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

Kseniia Petrushina

Moscow Institute of Physics and Technology

Expert: Vadim Strijov
Consultant: Andrey Grabovoy

2022

Research goal

Goal: Adapting the model to more complex data.

Compare uniform initialization with one based on a previously trained model by differences in convergence rate, prediction variance, achieved quality, and stability of the model.

Literature

- Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feed- forward neural networks, 2010.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.
- David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information, 2016.

Problem statement

Hypothesis

Student models initialized by the result of applying the function φ to the weights of the pre-trained teacher model are more persistent and achieve higher accuracy than models with default weights.

$D_2 = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_1}$ is *more complex* than $D_1 = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_2}$

Optimal parameters $\hat{\mathbf{u}}$ of the teacher model g on D_1 dataset are obtained from

$$\hat{\mathbf{u}} = \arg\min_{\mathbf{u}} \mathcal{L}_g(\mathbf{u}, D_1),$$

$\mathcal{L}_g(\mathbf{u}, D_1)$ - cross-entropy loss on D_1.

Problem statement

Initialize the student model f weights as $w_1 = \varphi(\hat{\mathbf{u}})$.

Quality criterions:

- The value of the loss function on corrupted data.
- Accuracy of predictions on D_2.

Problem solution

Function for weights initialization:

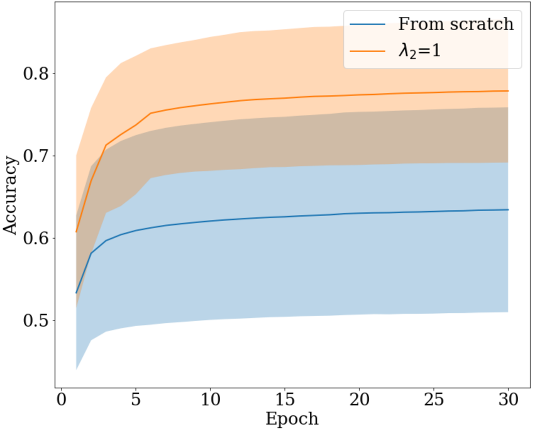
$$\varphi(\mathbf{u}) = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}),$$

$$\mathcal{L}(\mathbf{w}) = \lambda_1 \mathcal{L}_f(\mathbf{w}, D_1^*) + \lambda_2 \mathcal{L}_2(\mathbf{w}, \mathbf{u}) + \lambda_3 \mathcal{L}_3^\delta(\mathbf{w}, D_1^*) + \lambda_4 \mathcal{L}_4\left(\frac{\partial^2 \mathcal{L}_f}{\partial \mathbf{w}^2}\right).$$

- $\mathcal{L}_2(\mathbf{w}, \mathbf{u}) = \|\mathbf{u} - \mathbf{w}[\mathbf{u}]\|_2^2$
- $\mathcal{L}_3^\delta(\mathbf{w}, D_1^*) = \sum\limits_{(\mathbf{x}, y) \in D_1^*} \mathbb{E}_{\mathbf{x}' \in U_\delta(\mathbf{x})} \mathcal{L}_f(\mathbf{w}, \mathbf{x}', y)$
- $\mathcal{L}_4\left(\frac{\partial^2 \mathcal{L}_f}{\partial \mathbf{w}^2}\right) = \|\left(\frac{\partial^2 \mathcal{L}_f}{\partial \mathbf{w}^2}\right)\|_2^2$

Computational experiment

Goal: Compare the performance of models depending on the initialization of parameter



$$\mathbf{w}_1 = \varphi(\mathbf{u}^*)$$

$$\mathbf{w}_1 = (\mathbf{u}^*, \mathbf{w}_1'),$$

$$\mathbf{w}_1' \sim U[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$$

Conclusion

A new method for parameters initialization of fully-connected neural newtorks, which helps to achieve higher accuracy on *more complex* dataset.

Next: CNN and RNN, different default parameters initialization.