Abstract

Our experiment adapts several popular deep learning methods as well as some traditional methods on the problem of video emotion recognition. In our experiment, we use the CNN-LSTM architecture for visual information extraction and classification and utilize traditional methods such as for audio feature classification. For multimodal fusion, we use the traditional Support Vector Machine. Our experiment yields a good result on the AFEW 6.0 Dataset.

1. Introduction

The task of emotion recognition is attracting more and more attention from people with the fast developing of artificial intelligent and machine learning techniques. Emotion recognition from video plays an important role in lots of areas such as human-computer interaction and personalized advertising. Give a video with a people, it is important to use both spatial and temporal information provided in the entire video to curve the emotion of character in the video sequence. There might be some overlap between different emotion classes so it is sometimes quite hard to classify the given video into the labeled class. In this experiment, we adapt a deep learning based system to combine the different input modalities in the task of video emotion recognition and extract features from the long-term visual and audio signals in order to classify the emotion labels of a given video.

2. Related Works

It is always important to extract feature both in spatial dimension and temporal dimension in the area of video analysis. Traditional methods in video analysis research use hand-crafted features such as Histogram of Optical Flow (HOF) have the ability to extract local spatial and temporal information, which has been proved to work well.

As datasets are getting larger and larger, deep learning has become more and more powerful in the field of machine learning and made many breakthroughs in lots of computer vision problems [16, 15]. The deep hierarchical representation of the input data could extract information which improves the final result greatly.

Previous EmotiW challenge winners [5, 14, 17] have benefitted a lot from deep learning methods and the state-of-the-art results are achieved with the help of deep hierarchical representations. Most of the previous work used average aggregation after spatial features [14, 13]. In our experiment, we tried to adapt the LSTM model after appearance networks to utilize temporal information of the facial features provided in the video better. We also experiment with different types of fusion method for different input modalities.

To extract long-term information provided by the entire video, we tried to use the convolutional 3D framework to perform convolution over a stack of frames. Meanwhile, in order to aggregating spatial features extracted through Deep Conv-Nets, we tried to use recurrent neural networks. Original recurrent neural networks suffer a lot from problems like gradient vanishing [2]. Nevertheless, the Long Short Term Memory (LSTM) [12] clearly shows it power with the memory cell structure and has been proved to yield the state-of-the-art performances on many tasks such as emotion detection [21], handwriting recognition [9, 11] and speech recognition [10, 8]. So in our experiment, we adapt the LSTM architecture over the appearance features extract from convolutional neural networks to take long-term information into consideration.

3. Technical Approach

3.1. Visual

3.1.1 Convolutional 3D

In order to take motion information within the frames into account during the convolution process, it is a practical method to adapt the convolutional 3D method in the video emotion analysis problem. In 3 dimensional CNNS, convo-
lutional operations are applied on stacked frames input to obtain the features from both spatial and temporal dimensions. Thus, we chose to apply 3D Convolutional Neural Networks for the Facial Emotion Recognition problem. In this work, we choose to use the neural network with 5 convolution-pooling layers followed by 3 fully connected layers.

3.1.2 Bi-directional Long-Short-Term Memory (Bi-LSTM)

The traditional recurrent neural network (RNN) can model the temporal information. Given the input sequence \((x_1, x_2, \ldots, x_t)\), RNN computes the output sequence via the following equation:

\[
h_t = g(W_xh x_t + W_hh h_{t-1} + b_h) \\
z_t = g(W_hz x_t + b_z)
\]

Although RNN has a wide range of applications, such as speech recognition and handwriting recognition, there are still problems such as gradient disappearance or gradient explosion in learning long-term dependence. LSTM \([12]\) as a special RNN, can effectively solve the above-mentioned long-term memory problems.

The key of the LSTM network is that it can remember any length of time. LSTM unit gate mechanism: Input Gate determines whether the input is important to be remembered, Output Gate determines if the cell should output the value, and Forget Gate decides to keep or forget the value. We construct the LSTM unit as shown in Figure 1, which iterates each time as follows:

\[
i_t = \sigma(W_{xi} x_t + W_{hi} h_{t-1} + b_i) \\
f_t = \sigma(W_{xf} x_t + W_{hf} h_{t-1} + b_f) \\
o_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + b_o) \\
g_t = \phi(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \\
c_t = f_t \odot c_{t-1} + i \odot g_t \\
h_t = o_t \odot \phi(c_t)
\]

where \(\sigma\) is the sigmoid function and \(\phi\) is the hyperbolic tangent function.

The basic idea of bi-directional long-short-term memory model (Bi-LSTM) \([19]\) is that for each training sequence forward and backward are two LSTMs, respectively, which enable the complete past and future of each point in the output layer input sequence Context information. Figure 2 shows Bi-LSTM along time. It should be noted that for Bi-LSTM, there is no information flow between the forward layer and the backward layer, i.e. the unwrapping pattern is acyclic.

3.2. Training Techniques

We adapt some practical training techniques in our network training process to avoid the severe overfitting problem that we may face due to the limited training data.

Network Pretraining It is usually a good idea to initialize deep models with weights trained on some bigger datasets to solve the problem of insufficient training data \([1]\). The Deep Convolutional neural network takes grayscale facial images as input, so we choose to use model weights trained on the VGG-Face Dataset \([18]\) to be our initialization. Meanwhile, the Convolutional 3D network uses parameters pre-trained from the Sports 1M Datasets for initialization.

Data augmentation Data augmentation is an efficient and powerful technique which could boost the performance of the deep model. In our experiment, we not only adapt the widely used cropping and flipping augmentation skills, but adapt the corner cropping and scale jittering technique \([20]\) used in action recognition as well. Finally, one single image input could generate 50 augmented images, which would clearly avoid the severe problem of overfitting in small datasets.

3.3. Multimodal Fusion

In many discriminative tasks, the fusion of predictions or representations from models trained using different input
modalities yields a significant improvement. In this task, we decide to fuse the output of the convolutional neural networks (CNNs) and RNNs by weighted averaging their softmax scores. We hope the fusion of different model could perform better performance.

4. Experiments and Results

We tried several approaches to fully use the VGG network and audio features and did many experiments. The implementation details and results are listed below.

4.1. Dataset Overview

We conduct experiments on a facial emotion dataset, namely AFEW6.0 [5], which contains short video clips extracted from Hollywood movies. The AFEW6.0 dataset contains 7 action classes and 774 video clips for training and other 383 video clips for validating. The task is to predict one of seven emotion labels: angry, disgust, fear, happy, sad, surprise and neutral. Video clips present emotions with a high degree of variation, e.g. actor identity, age, pose and lighting conditions. We follow the evaluation scheme of the EmotiW2016 challenge [4] and adopt the training/validating splits for evaluation. Our experiments follow the original evaluation scheme using three training/testing splits and report average accuracy over these splits.

4.2. Audio System

4.2.1 Audio Feature Extraction

The audio features are extracted by the openSMILE toolkit with the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [6]. The openSMILE, an open-source Speech & Music Interpretation by Large Space Extraction, is a fast and flexible audio feature extractor. It can extract different types of low-level descriptors (LLD), and apply various functionals to these descriptors via a text-based feature set. The feature set eGeMAPS consists of Minimalistic Parameter Set and Extended Minimalistic Parameter Set. The former contains 18 LLDs, a total of 62,624 parameters, and the latter contains 7 LLDs, a total of 26 parameters. In total, the eGeMAPS contains 88 parameters. Thus, we get an 88-dimensional audio feature for each video.

4.2.2 Audio Classifier

We choose several traditional machine learning classes for audio classification, where the features are extracted as shown in the previous subsection. For the random forest Classifier, the number of trees in the forest is 500 and the minimum number of samples required to split an internal node is 28. For the support vector machine classification, the penalty parameter $c$ is set to 2.7e-6 and the tolerance for stopping criterion is set to 1e-3. The above classifiers are implemented by scikit-learn.

4.3. Visual System

4.3.1 Convolution 3D

To learn the Convolutional 3D network parameters, we use the stochastic gradient descent algorithm with a batch-size of 10 and a momentum of 0.9. The network parameters are initialized with models from the Sports 1M datasets. The network would be trained for 20,000 iterations with a learning rate of 0.0001. The learning rate would decrease to its 0.1 in iteration 5,000, 10,000 and 15,000.

4.3.2 Deep CNN

For Deep Convolutional Neural Network Architecture, we choose to use the VGG-16 architecture. We expand the training dataset with FER2013 datasets, which has 30,000 human faces with the same class labels to avoid the overfitting problem. To learn the CNN model weights, we choose to use stochastic gradient descent algorithm with a batch-size of 256 and a momentum of 0.9. The network would be trained for 10,000 iterations with a learning rate of 0.001. The learning rate would decrease to its 0.1 in iteration 4,000 and 8,000. The network parameters are initialized with models from VGG-Face datasets with 2.6 million human faces. We choose to use several data augmentation skills to get rid of the overfitting problem, which would be detailed in Section 3.2.

4.3.3 Convolutional-recurrent network

The fine-tuned VGG16-Face model are fed by preprocessed frames. We take the output of the fc6 layer as the input of our temporal model. Although the fc7 layer feature is more commonly used as the VGG filter. However, our experimental result shows that fc6 layer slightly better than the fc7 layer. The Bi-LSTM model used to capture temporal information has 1024 hidden nodes. We have tried a variety of structures and different parameters of the LSTM. We found that the number of nodes in the hidden layer has its saturation. If the number of hidden layer nodes is too large and training set is too small, overfitting occurs. Considering the limits of our datasets, we set the dropout equal to 0.85 lstm and fc layer. In respect of the average length of the video, we set timesteps as 64. The overview of the convolutional-recurrent network is shown in Figure 3.

4.4. multi-modal fusion

In the part of audio-visual models fusion, we have tried several methods, including the average of the probability distributions of each model, the maximum and so on. We
also tried to train random forests, support vector machines, adaboost and so on. Finally, we find that the support vector machine with the penalty parameter of 18.5 works best.

5. Results

Visual Method The result only using video information is listed below(Table 1). We could see that the VGG with Bi-direction LSTM architecture could achieve the best performance. It clearly make sense because the model utilize both spatial and temporal information provided by the entire video. The corresponding confusion matrix is in Figure 5.

| Method                | result |
|-----------------------|--------|
| VGG                   | 44.1%  |
| Convolutional 3D      | 32.1%  |
| VGG+Bi-direction LSTM | 51.9%  |

Table 1. Visual models results on the AFEW 6.0 dataset

Audio Method The result obtained only using audio method is listed here in Table 2. From the result, we could see that the audio classification result is not as good as those obtained by visual features. We infer this phenomenon results from the fact that audio features alone is kind of insufficient to stand for the whole video. We need to fuse the audio and visual classifiers, which is stated in the following section. The corresponding confusion matrix is in Figure 6.

| Method                          | result |
|---------------------------------|--------|
| Random Forest                   | 34.64% |
| Support Vector Classifier       | 39.64% |

Table 2. Audio models results on the AFEW 6.0 dataset

Overall Performance Our final fusion models achieve state-of-the-art performance on AFEW6.0 dataset (Table 3). They outperforming the EmotiW 2016 competition winner [7]. This performance shows it important using both spatial and temporal models to utilize the information provided by the entire video. To illustrate the performance of our model better, we computed the confusion matrices for our final model. The corresponding confusion matrix is in Figure 7.

| Method                                         | fusion method | result |
|-----------------------------------------------|---------------|--------|
| VGG+LSTM+C3D+RandomForest+SVC                 | SVM           | 47.3%  |
| VGG+LSTM+RandomForest+SVC                     | Average       | 50.13% |
| VGG+LSTM+RandomForest+SVC                     | Highest-Confidence | 41.78% |
| VGG+LSTM+RandomForest+SVC                     | SVM           | 55.3%  |
| EmotiW 2016 Winner                            | unknown       | 51.9%  |

Table 3. Multimodal model overall performance on the AFEW 6.0 dataset
6. Conclusion

We presented a multimodal emotion video-classification methods which could utilize the visual and audio information provided in the video. Our method is able to learn the weights of different model prediction so that the fusion of multimodal would clearly make sense. The resulting multimodal architecture achieves state-of-the-art performance on AFEW6.0 benchmarks, the accuracy is higher than the EmotiW2016 winner [7]. Meanwhile, we explore that for the AFEW6.0 dataset, in order to obtain state-of-the-art results, it is better not to use the Convolutional 3D method, the reason is that the dataset is too small for C3D not to overfit. Moreover, using LSTMs on CNN outputs alone yields the highest published performance measure for the AFEW6.0 benchmark. In this models, the video input and the audio input are treated using complete different models. In the future, it would be interesting to design a neural network whose input could be both video and audio information, where front fusion could be employed.
References

[1] D. Annane, J. C. Chevrolet, S. Chevret, and J. C. Raphal. Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems, 1(4):568–576, 2014.

[2] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–66, 1994.

[3] A. Dhall et al. Collecting large, richly annotated facial-expression databases from movies. 2012.

[4] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, pages 427–432. ACM, 2016.

[5] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In Proceedings of the 15th ACM on International conference on multimodal interaction, pages 509–516. ACM, 2013.

[6] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2):190–202, 2016.

[7] Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, pages 445–450. ACM, 2016.

[8] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In International Conference on Machine Learning, pages 1764–1772, 2014.

[9] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. Pattern Analysis & Machine Intelligence IEEE Transactions on, 31(5):855–868, 2009.

[10] A. Graves, A. R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. 38(2003):6645–6649, 2013.

[11] A. Graves and J. Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December, pages 545–552, 2008.

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735–1780, 1997.

[13] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, et al. Emonec: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces, pages 1–13, 2015.

[14] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülcehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. In Proceedings of the 15th ACM on International conference on multimodal interaction, pages 543–550. ACM, 2013.

[15] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188, 2014.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[17] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In Proceedings of the 16th International Conference on Multimodal Interaction, pages 494–501. ACM, 2014.

[18] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In British Machine Vision Conference, 2015.

[19] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11):2673–2681, 1997.

[20] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: towards good practices for deep action recognition. In European Conference on Computer Vision, pages 20–36. Springer, 2016.

[21] M. Willmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. Image & Vision Computing, 31(2):153–163, 2013.