Learning Supervised Scoring Ensemble for Emotion Recognition in the Wild

Ping Hu
Intel Labs China
Beijing, 100190, China
ping1.hu@intel.com

Dongqi Cai
Intel Labs China
Beijing, 100190, China
dongqi.cai@intel.com

Shandong Wang
Intel Labs China
Beijing, 100190, China
shandong.wang@intel.com

Anbang Yao
Intel Labs China
Beijing, 100190, China
anbang.yao@intel.com

Yurong Chen
Intel Labs China
Beijing, 100190, China
yurong.chen@intel.com

ABSTRACT

State-of-the-art approaches for the previous emotion recognition in the wild challenges are usually built on prevailing Convolutional Neural Networks (CNNs). Although there is clear evidence that CNNs with increased depth or width can usually bring improved predication accuracy, existing top approaches provide supervision only at the output feature layer, resulting in the insufficient training of deep CNN models. In this paper, we present a new learning method named Supervised Scoring Ensemble (SSE) for advancing this challenge with deep CNNs. We first extend the idea of recent deep supervision to deal with emotion recognition problem. Benefiting from adding supervision not only to deep layers but also to intermediate layers and shallow layers, the training of deep CNNs can be well eased. Second, we present a new fusion structure in which class-wise scoring activations at diverse complementary feature layers are concatenated and further used as the inputs for second-level supervision, acting as a deep feature ensemble within a single CNN architecture. We show our proposed learning method brings large accuracy gains over diverse backbone networks consistently. On this year's audio-video based emotion recognition task, the average recognition rate of our best submission is 60.34%, forming a new envelop over all existing records.

CCS CONCEPTS

• **Computing methodologies** → **Appearance and texture representations**; *Neural networks*;

KEYWORDS

Emotion Recognition; EmotiW 2017 Challenge; Deep Learning; Convolutional Neural Networks; Supervised Learning

ACM Reference Format:
Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen. 2017. Learning Supervised Scoring Ensemble for Emotion Recognition in the Wild. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3136755.3143009

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3143009

1 INTRODUCTION

The Emotion Recognition in the Wild (EmotiW) challenge has been successfully held for five years [5–9], and audio-video based emotion recognition is always one of the main tasks. The task is to assign a single emotion label to the video clip from the six universal emotions (Anger, Disgust, Fear, Happy, Sad and Surprise) and Neutral. Competitors are required to use their methods to do this automatically, and overall classification accuracy is the comparison metric. The dataset provided by this sub-challenge contains a wide variety of facial expressions in many different situations, such as diverse illumination, occlusion, and viewpoint variations. Since emotion recognition in unconstraint conditions is critical for many applications, for example, home robotics, human-machine interaction and artificial intelligence, EmotiW challenge series has attracted growing interest both in academia and industry.

As Convolutional Neural Networks (CNNs) based methods have made breakthroughs in many computer vision tasks [20, 27, 29], they have become the prevalent entry solutions in recent EmotiW challenges [12, 19, 33]. HoloNet [33] proposes three novel feature blocks to reduce redundant filters and enable multi-scale feature extraction. 2016 winner team [12] utilizes features from a hybrid network which combines Recurrent Neural Network (RNN) [18] and 3D Convolution Networks (C3D) [30]. We note that their use of CNNs is mostly focused on feature layer enhancement, while neglecting a careful consideration of supervised knowledge. In this paper, we explore the problem of how to design an ensemble learning strategy by deeply supervised scoring connection for advancing emotion recognition in the wild. Our method is inspired by two important conclusions. First, connection operation is proved to be effective to enhance the representation capability of features. DenseNet [17] utilizes dense connections between feature layers to ensure maximum information flow between layers in the network. The residual block in ResNet [14] adds stacked layer's output with its input by a shortcut connection and element-wise addition. Inception-ResNet [28] is proposed by combining the connections of Inception and residual blocks. Dual Path Network (DPN) [3] inherits both advantages of residual and densely connected paths, enabling effective feature re-usage and re-exploitation. Although

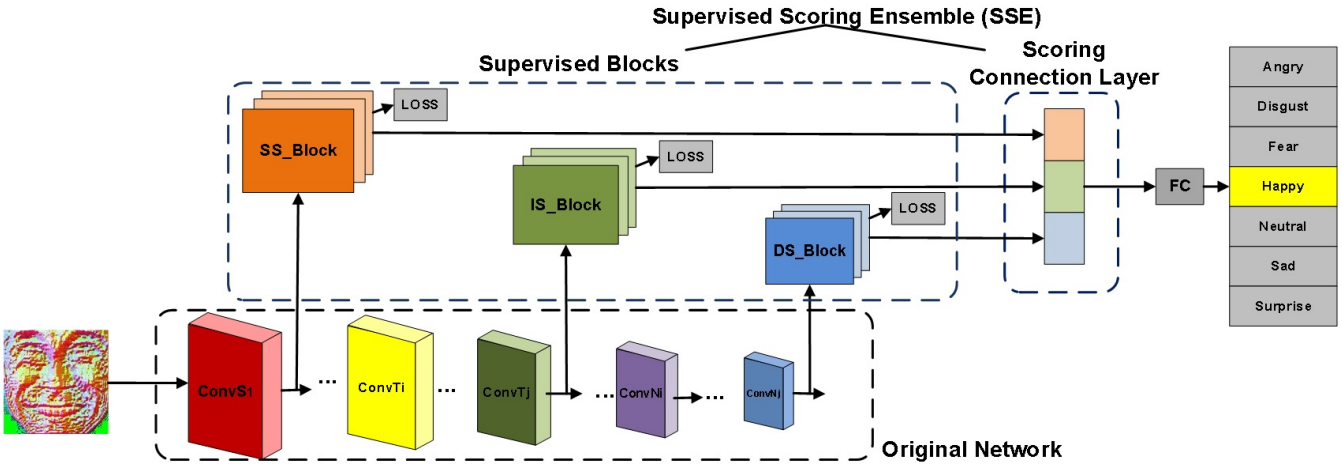

Figure 1: The Flowchart of Our SSE Learning Framework.

the network architectures are designed to be wider and deeper [4, 14, 16]by feature map connections, the training with supervision information is insufficient. Second, supervised learning can benefit training process, especially for shallow and intermediate layers. Deeply-Supervised Nets (DSN) [21] introduces several objective functions at each individual hidden layer, in addition to the overall objective function at the output layer. Deeply supervised learning has been explored in many other fields, like object detection [15], or 3D supervised networks [10]. However, the supervised information utilized from each layer is isolated or implicit. To explore the capability of supervised learning explicitly, we present an end-to-end deeply **S**upervised **S**coring **E**nsemble (SSE) learning strategy which incorporates following innovations: (1) To ease the training of any deep mainstream CNN architecture, supervised mechanism is imported in diverse layers from shallow to deep. The supervision blocks are designed differently according to the layer-wise feature description ability of the original networks prior to their derived layers. (2) To deeply mine the supervised ability in class-wise branch ensemble, we design the scoring connection layer and the accompanying second-level supervision that concatenates various probabilities from the corresponding supervised blocks. (3) The above two parts form the main body of our SSE learning strategy, and can be added to any CNN architectures, such as ResNet, DenseNet and so forth. In this way, class-wise scoring activations at diverse complementary feature layers are effectively combined. The main benefit of providing class-wise ensemble supervisions to various feature layers is that these extra regulation signals can flow back into the trunk of the network and thus alleviate the vanishing-gradient problem and strengthen feature propagation. Furthermore, we can also make an extension of SSE, in which features from different modalities can be naturally combined for jointly optimizing the loss function. Extensive experiments on EmotiW 2017 challenge well testify the efficacy of our method.

In the following sections, we will detail SSE architecture and describe its performance on the EmotiW 2017 audio-video based emotion recognition sub-challenge.

2 THE PROPOSED METHOD

The flowchart of our SSE learning framework is shown in Figure 1, better view in color mode. Our SSE learning strategy can be added to any original networks. It consists of two parts: several supervised blocks and one scoring connection layer. In what follows, we will start with the data preparation first, then describe detailed construction for each part of SSE learning.

2.1 Data Preparation

Our data preparation process is divided into four steps. First, an Adaboost based multi-view face detector [31] is used to locate target faces in the first frames of video clips. Then, Supervised Descent Method (SDM) [32] is applied to track facial features over time. After that, we cut face regions, scale them to the same size and do face frontalization [13]. Then we rescale the frontalized face images to a resolution of 128×128 pixels. At last, illumination effect is removed by a popular Discrete Cosine Transform (DCT) based method [2]. Since the tracked faces temporally disappear or are heavily occluded in some video frames, we only use the frames with clear faces. Further, for each video clip, we evenly sample at most 16 frames with an adaptive frame interval. As a result, the number of frames per video sequence we used for training and prediction varies from 3 to 16.

After above data preparation process, we combine the gray-scale face image with its corresponding basic Local Binary Patterns (LBP) and mean LBP feature maps [24, 25] to form a three-channel input. It should be highlighted that feature maps of other types can also be used as the inputs of our network.

2.2 Supervised Blocks

The effectiveness of deeply supervised learning has been demonstrated in many works [10, 15, 21]. The main point is to supervise earlier hidden layers, besides the output layer. Inspired by this, we tactfully bridge it with the characteristics of emotion representation, and propose three kinds of supervised blocks for shallow, intermediate and deep layers respectively. Figure 2 shows an example and the detail is described as below.

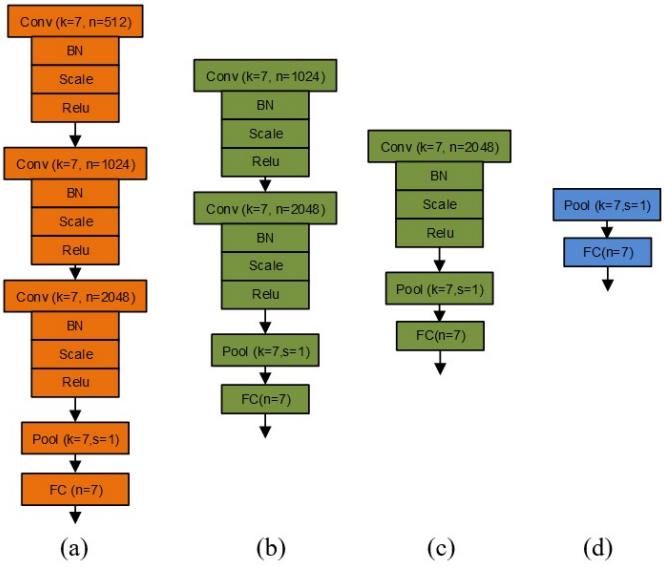

Figure 2: Supervised Blocks. (a) is an example for SS_Block. (b) and (c) are two examples for IS_Blocks. (d) is for DS_Block.

2.2.1 Shallow Layers Supervision.

We find that distinctions of emotions in shallow layers are slight. But these small differences are very important, especially for similar emotions. Therefore, we design supervised blocks for shallow layers to capture these minor changes. Considering that feature maps from shallow layers are not sufficiently discriminative, we design relevant blocks that introduce stacked dimensional reduction and non-linear mapping before classification.

As shown in Figure 1, there is only one supervised block for shallow layers, which is in orange color, denoted as SS_Block. Loss is calculated between the supervised information and the predicted values, then cost is passed back to SS_Block. The shallow layer ConvS1 just stands for a certain one, not necessarily the first layer of the original network. One example of the detailed design of SS_Block is presented in Figure 2 (a). There are three convolutional and relative layers (where the kernel size of filters is 7×7, and the length of feature maps is shown as n, from 512 to 2048) to deeply supervise the feature of shallow layers. The last layer is a Fully Connected (FC) layer, which produces the probability result with objective function.

2.2.2 Intermediate Layers Supervision.

The middle layers of networks are a connection to link shallow layers and deep layers. They present the hierarchical abstraction of input images. So the intermediate supervised blocks should be appropriately designed according to the feature description ability of their input layers in original networks. Choosing appropriate ensemble of intermediate supervision blocks can enhance the representation power of middle and bottom layers for emotion expression.

As shown in Figure 1, supervised blocks for intermediate layers are denoted as IS_Block in green color. ConvTi and ConvTj stand for different intermediate layers. There can be more than one IS_Block, which are the same or different based on their input layers from the original networks.Ensemble of the last prediction layers from them is described in section 2.3. In Figure 2, we design two different supervised blocks in (b) and (c). There are two convolutional and relative layers in Figure 2 (b), and only one convolutional and relative layers in Figure 2 (c).

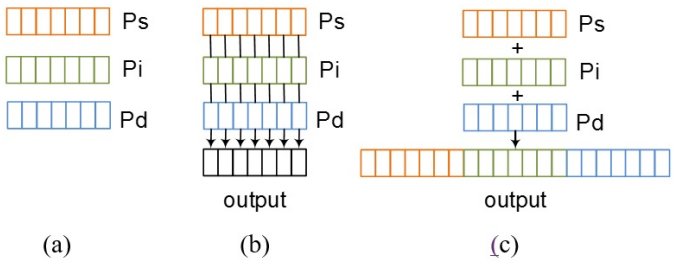

Figure 3: Scoring Connection Layer Design.

2.2.3 Deep Layers Supervision.

Deep layers represent high level semantic information, thus they are more concentrated on global feature description and more abstracted than shallow and intermediate layers. As different feature maps in deep layers have responded to different kinds of recognition classes or regions [1] already, the layers in deep supervised blocks should be less. The ensemble of them can flourish the discrimination of various emotion classes.

As shown in Figure 1, supervised blocks for deep layers are indicated as DS_Block in blue color. ConvNi and ConvNj stand for different deep convolutional layers. There may be more than one DS_Block, and they can be the same or different based on their input layers from the original networks. In Figure 2 (d), a deep supervised block DS_Block is built with a pooling layer and an FC layer.

The layers of state-of-the-art networks are often numerous, such as Resnet-101, Densenet-121 and so on. These deep networks are usually divided into several stages. Therefore, supervised blocks can be stretched out from shallow, intermediate and deep stages simultaneously. In this way, subtle supervised information, middle level supervised features and semantic supervised predictions can accumulatively join together, and further increase original network's discrimination performance. We can see that, the last layer of supervised block is the probabilistic values predicting the similarities to ground truth. This layer can be an FC layer or a convolutional layer. They are concatenated together for the scoring connection layer construction in section 2.3.

2.3 Scoring Connection Layer Design

In existing mainstream methods[3, 17, 28], they design various strategies to connect feature layers. Nevertheless, supervision information in these networks is used implicitly or insufficiently. In this part, we create the scoring connection layer for the first time, which assembles probabilities from diverse supervised blocks (defined in section 2.2) explicitly. In the scoring connection layer, class-wise scoring activations at diverse complementary feature layers

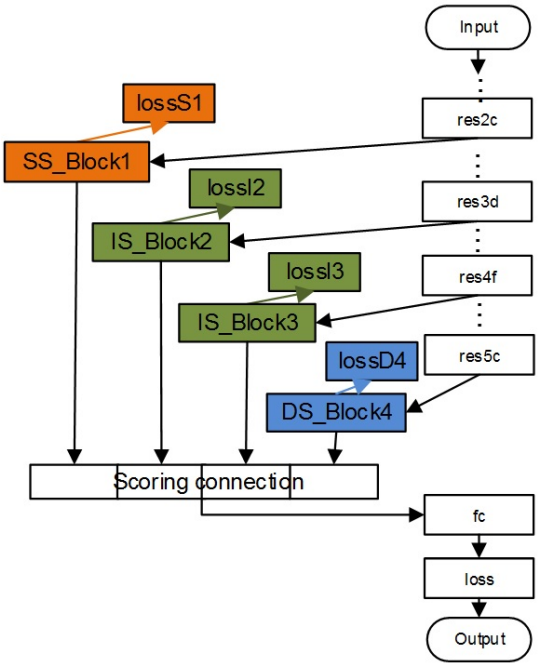

Figure 4: SSE-ResNet-50 Design.

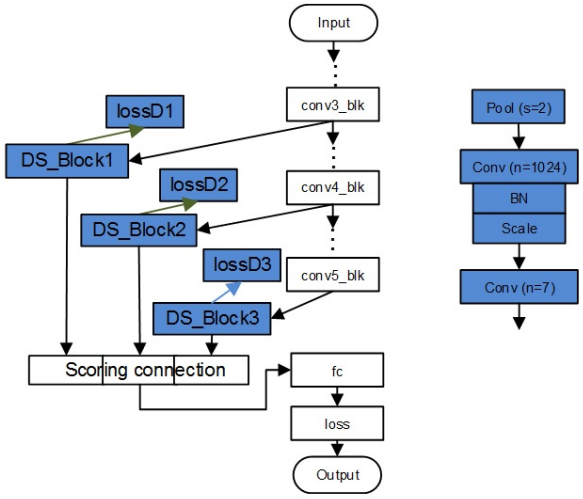

Figure 5: SSE-DenseNet-121 Design.

are aggregated and further used as the inputs for the final-level supervision, as shown in Figure 1.

Since there are N kinds of emotions, we use an N dimensional vector P to represent the probabilistic values of the layer from each supervised block. Taking the supervised blocks of Figure 1 as examples, we represent the vector P from SS_Block as Ps, the vector P from IS_Block as Pi, and the vector from DS_Block as Pd, as shown in Figure 3 (a). Two kinds of connection strategies are proposed as below.

2.3.1 *Element-wise Scoring Connection.*

One kind of ensemble is connecting probabilistic vectors with element-wise operation. This operation can be summation, product, or maximum, each of which is executed on corresponding element. Summation can be added with weights. This kind of connections changes the input probabilities with an operation, and outputs the element-changed result. The input of this operation (Ps, Pi and Pd) should have the equal length and it exports the output with the same length. Figure 3 (b) is an illustration.

2.3.2 *Concatenative Scoring Connection.*

This operation concatenates multiple input vectors to one single output vector, just as shown in Figure 3 (c). Although this connection is simple, it retains all the probabilistic values information for final prediction and doesn't require input Ps, Pi and Pd to have the same length.

3 OUR NETWORKS

As described above, the construction method of SSE learning has been clearly clarified. Next, we will detail how to integrate it into existing mainstream networks. Here, we choose ResNet, DenseNet and HoloNet as the test case examples.

3.1 SSE-ResNet

Taking into account the size of emotion dataset and the training time cost, we choose Resnet-50 among diverse variants of ResNets as our baseline model to rebuild. The building block of Resnet-50 is compact, therefore we need to design relatively complex supervised blocks to improve its performance.

Figure 4 shows where and what kind of supervised blocks are derived for ResNet-50. There are one shallow supervised block SS_Block1 from convolution layer res2c, two intermediate supervised blocks IS_Block2 and IS_Block3 respectively from convolution layer res3d and res4f, and one deep supervised block DS_Block4 from convolution layer res5c. The detailed design of them is the same as the examples in Figure 2. The last layer of each supervised block is an FC layer, and it is the probability of the supervised classification results. Connecting all the probabilities of FC layers forms the scoring connection layer for ResNet-50. It forwards supervised information (derived from shallow to deep layers) to the final output, and backwards the final supervised differences to shallow, intermediate and deep layers respectively.

3.2 SSE-DenseNet

DenseNet utilizes dense connections between feature layers to enhance the expression of features. Here, a comparatively small DenseNet, DenseNet-121 is selected as our test example. Since the building block of DenseNet strengthens feature propagation, the supervised blocks can be designed short and pithy.

As shown in Figure 5 (a), three supervised blocks are branched out from deep layers. They are DS_Block1 from convolutional layer conv3_blk, DS_Block2 from convolutional layer conv4_blk and DS_Block3 from convolutional layer conv5_blk. As both conv3_blk and conv4_blk can generate good semantic features, we design DS_Block1, DS_Block2 the same as DS_Block3, the detailed structure is shown in Figure 5 (b). After pooling and one convolutional layers with 1024 feature maps, we use another convolutional layer as the last probabilistic layer to achieve the same classification

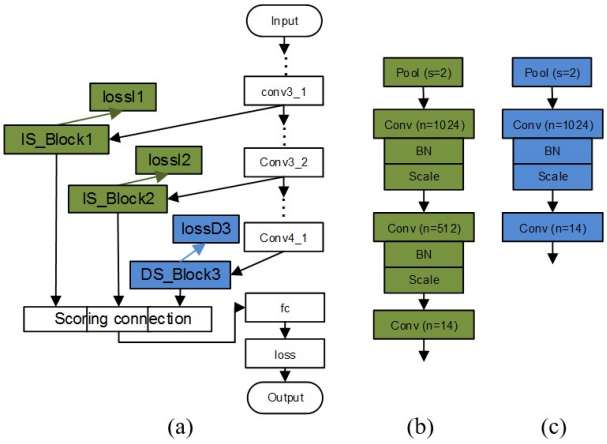

Figure 6: SSE-HoloNet Design.

effect [22] with reduced parameters. The outputs of DS_Block1, DS_Block2 and DS_Block3 are aggregated to form scoring connection layer.

3.3 SSE-HoloNet

HoloNet [33] is an efficient architecture in EmotW challenge 2016. It has only about 20 layers, but fully mines the properties of emotion and scale transformation. Unlike the above two networks that need pre-training, HoloNet can be directly trained on small data set, and has nice results. Thus, we apply various strategies of supervised learning to HoloNet, to further improve its performance.

Figure 6 shows an example. IS_Block1, IS_Block2 and DS_Block3 are derived from convolutional layers conv3_1, conv3_2 and conv4_1 respectively, as shown in Figure 6 (a). The supervised block of IS_Block1 and IS_Block2 is shown in Figure 6 (b). It has three convolutional layers. Compared to intermediate supervised blocks, we use relatively less layers for deep supervised block DS_block3 as shown in Figure 6 (c), because the features from conv4_1 are more dense and discriminative. The probabilistic layers of IS_Block1, IS_Block2 and DS_Block3 are connected to form the scoring connection layer. This layer concentrates the supervised information from intermediate to deep layers, and gets the final outputs with supervised ensemble learning.

4 MODEL FUSION

There is clear evidence that hand-crafted features have shown to be useful for improving the accuracy of CNN models when handling video based emotion recognition task [34]. Thus, we use the baseline of [34] as our hand-crafted model. Besides above visual models, similar to all previous participants, we also train an audio model to describe acoustic context cues.

We denote outputs from different models as $\phi_1, \phi_2, \cdots, \phi_N$, if there are N kinds of models. The final result from our fused models can be formulated as:

$$\phi = \sum_{i=1}^{N} \alpha_i \phi_i \tag{1}$$

$$\sum_{i=1}^{N} \alpha_i = 1, 0 \leq \alpha_i \leq 1$$

Here, α_i is the weight for model ϕ_i. We select α_i according to the performance on the validation set by greedy search.

5 EXPERIMENTS

In this section, we testify the efficacy of our proposed method through comprehensive experiments.

5.1 Parameter Settings

Dataset. We evaluate our SSE learning strategy on the audio-video based emotion recognition sub-challenge of EmotiW 2017. In this sub-challenge, the performers are asked to use their methods to automatically identify unique emotion labels for the respective videos from seven basic emotion categories including Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise. The video clips are collected from either real movie records or TV records. Similarly as previous years, the whole dataset is split into three sets: training set (773 video clips), validation set (383 video clips) and testing set (653 video clips). It should be emphasized that both training and validation sets are mainly composed of real movie records, however 114 out of 653 video clips in the testing set are real TV clips. This increases the difficulty of the challenge to some extent. In all our experiments described below, we strictly follow the evaluation protocols defined by the organizers of this challenge, and we do not use external face or emotion data for augmentation.

Implementation Details. Our visual models are mainly based on SSE-ResNet, SSE-DenseNet and SSE-HoloNet. We train several variant models for each of them, in which different supervised ensemble strategies are adopted for comparison and analysis. The popular Caffe tool [26] is used to train these models. The face images we used are processed with the steps descried in section 2.1. Each face image has a resolution of 128×128 pixels. Flip operation is used for data augmentation. For SSE-ResNet-50 and SSE-DenseNet-121, we use the public pre-trained models [14, 17] to initialize respective trainings with our proposed SSE. Besides ImageNet classification dataset, the pre-trained models are not further trained on external face or emotion dataset but just fine-tuned with given emotion dataset only. For SSE-HoloNet, we initialize the model as in [33]. All training and testing are performed on NVIDIA Titan X 12G GPUs. We use a weight decay of 0.0002 and a momentum of 0.9 for all three kinds of networks. For SSE-ResNet-50, SSE-DenseNet-121 and SSE-HoloNet, the respective batch size is set to 5, 32 and 128. The learning rate begins from 0.01 and is decreased by ten times per 5000/20000/20000 iterations, respectively. Totally, the models are trained for up to 140000/70000/70000 iterations, accordingly. For each testing video clip, the emotion score of a visual feature based model is obtained in two steps. First, it sequentially operates on every sampled video frame. Then the summation of the predicted emotion scores over all frames is used as the final emotion score of this model for this video. Since hand-crafted features are shown to be complementary to CNNs models, we also use the baseline proposed in [34] as our hand-crafted model. In addition to above visual models, similar to many previous teams, we also adopt an audio model to capture acoustic characteristics. We use openSMILE tool [11] to extract 1582-D acoustic feature over each

video clip, and an SVM model with a radial basis function kernel is trained for classification. Here, we set c=2.4428 and g=0.0025.

5.2 Does SSE Work Well?

We first use validation set to test the effectiveness of our SSE learning strategy. In order to verify the validity of supervised blocks, we first train the original networks of ResNet-50, DenseNet-121 and HoloNet without SSE learning strategy. ResNet-50 and DenseNet-121 are fine-tuned with their public pre-trained models [14, 17] on the emotion dataset. Then two groups of supervised blocks are designed respectively for ResNet-50, DenseNet-121 and HoloNet. One of them has three relatively simple supervised blocks, denoted as 3conv-s. And the other include four complex supervised blocks, as described in section 2, denoted as 4conv-c. It should be noticed that, although the shorthand is the same, the specific details are different for the three networks according to their original structures. To compare the results of two scoring connection ways, we fix the other parameters and only change connection methods. Table 1 provides a brief result summary of our models trained with different SSE settings. Here, we abbreviate Element-wise Scoring Connection as Eltwise and write Concatenative Scoring Connection as Concat. All the models are used to calculate the recognition accuracy on the validation set. From the results, we can see that, all the recognition rates with SSE learning strategies are higher than the corresponding ones from original model. This is a strong proof of our method. For DenseNet-121 and Holonet, the 3conv-s SSE strategy can get better results than the 4conv-c, in both scoring connection ways. This gives us a revelation that, more supervised blocks are not necessary for some networks. If the network itself has a rich feature expression, supervised blocks with less layers may be a better choice. For the results between two connection ways, we find that, sequential connection of them can bring higher results than element wise operation, which shows that concatenation of the supervised probabilities can better improve model's performance. According to above result analysis, we add 4conv-c to ResNet when using SEE for training, again the result is better than plain counterpart.

5.3 Results on EmotiW 2017

Now, we describe our results on the audio-video based emotion recognition sub-challenge of EmotiW 2017. To achieve competitive results on this sub-challenge, we test diverse ensembles of our trained models. Recall that we use a greedy parameter searching over validation set to determine the contribution portions of individual component models contained in respective ensemble. The contribution portions of individual component models determined over validation set are directly applied to testing set for third party evaluation. Our first 5 of 7 submissions are from an ensemble set containing the completely same component models: (1) 3 best versions of SSE-ResNet-50, SSE-DenseNet-121 and SSE-HoloNet discussed in Section 5.2; (2) 1 SVM classifier trained on audio feature set. Specifically, on the validation set, the highest overall recognition rate of our first 5 submissions is 55.09% and the lowest overall recognition rate is 53%. Accordingly, on the testing set, the highest/lowest overall recognition rate from the first ensemble set is 56.66%/54.21%. Considering that deep CNN features and traditional hand-crafted features have been proven to be complementary to each other in a variety of existing works [5, 8, 23, 34], we also incorporate one baseline model [34] into our last 2 submissions. Comparatively, we obtain consistently better accuracy when performing the ensemble jointly with our audio, deep CNN and hand-crafted models. As a result, we achieve the best overall recognition rate of 59.01%/60.34% on the validation/testing set, outperforming official baseline accuracy significantly. Detailed results are summarized in Table 2. The confusion matrices of our best submission both on the validation and testing data sets are shown in Figure 7. It can be seen that our solution performs well in recognizing emotion categories of Angry, Happy and Neutral, but it shows poor capability in identifying Disgust and Surprise. Similar conclusions are also reported in the methods of previous top teams [12, 34]. We believe this shall be mainly attributed to the intrinsically serious ambiguities.

5.4 Discussion

Our SSE explicitly inherits the advantage of deep supervision [21], namely providing dense supervision to diverse feature layers jointly rather than the standard method providing direct supervision to the top layer only, and hence it naturally eases the training of deep CNN models and improves the final predication accuracy. Beyond this advantage, our SSE also benefits from a new fusion structure in which class-wise scoring activations at diverse complementary feature layers are concatenated and further used as the inputs for second-level supervision, acting as a deep feature ensemble within a single CNN architecture. Although we clearly show the promising performance of our SSE, its performance can be further improved. First, existing evidence [30] shows that deep features smoothed over time space are critical for accurate emotion recognition, we believe a careful association of the proposed SSE and temporal sampling method may be beneficial to final predication accuracy. Second, it would be interesting to explore its efficacy in other contexts such as audio modality with deep neural networks.

6 CONCLUSION

In this paper, we present the method regarding our submissions to the audio-video based emotion recognition task of EmotiW 2017. In sharp contrast to the previous top teams that directly adopt deep CNNs with top-layer supervision to handle the problem of the emotion recognition, we propose SSE which provides dense supervision to diverse feature layers first and then bridges class-wise scoring activations for second-level supervision. We show that our SSE achieves significant accuracy improvements on several top CNN architectures compared with the standard training counterpart. Experiments also demonstrate that a simple ensemble of our final models obtains so far best accuracy on the audio-video based emotion recognition task. We hope our method can inspire future research in the related field.

REFERENCES

[1] D. Bau, B. Zhou, A. Khosla, and A. Torralba. 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations. *Proceedings of the IEEE Computer Vision and Pattern Recognition.* (2017).

Table 1: Result comparison of our networks on the validation set (%).

Method	Original	With SSE Learning Strategy			
		3conv-s + Eltwise	3conv-s + Concat	4conv-c + Eltwise	4conv-c + Concat
DenseNet-121	41.3594	44.1253	**45.6919**	43. 5625	44.6719
HoloNet	40.9922	44.2839	**46.4752**	41.4308	43.6031
ResNet-50	41.7755	-	-	-	**42.5587**

Table 2: Final recognition accuracy of our top 4 submissions to EmotiW 2017, both on the validation and testing sets (%).

Validation (%)	Test (%)	Method
54.57	55.74	3rd Fusion of 1 SSE-ResNet + 1 SSE-DenseNet + 1 SSE-HoloNet + 1 audio model
55.09	56.66	5th Fusion of 1 SSE-ResNet + 1 SSE-DenseNet + 1 SSE-HoloNet + 1 audio model
56.14	57.58	6th Fusion of 1 SSE-ResNet + 1 SSE-DenseNet + 1 SSE-HoloNet + 1 hand-crafted model + 1 audio model
59.01	60.34	7th Fusion of 1 SSE-ResNet + 1 SSE-DenseNet + 1 SSE-HoloNet + 1 hand-crafted model + 1 audio model

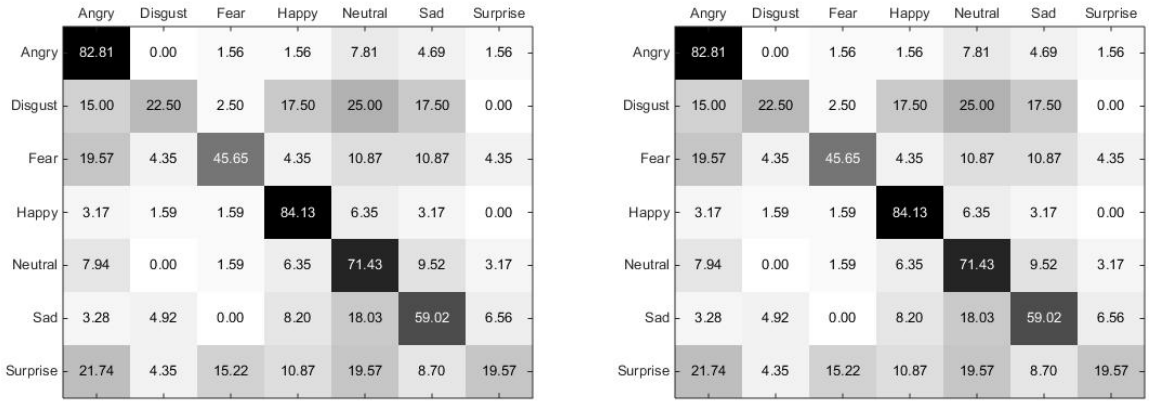

Figure 7: Confusion matrices of our best submission to EmotiW 2017. For the results on the validation set (left), the models are trained with given training data. For the results on the testing set (right), the models are trained with the union of given training and validation sets. In the figures, the darker the grid, the higher the recognition rate (%).

[2] W. Chen, JE. Meng, and S. Wu. 2006. Illumination compensation and normalization using logarithm and discrete cosine transform. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics).* 36, 2 (2006), 458–466.

[3] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng. 2017. Dual Path Networks. *arXiv preprint arXiv:1707.01629.* (2017).

[4] F. Chollet. 2016. Deep Learning with Separable Convolutions. *arXiv preprint arXiv:1610.02357.* (2016).

[5] A. Dhall, R. Goecke, T. Gedeon, and N. Sebe. 2013. Emotion recognition in the wild. *Proceedings of the 15th ACM on International conference on multimodal interaction.* (2013), 509–516.

[6] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon. 2017. From Individual to Group-level Emotion Recognition: EmotiW 5.0. *Proceedings of the 19th ACM International Conference on Multimodal Interaction.* (2017).

[7] A. Dhall, R. Goecke, J. Joshi, and T. Gedeon. 2016. EmotiW 2016: video and group-level emotion recognition challenges. *Proceedings of the 18th ACM on International Conference on Multimodal Interaction.* (2016), 427–432.

[8] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. 2014. Emotion Recognition In The Wild Challenge 2014: Baseline, Data and Protocol. *Proceedings of the 16th ACM on International Conference on Multimodal Interaction.* (2014), 461–466.

[9] A. Dhall, O.V.R. Murthy, R. Goecke, and et al. 2015. Video and Image based Emotion Recognition Challenges in the Wild: EmotiW 2015. *Proceedings of the 17th ACM on International Conference on Multimodal Interaction.* (2015), 423–426.

[10] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.A. Heng. 2016. 3D Deeply Supervised Network for Automatic Liver Segmentation from CT Volumes. *International Conference on Medical Image Computing and Computer-Assisted Intervention.* (2016), 149–157.

[11] F. Eyben, M. Wollmer, and B. Schuller. 2010. opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. *Proceedings of the 18th ACM on International Conference on Multimedia.* (2010), 1459–1462.

[12] Y. Fan, X. Lu, D. Li, and Y. Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. *Proceedings of the 18th ACM on International Conference on Multimodal.* (2016), 445–450.

[13] T. Hassner, S. Harel, E. Paz, and R. Enbar. 2015. Effective Face Frontalization in Unconstrained Images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (2015), 4295–4304.

[14] K. He, X. Zhang, S. Ren, and J. Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385v1.* (2015).

[15] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. 2016. Deeply supervised salient object detection with short connections. *arXiv preprint arXiv:1611.04849v2.* (2016).

[16] A.G. Howard, M. Zhu, B. Chen, and et al. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861v1.* (2017).

[17] G. Huang, Z. Liu, K.Q. Weinberger, and L. Maaten. 2016. Densely Connected Convolutional Networks. *arXiv preprint. arXiv:1608.06993.* (2016).

[18] S.E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. 2015. Recurrent Neural Networks for Emotion Recognition in Video. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction.* (2015), 467–474.

[19] S.E. Kahou, C. Pal, X. Bouthillier, and et al. 2013. Combining modality specific deep neural networks for emotion recognition in video. *Proceedings of the 15th ACM on International conference on multimodal interaction.* (2013), 543–550.

[20] A. Krizhevsky, I. Sutskever, and G.E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems.* (2012), 1097–1105.

[21] C. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. 2014. Deeply-Supervised Nets. *arXiv preprint. arXiv: 1409.5185.* (2014).

[22] M. Lin, Q. Chen, and S. Yan. 2013. Network In Network. *arXiv preprint arXiv:1312.4400.* (2013).

[23] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. 2014. Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild. *Proceedings of the 16th International Conference on Multimodal Interaction.* (2014), 494–501.

[24] T. Ojala, M. Pietikainen, and T. Maenpaa. 2002. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 24, 7 (2002), 971–987.

[25] M. Pantic, Z. Zeng, T.S. Huang, and G.I.Roisman. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 31, 1 (2009), 39–58.

[26] E. Shelhamer, J. Donahue, J. Long, Y. Jia, and R. Girshick. 2014. DIY Deep Learning for Vision: a Hands-On Tutorial with Caffe. *Proceedings of the 13th European Conference on Computer Vision.* (2014).

[27] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556.* (2014).

[28] C. Szegedy, S. Loffe, V. Vanhoucke, and A. Alemi. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261.* (2016).

[29] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. 2014. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (2014), 1701–1708.

[30] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. 2014. Learning Spatiotemporal Features with 3D Convolutional Networks. *arXiv preprint arXiv: 1412.0767.* (2014).

[31] P. Viola and M. Jones. 2001. Rapid Object Detection using a Boosted Cascade of Simple Features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (2001), 511–518.

[32] X. Xiong and FDL. Torre. 2013. Supervised Descent Method and Its Applications to Face Alignment. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (2013), 532–539.

[33] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen. 2016. HoloNet: towards robust emotion recognition in the wild. *Proceedings of the 18th ACM International Conference on Multimodal Interaction.* (2016), 472–478.

[34] A. Yao, J. Shao, N. Ma, and Y. Chen. 2015. Capturing AU-Aware Facial Features and Their Latent Relations for Emotion Recognition in the Wild. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction.* (2015), 451–458.