Radial Softmax: A Novel Activation Function for Neural Networks to Reduce Overconfidence in Out-Of-Distribution Data

Rain Vagel, Institute of Computer Science, University of Tartu

Introduction

Neural networks are used widely and give state-of-the-art results in fields such as image classification and machine translation. Wider adaption has drawn attention to their inability to distinguish out-of-distribution (OOD) data from in-distribution data[1]. This leads to high confidence predictions for data that the model has never seen. We created a modified version of softmax to reduce areas of confidence and distinguish OOD samples. Figures 1a to 1d visualise a two layer softmax model unable to distinguish OOD samples and how one class will dominate for OOD data.

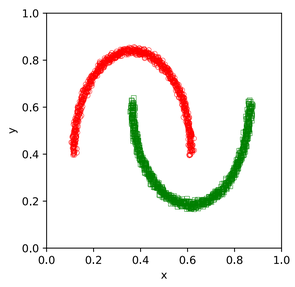

Figure 1a. Two moons dataset with two classes

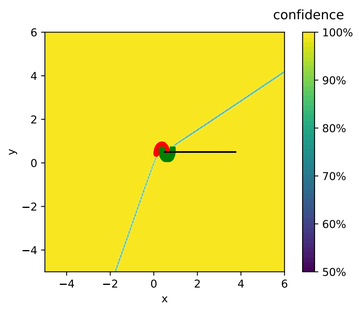

Figure 1b. Two moons area of confidence

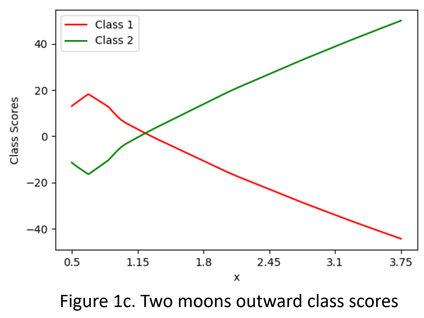

Figure 1c. Two moons outward class scores

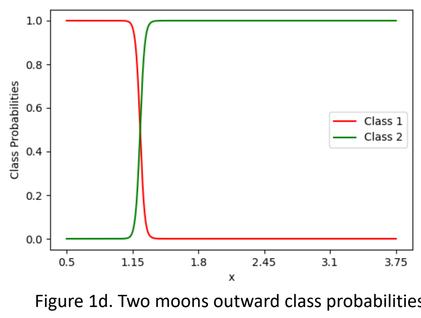

Figure 1d. Two moons outward class probabilities

Figure 1. Two moons softmax model confidence areas, class scores and probabilities visualisations after 400 training epochs

We propose a novel activation function that prioritises samples close to the training data, is able to learn the maximum confidences for each class and learn the approximate class distribution which it will predict for OOD samples.

Radial Softmax

Radial softmax is a novel activation function based on the regular softmax in Equation 1.

$$s(\boldsymbol{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (1)$$

The radial softmax defined in Equation 2 has the following features:

1. Class score $-|z_i|$ gives higher probabilities to samples close to the training set.
2. Parameter c_i specifies the maximum confidence for each class.
3. Parameter b_i specifies the approximate class distribution to predict for OOD samples.

$$s(\boldsymbol{z})_i = \frac{e^{c_i - |z_i|} + e^{b_i}}{\sum_{j=1}^{K} (e^{c_j - |z_j|} + e^{b_j})} \quad (2)$$

Methods

We demonstrate the effects of radial softmax in comparison to regular softmax on an eight moons dataset based on the two moons dataset. We show the areas of confidence, class scores and probabilities for OOD samples picked from around the training set.

To see how it performs for real-life data, we train LeNet [2] on MNIST and Small Resnet [3] on SVHN, CIFAR-10 and CIFAR-100 datasets each for 100 iterations and five times. In Table 1 we have brought out which datasets are used for OOD data for each of the training sets.

Table 1. Matrix of datasets used to sample OOD data from for each in-distribution dataset

	Testing Dataset							
Trained on	MNIST	EMNIST	FashionMNIST	SVHN	CIFAR-10	CIFAR-10 Grayscale	CIFAR-100	LSUN Classroom
MNIST	X	X	X			X		
SVHN				X	X		X	X
CIFAR-10				X	X		X	X
CIFAR-100				X	X		X	X

We have used the following metrics for the real-life datasets:

1. Test Error (TE) – Defined as 1 - accuracy.
2. Mean Maximum Confidence (MMC) – The mean of maximum confidence values for each data point scored by the model.
3. Area Under the Receiver Operating Characteristics (AUROC) - Represents a measure of separability as in how good the model is in predicting the correct class.
4. False positive rate (FPR) at 95% true positive rate (TPR) (FPR@95) - Defined as the FPR when the TPR is at 95%.

For radial softmax to be successful it needs to have similar TE and MMC for in-distribution data. Lower MMC, higher AUROC and lower FPR@95 on OOD samples.

Results

Radial softmax was successful in bringing the area of high confidence closer to the training dataset in Figures 2a and 2b. It was also successful on predicting approximately class distribution for OOD samples in Figure 2c, that were gathered from the black box in Figure 2b.

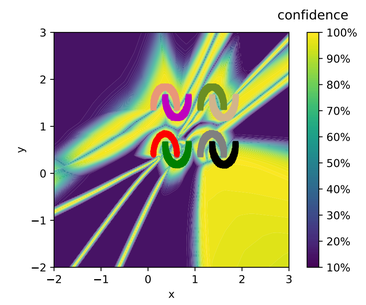

Figure 2a. Eight moons confidence area

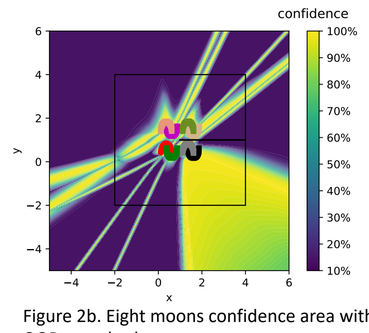

Figure 2b. Eight moons confidence area with OOD samples box

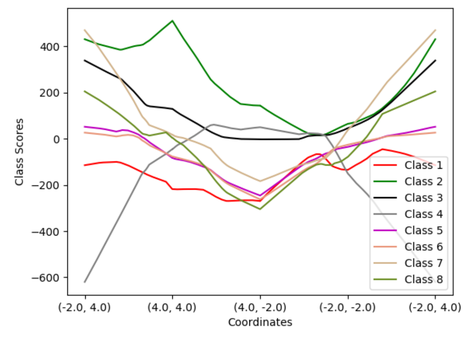

Figure 2c. Class scores for eight moons

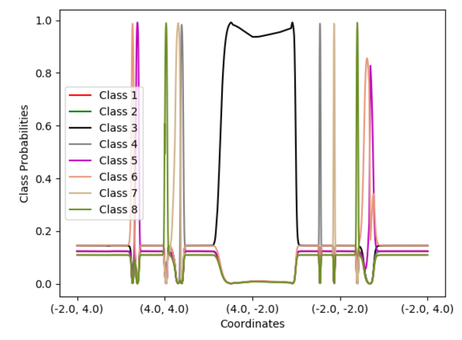

Figure 2d. Class probabilities for eight moons

Figure 2. Eight moons radial softmax model confidence areas, class scores and probabilities visualisations after 400 training epochs

In Table 2 we have presented the results of running the LeNet and Small Resnet models on real-life datasets. Bold represents metrics where radial softmax outperformed regular softmax.

Table 2. LeNet and Small Resnet models trained on real-life datasets with softmax and regular softmax. Bold represents comparison figures where radial softmax outperformed regular softmax.

Trained on **MNIST**	Softmax (TE: 0.56%, TL: 0.068) MMC	AUROC	FPR@95	Radial Softmax (TE: **0.55%**, TL: 0.071) MMC	AUROC	FPR@95
MNIST	0.99	-	-	0.989	-	-
EMNIST	0.818	0.889	0.351	0.82	**0.89**	0.358
FMNIST	0.69	0.968	0.142	**0.68**	**0.975**	**0.13**
CIFAR-10Grayscale	0.494	0.995	0.006	**0.492**	0.995	**0.001**
Trained on **SVHN**	Softmax (TE: 3.57%, TL: 0.23) MMC	AUROC	FPR@95	Radial Softmax (TE: **3.48%**, **TL: 0.21**) MMC	AUROC	FPR@95
SVHN	0.978	-	-	0.975	-	-
CIFAR-10	0.722	0.939	0.339	**0.665**	**0.945**	**0.278**
CIFAR-100	0.723	0.935	0.342	**0.664**	**0.943**	**0.281**
LSUNClassroom	0.728	0.938	0.348	**0.663**	**0.951**	**0.271**
Trained on **CIFAR-10**	Softmax (TE: 8.48%, TL: 0.42) MMC	AUROC	FPR@95	Radial Softmax (TE: 8.77%, TL: 0.43) MMC	AUROC	FPR@95
CIFAR-10	0.949	-	-	0.941	-	-
SVHN	0.743	0.887	0.683	0.75	0.869	0.732
CIFAR-100	0.767	0.885	0.712	**0.746**	0.845	**0.709**
LSUNClassroom	0.735	0.878	0.674	**0.687**	**0.888**	**0.61**
Trained on **CIFAR-100**	Softmax (TE: 31.89%, TL: 1.52) MMC	AUROC	FPR@95	Radial Softmax (TE: 32.67%, TL: 1.54) MMC	AUROC	FPR@95
CIFAR-100	0.748	-	-	0.701	-	-
SVHN	0.571	0.701	0.837	**0.455**	**0.755**	**0.826**
CIFAR-10	0.562	0.715	0.852	**0.5**	0.713	0.854
LSUNClassroom	0.549	0.725	0.849	**0.477**	**0.734**	**0.848**

We see that radial softmax outperformed regular softmax most of the times. As such radial softmax is suited for easy wins in distinguishing OOD samples. Parameter tuning can be used to further find confidence and distribution parameters for better results.

Conclusion

In this work we present a novel activation function called radial softmax that can be used in neural networks to distinguish OOD samples from in-distribution data. The function offers improvements on both toy datasets and real-life datasets, while being easy to use and not requiring custom training cycles or major changes to existing architectures.

References

[1] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. arXiv:1812.05720 [cs, stat], May 2019.

[2] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov./1998.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv:1706.06083 [cs, stat]*, September 2019

Acknowledgements

The author would like to thank Meelis Kull, Ardi Tampuu and Raul Vicente for their help during the project. The authors' studies are supported by the IT Academy programme.

Additional Information

Source: https://github.com/RainVagel/confidence-thesis/tree/master

UNIVERSITY OF TARTU
Institute of Computer Science
1632

IT Akadeemia