IM-JAIC at MediaEval 2018 Emotional Impact of Movies Task

Chloe Loughridge[1], Julia Moseyko[2]
[1]Punahou School, Honolulu, Hawaii, United States
[2]Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

ABSTRACT

In this paper, we describe our approach to subtask 2 of the Emotional Impact of Movies task from the Mediaeval 2018 Challenge. We compared the performances of LSTM ensembles to single LSTM models for predicting the fear-inducing seconds in movies. We also compared the performance of an LSTM model trained on audio feature data to the performance of an LSTM model trained on the outputs of a pretrained VGG16 model. Ultimately, we found that a single LSTM trained on VGG16 outputs achieved the highest F1 score on the test set.

1 INTRODUCTION

The Mediaeval emotional impact of movies task contains two subtasks: 1) valence and arousal score prediction, and 2) fear prediction. More information can be found in [1]. In this paper we describe our work on subtask 2, fear prediction.

2 RELATED WORK

This task is a sequel to the 2017 emotional impact of movies task, so there is a sizable body of related work from last year. Support vector regression algorithms were used in [2] to predict fear. A random forest algorithm was used in [7] to predict fear with considerable success. In the task of video action classification, LSTMs have achieved notable results when trained on the outputs of the AlexNet model and the GoogleLeNet model [3]. We aim to build on the previous work done for subtask 2 by implementing LSTM models for fear prediction.

3 APPROACH

3.1 Overview

To address subtask 2, fear prediction, we trained three Long Short-Term Memory (LSTM) ensembles and two single-layer LSTM models. LSTMs are known for their effectiveness at modeling time-series data and capturing long-term dependencies in this type of data [4].

3.2 LSTM Model Architecture

We set aside the last 12 movies from DevSet part 2 as our cross-validation set so we could compare LSTM model architecture variations. A simple single-layer LSTM with batch normalization [5] trained on the fc6 feature data from DevSet 1 achieved the best results on this cross-validation set. The LSTM model with a 1D temporal convolutional layer (trained on the same feature data from DevSet 1) performed slightly worse, but still achieved non-zero F1 scores. We ultimately used both model architectures in our ensembles, though we included more of the single-layer LSTMs in each ensemble.

3.3 Preprocessing the feature data

To train our models, we used the pre-extracted audio features and VGG16 fc6 layer visual features from the Liris-Accede dataset. The audio features were extracted using the openSmile toolbox, and the fc6 features were extracted with the Matlab neural networks toolbox [1]. These features have been the most useful in past papers [2, 7]. To test whether the same applied for our data, we trained multiple LSTM models on the visual features provided in the Liris Accede dataset (i.e., the fc6 feature data and the other visual features extracted using the LIRE library) [1]. Each LSTM was trained on 4 movies and tested on 3 movies from DevSet part 1. In this testing, only the models trained on audio features and fc6 visual features produced nonzero F1 scores.

The fc6 and audio feature data were compiled into matrices and padded so that the max number of timesteps for each movie was 6262 seconds. To reduce memory requirements, we chose a window size of 101 seconds to slide over the time series data with no seconds of overlap. To create labels, we converted the fear annotations into one-hot vectors for each movie. Each element in a movie's one-hot vector represented one second: fear-inducing seconds were ones while non-fear-inducing seconds were zeros. Finally, the training data for our models were handpicked so that about 20% of all timesteps fed into the model were fear-inducing.

4 RESULTS AND ANALYSIS

We submitted the following five runs:

Run 1: Ensemble of LSTMs + fc6 and Audio features
Run 2: Ensemble of LSTMs + Audio features
Run 3: Ensemble of LSTMs + fc6
Run 4: Single-layer LSTM + fc6
Run 5: Single-layer LSTM + Audio

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

Each ensemble consisted of four LSTM models trained on different subsets of DevSet part 1 and DevSet part 2 data. For the first run, two single-layer LSTM models were trained using fc6 feature data, and two single-layer LSTMs were trained using audio feature data. For run 2, all four LSTM models were trained using fc6 data. Three of these models were single-layer LSTMs and one was a single layer LSTM with a 1D convolutional layer attached. For run 3, all four LSTM models (again three of which were single-layer LSTMs and one of which contained a 1D convolutional layer) were trained using audio feature data. The results from these runs are listed in Table 1.

Table 1: Results from the Fear Subtask

Runs	Intersection_ Over_Union
Run 1	0.06496
Run 2	0.07507
Run 3	0.08742
Run 4	0.11992
Run 5	0.09874

The predictions of the individual models in each ensemble were averaged together to produce the ensemble's final output. On the whole, the ensembles performed worse than the single LSTM models. This could be due in part to the fact that predictions in the ensembles were joined via a simple average function, not a weighted average function. Results might have improved if the models with higher F1 scores were given greater influence over the final decision of the ensemble.

Between the two individual models in runs 4 and 5, the single layer LSTM trained on fc6 data (run 4) performed the best. This may suggest that visual features were more relevant to predicting fear-inducing segments than audio features.

5 CONCLUSIONS

In this paper, we described our approach to addressing the fear prediction subtask using LSTM models. We compared the performances of LSTM ensembles to single LSTMs trained on either fc6 or audio feature data.

There are a number of interesting future research avenues to explore and ways to improve what has been shared here. First, it may be beneficial to decrease the length of the sliding windows from 101 seconds and introduce a greater amount of overlap between them. Because of our design choice in this paper, we faced a class imbalance issue in our training data, which was heavily skewed towards non-fear inducing seconds. Decreasing the window size would make it easier to handpick a set of training data with a higher ratio of fear-inducing seconds to non-fear inducing seconds. Training the LSTM models on a balanced dataset may improve their performances.

Another potentially helpful strategy for dealing with the bias in the fear prediction dataset is to weight the cost function so the models are penalized more heavily for predicting zeros (non-fear inducing seconds) when they should be predicting ones (fear inducing seconds).

In terms of feature data, AlexNet and GoogleLeNet outputs could be promising to work with in the future. LSTMs trained on this feature data have achieved notable results for action recognition in videos [3], a task that seems related to fear prediction in movies.

Finally, the Phased LSTM [6] is a relatively recent model architecture innovation that could improve the accuracy scores of LSTMs when it comes to predicting irregular events in long sequences. Given the infrequency of fear-inducing seconds in the training data, adopting a Phased LSTM architecture could be promising.

ACKNOWLEDGMENTS

This work was supported in part by the AI Grant (now the Pioneer Fund).

REFERENCES

[1] E. Dellandrea, Martijn Huigsloot, L. Chen, Y. Baveye and M. Sjoberg, The MediaEval 2018 Emotional Impact of Movies Task, In MediaEval 2018 Workshop, Sophia Antipolis, France, 29-31 October 2018.

[2] Yang Liu, Zhonglei Gu, and Tobey H. Ko. 2017. HKBU at MediaEval 2017 Emotional Impact of Movies Task. *In Proceedings of MediaEval 2017 Workshop.* Dublin, Ireland.

[3] Ng, J., Hausknecht, M., Vijayanarasimhan, S., Rajat Monga, O. and Toderici, G. 2015. Beyond Short Snippets: Deep Networks for Video Classification. arXiv: 1503.08909. Retrieved from https://arxiv.org/abs/1503.08909

[4] Hochreiter, S. and Schmidhuber, J., 1997. LONG SHORT-TERM MEMORY. *Neural Computation*, 9(8), pp.1735-1780.

[5] Sergey Ioffe, Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv: 1502.03167. Retrieved from https://arxiv.org/abs/1502.03167

[6] Daniel Neil, Michael Pfeiffer, Shih-Chii Liu. 2016. Phased LSTM: Accelerating Recurrent Network Training for Long or Event-based Sequences. arXiv: 1610.09513. Retrieved from https://arxiv.org/abs/1610.09513

[7] Zitong Jin, Yuqi Yao, Ye Ma, and Mingxing Xu. 2017. THUHCSI in MediaEval 2017 Emotional Impact of Movies Task. *In Proceedings of MediaEval 2017 Workshop.* Dublin, Ireland.