ActivityNet Speech: A New Benchmark for Audio-Video Tasks

Anonymous

1. Dataset Introduction

In the recent years, with the continuous development of multimedia technology, audio and speech has gradually become a common modality as important as text in various interactions between intelligent agent and human user. Therefore, it's quite essential for researchers to develop the cross-modal architectures between audio and other modalities. However, compared with the urgent need for this kind of applications, there is still a large storage of corpus for training appropriate models or acting as credible benchmarks to evaluate the performances of them.

Considering this, we propose a new dataset, namely ActivityNet Speech, to explore the field of cross-interaction between speech and video. To be specific, we construct this dataset by re-annotate the videos and clips in the ActivityNet Caption dataset. ActivityNet Caption was proposed by Krishna *et al*. [2], which is regarded as an important benchmark in video-text tasks. There are totally 100k sentences connected with 20k videos which was introduced in Caba Heilbron *et al*. [1]. Each video is annotated with 3.65 natural language descriptions on average and marked with aligned temporal boundary timestamps and the average duration of videos is 117.74 seconds. On the basis of it, we transform the linguistic sentence annotation into speeches via a group of real human speakers. Consequently, the connections between texts and videos will be inherited naturally in our dataset.

2. Dataset Statistics

In order to help researchers get the knowledge of the properties in this dataset. We present some useful statistics information in this chapter. Specifically, there are totally 58 speakers participating in the construction and annotation of this dataset, which includes 30 male speakers and 28 female ones. We try to keep the balance between different genders in order to avoid this becoming a confounding factor to interfere the model reasoning process. From the overall perspective of this dataset, the length of audio varies from 1.62s to 36.5s. The most ones in the corpus last between 3 seconds and 10 seconds taking up 89.3% of all, which can be infered in Figure 1. Besides, the split of dataset for training and validation keeps consistent with the ActivityNet Caption dataset. The statistics of each part is shown in Table 1, respectively.

split	train	val_1	val_2
# of items	37421	17505	17031
Avg. length of audio	6.36s	6.38s	5.74s

Table 1. The statistics of the Activity Speech dataset

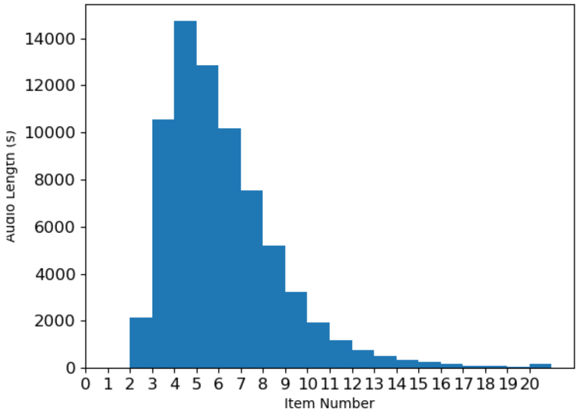

Figure 1. The histogram of audio length in the ActivityNet Speech dataset

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1

[2] 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. 1