Scene Segmentation Using the MovieScenes Dataset

Introduction

Scenes are critical units in movies, containing a composition of people, places, objects, and actions/activities. Different from individual tasks in a single domain, scene segmentation involves rich semantics and complex temporal interactions among different domains, making it much more challenging. At Eluvio, scene-segmentation is a crucial part of our workflow. It forms a basis for the deep semantic video/movie understanding, which plays a significant role in the intelligent content distribution in the Eluvio Content Fabric. Practically, accurate scene segmentation enables sufficient granularity in search and ad insertion. It could also provide a framework for all clip-level services.

One scene contains a series of consecutive shots. To define the problem, we first carry out shot detection for movies and group shots afterward to form scenes, where the scene boundary detection could be regarded as a binary classification problem on shot boundaries.

In this challenge, your task is to predict the scene segmentation for each movie given features for each shot. To facilitate candidates' progress, we have provided preliminary scene transition predictions. One can also directly optimize the preliminary predictions, instead of training on shot features from scratch, especially if computing resources are limited.

Data

The dataset [link], including 64 .pkl files corresponding to 64 movies, provides all the information required for the coding challenge.

1. Movie-level: movie id ('imdb_id').
2. Shot-level: four features ('place', 'cast', 'action', and 'audio'). These features are preprocessed and encoded as two-dimensional tensors, with the first dimension indicating the number of shots within a movie, and the second dimension specifying feature vectors, in this case, 2048, 512, 512, and 512, respectively. For those who care about how we conduct feature extraction, details of extraction models can be found in this paper (Rao et al.) [link].
3. Scene-level:
    a. Ground truth ('scene_transition_boundary_ground_truth') is a boolean vector labeling scene transition boundaries.
    b. Preliminary scene transition prediction ('scene_transition_boundary_prediction') is a prediction template, meaning your final output should resemble the format and calculate the probability of a shot boundary being a scene boundary, i.e.

contain one score (between 0 and 1) for each shot transition. You're also welcome to utilize the preliminary results for your predictions.

c. The 'shot_end_frame' is the end frame index for each shot, which is provided for evaluation only.

Evaluation

We take two commonly used metrics:

1. Mean Average Precision (mAP) -- AP measures the quality of scene transition predictions in a movie. Specifically, it is the mean of precisions achieved at each threshold, weighted by the increase in recall. Then mAP is the mean of all AP scores.
2. Mean Maximum IoU (mean Miou) -- Miou measures how well the predicted scenes and the ground-truth scenes overlap in a movie. Then we take the mean of all Miou scores as the final score. Note the size of intersection or union is measured by the number of frames. The metric is proposed in this paper (Baraldi et al.)[link], and more detailed explanations can be found there.

Implementation of these metrics is available at our team's github repo [link].

Notes

1. Any analytical insights are appreciated despite it's a predictive optimization problem.
2. The idea in Section 4.4 of the paper (Rao et al.)[link] is one candidate (but not the only) solution.
3. Check out our FAQ page [link] for more information, and please email us if you have further questions.

Reference

Baraldi, Lorenzo, et al. "A Deep Siamese Network for Scene Detection in Broadcast Videos."

arXiv.org, 2015, https://arxiv.org/abs/1510.08893.

Rao, Anyi, et al. "A Local-to-Global Approach to Multi-modal Movie Scene Segmentation."

arXiv.org, 2020, https://arxiv.org/abs/2004.02678.