Published in DAIR.AI · Follow

elvis · Follow
Jun 19, 2018 · 5 min read

State of the art Multimodal Sentiment Classification in Videos

This paper proposes a novel method for conducting multimodal sentiment classification from user-generated videos. Multimodal methods comprise of combining various modes of information such as audio, video, and text. The framework is mainly based on a long short-term memory (LSTM) model that enables utterances (units of speech bound by breathes or pauses) to capture contextual information.

What is Sentiment Analysis?

A sentiment analysis task involves many NLP sub-tasks and most commonly aims to detect polarity (positive/negative sentiment) in text. Emotion recognition is a derivative task in which the aim is to predict fine-grained emotions (e.g., fear and joy).

Why Multimodal information?

By combining vocal modulations and facial expressions with textual information, it is possible enrich the feature learning process to better understand affective states of opinion holders. In other words, there could be other behavioral cues in vocal and visual modalities that could be leveraged.

Contributions

The proposed framework considers the order, inter-dependencies, and relations that exist among utterances in a video, where others treat them independently. In other words, surrounding context should help to better classify the sentiment conveyed by utterances. In addition, audio, visual, and textual information are combined to tackle both sentiment and emotion recognition tasks.

Example

Consider the following utterance found in a review: “The Green Hornet did something similar”. Without any context, we may perceive this utterance as conveying negative sentiment. What if we included the nearby utterances: “It engages the audience more” and “I just love it”. Would the sentiment change to positive? You be the judge of that! Note that it is highly subjective but we can train a machine to detect these correlations automatically.

Models

Two main types of feature extraction methods are proposed:

F1: Context-Independent Features (a.k.a unimodal features for each modality)

Textual feature extraction is performed using a convolutional neural network (CNN) where the input is the transcription of each utterance, which is represented by the concatenation of corresponding word2vec word vectors. (See paper for more details of CNN)

Audio feature extraction is performed using the openSMILE open-source software, where low-level features, such as voice intensity and pitch, are obtained. (See paper for more details on audio features)

Visual feature extraction is performed using a 3D-CNN, where frame-level features are learned. (See paper for more details of 3D-CNN)

F2: Contextualized Features

An LSTM-based network is adopted to perform context-dependent feature extraction by modeling relations among utterances. Basically, unimodal features are fed as input to a LSTM layer that produces contextualized features as shown in diagram below.

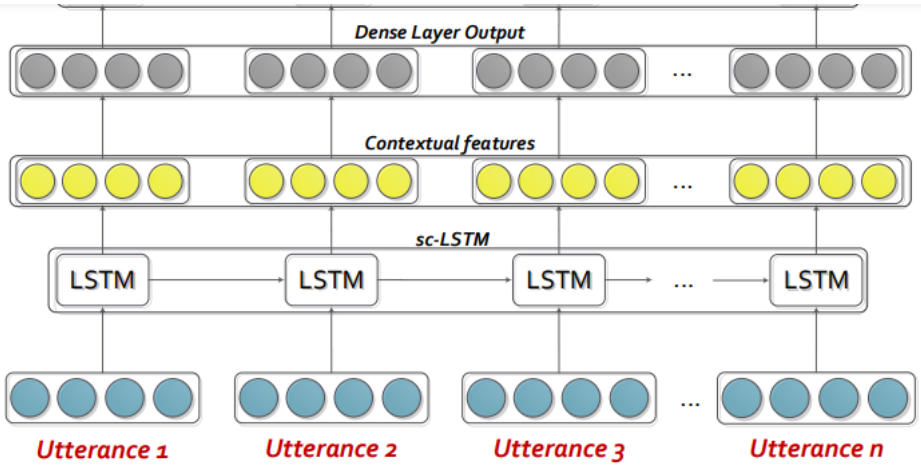

Figure 1: Contextual LSTM network: input features are passed through an unidirectional LSTM layer, followed by a dense and then a softmax layer. The dense layer activations serve as the output features.

Different variants of the LSTM model are experimented with, such as sc-LSTM (unidirectional LSTM cells), h-LSTM (dense layer ignored), bc-LSTM (bidirectional LSTMs), and uni-SVM (*unimodal features are used directly with SVM for classification*).

Fusing Modalities

There are basically two frameworks for fusing modalities:

- ***Non-hierarchical Framework*** — unimodal features are concatenated and fed into the various contextual LSTM networks proposed above (e.g., h-LSTM).
- ***Hierarchical Framework*** — The difference here is that we don't concatenate unimodal features, we feed each unimodal feature into the LSTM network proposed above. Think of this framework as having some hierarchy. In the first level, unimodal features are fed individually to LSTM networks. The output of the first level are then *concatenated* and fed into another LSTM network (i.e., second level). (Check diagram below for overview of hierarchy or see paper for all the details)

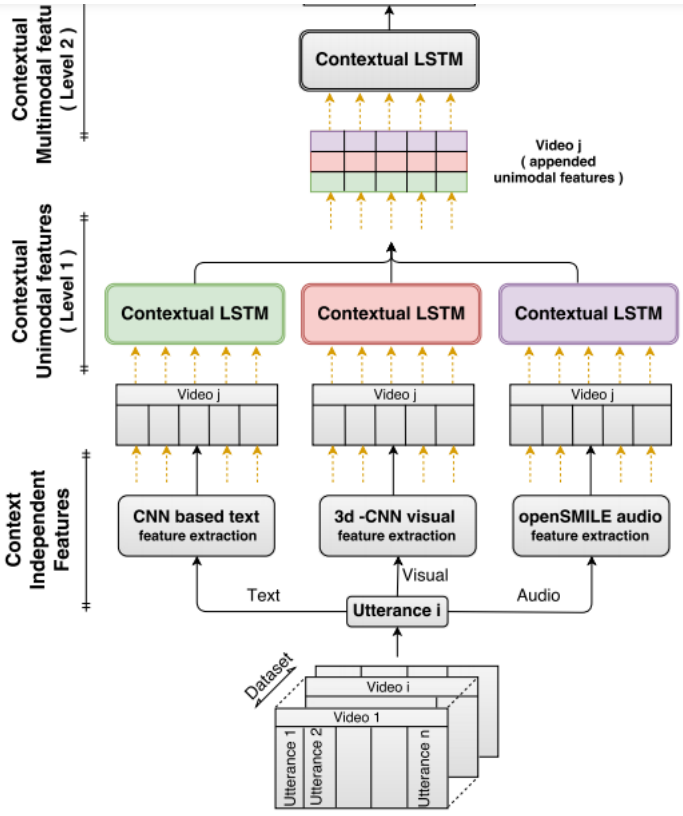

Figure 2: Hierarchical architecture for extracting context-dependent multimodal utterance features (see Figure 1 for the LSTM module).

Datasets

An important consideration in multimodal sentiment analysis is that person-independent datasets must be designed. In other words, train/test splits are disjoint with respect to speakers. The following datasets were used for the experiments:

- MOSI — contains video-based topic reviews annotated by sentiment polarity
- MOUD — contains product review videos annotated by sentiment polarity
- IEMOCAP — contains scripted affect-related utterances annotated by emotion categories

Main Findings

Modality	MOSI					MOUD					IEMOCAP				
	hierarchical (%)				non-hier (%)	hierarchical (%)				non-hier (%)	hierarchical (%)				non-hier (%)
	uni-SVM	h-LSTM	sc-LSTM	bc-LSTM		uni-SVM	h-LSTM	sc-LSTM	bc-LSTM		uni-SVM	h-LSTM	sc-LSTM	bc-LSTM	
T	75.5	77.4	77.6	**78.1**		49.5	50.1	51.3	**52.1**		65.5	68.9	71.4	**73.6**	
V	53.1	55.2	55.6	**55.8**		46.3	48.0	48.2	**48.5**		47.0	52.0	52.6	**53.2**	
A	58.5	59.6	59.9	**60.3**		51.5	56.3	57.5	**59.9**		52.9	54.4	55.2	**57.1**	
T + V	76.7	78.9	79.9	**80.2**	78.5	50.2	50.6	51.3	**52.2**	50.9	68.5	70.3	72.3	**75.4**	73.2
T + A	75.8	78.3	78.8	**79.3**	78.2	53.1	56.9	57.4	**60.4**	55.5	70.1	74.1	75.2	**75.6**	74.5
V + A	58.6	61.5	61.8	**62.1**	60.3	62.8	62.9	64.4	**65.3**	64.2	67.6	67.8	68.2	**68.9**	67.3
T + V + A	77.9	78.1	78.6	**80.3**	78.1	66.1	66.4	67.3	**68.1**	67.0	72.5	73.3	74.2	**76.1**	73.5

- **Hierarchy vs Non-Hierarchy:** From the results in the table above we can observe that hierarchical model significantly outperform the non-hierarchical frameworks (highlighted in green).

- **Modalities:** In general, unimodal classifiers trained on textual information perform best as compared to other individual modalities (results highlighted in blue). The exception was the MOUD dataset, which involved some translation. However, combining the modalities tend to boost the performance, indicating that multimodal methods are feasible and effective.
- **Generalizability:** To test for generalizability, the models were trained on one dataset (MOSI) and tested on another (MOUD). Individually, the visual modality caries the more generalized information. Overall, fusing the modalities improved the model.

(See paper for more qualitative analysis on the importance of contextualized information for multimodal sentiment classification.)

Call for Research

Here are a few ideas you can try to improve the current work:

- Currently, this work aims to evaluate methods on benchmark datasets, which are somewhat clean. You can try to collect your own datasets and label them automatically, rendering large-scale datasets. Also, keep in mind the domain; i.e., you can try to work on a different type of dataset that doesn't include reviews.
- It would be interesting to see more cases where contextualized information helps with sentiment classification.
- Also, a more advanced idea includes the fusion part of the framework. You can try to experiment with more sophisticated fusion techniques, such as those used here.

Software: openSMILE — Software for extracting acoustic features from audio

Dataset: MOSI

Paper: Context-Dependent Sentiment Analysis in User-Generated Videos

Presentation: Video Clip

Have any other questions regarding this paper? Send me a DM @omarsar0.