Anti-Distillation: Knowledge Transfer from a Simple Model to the Complex One

Kseniia Petrushina

Moscow Institute of Physics and Technology

Coauthors: Oleg Bakhteev, Andrey Grabovoy, Vadim Strijov

2022

Motivation

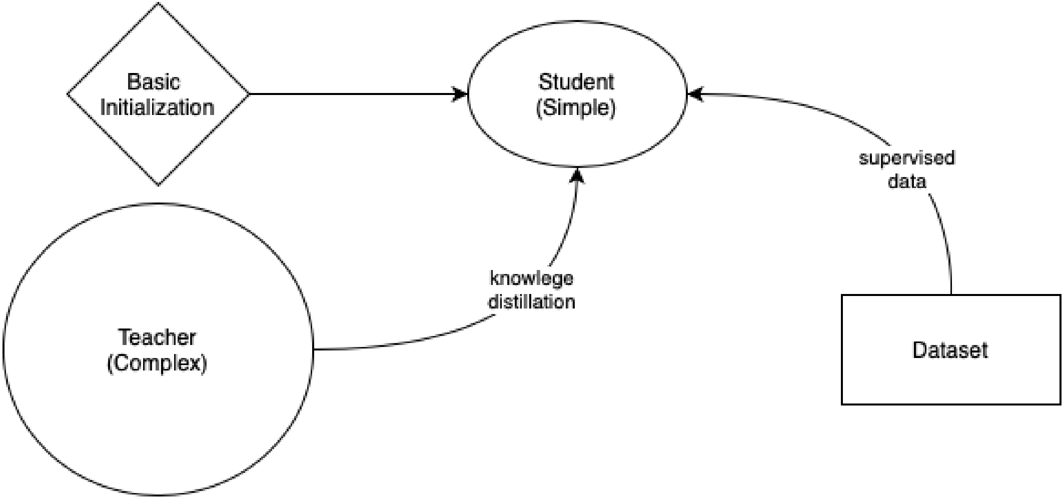
Goal: Adapting the *student* model to more complex data using information from *teacher* model.

Challenge: Existing *teacher* model can be not suitable for new data.

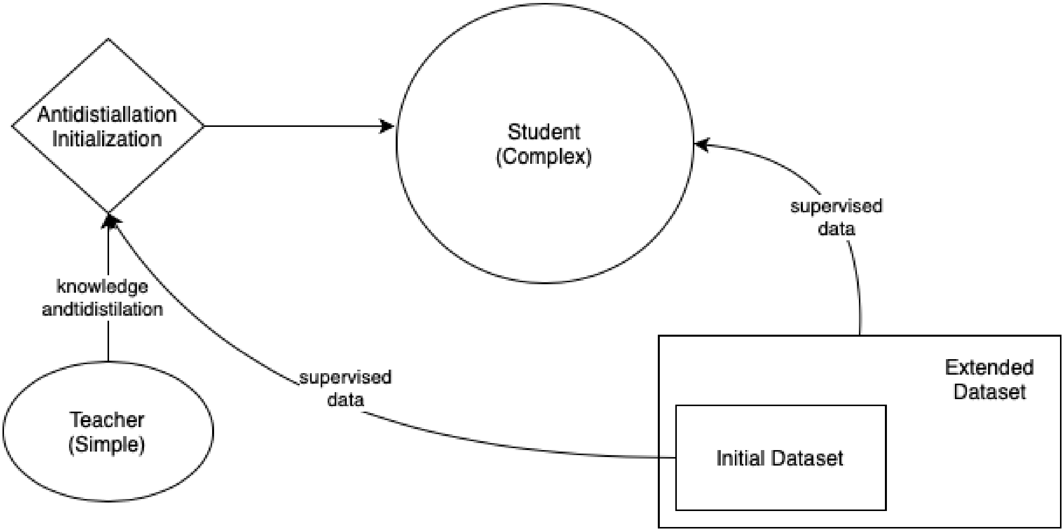
Solution: Extend the *teacher* model to get better initialization of the new model (*Anti-Distillation*[1]).

[1] Distillation - process of transferring knowledge from a complex model to the simple model

Distillation



Anti-Distillation



Problem statement

Consider two datasets

$$\mathfrak{D}_1 = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_1},\ \mathbf{x}_i \in \mathbb{R}^n,\ y_i \in C_1 = \{1, \dots, c_1\},$$

$$\mathfrak{D}_2 = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_2},\ \mathbf{x}_i \in \mathbb{R}^n,\ y_i \in C_2 = \{1, \dots, c_2\}.$$

$\mathfrak{D}_2$ is more complex than $\mathfrak{D}_1$.

Optimal parameters $\hat{\mathbf{u}}$ of the teacher model g on $\mathfrak{D}_1$ dataset are obtained from

$$\hat{\mathbf{u}} = \arg\min_{\mathbf{u}} \mathcal{L}_{\text{ce}}(\mathbf{u}, \mathfrak{D}_1),$$

$\mathcal{L}_{\text{ce}}(\mathbf{u}, \mathfrak{D}_1)$ - cross-entropy loss on $\mathfrak{D}_1$.

Problem statement

Student model is

$$\mathbf{f}_{\text{st}} : \mathbb{R}^n \to \Delta^{c_2}, \quad \mathbf{f}_{\text{st}}(\mathbf{x}) = \mathbf{f}(\mathbf{x}, \hat{\mathbf{w}}),$$

Optimization problem:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathcal{L}_{\text{ce}}(\mathbf{w}, \mathfrak{D}_2^{\text{val}}),$$

Function

$$\varphi : \mathbb{R}^{N_{\text{tr}}} \to \mathbb{R}^{N_{\text{st}}}$$

maps the teacher model parameters to student initial parameters $\mathbf{w} = \varphi(\hat{\mathbf{u}})$.

Hypothesis

The student models initialized by the result of applying the function φ to the parameters of the pre-trained teacher model is more persistent and achieve higher accuracy than models with default parameters.

Problem solution

Function for weights initialization:

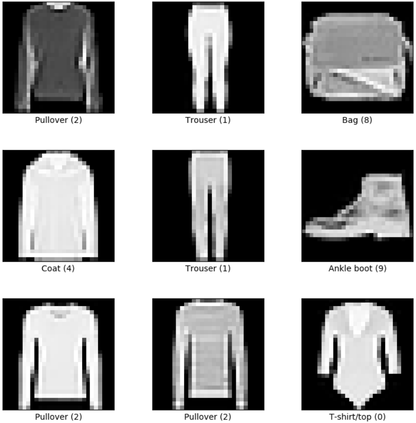
$$\varphi(\mathbf{u}) = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}),$$

$$\mathcal{L}(\mathbf{w}) = \lambda_1 \mathcal{L}_{\text{ce}}(\mathbf{w}, \mathfrak{D}_1) + \lambda_2 \mathcal{L}_2(\mathbf{w}, \mathbf{u}) + \lambda_3 \mathcal{L}_3^{\delta}(\mathbf{w}, \mathfrak{D}_1) + \lambda_4 \mathcal{L}_4 \left(\frac{\partial^2 \mathcal{L}_{\text{ce}}}{\partial \mathbf{w}^2} \right)$$

- $\mathcal{L}_2(\mathbf{w}, \mathbf{u}) = \|\mathbf{u} - \mathbf{Pr}[\mathbf{w}]\|_2^2$ – small difference between common weights in student and teacher.
- $\mathcal{L}_3^{\delta}(\mathbf{w}, \mathfrak{D}_1) = \sum\limits_{(\mathbf{x}, y) \in \mathfrak{D}_1} \mathbb{E}_{\mathbf{x}' \in U_{\delta}(\mathbf{x})} \mathcal{L}_{\text{ce}}(\mathbf{w}, (\mathbf{x}', y))$ – robustness to noise with respect to input data.
- $\mathcal{L}_4 \left(\frac{\partial^2 \mathcal{L}_{\text{ce}}}{\partial \mathbf{w}^2} \right) = \text{tr} \left(\frac{\partial^2 \mathcal{L}_{\text{ce}}}{\partial \mathbf{w}^2} \right)$– robustness to noise with respect to model parameters.

Computational experiments

Data: Fashion-MNIST dataset



Model: Multilayer perceptron

Baselines:

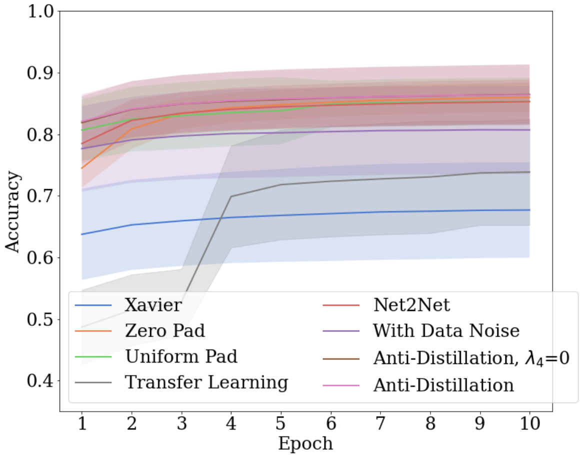
- Xavier initialization
- Transfer Learning
- Net2Net

Quality criteria

- Accuracy on validation set.
- Accuracy on validation set corrupted by FSGM-attack.
- Accuracy on validation set, provided that the model parameters are corrupted with noise: $\mathbf{w}_\varepsilon = \mathbf{w} + \varepsilon \boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Hyperparameter optimization

Accuracy on validation set

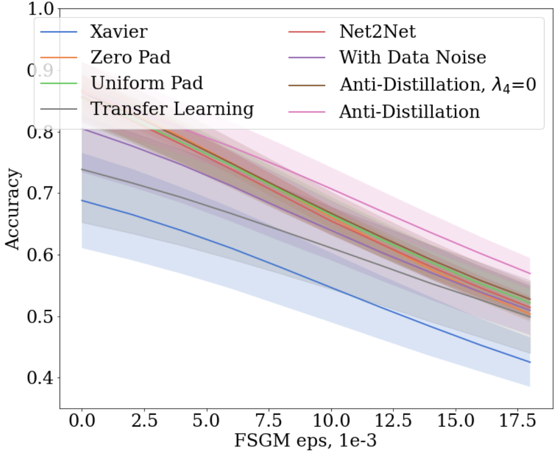


Anti-Distillation outperformes other methods.

Xavier	Zero	Uniform	Transfer
0.68	**0.86**	0.85	0.74

Net2Net	Noise	AD, $\lambda_4=0$	AD
0.85	0.81	**0.86**	**0.86**

Robustness to FSGM-attack

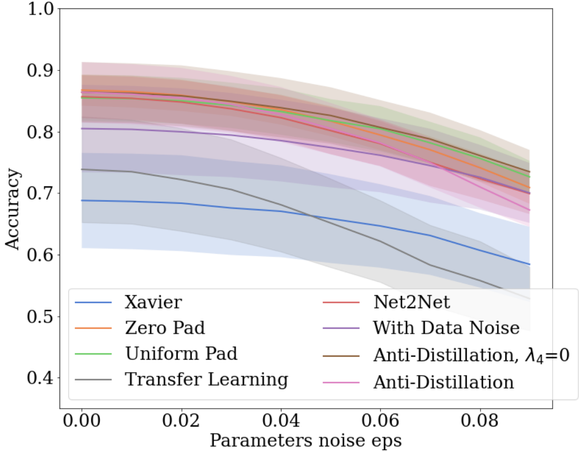


Anti-Distillation outperformes other methods.

Xavier	Zero	Uniform	Transfer
0.42	0.50	0.52	0.50

Net2Net	Noise	AD, $\lambda_4=0$	AD
0.51	0.51	0.53	**0.57**

Robustness to noise in model parameters



Anti-Distillation outperformes other methods.

Xavier	Zero	Uniform	Transfer
0.58	0.71	**0.73**	0.53

Net2Net	Noise	AD, $\lambda_4=0$	AD
0.70	0.70	**0.73**	0.67

Conclusion

We:

- considered the problem of the model extension to a new dataset.
- proposed the method for knowledge transfer from a simple model to a more complex model.

Anti-Distillation:

- achieved higher accuracy on *more complex* dataset.
- showed robustness of the Anti-Distillation to adversarial noise in input data and normal noise in model parameters.

Next: other datasets, other neural network architectures.

Literature

- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.
- Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feed- forward neural networks, 2010.
- Tianqi Chen, Ian Goodfellow, Jonathon Shlens. Net2Net: Accelerating Learning via Knowledge Transfer, 2015.