ACM Multimedia 2016 Grand Challenge
MSR Video to Language Challenge

Challenge Overview

Video has become ubiquitous on the Internet, broadcasting channels, as well as personal devices. This has encouraged the development of advanced techniques to understand video content for a wide variety of applications. Recognition of videos has been a fundamental challenge of computer vision and multimedia for decades. Previous research has predominantly focused on recognizing videos with a predefined yet limited set of individual words. **In this grand challenge, we go one step further and target at translating video content to a complete and natural sentence, which can be regarded as one of the ultimate goals of video understanding.** This challenge will bring together diverse topics in the areas of multimedia, computer vision, natural language processing and machine learning, as well as multiple modalities (textual, visual, and aural modalities) and multiple ways of understanding and analyzing video content.

To further motivate and challenge the academic and industrial research communities, we are releasing "Microsoft Research - Video to Text" (MSR-VTT), a large-scale video benchmark to public for video understanding. The dataset contains 41.2 hours and 200K clip-sentence pairs in total, covering the most comprehensive categories and diverse visual content, and representing the largest dataset in terms of sentence and vocabulary in the research communities. The sentences were annotated by the Amazon Mechanical Turk workers in a quality-control protocol. The dataset can be used to train and evaluate video to language (a.k.a. video captioning) tasks, and other tasks (e.g., video retrieval, event detection, video categorization, etc.) as well in the near future. All the datasets (including training, validation, and testing) can only be used for research purpose.

By participating in this challenge, you can:

- Leverage MSR-VTT benchmark to boost research on an emerging task of video to language;
- Try out your video to language system using real world data;
- See how it compares to the rest of the community's entries;
- Get to be a contender for ACM Multimedia 2016 Grand Challenge;

Task Description

This year we will focus on video to language task. Given an input video clip, the goal is to automatically generate a complete and natural sentence to describe video content (a.k.a. video caption), ideally encapsulating its most informative content and dynamics.

The contestants are asked to develop their video to language systems based on the MSR-VTT dataset provided by the Challenge (as training data) and any other public/private data (as training data) to recognize a wide range of object, scene, event, etc., in the images and/or videos. For the evaluation purpose, a contesting system is asked to produce at least one sentence of each test video. The accuracy will be evaluated against human pre-generated sentence(s) during evaluation stage. We will also provide

subjective evaluation of sentence quality by human judgers. The contestants need to introduce their systems and datasets in the conference.

Dataset

The dataset is based on MSR-VTT and we split the data according to 60%:30%:10% in the training, testing and validation sets, respectively. The table below shows the statistics of MSR-VTT dataset, which has been accepted to CVPR 2016.

Dataset	Context	Sentence source	#Video	#Clip	#Sentence	#Word	Vocabulary	Duration (hr)
MSR-VTT	20 categories	AMT workers	7,180	10,000	200,000	1,856,523	29,316	41.2

* In the MSR-VTT dataset, we provide the category information for each video clip. The video clip contains **audio** information as well.

Reference: Jun Xu, Tao Mei, Ting Yao, and Yong Rui, MSR-VTT: A Large Video Description Dataset for Bridging Video and Language, IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

Submission Format

Each team is allowed to submit the results of at most three runs and selects one run as the primary run of the submission (we do not guarantee to evaluate second and third runs), which will be measured for performance comparison across teams.

Each run must be formatted in a Jason File as follows.

```
{
"version": "VERSION 1.2",
"result":[
        {
        "video_id": "video2218",
        "caption": "a panel of people talk about things"
        },
        ...
        {
        "video_id": "video1835",
        "caption": "a person is playing the piano"
        }
],
"external_data":{
```

"used": "true", # Boolean flag. True indicates used of external data.
"details": "First fully-connected layer from VGG-16 pre-trained on ILSVRC-2012 training set" # String with details of your external data.
}
}

Note: comments in blue are illustrative and help us to provide inline detailed explanations. Please avoid them in your submissions. Participants please strictly follow the submission format.

Evaluation Metric

The evaluation provided here can be used to obtain results on the testing set of MSR-VTT. It computes multiple common metrics, including BLEU@4, METEOR, ROUGE-L, and CIDEr.

In addition, we will carry out the human evaluation of the systems submitted to this challenge on a subset of the testing set. Human judgers will be asked to rank the generated sentences of the primary run from each team and a reference sentence from 1 to 5 (higher is better) with respect to the following criteria.

- Coherence: judge the logic and readability of the sentence.
- Relevance: whether the sentence contains more relevant and important objects/actions/events in the video clip?
- Helpful for blind (additional criteria): how helpful would the sentence be for a blind person to understand what is happening in this video clip?

Participation

The Challenge is a team-based contest. Each team can have one or more members, and an individual cannot be a member of multiple teams.

At the end of the Challenge, all teams will be ranked based on both objective evaluation and human evaluation described above. The top three performing teams will receive award certificates and/or cash prizes (prize amounts TBD). At the same time, all accepted submissions are qualified for the conference's grand challenge award competition.

Timeline

- April 15, 2016: Dataset available for download (training and validation set)
- May 31, 2016: Test set available for download
- June 10, 2016: Results submission
- June 11 - June 25, 2016: Objective evaluation and human evaluation
- June 30, 2016: Evaluation results announce.
- July 6, 2016: Paper submission deadline (please follow the instructions on the main conference website).

Paper Submission

Please follow the guideline of ACM Multimedia 2016 Grand Challenge for the paper submission.

Related Information

Please find more information as follows.

- MSR Multimedia Challenge: http://ms-multimedia-challenge.com/
- MSR Video and Language Project: http://research.microsoft.com/en-us/projects/video-language/

Grand Challenge Contacts

Tao Mei (tmei@microsoft.com)

Ting Yao (tiyao@microsoft.com)

Yong Rui (yongrui@microsoft.com)