Residual Attention-based Fusion for Video Classification

Samira Pouyanfar, Tianyi Wang, Shu-Ching Chen
School of Computing and Information Sciences,
Florida International University, Miami, FL 33199, USA
{spouy001, wtian002, chens}@cs.fiu.edu

Abstract

Video data is inherently multimodal and sequential. Therefore, deep learning models need to aggregate all data modalities while capturing the most relevant spatio-temporal information from a given video. This paper presents a multimodal deep learning framework for video classification using a Residual Attention-based Fusion (RAF) method. Specifically, this framework extracts spatio-temporal features from each modality using residual attention-based bidirectional Long Short-Term Memory and fuses the information using a weighted Support Vector Machine to handle the imbalanced data. Experimental results on a natural disaster video dataset show that our approach improves upon the state-of-the-art by 5% and 8% regarding F1 and MAP metrics, respectively. Most remarkably, our proposed residual attention model reaches a 0.95 F1-score and 0.92 MAP for this dataset.

1. Introduction

Multimodal data analytics has recently attracted significant attention in the deep learning and computer vision community. One of the useful yet challenging tasks in deep learning is video content analysis and understanding [7]. Since video data includes visual, audio, metadata and text description, it can provide a great opportunity in the multimodal deep learning area. One of the main challenges in video processing is how to integrate the information from multiple data modalities to effectively gain insight from the video. To tackle this challenge, many researchers have proposed various data fusion techniques using deep learning [3]. It is also important to automatically learn the significance of each data modality during the fusion step instead of simply concatenating them. Besides, due to the spatio-temporal nature of video, it is imperative to take both static and temporal information into account. To overcome these challenges, this paper presents a new framework using Convolutional Neural Networks (CNNs) and Long short-term memory (LSTM) for multimodal spatio-temporal feature extraction and fusion.

In deep learning research, "Attention" mechanism [9] has been introduced and used in recent years for various sequence-based tasks such as machine translation [2]. It also shows promising results in visual data analytics such as image classification [5]. In this paper, attention is used and followed by temporal layers to not only allow the network to pay attention to the parts of the video sequences that are required, but also diminishing the irrelevant information or noise. This is similar to human perception which concentrates on only a subset of the whole information it receives. We also incorporate the shortcut path or residual mapping [4] to the attention-based recurrent layers to further enhance the performance of the video classification.

This work is an extension of our previous work on multimodal deep learning for natural disaster management [7]. Specifically, in this work, we investigate the importance of residual connection and attention mechanism in LSTM for multimodal data fusion. In particular, we proposed residual attention for multimodal temporal feature extraction and fusion. The experimental results illustrate the significance of the residual attention connection in LSTM. Finally, we utilized a Weighted Support Vector Machine (WSVM) for the imbalanced video classification.

2. Proposed Framework

The proposed framework starts with static multimodal feature extraction followed by temporal feature analysis and fusion modules as explained below.

For static multimodal feature extraction, the state-of-the-art pre-trained models are employed for each data modality. For visual data, the last pooling layer of the Inception-V3 [8] is used to extract the features from video frames using transfer learning. Audio features are extracted using the last convolutional layer of SoundNet [1] which utilizes the natural synchronization between visual and audio data. Finally, text features are automatically obtained using GloVe [6]. After each feature set is generated, they are combined using the proposed spatio-temporal RAF module.

Figure 1 shows various residual attention mechanisms

used in this work for spatio-temporal feature extraction from each data modality. The proposed spatio-temporal feature extraction module generates the input for the fusion module. The fused feature set is constructed by stacking several Residual Attention Bi-directional LSTM (RABL) blocks. Each RABL block takes the output of the previous block as the input and then passes it to the first bidirectional LSTM (BiLSTM) layer. Let $c_t^{(i)}$ be the i^{th} temporal feature vector generated by the BiLSTM at time step t. The attention layer constructs a context vector h_t for $c_t^{(i)}$ at time step t by assigning the attention weights $a_t^{(i)}$. The context vector can be calculated as:

$$h_t = \sum_{i=1}^{M} a_t^{(i)} c_t^{(i)} \tag{1}$$

where M is the total number of features. The hidden state h_t from the first BiLSTM layer is fed into an activation function to generate the relevant score $s_t^{(i)}$:

$$s_t^{(i)} = tanh(W h_t + b) \tag{2}$$

where $s_t^{(i)}$ is the relevant score for feature i in time step t. W and b are the weight and bias parameters that are learned by the model. $tanh()$ is the hyperbolic tangent function (activation function). The attention module then generates the attention weight $a_t^{(i)}$:

$$a_t^{(i)} = \frac{exp(w_t^{(i)} s_t^{(i)})}{\sum_{j=1}^{M} exp(w_t^{(j)} s_t^{(j)})} \tag{3}$$

where $w_t^{(i)}$ is the learned model weight for feature i in time step t. The denominator calculates the sum of the product of the weight and the relevant score of all features in time step t. The residual unit is formed by creating shortcuts between each BiLSTM and attention layer. It helps the network minimize information loss by combining the learned non-linear mapping $F(x)$ with the identity mapping x:

$$Y = F(x) + x \tag{4}$$

where x and Y are the input and output of the residual block. In this work, we investigate different combinations of the attention and residual components in BiLSTM. Figure 1 shows these combinations including a late attention module (applying attention after a series of residual BiLSTM), a fully residual attention module (applying residual attention components after each BiLSTM), and finally a late residual attention component (applying residual attention after the final BiLSTM layer). The outputs of the residual attention modules for each modality are then fed into a fully connected layer to generate the final features, which are concatenated as the joint representation:

$$v_{c,t} = [W_{v,t} c_{v,t}, W_{a,t} c_{a,t}, W_{k,t} c_{k,t}] \tag{5}$$

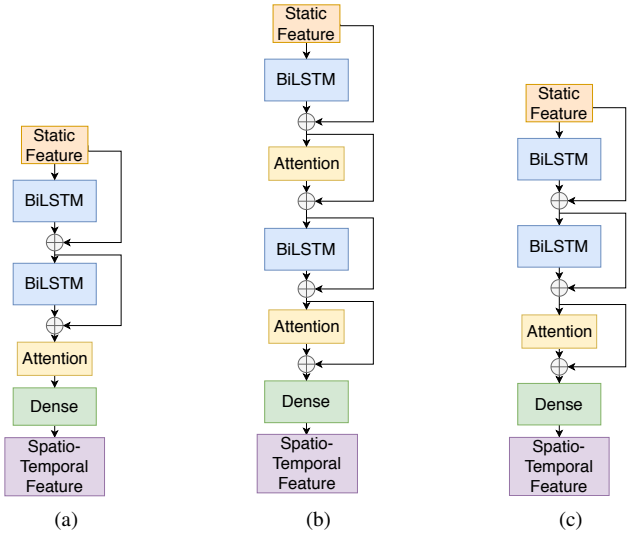

Figure 1: The residual attention modules. (a) late attention; (b) fully residual attention; (c) late residual attention.

where $v_{c,t}$ is the joint representation vector at time step t. $W_{v,t}$, $W_{a,t}$ and $W_{k,t}$ are the learned weight parameters for visual, audio and text features at time step t, respectively.

The weight for each feature in the joint vector is learned automatically using a Weighted SVM that compensates the class imbalance problem by penalizing the misclassification of instances that belong to the minority classes.

3. Experimental Analysis

In this work, a natural disaster video dataset [7] containing 1540 video clips and seven concepts (shown in Figure 2) is used for evaluation purposes. The performance metrics include micro F1 and Mean Average Precision (MAP) which are the proper metrics for imbalanced data classification. Table 1 shows the performance comparison between the baselines and our proposed framework. The first three rows show the performance results of single models in which only one modality is used for video classification. It can be seen from the table that the audio model provides less information than the visual and textual models. On the other hand, the textual model performs better than all of the other single modality models regarding the F1 score and MAP. The next model is the early fusion model that combines static features from all data modalities and then applies several BiLSTM layers which are followed by a dense layer for classification, while the late fusion model concatenates the BiLSTM features before applying the classification layer. The results show the superiority of late fusion compared to the early fusion model. Finally, the last three rows show the performance of the proposed RAF techniques (please refer to Figure 1 for the details of each method). It can be seen

(a) (b) (c) (d) (e) (f) (g)

Figure 2: Dataset samples. (a) flood; (b) damage; (c) emergency response; (d) demo; (e) victim; (f) briefing; (g) human relief.

Table 1: Evaluation results on the disaster test dataset

Model	F1	MAP
Audio model	0.502	0.420
Visual model	0.677	0.602
Textual model	0.779	0.695
Early fusion	0.812	0.735
Late fusion	0.902	0.841
Proposed framework (late attention)	0.933	0.891
Proposed framework (fully residual attention)	0.947	0.910
Proposed framework (late residual attention)	**0.953**	**0.920**

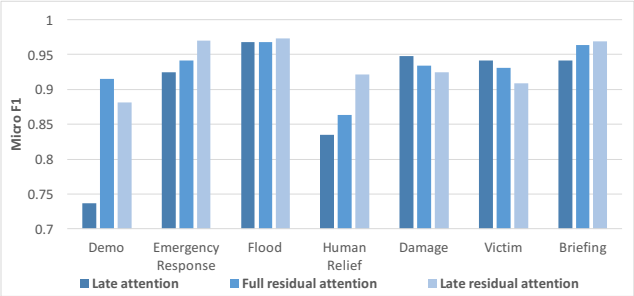

Figure 3: Micro F1 comparison of different proposed residual attention fusions for each disaster class

that the proposed technique improves both F1 and MAP between 0.3-0.5 and 0.5-0.8, respectively, compared to the best previous results (late fusion). In particular, the late residual attention outperforms the late attention and fully residual attention regarding the F1 and MAP scores. The detailed comparison results between the RAF methods are shown in Figure 3. This figure shows the F1 score for each proposed fusion method separated by each disaster class. Although late attention performs better for two concepts (e.g., "damage" and "victim") compared to the residual attention methods, it performs poorly on other concepts (e.g., "demo" and "human relief"). It can be concluded that residual attention connections are helpful in multimodal temporal data analysis.

4. Conclusion

This paper studies the impact of residual attention connections in BiLSTM for multimodal deep learning. For this purpose, a disaster video dataset including audio, image frames, and text is utilized to evaluate the proposed multimodal fusion technique. The experimental results demonstrate the significance of the residual attention connections when concentrating on specific times and modalities for video classification.

Acknowledgment

This project is supported in part by NSF CNS-1461926 and the Dissertation Year Fellowship award from Florida International University's Graduate School.

References

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. SoundNet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900, 2016.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *arXiv preprint arXiv:1409.0473*, 2014.

[3] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *IEEE International Conference on Computer Vision*, pages 2612–2620, 2017.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[5] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, pages 2204–2212, 2014.

[6] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[7] Samira Pouyanfar, Yudong Tao, Haiman Tian, Shu-Ching Chen, and Mei-Ling Shyu. Multimodal deep learning based on multiple correspondence analysis for disaster management. *World Wide Web*, 2018.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.