Overview of Tencent Multi-modal Ads Video Understanding Challenge

Zhenzhi Wang
State Key Laboratory for Novel Software Technology, Nanjing University, China
zhenzhiwang@outlook.com

Liyu Wu
Peking University Shenzhen Graduate School, China
lywu420@hotmail.com

Zhimin Li
Huazhong University of Science and Technology, Wuhan, China
lizm@hust.edu.cn

Jiangfeng Xiong
Tencent Data Platform, Shenzhen, China
jefxiong@tencent.com

Qinglin Lu
Tencent Data Platform, Shenzhen, China
qinglinlu@tencent.com

ABSTRACT
Multi-modal Ads Video Understanding Challenge is the first grand challenge aiming to comprehensively understand ads videos. Our challenge includes two tasks: video structuring in the temporal dimension and multi-modal video classification. It asks the participants to accurately predict both the scene boundaries and the multi-label categories of each scene based on a fine-grained and ads-related category hierarchy. Therefore, our task has four distinguishing features from previous ones: ads domain, multi-modal information, temporal segmentation, and multi-label classification. It will advance the foundation of ads video understanding and have a significant impact on many ads applications like video recommendation. This paper presents an overview of our challenge, including the background of ads videos, an elaborate description of task and dataset, evaluation protocol, and our proposed baseline. By ablating the key components of our baseline, we would like to reveal the main challenges of this task and provide useful guidance for future research of this area. In this paper, we give an extended version of our challenge overview. The dataset will be publicly available at https://algo.qq.com/.

CCS CONCEPTS
• Computing methodologies → Activity recognition and understanding: Video segmentation.

KEYWORDS
Multi-modal Video Analysis, Temporal Segmentation, Multi-label Classification.

ACM Reference Format:
Zhenzhi Wang, Liyu Wu, Zhimin Li, Jiangfeng Xiong, and Qinglin Lu. 2021. Overview of Tencent Multi-modal Ads Video Understanding Challenge. In Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3479222

1 INTRODUCTION
As a leading marketing platform in China, Tencent Advertising Group gathers massive business data, marketing technologies and professional service capabilities. The goal of advertising is to convey the right information to the target customer with an appropriate price. Due to the richer contents, better delivery and more persuasive effects of the video-form ads than the image/text-form ads, the percentage of video-form ads has an explosive growth at the era of 5G. Therefore, the understanding of ads videos becomes significant and urgent. In addition to a simple video classification task, a comprehensive ads video analysis temporally has an even more significant impact on many chains of the ads system. For example, 1) By understanding the detailed content of ads videos, a single ads video can be automatically edited to have various lengths in order to fit different ads platforms; and 2) The fine-grained categories of each video segment rather than the coarse video-level categories can better serve ads recommendation. By inspecting the statistics of the characteristics in the massive ads videos from the perspective of presentation forms or narratives, we build a comprehensive taxonomy for categories in ads. For example, some characteristics focus on temporal information such as ‘golden 3 second’; while some focus on audio information such as ‘pop music’ and ‘single vocal’; the others focus on semantic information such as ‘ancient china’ and ‘indoor scenes’. These complex semantic representations motivate us to build a class hierarchy with four dimensions: ‘presentation form’, ‘style’, ‘place’ and ‘narrative’: classes in presentation forms mainly describe the ads videos from the perspective of video making, audio or typesetting, while those in the style from the perspective of age, emotions, theme or character relationships. Other categories in place are about locations and those in narratives are about the temporal structure of ads.

Even after the analysis of characteristics and abstraction of ads videos’ paradigm, ads video understanding is still too complex for deep models to learn and predict. To make the goal of ads video understanding more specific, we remove the categories which are too abstract and let our challenge focus on three key perspectives: temporal segmentation, multi-modal information, and multi-label...
Figure 1: The visualization of our Ads Video Structuring task. By defining semantically-consistent segments as ‘scene’, our task requires precisely segmenting videos as scenes and predicting scene-level multi-label categories.

Figure 2: The visualization of video classification with multi-modal information. Ads videos commonly contain three modalities: RGB frames, audio, and text (extracted from ASR/OCR). A simple baseline is to extract features from each modality and perform a classification after modality fusion.

classification. Ideally, we want to know the details of each frame in the videos. Yet, the properties of ads videos, such as fast-pace, a large amount of information yet short length, make the gathering of all fragments’ details impossible to guide the production of future ads videos. Therefore, we define the concept of ‘scene’, which is a ‘super shot’ in the semantic-level. A scene commonly contains one or some consecutive shots and conveys high-level semantics. For example, the first scene in Fig. 1 consists of 4 shots, shifting between the couples in the street and the customer service. However, all the 4 shots are about making a phone call, so we group them into a scene. By detecting the boundary of our defined scene, a useful structure of ads video is obtained. Meanwhile, a comprehensive category hierarchy is built, where different categories are related to different modalities. The most important modalities for our task are video frames, audios, and texts obtained from ASR/OCR.

We propose the multi-modal ads video structuring task as the main track of our ACM Multimedia 2021 grand challenge ‘Multimodal Ads Video Understanding’. It has three main characteristics: 1) accurate temporal localization of scene boundaries; 2) multimodal information is essential; 3) comprehensive category hierarchy from four different perspectives for multi-label scene-level classification. We believe that our task can be a good representative of ads video understanding in the ads system.

2 RELATED WORKS

Due to the fact that the multi-modal video structuring task is motivated by the need of real-life applications, this task has no yet been explored by the computer vision community before. The most related task is the temporal action segmentation task, yet we have three major differences from it: 1) multi-modal information instead of only RGB frames, 2) there are shot changes in ads videos, and 3) multi-label classification with a long-tailed distribution of categories. In the following subsections, we will first introduce recent progress in temporal video segmentation tasks, and some recent video classification methods as well as utilizing multi-modal information in videos. Then we will analyze the feasibility of several choices for our task’s baseline.

2.1 Temporal Video Segmentation

Temporally segmenting videos to get the video parts of interest has been explored by the computer vision community for several years. A typical setting of video segmentation is temporal action segmentation task which aims to parse the whole video into different human actions. Previous datasets commonly contains instructional videos [27, 45, 48] with human performing diverse actions. Temporal action segmentation requires the model to predict a action label for each frame with only visual information. Recently, another setting of video segmentation is proposed to tackle the movie segmentation problem [34]. Different from action segmentation task, movies are usually much longer (e.g., 2 hours) than instructional videos collected in YouTube (e.g., 5 min), yet temporal segmentation of movies is based on the pre-extracted shots and the evaluation of ‘scene’ segmentation is shot-level instead of frame-level.

Enables the exploration of our proposed task. Our challenge also has a simplified version as track 2 of our challenge, i.e., video-level multi-label classification task. Next we will mainly introduce track 1 (video structuring task) due to that track 2 is a subset of track 1.

We will give an elaborate description of our proposed benchmark and the baseline model in the following sections: In Sec. 2, we will review the topics related to our task and analyze the feasibility of the baseline model. In Sec. 3, we will give a detailed description of taxonomy and annotation process of our Tencent-AVS Dataset, and also analyze the proper metrics for our task. In Sec.4 and Sec.5, we propose our used baseline model and ablate its key components.
2.1 Temporal Action Segmentation. They usually model the temporal relations and predict action categories for each frame upon densely sampled features extracted by off-the-shelf classification backbones (e.g., I3D [7]). For example, ED-TCN [28] is a temporal convolution network for action segmentation with an encoder-decoder architecture. TDRN [29] introduced deformable convolutions and a residual connection into [28]. MS-TCN [16] utilized dilated temporal convolution in action segmentation for capturing long-term dependencies and operated it on the full temporal resolution. These works mainly focused on improving receptive field for modeling long-term dependency with encoder-decoder structure [28, 29], dilated convolution [16]. Recently, BCN [54] and ASRF [25] propose to use detected action boundaries to handle over-segmentation errors. Our proposed ads video structuring task contains multi-modal information which cannot be sampled densely, therefore the per frame classification scheme in action segmentation could not apply to ours. Besides, ads videos have a lot of shot change, which may bring more challenges to temporal segmentation methods.

2.1.1 Scene Segmentation in Movies. Scene segmentation (or scene boundary detection) with supervised setting is proposed by Baraldi et al. [4] and Rotman et al. [36] in short videos or documentaries, and scaled up to shot-based long-form movies by Rao et al. [34]. In addition, Chen et al. [9] used unsupervised contrastive learning to obtain more discriminative features for scene boundary detection. This task requires temporal models to predict whether each shot boundary is a scene boundary, therefore it is a binary classification problem. This task also involves multi-modal information, such as place, actor, action and audio. Its evaluation is mAP for binary classification without any pre-defined categories. Therefore our task differs from it due to our multi-label classification with a comprehensive category hierarchy from three perspectives.

2.2 Video Classification

2.2.1 Action Recognition with Single-modality. Video classification is usually investigated in the form of action recognition. Typical methods mainly lie in two categories: 1) two-stream network with RGB frame and optical flow as input to capture appearance and motion information respectively [42, 53]; 2) 3D convolution networks to directly model appearance and motion with 3D convolution filters, such as I3D [7], S3D [55], and X3D [17]. Recently, methods with self-attention or transformers [15, 32] show advantages in modeling long-range dependency and also achieve state-of-the-art performance on standard benchmarks. However, they mainly focus on human action and model temporal relations via only single modality, i.e., visual features.

2.2.2 Multi-modal Information in Videos. Recently, multi-modal data is exploited to enrich video representations due to the multi-modal nature of videos. MMT [18] uses a multi-modal transformer to jointly encode different modalities in videos, which shows superior performance than previous methods with simple concatenation or channel-wise product as multi-modal fusion. Other methods utilize multi-modal data as supervision signals for pre-training [1, 3, 46]. Different from previous action recognition benchmarks, our proposed ads video structuring dataset has rich multi-modal information such as audio and text extracted from both ASR and OCR. How to effectively utilize multi-modal data in ads videos becomes an challenging problem and deserves further investigations.

2.3 Feasibility Analysis

Based on the analysis above, we conclude two possible methods for our ads video structuring task.

2.3.1 End-to-end Methods. End-to-end methods ignore the shot change and directly segment scenes and also classify them with the video features captured by 3D-CNNs [7, 51] or two-stream network [53]. This approach is more practically direct and conceptually simple. Yet it still has some problems: 1) A dense sampling of locations cannot enable the employment of multi-modal information due to the minimum length required by audios or the inaccurate ASR/OCR timestamps, while a sparse sampling may lead to bad performance due to the approximation errors in training and the sparse predictions in inference. The dense sampling of sliding windows will also lead to huge computational costs; 2) The problem of long-tailed distribution becomes even more serious due to the segmented units for classification; 3) Whether the video feature backbone can capture useful informations when crossing the shot change is not sure.

2.3.2 Multi-stage Methods. Due to a lot of shot changes in ads videos, video classification backbones [7, 51, 53] may be confused when crossing the shot boundaries. To ensure a more stable performance, we decompose the task into three parts and optimize them respectively: 1) shot boundary detection, 2) scene segmentation (a.k.a., scene boundary detection) by aggregating shots, and 3) scene-level classification.

Firstly, shot boundary detection methods [40, 43, 44] determine whether each frame is a shot change by comparing the difference between adjacent frames. The major difficulty of shot detection is that the gradual transition effects in ads videos are hard to detect. Secondly, scene segmentation methods [34] commonly use a temporal modeling network such as convolution network or a Bi-LSTM [38] to capture high-level semantics to aggregate shots as complete semantic units (i.e., ‘scene’ in our task). Different from end-to-end methods, shot-level multi-modal information can be utilized in the scene segmentation stage now due to larger length of the minimum units (i.e., a shot rather than a single frame).

Finally, a multi-label classification method can be used to predict the categories of each segmented ‘scene’ with multi-modal features. As for multi-modal modeling, a BERT family of models [13, 37] has been proven as effective to capture semantics of texts; VGGish [22] uses a CNN architecture to model audio waveforms and achieve promising performance; various video feature backbones mentioned above [7, 51, 53] can effectively capture both the appearance and motion cues. The features extracted from each modality will be fused in many ways, such as concatenation or cross-attention [52], to perform the final classification. The major difficulty for scene-level classification is the long-tailed distribution of the categories from four different perspectives. To decouple the influence of different stages and get a more stable performance for...
our proposed new task, we adopt the multi-stage approach as the baseline of our challenge.

3 BENCHMARK

As mentioned in Sec. 1, we propose the ads video structuring and multi-modal video classification tasks, which we believe is a good representative of ads video understanding. In this section, we will introduce our proposed Tencent Ads Video Structuring (Tencent-AVS) benchmark. To make it solid, we collect large-scale ads videos from real business data and annotate the scene boundaries and the categories with careful quality controls. We also examine many previous metrics and choose proper ones to evaluate methods’ performance on our benchmark. We will introduce the details of our dataset in Sec.3.1 and metrics in Sec.3.2.

3.1 Tencent Ads Video Structuring Dataset

Based on our defined ‘scene’ in Sec. 1, we build our large-scale Tencent-AVS Dataset with accurate annotations for scene boundaries and scene-level multi-label categories. We will introduce the details of our dataset and the annotation process as follows.

**Taxonomy.** By exploiting the characteristics of ads videos and collecting the advertiser-uploaded key words, we build a three-layer hierarchy for the categories of ads videos from four perspectives: ‘presentation form’, ‘style’, ‘place’ and ‘narrative’, as shown in Fig. 3.1. Then we consult the experts of advertisement and add the missing classes into our category hierarchy to form 233 classes. Finally, we remove the classes with a very small number of instances (i.e., ≤ 100) after finishing the annotations of our dataset to make the final dataset have 82 high-frequency classes. The number of category with the presentation, style, and place are 25, 34, and 23, respectively. In the presentation category, there are ad videos covering grid layout, dubbing, and padding. In the style category, ad videos contain emotion, subject, and relationship contents. In the place category, there are locations including home, office, and school. Fig. 5 shows the label distribution of presentation, style, and place. We can observe the style contains a large amount of labels and the place contains the least. Besides, the labels among these three main categories suffer form long-tail distributions. An example is showed in the place category where the majority of labels exist within indoor, outdoor, home, and office subcategories.

**Annotation process and quality control.** Dissimilar to the existing action localization or segmentation task whose main challenge of annotation process is the ambiguity of action boundaries, human-edited ads videos have many shot changes as clear demarcations. Therefore, the main challenge of our dataset is the annotations of our multi-label fine-grained categories. Specifically, a scene has about 6 classes on average, so annotators are easy to miss some classes instead of annotating a wrong label for a specific class. Based on these concerns, we write a detailed handbook for annotators, which contains a detailed description and a sample video for each class, and ask them to pay attention to the easily-missed classes. Our handbook also describes when scene changes and in which situation the scene crosses multiple shots. Our annotation process consists of four steps: 1) trial annotation, 2) annotation, 3) sampled inspection and revision, and 4) post-processing. In step 1, we let 7 annotating companies to try the annotation and select top 3 companies to continue step 2 according to their accuracy. After each batch of data is annotated, we will randomly sample about 1/10 videos and check the quality of annotations, especially the missing classes. We will approve the batch if the accuracy of annotation is ≥ 90% and otherwise let the annotators to revise their annotations for the whole batch. Each video’s annotations will be revised for about 3 times on average. Therefore, we believe that the quality of our dataset is carefully controlled. Finally, as scene boundaries annotated by humans are hard to reach the per-frame precision, we use a shot detection method TransNet v2 [43] to refine the little shifting of scene boundaries as a post-processing, i.e., we use the nearest shot boundary to substitute the original scene.
boundary annotation if the distance between them is small (i.e., ≤ 0.1s).

**Splits and statistics.** Ads videos in Tencent-AVS dataset come from our real business data. It contains 12k Ads videos (totally 142.1 hours) and is divided into 5,000 for training set, 2,000 for validation set and 5,000 for test set, as shown in Table 1. In Fig. 4, we demonstrate the statistic of the video duration (left) and label instances per video (right) in Tencent-AVS dataset. Most of video duration in Tencent-AVS are between 25 and 60 seconds, with average length as 45.5 sec. As shown in Table 2, Tencent-AVS is comparable to other video segmentation datasets without class annotations (e.g., MovieScenes and Kinetics-GEBD only has scene boundary annotations) and is much larger than video segmentation datasets with class annotations (e.g., 50Salads and Breakfast). Furthermore, our dataset covers a wider range of real-world scenes instead of a very small domain (e.g., cooking), which means our class hierarchy is more diverse than previous datasets. In Table 3, we compare Tencent-AVS with other multi-label video classification datasets: Firstly, The number of instances in our dataset shows a significant advantage than previous datasets. In addition, our dataset contains different types of videos from many industries such as education, games or E-commerce. Finally, our dataset contains not only human actions but also other elements such as places and objects. However, contents from other datasets are from single domain and focus on human actions.

**Access to our Tencent-AVS dataset.** Our dataset will be publicly available at https://algo.qq.com: Videos from all three splits and annotations of train/val sets can be downloaded by filling an application sheet. We will maintain an online server for evaluating the performance on the test set and the leaderboard will be updated on the website.

### 3.2 Metrics

**Ads Video Structuring.** Due to the properties of the proposed task, e.g., multi-label and temporal segmentation, many metrics adopted by existing tasks cannot directly be used in our task. Here we review some related metrics and introduce the metrics we adopted. The most related action segmentation task [16, 28, 54] adopts per-frame accuracy, an edit-distance based score and F1-score with IoU score where

\[ \text{IoU} = \frac{|A \cap B|}{|A \cup B|} \]

and take an average of mAPs from IoU 0.5 to 0.95 with stride 0.05 as our final mAP evaluation result; for F1@0.5s, we use 0.5s instead of 5% of the length of segment in [39] to avoid the influence from long segments to their neighboring short segments. Formally, the average mAP we use is

\[ mAP[\text{IoU} @ 0.5 : 0.95] = \frac{1}{10} \sum_{Iou=0.5}^{0.95} \frac{1}{N} \sum_{i=1}^{N} \text{AP}^i \text{IoU}, \]

where \( N \) is the total number of categories, and AP is the Average Precision commonly used in detection tasks. Our F1@0.5s matches each predicted boundary to the nearest ground-truth scene boundaries and regard it as True Positive (TP) if their distance is less than 0.5s, otherwise it is a False Positive (FP). The ground-truth scene boundaries will be deleted once it is positively matched (i.e., ≤ 0.5s), and the ground-truths which do not have any positive matches to be False Negatives (FN). Therefore, the precision is \( TP/(TP + FP) \) and recall is \( TP/(TP + FN) \). The final F1@0.5s is computed by

\[ F1@0.5s = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \]

**Video-level Multi-label Classification.** For video-level multi-label classification, we also evaluate the performance by Global Average Precision (GAP) [26]. It calculates the average precision based on top 20 class predictions for each video:

\[ \text{GAP} = \frac{1}{n} \sum_{j=1}^{n} p(j)(r(j) - r(j - 1)) \]

where \( n \) is the number of true positive predictions, \( p(j) \) is the precision and \( r(j) \) is the recall at top \( j \).

### 4 Baseline

As mentioned in Sec.2.3.2, we choose the multi-stage method as our baseline, which contains three stages: shot detection, scene segmentation, and scene-level multi-label classification as shown in Fig. 6. We will introduce the details as follows.

#### 4.1 Shot Detection

We observe that the traditional shot detection method [8] shows the bad performance in ads video domain, e.g., having similar colors, gradual transition effects or static images as padding in border regions. Therefore, we use the learning-based method TransNet v2 [43] and optimize its performance for situations such as gradual transition effects or static paddings by adding more training data and a spatial SE [23] module. All the following models are based on the split shots and the scene boundaries are the subset of shot boundaries, so that performance of scene segmentation is bounded by shot detection. Please refer to the TransNet v2 paper [43] for more implementation details.

#### 4.2 Scene Segmentation

After splitting the videos into a sequence of shots, we adopt LGSS [34] to predict the scene boundaries, which is a binary classification problem based on the shot boundaries. Specifically, due to ads videos’ shorter duration (~30s) than the movies in LGSS, we only use the Bi-LSTM [38] part to model the relations between shots with multi-modal features. The shot-level video and audio features are extracted by pre-trained off-the-shelf feature extractors, i.e., ISD [7] and VGGish [22]. Then the multi-modal features are concatenated
Anno. Type | Scene segmentation |
---|---|
MPII-Cooking [35] | Action segmentation |
YouCook [12] | |
50Salads [45] | |
Breakfast [27] | |
“5 tasks” [2] | |
Ikea-FA [50] | |
YouCook2 [58] | |
Epic-Kitchens [11] | |
COIN [49] | |
OVSD [36] | Scene segmentation |
BBC [4] | |
MovieScenes [34] | |
Tencent-AVS | Scene structuring |

| Dataset | #Labels | #Videos | #Labels per video | #Classes | Domain |
|---|---|---|---|---|---|
| Charades [41] | 67k | 9.8k | 6.8 | 157 | Actions at Home |
| ActivityNet [6] | 39k | 26k | 1.4 | 203 | Diverse Actions |
| MPI-Cooking [35] | 13k | 273 | 46 | 78 | Cooking |
| MultiTHUMOS [56] | 38.7k | 400 | 10.5 | 65 | Sports |
| Tencent-AVS | 168.5k | 12k | 14.1 | 82 | Ads |

4.3 Multi-modal Classification.

4.3.1 Feature Embedding Extraction. We use four different feature extractors to extract video, image, audio, and text embedding, respectively: 1) Video: We use a 1 fps sampling rate to process each video. The frames are then sent into ViT-Large [14] to be a 1024-d video-level feature vector. 2) Image: We take the middle frame of the video as the keyframe, which is processed by ResNet-50 [19] as a 2048-d feature vector. 3) Audio: The audio signal is extracted by FFmpeg and passed through VGGish [22] network to be a 128-d feature vector. 4) Text: Text information is obtained from OCR or ASR. For each sentence, a sequence of word is first tokenized and then fed into BERT [13] (chinese version). We utilize the 768-d ‘CLS’ token from the last layer of BERT as the sentence feature. The ResNet and Bert networks are finetuned in an end-to-end manner during training while other parameters of feature extractors are fixed.

4.3.2 Feature Aggregation. NeXtVLAD [30] is efficient and effective method for temporal feature aggregations in videos. By deciding frame-level importance score for a sequence of frames, this network aims to learn the global representations of this sequence. Here we use NeXtVLAD to learn compactly aggregated video representations for visual and audio features.

Specifically, given $M$ frames sampled in a video, we extract the frame-level features $x \in \mathbb{R}^{M \times N}$ via a pre-trained backbone. The features are aggregated as $K$ clusters by NeXtVLAD. Each feature
Ablation on modalities in scene segmentation. We ablate some important modalities in Table 3, which measures the similarity of $x_i$ to the cluster $k$, which describes the similarity of anchor points of cluster $k$ and $a_0(x_i)$ is an assignment function of $x_i$ to cluster $k$, which measures the similarity of $x_i$ and anchor points of cluster $k$.

In order to save computational cost, NeXtVLAD first expands the feature $x$ as $\tilde{x} \in \mathbb{R}^{M \times N}$ via a FC layer, where $\lambda$ is a width multiplier. Then a reshape operation is applied to transform $\tilde{x} \in \mathbb{R}^{M \times N}$ to $\tilde{x} \in \mathbb{R}^{M \times G \times \frac{N}{G}}$, where $G$ is the size of groups. This operation splits $\tilde{x}_i$ into lower-dimensional features $\{\tilde{x}_i^g | g \in \{1, \ldots, G\}\}$. Then the feature vector can be written as:

$$v_{ijk} = a_g(\tilde{x}_i)a_k(\tilde{x}_i)(\tilde{x}_i^g - c_{kj})$$

where $a_g(\tilde{x}_i)$ measures the similarity of $\tilde{x}_i^g$ to the cluster $k$ and $a_0(\tilde{x}_i)$ can be considered as an attention operation over the groups. $\tilde{x}$ and $\tilde{x}_i$ are the transformation of input feature $x$, depending on the size of groups $G$, and the width multiplier $\lambda$. The video-level descriptor $y$ can be obtained by aggregating frame-level features as:

$$y_{jk} = \sum_{i,g} v_{ijk}$$

After we concatenate the global representations of video and audio output by NeXtVLAD, the representation is amplified by a SE Context Gating operation. Finally, a Mixture-of-Experts [33] classifier with a Sigmoid activation is adopted for video-level multi-label classification.

4.3.3 Loss Function. The long-tailed distribution of our dataset brings class imbalance issues. To mitigate this, we follow recent works [10, 24, 31] to down-weight the loss values for the frequently appeared samples. We utilize asymmetric loss (ASL) [5] (a improved version of focal loss [31]) to reduce the contribution of negative samples when their predicted confidence are low. Meanwhile, we assign the asymmetric penalty for each positive and negative sample. The losses for positive/negative samples are:

$$L_+ = (1 - p)^y \log(p)$$

$$L_- = p_m \gamma \log(1 - p_m)$$

where $p$ is the network output and $\gamma$ is the penalty parameter for the positive and negative samples. When $\gamma = 0$, Eq. 7 degenerates to the binary cross-entropy loss. $p_m = \max(p - m, 0)$ is the threshold for discarding easy negatives, which is a lower bound of the confidence $p$. It is worth noticing that the mis-labels, easy negative samples, and dropout make the network converges efficiently.

5 EXPERIMENTS

To reveal the key challenges of our proposed task, we conduct some important experiments as follows. We hope these experiments could provide useful guidance for the future directions of ads video understanding. All experiments in Sec 5.1 and Sec 5.2 follow a standard machine learning pipeline: we use training set to train our model and adjust the hyper-parameters on the validation set, then report the performance on the test set.

5.1 Ads Video Structuring

Ablation on temporal modeling in scene segmentation. Although Convolution Neural Network has a dominating place in the computer vision community, we observe that LSTM can better perform in the shot-level temporal modeling in Table 4. We guess that the reason is the high-level extracted multi-modal features and shorter length sequences of the shot-based modeling rather than commonly-used raw images, so temporal modeling in our task is more similar to NLP tasks.

Ablation on modalities in scene segmentation. As a multi-modal task, modalities are crucial for recognizing information from different perspectives. We ablate some important modalities in Table 5 and Table 6 to reveal their impact on the final performance: 1)
Table 4: Ablation on temporal modeling in scene segmentation. The first two metrics are precision and recall of scene boundaries in the shot-level.

| Modality          | Precision | Recall | F1@0.5s | mAP  |
|-------------------|-----------|--------|---------|------|
| Image (ResNet-50) | 63.19     | 76.96  | 69.40   | 12.23|
| Audio (VGGish)    | 31.04     | 61.55  | 41.27   | 4.22 |
| Video (I3D)       | 80.59     | 76.25  | 77.83   | 3.88 |
| All modalities     | 82.32     | 74.93  | 78.45   | 20.14|

Table 5: Ablation of modalities in scene segmentation. The results are based on the best scene segmentation result.

| Modality                        | mAP  |
|---------------------------------|------|
| Image (ResNet-50)               | 14.53|
| Audio (VGGish)                  | 13.41|
| Video (Inception-v3 + NeXtVLAD) | 20.00|
| Text (chinese version BERT)     | 2.78 |
| All modalities                   | 20.14|

Table 7: Ablation of shot detection methods. † means our implementation optimized for ads videos.

| Modality          | GAP  |
|-------------------|------|
| Video (ViT + NeXtVLAD) | 76.6 |
| Audio (VGGish)    | 68.5 |
| Text (chinese version BERT) | 58.2 |
| Image (ResNet-50) | 63.4 |
| All modalities     | 77.1 |

Table 8: Ablation of modalities in ads video classification.

For segmentation, the appearance cues are more important, so the video modality achieves similar results with using all modalities, while merely using audio achieves a very poor result. We think the reason is that audio mainly contains BGM and human voice and it lacks of the discriminative ability to differentiate the scenes. 2) For classification, the text modality contributes the least and audio shows better performance than its in segmentation. Our categories are mainly about appearances or musics yet very few categories are about speech, so the alignment about our categories and text is not good enough.

Ablation on shot detection methods. Since both our temporal modeling and scene classification are based on the shot detection result, its performance has a large impact on our final result. The reason of the large gap between traditional method and deep-learning-based method in Table 7 is two-fold: 1) In training, inaccurate shot boundaries will bring noise to the ground-truth of scene boundary in the approximation from the true ground-truth to shot-level binary ground-truth. 2) In inference, models cannot predict a scene boundary if there is no shot boundary.

5.2 Video-level Classification (AP). The per-class AP is shown in Fig. 7, where all categories are shown in x-axis in a descedent order of per-class data size. We can find that although head classes tend to have better performance, many classes such as ‘zoom in’ and ‘painting’ do not follow this rule, indicating the difficulty of our Tencent-AVS dataset. We denote feature aggregation in 4.3.2 as FA, and FA with asymmetric loss described in 4.3.3 and data augmentation in ?? as the proposed method. The blue bar represents the performance of FA (i.e., baseline with GAP=69.1) and cyan bar is the improved version (i.e., baseline++ with GAP=78.8). We achieve a 14.0% relative gain in GAP by our proposed classification method
Ablation on additional components. Baseline++ contains several modules such as ViT backbone, MoE classifier, ASL loss and data augmentation. Table 9 shows how each module improves the baseline model (i.e., FA). Baseline with inception-v3 achieves GAP 69.1. By changing the backbone as ViT, the performance is improved to 73.8. The performance reaches 76.9 when we add ASL (mentioned in 4.3.3). We also use Mixture-of-Experts (MoE) [33] to improve the classifier, which leads to a 77.1 GAP.

Ablation on data augmentation. Data augmentation methods are proved to be effective for preventing overfitting and improving generalizations. We follow VideoMix [57] to conduct two types of data augmentation in videos: spatial mix and temporal mix, as shown in Fig. 8. We create a new training video which is constructed by replacing and stacking from one video to another. Specifically, spatial mix operation defines a binary mask tensor $M$, which signifies the replace location in two video tensors. Given two video samples $x_1$ and $x_2$, the new video can be obtained by

$$\hat{x} = M \odot x_1 + (1 - M) \odot x_2$$

where $\odot$ is the element-wise multiplication. Similarly, temporal mix operation aims to stack two video along the temporal dimension. By further adopting data augmentation methods such as temporal mix (TM) and spatial mix (SM), we achieve a further performance gain on all the additional components and lead to 78.8 GAP.

6 CONCLUSION

In this paper, we propose the multi-modal ads video structuring task and video classification task in our Multi-modal Ads Video Understanding Challenge to understand ads videos in an in-depth manner. We describe our proposed Tencent-AVS benchmark from the perspectives of class taxonomy, data annotation, and metrics. By ablating our proposed baseline for this task, we reveal the key challenges in ads video structuring and would like to provide useful guidance for future research in this area. We will continue to work with the multimedia community to contribute to related topics.

REFERENCES

[1] Hassan Akkari, Liangzhe Yuan, Rui Qian, Wei-Hong Huang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. arXiv preprint arXiv:2104.11178 (2021).

[2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. 2016. Unsupervised Learning from Narrated Instruction Videos. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 4575–4583. https://doi.org/10.1109/CVPR.2016.495

[3] Jean-Baptiste Alayrac, Adria Recasens, Rosalía Schneider, Bela Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-Supervised MultiModal Versatile Networks. NeurIPS 2, 6 (2020), 7.

[4] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. 2015. A deep siamese network for scene detection in broadcast videos. In Proceedings of the 23rd ACM international conference on Multimedia. 1199–1202.

[5] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihui Zelnik-Manor. 2020. Asymmetric loss for multi-label classification. arXiv preprint arXiv:2009.14119 (2020).

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

[7] João Carreira and Andrew Zisserman. 2017.quo vadis, action recognition? A new model and the kinetics dataset. In CVPR. IEEE Computer Society, 4744–4753.

[8] Brandon Castellano. 2018. Pyscenedetect: Intelligent scene cut detection and video splitting tool. https://pyscenedetect.readthedocs.io/en/latest/

[9] Shixing Chen, Xiaohao Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. 2021. Shot Contrastive Self-Supervised Learning for Scene Boundary Detection. CoRR abs/2104.13537 (2021).

[10] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 9268–9277.

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In Proceedings of the European Conference on Computer Vision (ECCV). 720–736.

[12] Pradip Das, Chenliang Xu, Richard F. Doelle, and Johnson J. Corso. 2013. A Thou-
sand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching. In 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 25-28, 2013. IEEE Computer Society, 2634–2641. https://doi.org/10.1109/CVPR.2013.340

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT (1). Association for Computational Linguistics, 4171–4186.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).

[15] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanzhe Li, Zhiyeng Li, Jianendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. arXiv preprint arXiv:2104.11227 (2021).

[16] Yuanzhe Fan and Jürgen Gall. 2019. MS-TCN: Multi-Stage Temporal Convolu-
tional Network for Action Segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. 3575–3584.

[17] Christoph Feichtenhofer. 2020. X3d: Expanding architectures for efficient video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 203–213.

[18] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-
modal transformer for video retrieval. In Computer Vision–ECCV 2020: 16th Euro-
pean Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[20] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understand-
ing. In CVPR. IEEE Computer Society, 961–970.

[21] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemaeneke, Aren Jansen, R Channning Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, and Hamid. 2021. Shot Contrastive Self-Supervised Learning for Scene Boundary Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 131–135.

[22] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F. Gemaeneke, Aren Jansen, R. Channning Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. 2017. CNN archi-
tectures for large-scale audio classification. In ICASSP. IEEE, 131–135.

[23] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In CVPR. IEEE Computer Society, 7132–7141.
