Applied Deep Learning

Sentiment Recognition Using Multimodal Deep Neural Network

Nagham Dahi[1]
YCH3Y3

[1] Eötvös Loránd University, Budapest, Hungary
ych3y3@inf.elte.hu

1 Introduction:

Human communication includes rich emotional content, so the development of multimodal emotion recognition plays an important role in communication between humans and computers. Because of the speaker's complex emotional characteristics, emotion recognition remains a challenge, particularly in picking up emotional cues via a variety of approaches, such as speech, facial expressions, and language.

Audio and visual cues are especially vital to the human observer in understanding emotions. However, most of the previous work on emotion recognition was based solely on linguistic information, which can overlook various forms of nonverbal information. In this paper, we present a new multimedia approach to emotion recognition that improves the BERT model of emotion recognition by integrating it with heterogeneous features based on language, audio, and visual modalities. Specifically, we improved the BERT model due to the heterogeneous features of the audio-visual methods. We present the self-multi-attention fusion module, the multi-attention fusion module, and the video fusion module, which are attention-dependent multimedia fusion mechanisms using a recently proposed transducer architecture. We explore optimal ways to combine fine-tuning of audio and visual features into co-embedding while combining a pre-trained BERT model with fine-tuning methods. In our experience, we evaluate commonly used CMU-MOSI, CMU-MOSEI and IEMOCAP datasets for multimodal sentiment analysis.

Ablation analysis indicates that the audio and visual components contribute significantly to the recognition results, indicating that these methods contain information highly complementary to sentiment analysis based on video input. Our method shows that we achieve cutting edge performance in the CMU-MOSI, CMU-MOSEI, and IEMOCAP dataset.

Recognizing Uman Procedures (HAR) in videos has gained more attention than it once was due to its additional importance and usefulness in various applications.

From the past decades, visualization-based deep learning methods have created a new era that offers satisfactory results in classification, recognition and detection tasks of static images. Inspired by these successes, various visual techniques were used to recognize the actions in the videos. In spite of remarkable achievements in the recognition of visual actions, the fundamental challenge still exists due to noise noise as well as redundant information in videos on the spatio-temporal dimension. Due to the noisy movement and the appearance of unimportant information, the distinction between convergent layers and the comprehensive distinction within the layer makes this task very difficult. Reducing the distinction between the interlayer feature and increasing the distinction between the layer features can be an effective solution to the issue in question.

Several action recognition techniques based on depth information were introduced along with RGB frames using the RGB-D dataset for 3D motion recognition, attention mechanism, and skeletal method to approach the problem addressed. These published methods require data to be pre-processed, which further complicates latency and time during forecasting. Inspired by working in data rather than pre-processing, we focus on incrementing and then coordinating feature discrimination according to the entropy of the input batch. To recognize the action in videos where the background is dynamic and heterogeneous, it is not possible to train a model using all possible backgrounds and motion. During training, the model learns appearance and temporal patterns with movement.

Due to the dynamic and heterogeneous background, during inference (also training) the network gets confused with the noisy movement of backgrounds and/or other unnecessary spatial cues. Our motivation comes from this observation. We train our model to overlook the background but in an efficient and smart way. Instead of using the attention mechanism and skeleton method, we take the entire background of any given video and send it to the network. Since the attention mechanism focuses on the movement of the object and the skeletal method completely ignores the background, we have come up with an optimal use of the background, in which some actions are determined with the help of the background. To solve the above issues,

We investigate a different approach to an efficient and effective batch entropy-monitored feature learning technique for batch inputs based on the assumption that the amount of feature discrimination is related to the corresponding batch input entropy.

2 Related Work

Understanding multimedia sentiment is a popular research area in recent years. Previous work used an early fusion approach to sequence input features from different modalities and then immediately perform a multimodal fusion. The decision-level fusion method trains different models for each method and then integrates the conclusion of each method to make a final decision. Weighted product base for audio and video resolution level integration. However, decision-level inclusion cannot explore the dynamics of joint modalities by design. Therefore, the following efforts have chosen to integrate multi-modal features as a means of integrating and train them into an integrated architecture to embed inter-media correlations in the learning process.

Harwath et al. A data set of images and audio combined in this way to associate the spoken words with their visual representation. Similarly, Zhou et al. Combining the features of text and voice method, he proposed a semi-supervised multipath generating neural network to better infer emotion.

Duong et al. Clustering is used to classify emotional states by Incorporating features in image and text modalities. while this efforts have focused on the use of binary fusion, and others have explored combine audio, visual and linguistic features together.

Explore multimodal sentiment analysis at the aspect level by proposing a multi-hop memory network to model cross-method and one-way interactions between the three feature areas. Suggest a tensor merger network that expresses multimodal fusion information using the product of visual, audio and visual features. He proposed a recurrent multi-stage fusion network (RMFN) that analyzes the multimodal fusion problem into multiple phases using LSTM to capture synchronous and asynchronous multimodal interactions. To mitigate the added computational cost due to consideration of all three methods, reduce the computational complexity of the parameters by implementing a low-rank multimodal fusion method that uses a low-rank tensor. The LSTM is separately applied to text, visual and audio first, and its extracted features are combined into a multi-level fusion learning architecture. Due to its effectiveness, attention the mechanism has recently attracted some in the field for its ability to combine multimedia features.

Suggested multi-interest recurring network framework learns features using attention for multimodal representation. Intermodal learning interactions were proposed by designing an attention-based multimodal architecture using multimodal adapters. Also, recently, learning transfer technologies that use pre-trained networks to extract features have advanced significantly.

The BERT model, a transformer-based model, showed improved performance by fine-tuning from previously trained weights for a specific downstream task. As shown here, the use of BERT with its wide availability of pre-trained weights, can save time and cost on a variety of tasks.

We present an effective fusion framework for fine-tuning by integrating heterogeneous nonverbal features that complement BERT's linguistic expressions. The following sections describe our proposed structure and training process.

To better understand the method of multimodal fusion and the significance of BERT, we performed ablation studies to understand the impact of our proposed model.

3 Dataset and Features and Methods

The goal: is to implement a multi modality model. The modalities can be the open face features, opensmile features, ...

so first step is to extract these features, for each video we get a csv, we get features there, then we pre-process, and create a model for each one

afterwards, we use a model fusion to concatenate the models into one, and add a shallow network on top of it.

I used 2 modalities openFace to extract visual features and opensmile to extract the audio features.

References

[1] H. IMTIAZ, S. ASHRAF, H. ALAMGIR, and H. DELOWAR : Batch Entropy Supervised Convolutional Neural Networks for Feature Extraction and Harmonizing for Action Recognition, 2020.

[2] L. SANGHYUN, K. DAVID, and K. HANSEOK, H. DELOWAR : Multimodal Emotion Recognition Fusion Analysis Adapting BERT With Heterogeneous Feature Unification, 2020.