Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 56-2 (2023) 8227–8230

The effect of feature space separation at different training states of CNN

N. Ding*. H. Arabian*. K. Möller*.

** Institute of Technical Medicine (ITeM), Furtwangen University, 78054, Germany (Tel: +49 (0)7720 307 4648; e-mail: din@hs-furtwangen.de).*

Abstract: Convolutional neural networks (CNNs) are successful in many different applications, however, such model decisions can be easily changed by slight modification on the inputs. The robustness needs to be guaranteed for the safety critical fields like medicine, therefore, it is necessary to understand the decision making procedure of CNN models. As the CNN model automatically extracts the image features and makes the corresponding predictions, observing the learned features space can approximately represent the decision boundary. In this paper, the use of linear interpolation to monitor the learned feature space is applied to analyze the separability property of a CNN model at different classes. By forcing the CNN to learn to separate the extracted features at different layer depths by adding the conformity loss, the classification distribution was more separable and stable to enhance the robustness of the model. The performance of linear interpolation showed the model had better classification abilities, where there are fewer perturbed classes appearing.

Keywords: adversarial attack, convolutional neural network, learned feature space, robustness

1. INTRODUCTION

Recently, convolutional neural networks (CNNs) have shown tremendous achievements in image classification applications as they can automatically extract features from the input image that are helpful to assign a proper class. But for an unknown reason, a slight human-imperceptible perturbation on benign samples can drastically change a well-trained CNN's prediction. This security weakness can be dangerous in security critical applications, especially in the medical area or other highly regulated domains (Ruan et al., 2018) (Ren et al., 2020). For instance, surgical tool recognition is one of the applications of CNN in the medical domain. Online surgical tool recognition is applied to provide support in developing context aware systems in modern operating rooms (Alshirbaji et al., 2020). When the CNN model makes wrong predictions, it can be harmful. For the purpose of safety, the robustness of such CNN models trained on surgical tool classification need to be carefully evaluated before being deployed.

Therefore, we need to understand why and how the CNNs are vulnerable, for the further step to improve the robustness and safeness. In this paper, the training procedure of a CNN model trained for classifying surgical tools is observed, to improve the learned feature space separation in order to enhance the decision robustness. Linear interpolation can be considered as an adversarial attack technique that intends to lead the model to inaccurate predictions by combining two features from different images. When the learned solutions are not robust enough, the CNN is not able to distinguish the mixed features from different images, and assigns the generated image to a class other than both image classes. To measure model robustness on different training states, the performance by linear interpolating randomly two legitimate images is evaluated. These adversarial samples facilitate the evaluation of the model performance. Furthermore, assessing the model robustness and its relation with training level.

1.1 Adversarial attack

There are many adversarial attack approaches to fool a CNN prediction. For instance, the adversarial perturbations can be generated in the backpropagation, by iteratively varying the input image, such as the fast gradient sign method (FGSM) (Goodfellow et al., 2014), the iterative fast gradient sign method (I-FGSM) (Kurakin et al., 2016), the momentum iterative fast gradient sign method (MI-FGSM) (Dong et al., 2018). Other perturbation method include the use of the forward derivative to construct adversarial saliency maps to produce desired perturbations (Papernot et al., 2016), the consideration of a deep neural network as a linear model to search for the minimum perturbation using the "deepfool" algorithm (Moosavi-Dezfooli et al., 2016), or by producing adversarial samples directly using existing network architecture e.g., generative adversarial networks (GANs) (Xiao et al., 2018).

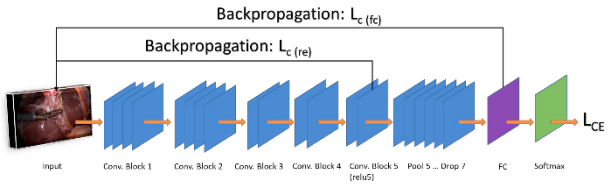

Fig. 1. AlexNet architecture and traditional cross-entropy loss with additional backpropagation optimization of conformity loss from layer "relu5" and the fully connected layer.

1.2 Adversarial defence

10.1016/j.ifacol.2023.10.1005

To mitigate the influence of adversarial perturbations, many defence methods to counter adversarial attacks are proposed. Adversarial training is an effective scheme by training the network both with clean and generated adversarial images, such as basic adversarial training (Goodfellow et al., 2014), max-margin adversarial training (MMA) (Ding et al., 2018); instance adaptive adversarial training (IAAT) (Balaji et al., 2019). In this paper, the focus is on the learned feature separation method proposed by (Mustafa et al., 2019). The method adds another loss function, in addition to the traditional cross-entropy loss, to separate the feature representations at multiple layer depths (Figure 1). However, in practice, the proposed prototype conformity loss was found not to converge during the training process. As a result, the use of another loss function to force the model to learn to separate the learned feature space was implemented. Additionally, the parameters are updated in the backward propagation form the specific layer where the conformity loss was calculated, instead of going through the whole model structure, to make sure the deeper layers would not be influenced by the feature separation ability of shallower layers.

2. METHODS

2.1 Material

A fine-tuned AlexNet (Krizhevsky et al., 2017) trained on laparoscopic video images for surgical tool classification was used. The original dataset of Cholec80 (Twinanda et al., 2016), which is a big dataset containing 80 laparoscopic videos was used, where there are 7 kinds of surgical tools used in the cholecystectomy procedure. To fulfil the task of using Softmax classification to classify these 7 surgical tools, one-class images are extracted from the Cholec80 to build a derived database. The derived dataset has 80,190 images in total, from which 25,000 were used for training and the remaining 55,190 used for testing.

To monitor the training process and the model performance in relation to the training states, the model was recorded when its training accuracy reached to 75%, 85%, 95%, and 99%, and snapshot models were named accordingly: model 75, model 85, model 95, model 99 (Ding and Möller, 2021) (Ding and Möller, 2022a). The performance of the CNN trained with the combination of traditional cross-entropy loss and conformity loss was also compared (Figure 1),the later trained model will be named as new model 75, new model 85, new model 95, and new model 99.

The performance was evaluated by the accuracy and f1-score, the latter is the harmonic mean of precision and recall.

2.2 Conformity loss

To restrict the overlapping between different classes to improve the separation ability of the model, an additional loss function was introduced to constrain the distance of the extracted feature to its true class region centre and increase the distance of extracted feature from other class region centres, in order to separate the different class regions. The conformity loss used is described in (1).

$$L_c(x,y) = \sum_i \frac{2*\|f_i - w_i\|_1}{\frac{1}{k-1}\sum_{j\neq i}(\|f_i - w_j\|_1 + \|w_i - w_j\|_1)} \quad (1)$$

For a given class i, f_i is the feature extracted from a specific layer, w_i is the current class centroid, and w_j the other class centroids. By converging the conformity loss in the training process, the numerator will be reduced and the denominator will be enlarged.

2.3 Adversarial evaluation

A small test set was created to evaluate the model robustness to adversarial perturbations. Two images were randomly chosen, and the adversarial images were generated using linear interpolation, as described in (2), with a fraction of $T \in [0,1]$.

$$x^* = (1-T)x_1 + Tx_2\,,\ T \in [0,1] \quad (2)$$

Where x^* is the generated image, x_1 and x_2 are the two random inputs. T was assigned with 100 values gradually increased from 0 to 1 with a step size of 0.01.

From previous experiments (Ding and Möller, 2022b), it was noticed that when a model is not robust, there would be some perturbed classification during interpolation, which represents a disturbed decision boundary (Figure 2). These pop-up interference classes were integrated as another evaluation metric, to present the classification stability, i.e. the robustness of a trained model.

$$Area_i = \int_0^{100} P_i(n_t)dn_t \ (3)$$

$P_i(n_t)$ is the probability of class i of the n_{th} generated image x^* at $T=t$. $Area_i$ indicates the class i probability integration during the whole interpolation process (Ding and Möller, 2022b).

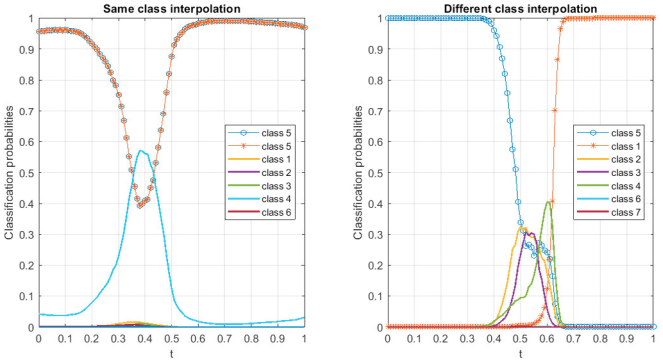

Fig. 2. Linear interpolation of two images. The classification probability trend shows a distorted decision boundary. The area under these pop-up interference classes(the peaks showed up in the middle of t) will be computed as a robustness index for classification stability.

2.4 Training algorithm

The training algorithm was summarized as below, the conformity loss was included for training after few iterations, before that, only the class centroids were updated.

Algorithm 1: Model training with conformity loss

Input: pre-trained model f, model parameters θ, model parameters θ_1 till layer l_1, model parameters θ_2 till layer l_2, training set{x,y}, maximize epoch T, stopping criterion with different training states.
Output: fine-tuned model F_θ .
For epoch < T:
  If iteration < 15,
    Update θ with L_{CE};
    Update the feature centres of each class i;
  Else iteration >= 15,
    Compute the conformity loss at layer l_1 and l_2;
    Compute the gradients from L_{c1} and L_{c2} separately; update parameters θ_1 and θ_2;
    Update θ with L_{CE};
    Update the feature centres of each class i;
Return: trained model F_θ meet training stop criterion.

3. RESULTS & DISCUSSION

In the experiments, the conformity loss at layer "relu5" and fully connected (fc) layer was calculated. The model parameters were updated from converging both the conformity loss and the cross-entropy loss. However, as the conformity loss only depends on the feature extraction ability from previous layers, only the model parameters from corresponding layer and shallower layers are updated. For instance, the conformity loss calculated from "relu5" layer can only influence the parameters updating before "relu5".

With the new training procedure, the new model are recorded at the same training states, to compare the performance with traditional trained models. The accuracy evaluation showed that the new models have a slightly worse performance, with a decrease in accuracy of approximately 2~3%, except for model 75 where an increment of 1.2% was seen. Similarly, except for model 75, the other models had a 4~6% decrement in the F1-score.

Table 1. Performance evaluations and comparisons.

Model	Acc.	F1-score	Model	Acc.	F1-score
75	81.1%	36.29%	New 75	82.3%	44.28%
85	90.3%	68.16%	New 85	87.0%	61.86%
95	94.0%	81.87%	New 95	92.2%	77.68%
99	95.2%	86.89%	New 99	93.3%	81.45%

However, by observing the feature space separation, it was noticed that the classification distribution showed visible improvements compared to the traditional training criterion. Figure 3 shows the extracted feature distribution of the test set by different layers and different models. The "relu5" feature distribution has slightly improved in the new model 99. Meanwhile, the fc layer has clearly improved classification distribution, as most of the samples are gathered to a smaller classification region.

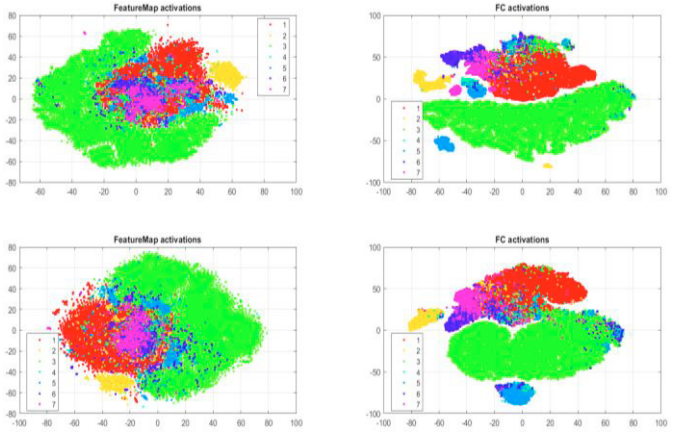

Fig. 3. Feature activation distribution from "relu5" layer (1st column) and fully connected (FC) layer (2nd column). The 1st row shows the extracted feature distribution from the original model 99, the 2nd row shows the extracted feature distribution from the new model 99 trained with conformity loss.

For the adversarial robustness evaluation, 3 images of each class were selected to perform the linear interpolation test. The area under curve was computed for the interference classes shown up during the interpolation. Figure 4 shows the comparison of the original models and the new trained models with additional conformity loss. Less interference classes were seen, this indicates the separation ability improved on distinguishing features extracted from different images.

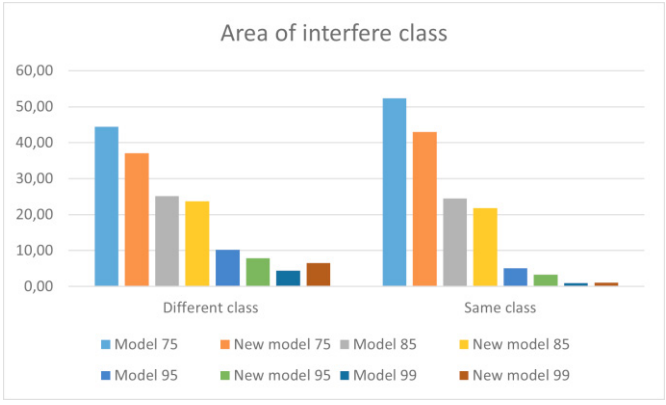

Fig. 4. Area of interfere classes were generally diminished in the new trained model with conformity loss at different training states.

It was noticed that, even though a better classification distribution was achieved, the new trained model gained lower accuracy on the test set. New model 75 had better performance than the original model 75, but with longer training iterations. This indicates the performance was improved by presenting more training samples.

On the other hand, the optimization with conformity loss has shown its effect on the separation ability of different classes. With less interference classes showing up in the linear interpolation, the model has gained more ability to distinguish the different features. The feature distribution was improved with less overlapping. Nevertheless, the test set is too small,

potentially leading to the contingency of statistic evaluation. In the further experiments, more samples as evaluation objects will be considered.

4. CONCLUSIONS

In this research, the CNN models classification accuracy and classification robustness with two training criterions were compared. The model trained with traditional cross-entropy loss had a slightly better classification accuracy, but lacked classification stability and separable ability compared to the model trained with a combination of cross-entropy and conformity loss. In future work, the focus will be on the trade-off problem of accuracy and robustness.

ACKNOWLEDGEMENT

This work was supported by the German Federal Ministry of Research and Education (BMBF under grant CoHMed/PersonaMed KFZ 13FH5I06IA).

REFERENCES

Alshirbaji, T.A., Ding, N., Jalal, N.A. and Möller, K. 2020. The effect of background pattern on training a deep convolutional neural network for surgical tool detection. Proceedings on Automation in Medical Engineering. 1(1), pp.024–024.

Balaji, Y., Goldstein, T. and Hoffman, J. 2019. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. arXiv preprint arXiv:1910.08051.

Ding, G.W., Sharma, Y., Lui, K.Y.C. and Huang, R. 2018. Mma training: Direct input space margin maximization through adversarial training. arXiv preprint arXiv:1812.02637.

Ding, N. and Möller, K. 2021. Generating adversarial images to monitor the training state of a CNN model. Current Directions in Biomedical Engineering. 7(2), pp.303–306.

Ding, N. and Möller, K. 2022a. Robustness evaluation on different training state of a CNN model. Current Directions in Biomedical Engineering. 8(2), pp.497–500.

Ding, N. and Möller, K. 2022b. Using linear interpolation to monitor the training state of a CNN model In: Singapore.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X. and Li, J. 2018. Boosting adversarial attacks with momentum In: Proceedings of the IEEE conference on computer vision and pattern recognition., pp.9185–9193.

Goodfellow, I.J., Shlens, J. and Szegedy, C. 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. 2017. Imagenet classification with deep convolutional neural networks. Communications of the ACM. 60(6), pp.84–90.

Kurakin, A., Goodfellow, I. and Bengio, S. 2016. Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

Moosavi-Dezfooli, S.-M., Fawzi, A. and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks In: Proceedings of the IEEE conference on computer vision and pattern recognition., pp.2574–2582.

Mustafa, A., Khan, S., Hayat, M., Goecke, R., Shen, J. and Shao, L. 2019. Adversarial defense by restricting the hidden space of deep neural networks In: Proceedings of the IEEE/CVF International Conference on Computer Vision., pp.3385–3394.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B. and Swami, A. 2016. The limitations of deep learning in adversarial settings In: 2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, pp.372–387.

Ren, K., Zheng, T., Qin, Z. and Liu, X. 2020. Adversarial attacks and defenses in deep learning. Engineering. 6(3), pp.346–360.

Ruan, W., Wu, M., Sun, Y., Huang, X., Kroening, D. and Kwiatkowska, M. 2018. Global Robustness Evaluation of Deep Neural Networks with Provable Guarantees for the $ L_0 $ Norm. arXiv preprint arXiv:1804.05805.

Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M. and Padoy, N. 2016. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging. 36(1), pp.86–97.

Xiao, C., Li, B., Zhu, J.-Y., He, W., Liu, M. and Song, D. 2018. Generating adversarial examples with adversarial networks. arXiv preprint arXiv:1801.02610.