StressNet: A Dynamic Dropout Layer based Neural Network for Stress Recognition

Hao Wang

Research School of Computer Science
Australian National University
u7195922@anu.edu.au

Abstract. Stress is a body response to the changing of environmental conditions, such as facing time pressure, threats, or scary things. Being in a stressful state for a long time will seriously affect our physical and mental health. Therefore, it is of importance for human to monitor the stress situation in time. In this paper, we propose a deep neural network with novel dynamic dropout layers to address the stress recognition task through thermal images. Dropout regularization has been widely used in various deep neural networks for combating overfitting. In the stress recognition task, overfitting is a common phenomenon. Our experiments show that the our proposed dynamic dropout layers could not only speed up the training process and alleviate overfitting, but also make the network focus on the important features and ignore unimportant feature information at the same time. The proposed approach was evaluated in comparison with the baseline models[5] [10] over the ANUStressDB dataset. The experiment results show that our model has achieved 95.8% classification accuracy on the test set.

Keywords: Stress Recognition · Deep Neural Network · Dropout · Dynamic Masks

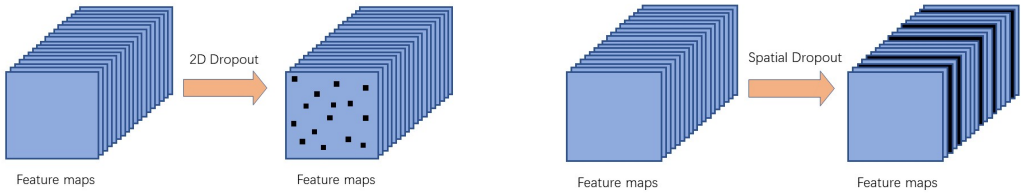
1 Introduction

We experience stress every minute of our lives. Whether slight or intense, it always there and affects our health conditions. Stress can affect our central nervous system (CNS) [6] and has destructive effects on our memory, cognition and learning systems [14]. Traditional technologies for stress detection are mainly contact-based such as face-to-face stress consulting by a psychologist. Self-reporting systems are the alternative way but may not be able to detect stress situations in a short period of time. [5]. Nowadays, the world has come to a standstill due to the COVID-19 pandemic [9]. It is imperative to the need of reliable contact-less monitoring systems that could detect the changing of stressful situations in our daily life. Stress is highly correlated with our physiological response that can be collected by some sensors such as an RGB or a thermal camera. If we are stressed, the features extracted from images produced by these cameras will be different from the features when we are calm. Researchers were attempting to detect and classify stress states from calm states by measuring physiological signals features extracted from the RGB or thermal images. Pavlidis et al. [7], [8] measured the stress by using a thermal sensor to detect the increase of blood flow in one's forehead region caused by stress conditions. Sharma et al. [10] apply the support vector machines on the Spatio-temporal features extracted from thermal images and achieved 86% accuracy. Shastri et al. [11] proposed a physiological function to measure the transient perspiration captured by thermal images. Irani et al. [5] extracted features from super-pixels by fusing the features extracted by RGB and thermal images and achieved 89% classification accuracy on the ANUstressDB dataset.
Dropout is a stochastic regularization technique [4][12] that has a good effect on combating with overfitting, so it is widely used in deep neural networks (DNN). Conventionally, it works mainly in fully-connected (FC) layers by randomly "dropping" out the activation of a neuron with a certain probability p for each training case. The essence of the dropout method is to randomly prun the neural network. The process has the effect of model averaging by simulating a large number of networks with different network structures, which, in turn, making node activations in the network more robust to the inputs.

1.1 Motivations

Inspired by the mechanism of dropout, other stochastic method were proposed to create the dynamic sparsity within the network. For example, spatial dropout [13] extended the dropout approach to the convolutional layer to remove the feature map activations in the object localization task, accounting for strong spatial correlation of nearby pixels in natural images. The standard dropout and spatial drop at the output of 2D convolutions are shown in **Fig.1**. The standard dropout is a pixel-wise zeroed approach while the spatial dropout is a channel-wise zeroed approach on the feature maps. The sparsity generated by these two approaches can induce more randomness to the network during training phase, which can improve the robustness of the model.
Overfitting is a common phenomenon on the stress recognition task since stress is a subjective response, which means that different people will response to the same situation differently. On the other hand, stress has a strong

coherence, and this coherence will make the body's response to changes in the external environment lagging behind. These problems make the detection of pressure very difficult. Overcoming the overfitting problem seems to be the top priority of the stress recognition task. The biggest problem in Dropout is that we need to set all fixed probability values for all dropout layers. These probability values are the hyperparameters of the model that need to tune. Although we can use cross validation or genetic algorithm(GA) to obtain the best hyperparameters combination on the validation set, this will greatly increase the training time. On the other hand, cross validation wastes some training data, which is undoubtedly worse for stress detection tasks that are prone to overfitting. Therefore, we propose a dynamic dropout layer that could generate masks to eliminate some elements in the feature maps. These masks generated by some independent convolutional layers whose parameters will update during training via back propagation. Unlike the traditional dropout layer, dynamic dropout layer does not retain the values that has not been zeroed, but scales them according to different parameters. The scaled parameters can still be updated during training.



(a) The standard dropout(2D) (b) The spatial dropout

Fig. 1: Two different dropout approaches. (a) Standard dropout will zero each element of its input by a fixed probability p. (b) The spatial dropout will zero each channel of its input by a fixed probability p.

1.2 Technical Contributions:

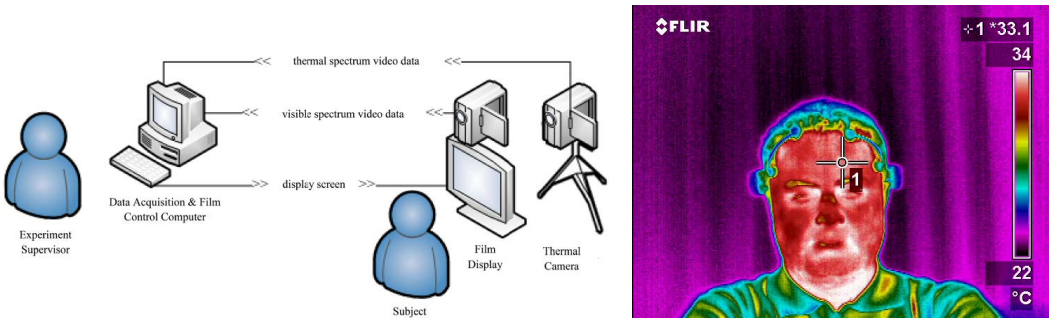
- A Deep Neural Network StressNet that achieved high accuracy on the ANUStressDB dataset.
- Dynamic dropout layers could speed up training and reduce overfitting.
- An exploration of model parameter settings that is different from traditional strategies such as cross-validation and genetic algorithms, but let the model learn the parameters during training.

The rest of this paper is organized as follows: Section 2 explains the task we designed and the data pre-processing method we applied, then we introduce the architecture of the proposed model. Section 3 shows the results of the experiments. Section 4 provides a discussion and ablation study and Section 5 includes the future work and concludes this paper.

2 Methodology

2.1 Dataset and Data Pre-processing

Dataset: ANUStressDB The dataset we used in this paper is the ANUStressDB dataset [5] containing thermal videos of 35 subjects watching stressed and not-stressed film clips validated by the subjects. Each video has 20 video clips that half of them belongs to the label "Stressful" and the other half of videos belongs to "Calm". Therefore, it is a well-balanced dataset. A schematic diagram of the experiment setup of ANUStressDB and the example image frame extracted from one of these videos are shown in **Fig.2** [10].



(a) Setup for the film experiment of ANUStressDB [10] (b) A sample image

Fig. 2: The setup process of ANUStressDB and one sample images extracted from one of the videos. The image has the size of 640 × 480.

Data Pre-processing The data pre-processing in this paper is shown as **Fig.3**. From the given information about the experiment generating ANUStressDB, there exists about 30 seconds for preparing on the beginning of each video and there exist about 6 seconds between each pair of clips. We use FFMPEG to segment each video into 20 video segments by the order of labels. Then for each video segment, we extract 30 frames evenly to form the image dataset. We randomly split the image dataset into training set and test set by the ratio of 80:20. This is a big dataset, so we also extract a smaller dataset eaxtracting images only from the first 6 labels of each video for evaluate our proposed approaches. For each image in the training set and test set, we apply center crop on it to get the image of size 480 × 480 and then resize it to 240 × 240.

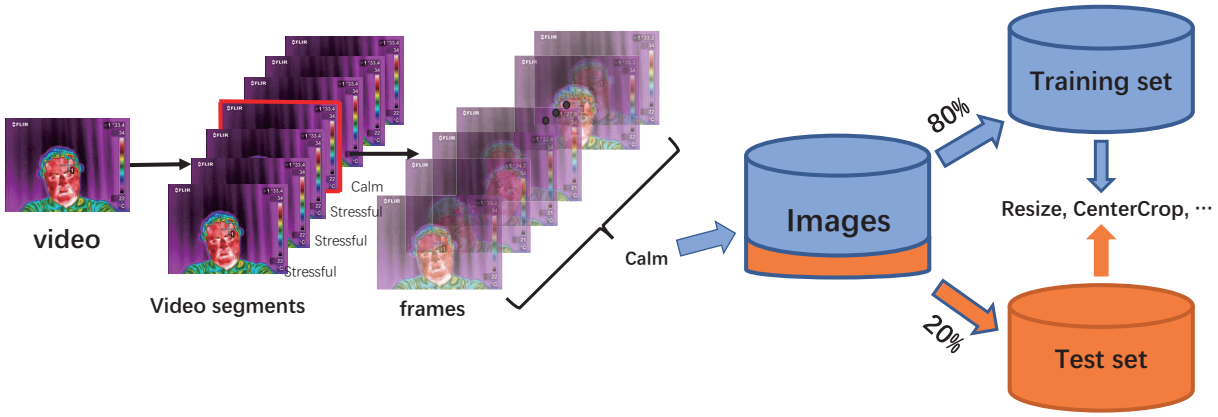


Fig. 3: Data Preprocessing

2.2 Network Architecture

The architecture of the proposed model is shown in Fig. 4. The model is mainly composed of a Shallow Convolutional Layer(SCL) and a Deep Convolutional Layer(DCL). There also exists a standard dropout Layer before the final fully connected layer. We used ReLU [2] as the activation function in the Dynamic Dropout Layer to generate the mask and the sigmoid activation function in the last fully connected layer. Other activation functions are all GELUs [3] which could generate more robust model and speed up training considerably. Our

experiments show that GELUs could reduce the triaining time about 40% compared with ReLU and 45% PReLU respectively.

$$\mathrm{GELU}(x) = xP(X < x) = x\Phi(x) = \frac{1}{2}x[1 + \mathrm{erf}(\frac{1}{\sqrt{2}})] \tag{1}$$

where $X \sim N(0,1)$ and $\Phi(x)$ its cumulative distribution function.

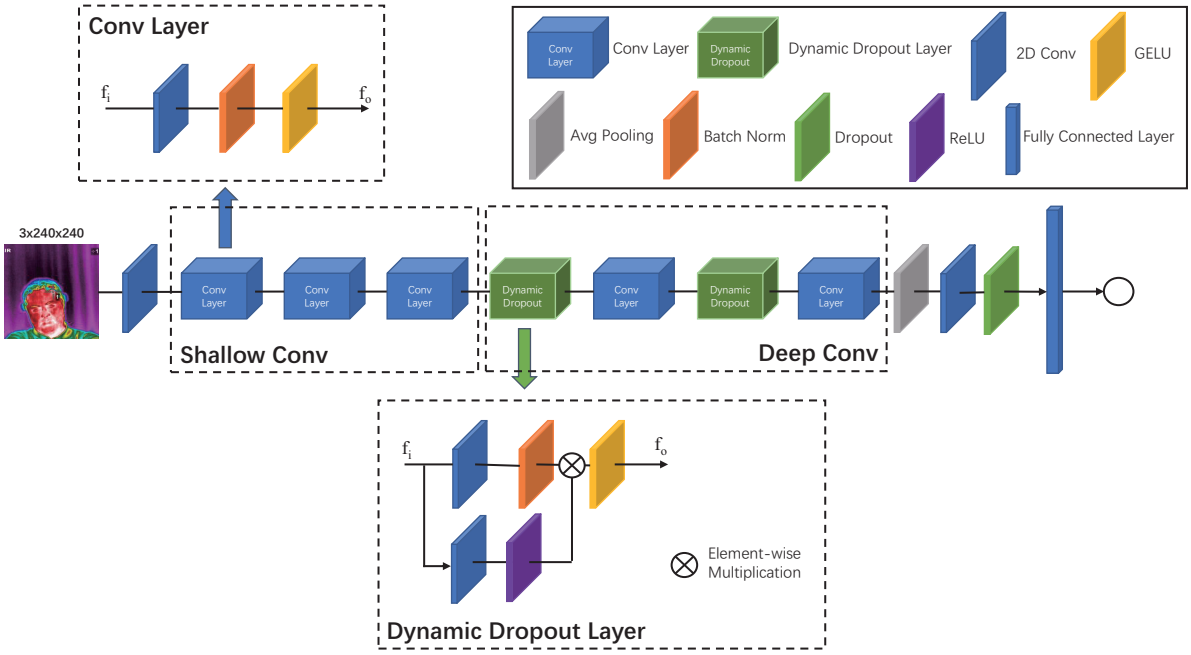


Fig. 4: Model Architecture.

Shallow Convolutional Layer(SCL) The SCL contains three convolutional layers. Each convolutional layer contains a 2D convolution with kernel size of 3 followed by a Batch Normalization [1] and a GELU activation function. We have the equation of the CL as

$$\mathbf{f_o} = \mathrm{CL}(\mathbf{f_i}) = \mathrm{GELU}(\mathrm{BatchNorm}(\mathrm{Conv}(\mathbf{f_i}))) \tag{2}$$

The reason why we do not use the Dynamic Dropout Layer(DDL) in SCL is that the network needs to fully extract the features of the image in the shallow layer. If dropout is added to the shallow network, it will cause serious interference to the network, so that the deep network cannot extract useful feature information, which will makes the training phase slow and even underfitting.

Deep Convolutional Layer(DCL) The DCL contains two convolutional layers and two Dynamic Dropout Layers(DDL). A DDL contains a same architecture as the convolutional layer as the main trunk. And it also has a side branch network including a convolution and a ReLU activation function. The side branch network can generate feature maps with the same shape as the output of the main trunk network. Then the feature maps generated by the backbone network and the masks are correspondingly multiplied, and the result goes through the GELU function to generate the output of DDL. Note that the convolutions in DDL have kernels with stride of 1 while the convolutional layer in DCL has kernel with stride of 2. The process of DDL can be written as

$$\mathbf{f_o} = \mathrm{DDL}(\mathbf{f_i}) = \mathrm{GELU}(\mathrm{BatchNorm}(\mathbf{Conv_1}(\mathbf{f_i})) \star \mathrm{ReLU}(\mathbf{Conv_2}(\mathbf{f_i}))) \tag{3}$$

Deep layers in the network will recieve the out feature maps from previous layers. Howerevr, we do not expect the neural network to learn all the feature information, because this can easily lead to overfitting. In the experiment section, we will show that if dropout is not added to the deeper layers, although the loss of the network in the training set continues to decline and the accuracy rate continues to increase, it will struggle in the test set, and

the accuracy rate will stay in a certain range for a long time and even drop. At the end of two nets, there exists a average pooling layer and a standard dropout layer. And the result will pass through a fully connected layer for classification. We use the Binary Cross Entropy loss

$$L_2 = -\frac{1}{N}\sum_{n=1}^{N}\{t_n log(y_n) + (1-t_n) log(1-y_n)\} \tag{4}$$

3 Experiments

We perform the experiment section by comparing the results of different model settings and we used the model of [5] and [10] as baselines.

3.1 Experimental Setup

Datasets As we discussed in section 2.1, we use the smaller image dataset containing the first 6 labels of each video to evaluate our proposed model and the whole image dataset to compare with the baseline.

Development Environment We use Pytorch 1.6.0 to construct our models. We use a NVIDIA RTX 2070 SUPER GPU to training our models on the Windows 10 platform.

Hyperparameters The hypaperameters used in this paper is shown in Table.1.

Table 1: Hyperparameters

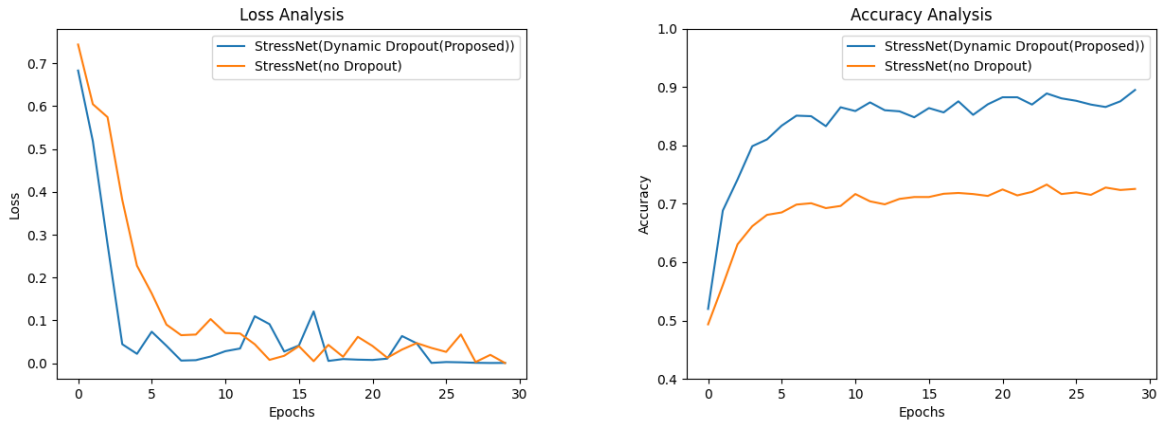
Dataset	Epochs	Learning rate	Bacth size	Optimizer
Small(6 labels)	30	5e-4	16	Adam
Whole(20 labels)	100	1e-3	16	Adam

3.2 Results and Discussion

The results of our experiments are shown in **Table.2**, **Fig.5** and **Fig.6**. Obviously, the Dynamic Dropout Layer could speed up training and alleviate overfitting phenomenon which is quite common in the stress recognition task. However, Batch Normalization is generally considered to have the same series of advantages as above, but why in our experiments, the model with batch normalization layers has a serious overfitting phenomenon? Let us try to analyze the mechanisms of Batch Normalization and dropout against overfitting. The mechanism for Batch Normalization to achieve good generalization is to try to find the solution of the problem in a smoother solution subspace. The emphasis is on the smoothness of the process of processing the problem. The implicit idea is that the smoother solution has better generalization ability, For dropout, robustness is the goal, that is, the solution is required to be insensitive to the disturbance of the network configuration, and the implicit idea is that the more robust generalization ability is better. From this mechanism, Dropout's control mechanism for over-fitting is actually more direct, superficial and simpler, while Batch Normalization is more indirect, but at a lower level and more essential. However, the mechanisms of the two are different and overlapped. The smoothness and robustness are often the same, but they are not completely consistent. It cannot be simply said which effect is better, but from the mechanism point of view, the Batch Normalization idea seems to be statistical It is better, but because Batch Normalization's method for constraining smoothness is not complete, it just chooses a control mode, so it cannot fully reflect its control smoothness idea, so Batch Normalization is not have much effect on the generalization performance compared with the dropout.

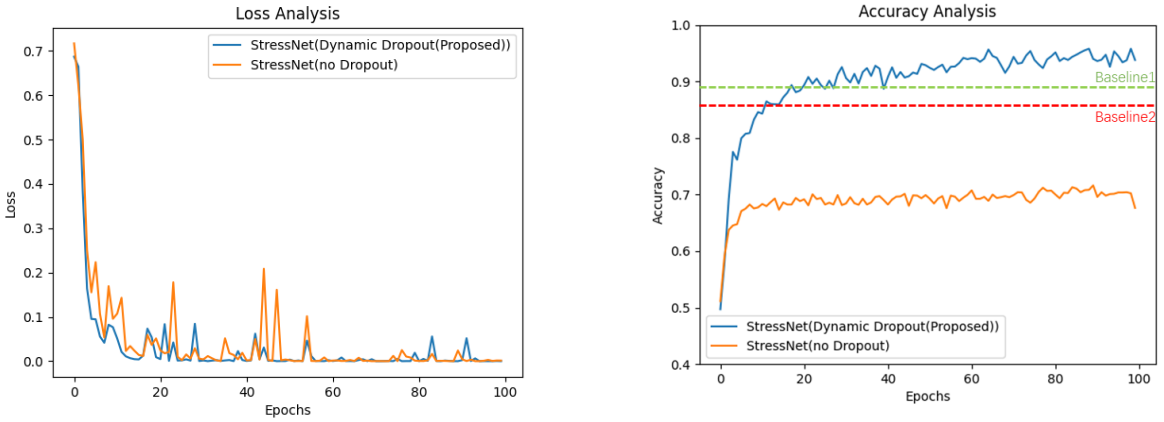
Table 2: Performances of our models on the test set.

	Baseline1[5]	Baseline2[10]	StressNet(Proposed)	StressNet(no dropout)
Small dataset(6 labels)	/	/	**89.7%**%	73.3%
Whole dataset(20 labels)	89.0%	86.0%	**95.8%**	71.4%



(a) Loss Analysis (b) Accuracy Analysis

Fig. 5: Results of experiment on the smaller dataset of the StressNet with Dynamic Dropout Layers and without any Dropout layers.



(a) Loss Analysis (b) Accuracy Analysis

Fig. 6: Results of experiment on the whole dataset of the StressNet with Dynamic Dropout Layers and without any Dropout layers.

4 Discussion

4.1 Visualization of learned masks

The masks generated by the first and second Dynamic Dropout Layers are shown in **Appendix Fig.8**. I randomly select 64 of these masks to visualize. As the figure shows, the Dynamic Dropout Layer could generate sparse masks. These masks will dropout the background part of the image that has less information about whether the person is stressed or calm and makes the network focus on the facial part in the image. In addition, the deep DDL will produce more sparse masks, and the non-zero part of these masks contains the key features to distinguish the participants from stressful or calm.

4.2 Ablation Study

We evaluate performances of different kinds of dropout layers in our StressNet.

Standard dropout v.s. Spatial Dropout v.s. DDL The performances of StressNet with different kinds of dropout layers on the smaller dataset are shown in **Fig.7**. It can be seen from the loss curves that the performance

of spatial dropout is the worst, because his zero-setting strategy is too aggressive, which will cause many useful feature maps to be discarded during forward propagation. This will greatly reduce the training speed of the network. The result of standard dropout is acceptable. Compared with the previous strategy without dropout, the introduction of dropout improves the generalization of the model. However, as we discussed above, the setting of hyper-parameters is a critical factor, and during the training process, the hyper-parameters are fixed, which brings many restrictions, and it may require multiple training processes to find the best combination of hyperparameters. The Dynamic Dropout Layer proposed in this paper has obvious advantages in speeding up the training process and improving the generalization ability of the model, and it also avoids the hyperparameter turning process.

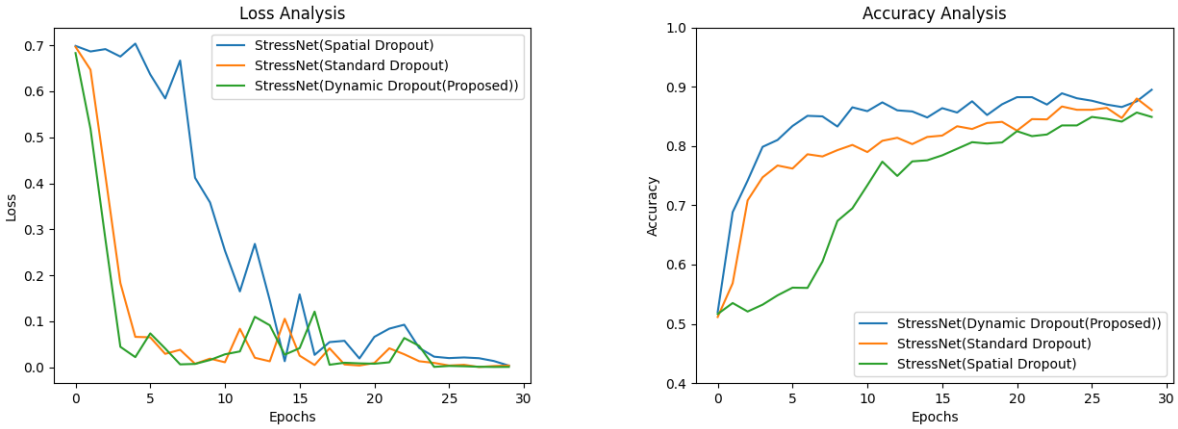


Fig. 7: Ablation study.

5 Conclusions and Future Work

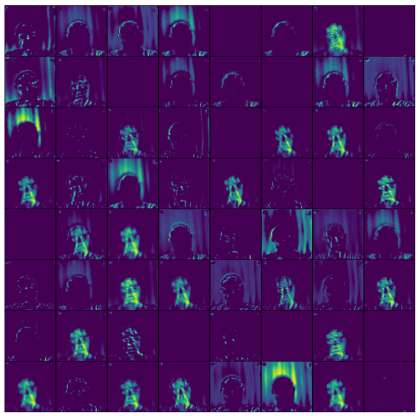
In this paper, we propose a A Dynamic Dropout Layer based Neural Network for Stress Recognition named StressNet to solve the stress recognition task. Our experiments prove that in the task of stress recognition, we need the dropout approach to combat the overfitting problem, especially in the deep part of the neural network. In addition, different from the traditional over-fitting method applied to the fully connected layer, this paper explored the impact of different types of dropout methods in the convolutional layer on the generalization of classification tasks. This paper proposes a dynamic dropout method to generate masks. Its parameters can be updated during the training process using backpropagation. Compared with the traditional dropout method, this method improves the training speed and the generalization ability of the model. Due to limited time, we did not explore the limitations of the model and optimize the model structure, or test the effect of the model on other data sets. Therefore, in the future, I may explore on other conventional RGB data sets, such as CIFAR-10, ImageNet, etc.

References

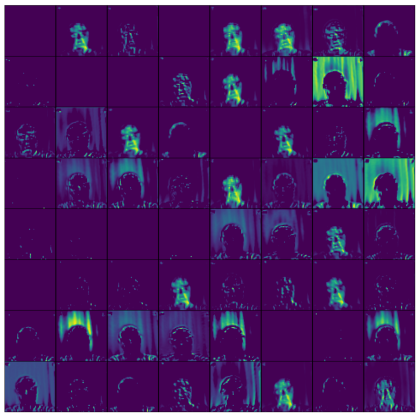
1. Ioffe, S., Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*(pp. 448-456). PMLR.
2. Glorot, X., Bordes, A., Bengio, Y. (2011, June). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 315-323). JMLR Workshop and Conference Proceedings.
3. Hendrycks, D., Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv*:1606.08415.
4. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:*1207.0580.
5. Irani, R., Nasrollahi, K., Dhall, A., Moeslund, T. B., Gedeon, T. (2016, December). Thermal super-pixels for bimodal stress recognition. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)* (pp. 1-6). IEEE.
6. Lupien, S. J., Lepage, M. (2001). Stress, memory, and the hippocampus: can't live with it, can't live without it. *Behavioural brain research*, 127(1-2), 137-158.
7. Pavlidis, I., Levine, J. (2002). Thermal image analysis for polygraph testing. *IEEE Engineering in Medicine and Biology Magazine*, 21(6), 56-64.
8. Pavlidis, I., Eberhardt, N. L., Levine, J. A. (2002). Seeing through the face of deception. *Nature*, 415(6867), 35-35.

9. Roser, Max, et al. "Coronavirus pandemic (COVID-19)." (2020) *Our world in data*.
10. Sharma, N., Dhall, A., Gedeon, T., Goecke, R. (2014). Thermal spatio-temporal data for stress recognition. *EURASIP Journal on Image and Video Processing*, 2014(1), 1-12.
11. Shastri, D., Papadakis, M., Tsiamyrtzis, P., Bass, B., Pavlidis, I. (2012). Perinasal imaging of physiological stress and its affective potential. *IEEE Transactions on Affective Computing*, 3(3), 366-378.
12. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929-1958.
13. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C. (2015). Efficient object localization using convolutional networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 648-656).
14. Yaribeygi, H., Panahi, Y., Sahraei, H., Johnston, T. P., Sahebkar, A. (2017). The impact of stress on body function: A review. *EXCLI journal*, 16, 1057.

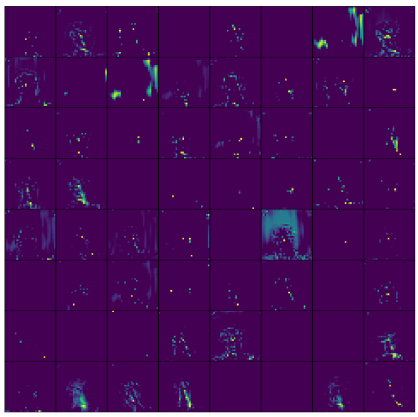
A Appendix



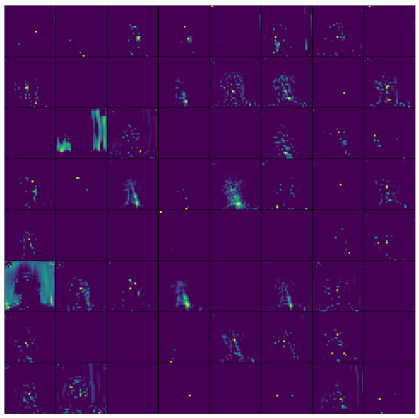
(a) 64 masks generated by the first DDL



(b) 64 masks generated by the first DDL



(c) 64 masks generated by the second DDL



(d) 64 masks generated by the second DDL

Fig. 8: Visualization of 64 masks generated by two Dynamic Dropout Layers.