Project proposal - Latent-Mixup : mixing latent variables to train robust classifiers

1 Motivation

Current Deep Neural Network can be very accurate on the training set and still miss-classify with high confidence samples that slightly differ from the samples seen during training [Ben+10], are underrepresented in the training set [Has+18], or are adversarial [Sze+13]. Recently, the MIXUP method has been proposed [Zha+17] and empirical evidence showed that classifiers trained with that method were more robust compared to models trained the standard way. The idea of MIXUP is to augment the dataset with new images that are convex combinations of two images from the dataset. The label associated with the new image is the convex combination of the labels of the two original images with the same mixing factor. Generalizing that idea lead to MANIFOLD MIXUP where the hidden representations (activation of intermediate layers) inside a neural network are mixed instead of the raw samples.

Building on that idea our aim is to mix the latent codes of two samples that we got using a generative model such as a GAN [Goo+14] in order to generate new samples, the label associated with the new sample is again the convex combination of the original labels. More precisely we randomly sample two images of the training dataset and find the latent vectors for which the GAN's generator can reasonably reconstruct the initial images. We then mix the latent codes using a convex combination with some $\lambda \in [0, 1]$, and use the GAN's generator to build an new image based on that mixed latent code.

2 Scope of the project

The goal of the project is to implement the idea described above. We will train or download GANs and find the latent vectors corresponding to samples using gradient descent with the visual features of the sample as the target, similar to the way it was done in [LT17] or [Ins19]. Another method to get the latent code faster is to initialize the search using a pre-trained model as described in [Ins19]. Using the newly generated latent vector we generate a new image using the generator of the GAN and assign it the mixed label. We then train a classifier using both the original dataset and the mixed samples with their corresponding labels.

In order to evaluate our method we will compare our training with : (i) the standard training (ii) MIXUP as in [Zha+17] (iii) MANIFOLD MIXUP as in [Ver+18] (iv) adversarial training techniques as presented in [Bai+21] and [LHL17].

We will evaluate the classifiers trained with the different techniques based on their robustness to simple operations such as blurring as well as on the robustness scores given by some more advanced tools such as DEEPFOOL [MFF16]. We will also test how the methods affect the accuracy of the classifier.

For practical reason we will train our model on the MNIST, Fashion-MNIST and CIFAR-10 datasets as these are relatively small.

If time permits we could also evaluate the training procedure using VAEs [KW13] instead of GANs to see how the nature of the generative model affects the robustness of the trained classifier.

References

[Ben+10] Shai Ben-David et al. "A theory of learning from different domains". In: *Machine learning* 79.1 (2010), pp. 151–175.

[KW13] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2013. DOI: 10.48550/ARXIV.1312.6114. URL: https://arxiv.org/abs/1312.6114.

[Sze+13] Christian Szegedy et al. *Intriguing properties of neural networks*. 2013. DOI: 10.48550/ARXIV.1312.6199. URL: https://arxiv.org/abs/1312.6199.

[Goo+14] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. DOI: 10.48550/ARXIV.1406.2661. URL: https://arxiv.org/abs/1406.2661.

[MFF16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[LHL17] Hyeungill Lee, Sungyeob Han, and Jungwoo Lee. *Generative Adversarial Trainer: Defense to Adversarial Perturbations with GAN*. 2017. DOI: 10.48550/ARXIV.1705.03387. URL: https://arxiv.org/abs/1705.03387.

[LT17] Zachary C. Lipton and Subarna Tripathi. *Precise Recovery of Latent Vectors from Generative Adversarial Networks*. 2017. DOI: 10.48550/ARXIV.1702.04782. URL: https://arxiv.org/abs/1702.04782.

[Zha+17] Hongyi Zhang et al. *mixup: Beyond Empirical Risk Minimization*. 2017. DOI: 10.48550/ARXIV.1710.09412. URL: https://arxiv.org/abs/1710.09412.

[Has+18] Tatsunori Hashimoto et al. "Fairness Without Demographics in Repeated Loss Minimization". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 1929–1938. URL: https://proceedings.mlr.press/v80/hashimoto18a.html.

[Ver+18] Vikas Verma et al. *Manifold Mixup: Better Representations by Interpolating Hidden States*. 2018. DOI: 10.48550/ARXIV.1806.05236. URL: https://arxiv.org/abs/1806.05236.

[Ins19] Arxiv Insights. *Editing Faces using Artificial Intelligence*. Youtube. 2019. URL: https://youtu.be/dCKbRCUyop8?t=938.

[Bai+21] Tao Bai et al. *Recent Advances in Adversarial Training for Adversarial Robustness*. 2021. DOI: 10.48550/ARXIV.2102.01356. URL: https://arxiv.org/abs/2102.01356.