commentary and contains more sentence components, including object (noun), action (verb), and modifier (adjective and adverb). As shown in Table 1, which illustrates the comparison of the average number of verbs, nouns, adjectives and adverbs per sentence of our dataset and traditional video captioning dataset MSR-VTT [3]. This aligns with our goal of providing vivid commentary on basketball highlights. For basketball highlights videos, our BH-Commentary dataset presents longer descriptive commentary and contains more sentence components, including object (noun), action (verb), and modifier (adjective and adverb). This aligns with our goal of providing vivid commentary on basketball highlights.

3.4 Dataset Samples

Several randomly selected samples of BH-Commentary dataset are shown in the Figure 4. It is evident that our BH-Commentary dataset is descriptive in nature, and the richness of its content poses certain challenges to the highlights commentary generation task, serves as inspiration for further research in the field.

3.5 Additional Qualitative Results

We present additional qualitative results in Figure 5 and Figure 6. For each highlight video, we show our prediction and the corresponding groundtruth commentaries, where the groundtruths are

more descriptive and challenging. Figure 5 illustrates our qualitative results on our dataset in English version. We can observe that our benchmark model performs well for basketball highlight videos. For example, our model is capable of recognizing different actions and generating corresponding terms, such as “dribble past the defense”, “clever pass”, “three-pointer”, “alley-oop”, and so on. In addition, our predictions can correctly convey the player’s actions, although the descriptive component is not as full of as in the groundtruth. This remains a goal we aim to achieve in future work. Figure 6 illustrates our qualitative results on our dataset in Chinese version. Our benchmark model can recognize the actions and terms, and generates semantically reasonable commentaries for the basketball videos. However, similar to its performance on the English version, our benchmark model faces a challenge with the Chinese version: while it accurately identifies actions and events, the descriptive elements are not as comprehensive as those in the groundtruth.

4 DISCUSSION

Video description promoted the accessibility of videos for users. This paper is dedicated to the task of sports highlights commentary generation and introduces a novel dataset as the cornerstone for this task. Although our method surpasses previous state-of-the-art approaches, it’s important to note that the model doesn’t always ensure perfect predictions. It is apparent that for sports highlight commentary generation task, given the richness and complexity of descriptive content, our current model still has notable shortcomings. We present our work as an inspiration, hoping that subsequent research efforts can address and improve upon this challenge.

REFERENCES

- [1] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [2] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Adv Neural Inform. Process. Syst.*
- [3] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE Conf. Comput. Vis. Pattern Recog.* 5288–5296.



English Commentary:

Player A executes a curved breakthrough against the defense and completes a left-handed dunk.

Chinese Commentary:

球员A弧线突破防守，上演左手扣篮。



English Commentary:

Player A avoids the defense in the air and scores with a reverse layup.

Chinese Commentary:

球员A空中避开防守反身上篮打进。



English Commentary:

Player A charges through the defense alone and completes a flying dunk over the opponent.

Chinese Commentary:

球员A单骑闯连营，完成飞身隔扣。



English Commentary:

Player A avoids the defense in the air and scores with a reverse layup.

Chinese Commentary:

球员A空中避开防守反身上篮打进。



English Commentary:

Player A delivers a brilliant bounce pass, and Player B finishes with a powerful two-handed dunk, tearing apart the opponent's defense.

Chinese Commentary:

球员A击地妙传，球员B完成双手暴扣，撕裂对方整条防线。

Figure 4: More samples of BH-Commentary dataset.

**Groundtruth:**

Player A drives to the basket and executes a perfect dunk for the score.

Ours:

Player A dribbles past the defense and executes a one-handed slam dunk.

**Groundtruth:**

Player A delivers a clever assist, and in the final moments, Player B's three-point shot pierces through the defense with pinpoint accuracy.

Ours:

Player A's clever pass sets up player b for a buzzer-beating three-pointer.

**Groundtruth:**

Player A throws the ball into the paint, and Player B seizes the opportunity with a violent dunk off the alley-oop.

Ours:

Player A's precise pass sets up Player B for an alley-oop dunk.

**Groundtruth:**

Player A launches a precise shot, hitting a deep three-pointer like an arrow piercing through clouds, sealing the game from miles away.

Ours:

Player A nails a long-range three-pointer from the arc.

**Groundtruth:**

Player A initiates a backboard pass, seamlessly connecting with Player B who finishes with a thunderous two-handed dunk.

Ours:

Player A's clever pass sets up Player B for a two-handed slam dunk.

Figure 5: Additional qualitative results generated by our benchmark model compared with groundtruth in English version of BH-Commentary dataset.



Groundtruth:

球员A抢断反击，上演单臂战斧劈扣。

Ours:

球员A抢断球员B，上演反击暴扣。



Groundtruth:

球员A后场扔出超远空接，助球员B腾空而起双臂扣篮。

Ours:

球员A送出空中连线，助球员B完成双臂灌筐。



Groundtruth:

球员A四人防守中背身妙传球员B，助后者三分一击中的。

Ours:

球员A突破分球助球员B三分命中。



Groundtruth:

球员A大步流星，直杀篮下，暴力劈扣。

Ours:

球员A单人突破，上演单臂劈扣。



Groundtruth:

球员A行进间眼观六路送出妙传，助球员B篮下拔地而起，力劈华山。

Ours:

球员A仙人指路，助球员B完成暴力劈扣。

Figure 6: Additional qualitative results generated by our benchmark model compared with groundtruth in Chinese version of BH-Commentary dataset.