Active Supervision

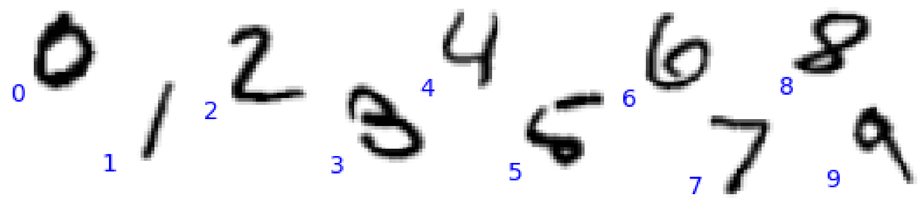
Ricardo Cruz • ASM Shihavuddin • Anders Dahl • Jaime S. Cardoso

This is a work in progress.

Abstract: The proposal is for the model to produce data near the decision boundary, allowing the user to label such data in order to indirectly nudge the decision boundary by incorporating the new data into the learning process. This is expected to improve the performance of the model.

Motivation

In **supervised learning,** the user labels each image prior to training the model.



Then, the user waits until the model is done training.

Sometimes **active learning** is used to collect and label more images of the harder cases.

Proposal

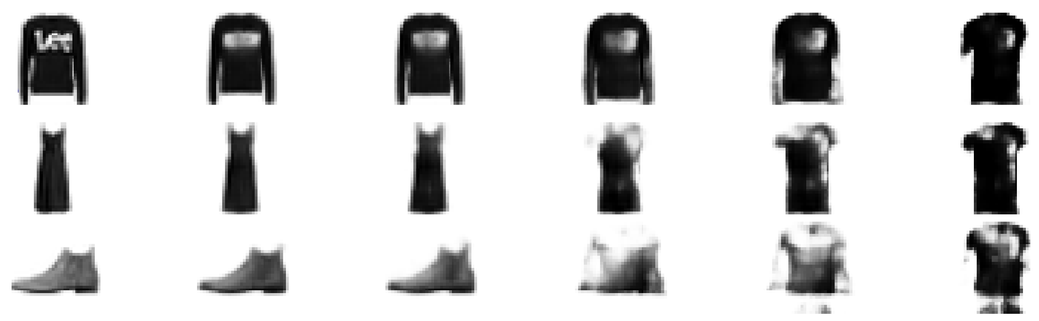
The distance of the decision boundary to the data has been found to be a major predictor of overfitting [1].

Active supervision: allow the user to adjust the decision boundary indirectly by pushing or pulling the phase transition after which the data is recognized as being of another class.

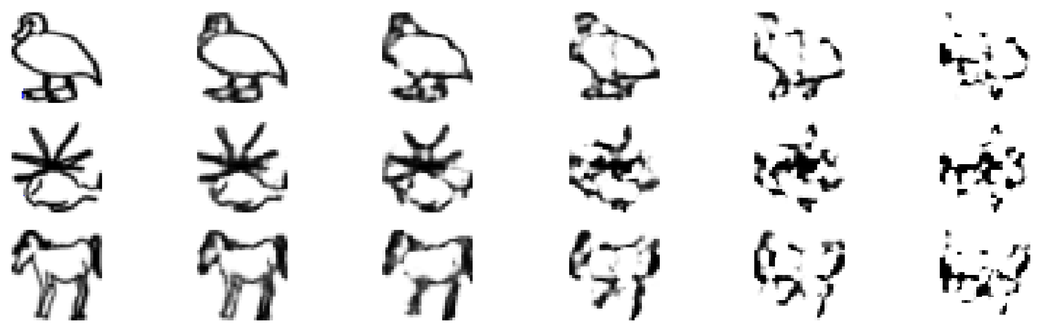
Preliminary Results

MNIST: transforming digits into 0

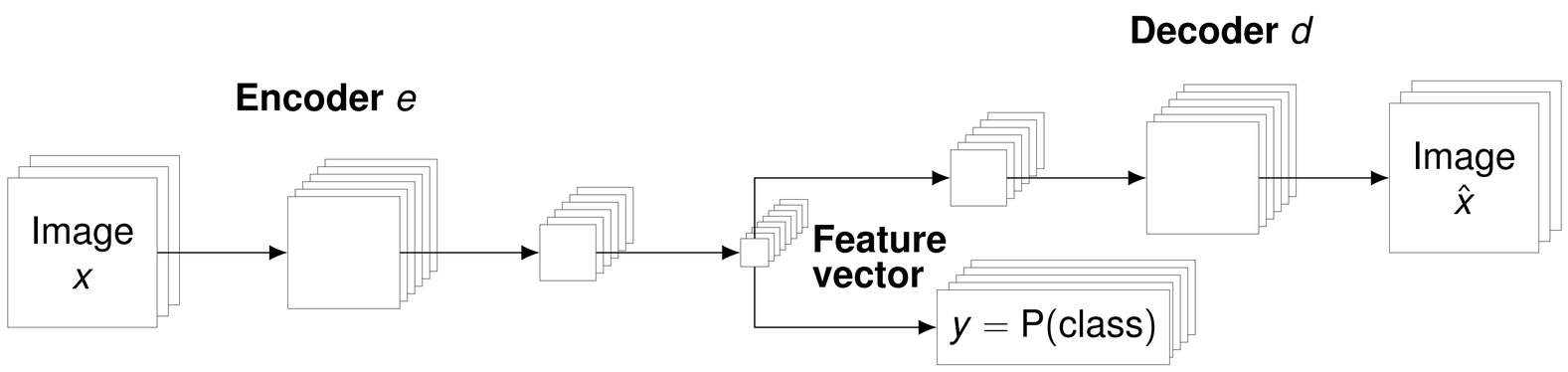
Fashion MNIST: transforming into a t-shirt



Sketches: transforming into an airplane



Method



The encoder is trained for the supervised task. The decoder is trained as an identity function.
$\mathcal{L}_{\text{encoder}} = \text{ce}(y, \hat{y})$
$\mathcal{L}_{\text{decoder}} = \text{ce}(x, \hat{x})$

While the model is trained, the feature vector of a random image (Step 1) is nudged in the direction of the decision boundary (Step 2) and a new image is produced (Step 3).

- Step 1. $f = e(x)$
- Step 2. $f' = f + \sigma \operatorname{sign}\left(\frac{\partial y}{\partial f}\right)$ [2]
- Step 3. $\hat{x} = d(f')$

Other methods could be used to instantiate the idea here proposed. For example, adversarial generation coupled with adversarial training [3].

References

[1] Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions. *arXiv preprint arXiv:1810.00113*, 2018.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[3] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *arXiv preprint arXiv:1904.12843*, 2019.