Visual Emotion Recognition Using ResNet

Azmi Najid, Dina Chahyati
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
azmi41@ui.ac.id, dina@cs.ui.ac.id

***Abstract*— Given an image, humans have emotional reactions to it such as happy, fear, disgust, etc. The purpose of this research is to classify images based on human's reaction to them using ResNet deep architecture. The problem is that emotional reaction from humans are subjective, therefore a confidently labelled dataset is difficult to obtain. This research tries to overcome this problem by implementing and analyzing transfer learning from a big dataset such as ImageNet to relatively small visual emotion dataset. Other than that, because emotion is determined by low-level and high-level features, we will make a modification to a pretrained residual network to better utilize low-level and high-level feature to be used in visual emotion recognition. Results show that general (low-level) features and specific (high-level) features obtained from ImageNet object recognition can be well utilized for visual emotion recognition.**

Keywords*— *visual emotion, transfer learning, convolutional neural network

I. Introduction

Visual emotion recognition aims to associate images with appropriate emotions [7]. Humans have an emotional reaction towards images they observed. Every human being have a different emotional reactions to an image. This reaction is very subjective, so it is difficult to collect or label images that correctly represent the emotions of an image. Because of this subjective nature, it is difficult for computers to recognize the emotion of an image.

Along with the development of technology, computing power and data sources are more and more accessible. This development encourages research related to image processing to use deep learning. Deep architecture like deep convolutional neural network needs a big dataset for it to works well [20]. However, this method is difficult to use in visual emotional recognition, since large and confidently labelled data are difficult to obtain. Although You et al., have built a Large-Scale Emotion Dataset [1] consisting of 23,000 images to encourage research related to visual emotion recognition, it is not as large as ImageNet dataset which contain 14 million images [1], [21].

Bigger dataset allows deep neural network to generalize the model to a new data better [19]. Deep convolutional neural network has a very deep number of layers and is able to solve complex problems like ImageNet Object Recognition [2], [19]. Deep architectures such as ResNet can give a good performance for a big dataset like ImageNet, but for a small dataset this model vulnerable to overfitting [4], [19]. To solve the overfitting problem, usually regularization method like dropout, weight decay, and data augmentation is used to reduce generalization error [19]. Another solution to improve performance on a small dataset is to use transfer learning [5], [6]. Transfer learning enable knowledge learned from a task to be used in another task. This knowledge is represented as a network model (pretrained) which was trained on very large dataset. This pretrained network can be used as a feature extractor or as new network parameter initializer that needs to be fine-tuned later [18]. This method has shown good result [5], and this approach began to be widely used for training relatively small datasets such as the Large-Scale Emotion Dataset [1], [7], [16].

The state-of-the-art method in visual emotion recognition fuses features from the previous layers of CNN to utilize the low-level and high-level feature extracted from the network, but this method needs to train the network from scratch. Training model from scratch especially on a deep CNN takes a long time, compared to using pretrained network. By using pretrained network we can get the generic CNN features that is ready to be used or trained (fine-tuned) in much less time for other task such as visual emotion recognition. Other than that, utilizing pretrained network also showed significant improvement in classification accuracy for small dataset like the emotion dataset [1], [16]. Therefore, to make use of these advantages we will explore on how to utilize low-level and high-level feature obtained from a pretrained deep neural network such as ResNet for our problem.

In this paper, we will report the experimental result of applying transfer learning from ImageNet dataset to Large-Scale Emotion Dataset. We also modify the ResNet architecture in order to better utilized the extracted features from the pretrained network to be used in visual emotion recognition. Important issues addressed is the imbalance class of the dataset, fine-tuning, dropout and model ensembling.

The rest of the paper is organized as follows. We present related works in Section 2. Section 3 describes the proposed method, and Section 4 presents the results of our experiments. Finally, section 5 concludes the works described in this paper.

II. Related Works

Visual emotion recognition has been researched extensively [1], [8], [7], [9], [16]. Generally, there has been two approaches on visual emotion recognition, which is using handcrafted features or Convolutional Neural Network (CNN) based features [7]. Both approaches are trying to recognize emotion based on various kinds of low-level and high level features as shown in Figure 1. By low-level features, we mean the saturation, brightness, texture, or dominant colours in the image. Example of high-level feature is the semantics of objects detected in the image (tiger swimming, human swimming, etc).

Rao et al , introduced Multi-Level Deep Representation Network (MldrNet) which uses CNN to obtain low-level feature such as texture and aesthetics using T-CNN and A-CNN [9]. The use of CNN to obtain low level feature is also used by Zhu et al, the difference is that the features are obtained from every convolution layer. The reason is that

978-1-5386-8402-3/18/$31.00 ©2018 IEEE

every layer in CNN extracts different feature. First few layers of CNN extracts low level feature such as colour and shape, while later layers extracts higher level feature such as semantics of the picture [7]. These feature extracted from each layer are inserted into a Recurrent Neural Network to exploit the dependency from those features.

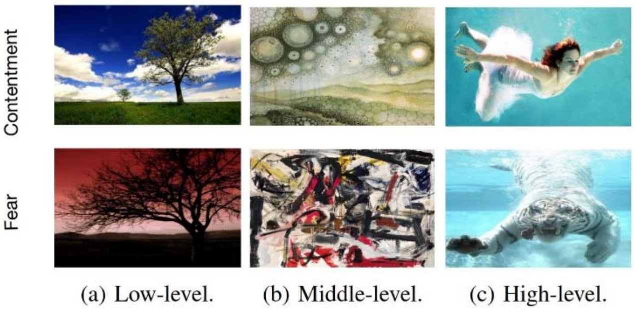


Fig. 1: The same emotion can be evoked from different emotion stimuli. The images in first row are from “Contentment”, and the images in second row are from ”Fear”. We can see image emotion is related to many factors, such as low-level features like color (a), middle-level features like texture (b), and high-level features like semantic content (c). [7].

The state of the art method on visual emotion recognition proposed by Zhu et al., incorporate low-level and high- level features obtained from a Convolutional Neural Network (CNN) to perform visual emotion recognition [7]. Features extracted from every layer of the network will then be used to make a prediction, so the network will have predictions based on each feature extracted from every layers of the network. Then these predictions will be combined to make the final prediction of a given image. This model gives the best result compared to other method despite only using shallow network (5 convolutional layer) and trained with random parameter initialization (without using pretrained network) however Yang et al., shows that Transfer Learning using a very deep convolutional neural network significantly improve accuracy in performing visual emotion recognition [7], [16]. Experiment conducted by Yang et al., compare the accuracy between training ResNet from scratch and using pretrained network. The experiment shows that training ResNet with a pretrained network gives 64.67% accuracy while training ResNet from scratch only gives 49.76% accuracy [16].

Training a model using Transfer Learning is easier/faster because we can use “off-the-shelf” CNN features to perform visual emotion recognition, moreover Yang et al., also shows that using a pretrained network can improve the performance of visual emotion recognition [16]. Visual emotion recognition as Zhu et al., described is determined by low-level and high-level features [7], so to recognize emotion of a given image, we will make a modification of a pretrained network to better utilize low-level and high-level features from a network.

III. Proposed Method

We proposed to incorporate low-level and high-level features obtained from a pretrained deep CNN. The architecture that we chose as the pretrained model is ResNet-101 [2] trained on ImageNet dataset. The model is evaluated using Hold-Out Validation [4]. Hold-Out Validation divides the dataset into three parts: training set, validation set, and test set. The ratio that we used is 80:5:15 respectively [1].

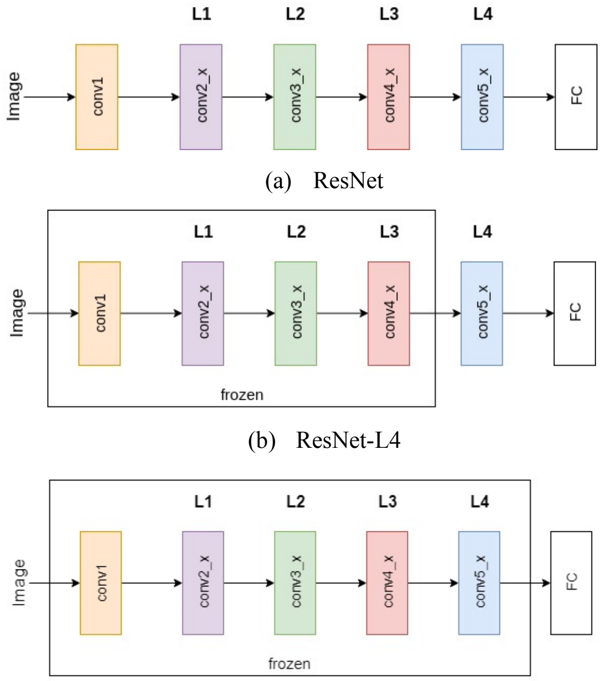
As mentioned briefly, we will explore four issues related to transfer learning. The first is the imbalance class of the dataset. We solve this issue using weighted sampling technique as described in part A. The second issue is about evaluating how far should we fine-tuned the pretrained model as described in part B, this result will tell us how important extracted feature on some part of the network for recognizing emotion. The third is to address the overfitting issue on a very deep neural network such as ResNet, to overcome this we will evaluate whether dropout will improve the recognition performance, as described in part C. The last issue is about ensembling the ResNet architecture, in this section we will introduce our method on how to incorporate low-level and high-level feature extracted from the pretrained network, this part will be described in section D.

A. Sampling

Imbalance class on the dataset causes classes with a small number of data to be recognized not as well as classes with bigger number of data. This experiment attempted to use weighted random sampling technique to load data into a mini-batch to solve the imbalance class problem. The difference between weighted random sampling and simply doing shuffling (random sampling) is, with shuffling the probability of sampling each image is the same regardless of the emotional class, while the weighted random sampling wants to match the ratio of each emotional class to the minibatch. This is done by making the sampling weight inversely proportional to the number of images in a class. We also employ data augmentation to the image so in one iteration the exact same images in class with a small number of data does not appear repeatedly [19], [11].

B. Fine-tuning ResNet

ResNet configuration is divided into 4 section of layers, namely L1, L2, L3, and L4. The goal of this experiments is to find out how useful extracted features on some parts of the layer are. We run three experiment on fine-tuning. In the first experiment, only the fully connected (FC) layer is fine-tuned, while the other layers are freezed with the pretrained model. In this model (ResNet-fc), the pretrained model is used as a full feature extractor. In the second experiment, FC and L4 are fine-tuned (ResNet-L4). In the last experiment all layers are finetuned (ResNet). These models are illustrated in Figure 2 below.



(a) ResNet

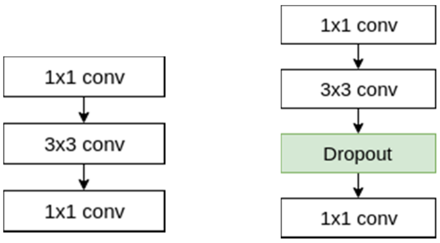
(b) ResNet-L4

(c) ResNet-FC

Fig. 2: Three models of fine-tuned experiments

C. Dropout

ResNet-101 is a ”very deep neural netwok” [2]. This deep architecture has many parameters so it will prone to overfitting and in the last experiments we will see the developed model is suffered from overfitting. Ebrahimi et al., make a modification to the original residual block by adding a dropout layer to prevent this [4], [12], [13]. The dropout layer added in residual block is placed between 3x3 convolution with 1x1 convolution layer as illustrated in Figure 3. In this experiment we compare ResNet-L4 with and without dropout layer. Dropout layer will randomly set entire feature map or channel to be 0 with a probability of 0.5.



(a) Originial residual block

(b) Residual block with dropout

Fig. 3: Dropout experiment

D. Ensemble Method

Transfer learning is proved to be effective because the extracted feature on pretrained network is transferable for different tasks [5], [6]. Features are transferable because CNN not only extracts specific features (high-level features) for a task but also extract more general features (low-level features) [5], [6]. Visual emotion recognition is also influenced by low-level and high-level feature as illustrated in Fig 1.

This experiment wants to utilize the extracted features in some parts of ResNet to perform visual emotion recognition. The idea is for the model to have different perspective from low-level and high-level feature on recognizing emotion from an image. To do this, branches from some parts of the network are created to extract different kind of features. Then, prediction from every branch is combined (stacked) by calculating the weighted average from all the predictions. This model is inspired by Zhu et al., the difference is that branches are created from every output of ResNet layers section (section L0, L1, L2, L3, L4) rather than from every convolution layer, because ResNet-101 has too many convolution layer [7]. We will call this model as ResNet-ensemble and is illustrated in Figure 4 below.

$$P(z=c|V)=\frac{\exp(W_c\,V)}{\sum_k \exp(W_k\,V)} \tag{1}$$

$$L(V)=-\log(P(z=c|V)) \tag{2}$$

Training ResNet-ensemble is done in two steps. The first step is to train every branch to predict emotion based on a given features. Output of every branch is formed into one dimensional vector V , then to make a prediction, softmax function is used, Equation 1 [3], [7]. Training is done by minimizing cross-entropy loss in Equation 2 [3], [7]. In this step the trained parameters are the ResNet parameter and the fully connected layers of every branch. In the second step the trained parameters are the the last fully connected layers and also the fully connected layers of every branch.

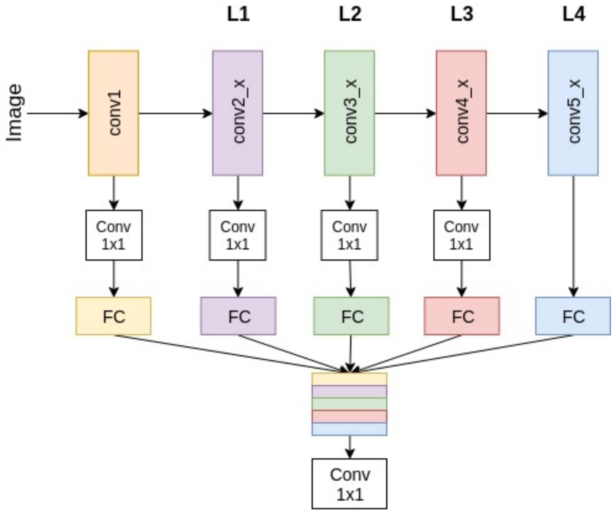


Fig. 4: ResNet-ensemble

For both steps optimization using Stochastic Gradient Descent (SGD) with mini-batch’s size 32. The learning rate of ResNet parameters are 1e-4, and it is smaller than other parameter's learning rate which is 1e-3, that is because we dont want to change the pretrained network parameters too much.

The other ensemble model we developed is done by combining feature extracted from layers section L0, L1, L2, and L3 (ResNet-skip). These features is combined by calculating the weighted average of the activation value on feature map [14, 20]. To calculate this we use ResNet's identity shortcut connection (skipping connection). This network structure can be seen in Figure 5. In order for the feature can be combined, dimension and number of channels (C×H×W) of a feature map must be the same. To do this we use 1×1 convolution to down-sample the feature as ResNet does on the first input of layers section.

The difference between this model and ResNet-ensemble model is that this model is trained in only one step, and also the parameters are trained only on the subsequent layers after combining feature map in the previous layer which is parameter in section L4 and in the fully connected layer (FC) section.

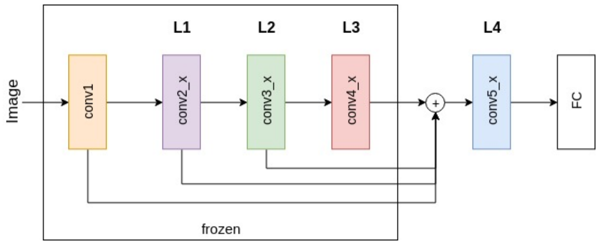


Fig. 5: ResNet-skip

IV. Results and Discussion

In this section, we will present the results of our experiments.

A. Experiment Detail

In the process of training, we use the same hyperparameter as the original ResNet parameter with a little adjustment on the hyperparameter value. The model is trained using SGD optimization with momentum = 0.9 (nesterov), weigh decay = 1e-4, learning rate = 1e-4, and mini-batch size of 32 for 50 epoch. Our model is implemented using deep learning framework PyTorch on one Nvidia GTX 1070. Performance metric used in this experiment is accuracy, but because there is an imbalance class problem on Large Scale Emotion Dataset [1], we also use macro average true positive rate to see how well the model can learn to recognize all eight emotion class including the emotion class with a small amount of sample.

B. Sampling

This sampling experiment was performed on ResNet by simply training the parameter on the replaced classifier. The result is shown in Table I, which tells that the model with weighted sampling gives almost the same accuracy as shuffling, but the macro average true positive rate (avg-recall) is significantly higher.

Two class that is greatly affected was Fear and Anger. In the dataset of 8 emotion classes, only 4.4% of the images are in Fear class, and only 5.4% is Anger. When we used mere shuffling, the number of images trained for this two classes are small, so their accuracy are only 5% and 10%. By using weighted sampling, the number of images for those two classes are added to match up the number of images in other classes. This method increased the accuracy of the two classes to 27% and 30%.

TABLE I. Sampling Experiment

	accuracy (%)	avg-recall (%)
Shuffling	62.04	51.04
Weighted	62.73	56.44

C. Fine-tuning ResNet

The result of fine-tuning is shown in Table II. The ResNet-fc model, only change the classifier of ResNet. ResNet-fc only make use of high-level feature which is extracted for ImageNet object recognition to do visual emotion recognition. That means the accuracy of 62.73 % is obtained using high-level feature for determining semantic content of the image. It shows that semantic content of an image has a big impact in recognizing emotion of an image. The ResNet-L4 and ResNet model give better evaluation result because these model have access to lower-level feature extracted from the earlier layer. So these model can utilize low-level feature like color, brightness, contrast, texture, etc for visual emotion recognition [4], [7], [10].

TABLE II. Fine-Tuning Experiment

	accuracy (%)	avg-recall (%)
ResNet-fc	62.73	56.44
ResNet-L4	64.95	60.30
ResNet	66.27	62.42

D. Dropout

The results of this experiment presented in Table III is contrary to Ebrahimi et al., results. Based on this result we see that dropout does not provide any significant improvement. This is because batch normalization has regularization capability and Ioffe et al., state that in batch normalized network, dropout can be either removed or reduced in strength [17].

TABLE III. Dropout Experiment

	accuracy (%)	avg-recall
ResNet-L4	64.95	60.2
ResNet-L4-dropout	65.04	60.29

E. Ensemble Method

The ResNet-ensemble model has more parameters to trained compared to ResNet-skip, so the training time of ResNet-ensemble also takes longer than ResNet-skip. Based on our experiment ResNet-ensemble gives a better result than

ResNet-skip as shown in Table IV Confusion matrix of ResNet-ensemble model can be seen in Figure 6..

TABLE IV. ENSEMBLE EXPERIMENT

	accuracy (%)	avg-recall
ResNet-ensemble	67.74	62.4
ResNet-skip	66.99	62.60

Based on all experiment, the best accuracy (67.74%) is achieved by ResNet-ensemble method using weighted sampling, but as shown in Table IV the avg-recall of ResNet-skip is slightly better than the ResNet-ensemble model. This means that ResNet-skip performs better on small sample emotion such as anger and fear. Confusion matrix for both model can be seen in Figure 6 and 7 below.

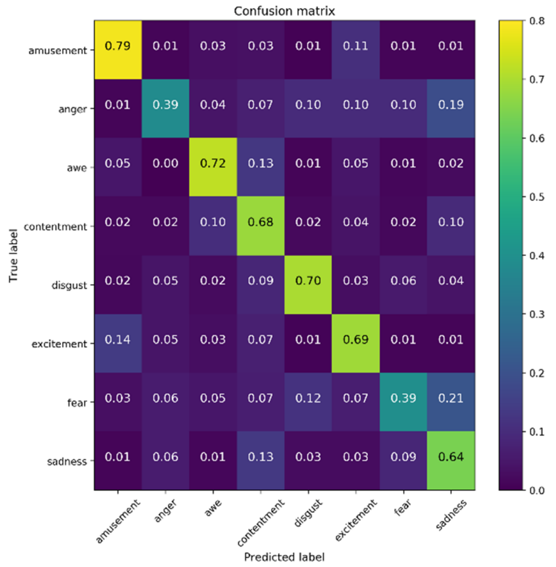


Fig. 6: Confusion matrix of ResNet-ensemble

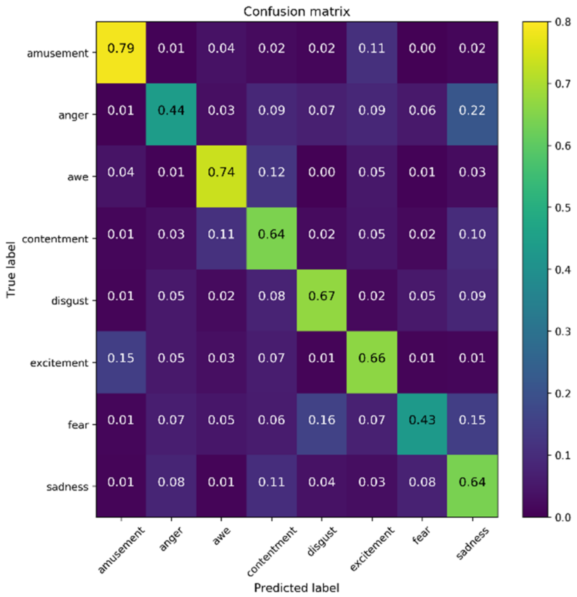


Fig. 7: Confusion matrix of ResNet-skip

All model developed in this experiments has different architecture and parameters to be fine-tuned. Because of this, these model will also have different training time compared to another. Table V present all models performance and their training time per epoch. As presented in Table V, we can see that ResNet-skip performs as good as ResNet-ensemble while spending much less time to trained. ResNet-skip was faster to train because the model only learn parameter on the new skipping connection and the parameter in section L4 of ResNet, while ResNet-ensemble fine-tuned all ResNet parameter in both training step.

TABLE V. COMPARED METHOD

	accuracy (%)	avg-recall	minute/epoch
ResNet-fc	62.73	56.4	4.86
ResNet-L4	64.95	60.3	4.74
ResNet	66.27	62.4	7.98
ResNet-skip	66.99	62.6	4.74
ResNet-ensemble	67.74	62.42	8.60

V. CONCLUSION

We present a modification of ResNet architecture (ResNet-ensemble and ResNet-skip) to use features from a pretrained network for visual emotion recognition. This modification allows low-level feature and high-level feature extracted from a pretrained network to be utilized better for visual emotion recognition. In our experiment we also show that weighted random sampling can help to improve performance on emotion dataset [1] which has an imbalance class problem. From the fine-tuning experiment we also find that semantic content of an image has a big influence on determining emotion of a given image.

REFERENCES

[1] You, Q., Luo, J., Jin, H., Yang, J. (2016). Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark. *AAAI Conference on Artificial Intelligence*, 30.

[2] He, K., Zhang, X., Ren, S., Sun, J. (2015). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 770-778.

[3] PyTorch documentation. (2018). Retrieved from https://pytorch.org/docs/ stable/index.html

[4] Chollet, F. (2018). *Deep Learning with Python*. Shelter Island, NY: Manning.

[5] Torrey, Lisa., Shavlik, J. (2009). Transfer Learning. *Handbook of Research on Machine Learning Applications*.

[6] Yosinski, J., Clune, J., Bengio, Y., Lipson, H. (2014). How transferable are features in deep neural networks. *Neural Information Processing Systems, 27*.

[7] Zhu, X., Li, L., Zhang, W., Rao, T., Xu, M., Huang, Q., Xu, D. (2017). Dependency Exploitation: A Unified CNN-RNN Approach for Visual Emotion Recognition. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.

[8] Machajdik, J., Hanbury, A. (2010). Affective Image Classification using Features Inspired by Psychology and Art Theory. *Proceedings of the 18th ACM international conference on Multimedia, 83-92*.

[9] Chen, M., Zhang, L., Allebach, J.P. Learning deep features for image emotion classification, 2015 *IEEE International Conference on Image Processing, 4491-4495*.

[10] Zeiler, M.D., Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *CoRR*.

[11] Sun, L. (2017). ResNet on Tiny ImageNet. *Stanford Report*.

[12] Ebrahimi, S.M., Abadi, H.K. (2016). Study of Residual Networks for Image Recognition. *Stanford Report*.

[13] Sritasvata, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*.

[14] Ju, C., Bibaut, A., van der Laan, M.J. (2017). The Relative Performance of Ensemble Methods with Deep Convolutional Neural Networks for Image Classification.

[15] Bamos. (2017). densenet.pytorch. Retrieved from https://github.com/bamos/densenet.pytorch/blob/master/compute-cifar10-mean.py

[16] Yang, J., She, D., Sun, M. (2017). Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.

[17] Ioffe, S., Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR*, abs/1502.03167.

[18] Stanford cs231n lecture notes. [site]. Retrieved from http://cs231n.github.io/

[19] Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. MIT Press.

[20] Engelbrecht, A.P. (2007). *Computational Intelligence: An Introduction*. Chichester, West Sussex: John Wiley.

[21] ImageNet. http://www.image-net.org/