Linear Probing of Deep Neural Networks with Minimax Risk Classifiers

Eugenio Cainelli[1], Santiago Mazuelas[1,2]

[1] BCAM - Basque Center for Applied Mathematics, [2] IKERBASQUE - Basque Foundation for Science

ecainelli@bcamath.org, smazuelas@bcamath.org

Abstract

Linear probing is one of the principal strategies in deep transfer learning, where a linear classifier is trained on top of the frozen layers of a pretrained neural network to adapt it to new tasks. In this work, we introduce a novel linear probing approach by employing minimax risk classifiers (MRCs), linear classifiers coming from the setting of robust risk minimization. Our method provides tight performance bounds while enabling robust and efficient learning. Additionally, since linear probing is often applied in settings with limited sample size, traditional model selection methods like cross-validation may lose reliability. Our method introduces an alternative validation procedure, leveraging the upper bounds provided by the MRCs during learning. This approach has been shown to be significantly faster and to yield similar outcomes compared to conventional techniques.

Problem setup

Deep Transfer Learning

Few samples from downstream task $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n) \sim \mathrm{p}^*$

Neural Network pre-trained on large dataset $(\tilde{x}_1, \tilde{y}_1), (\tilde{x}_2, \tilde{y}_2), \cdots, (\tilde{x}_N, \tilde{y}_N) \sim \tilde{\mathrm{p}}^*$

Fine-tune the general representation to obtain a classifier for downstream task

Key assumptions

$$N > n$$

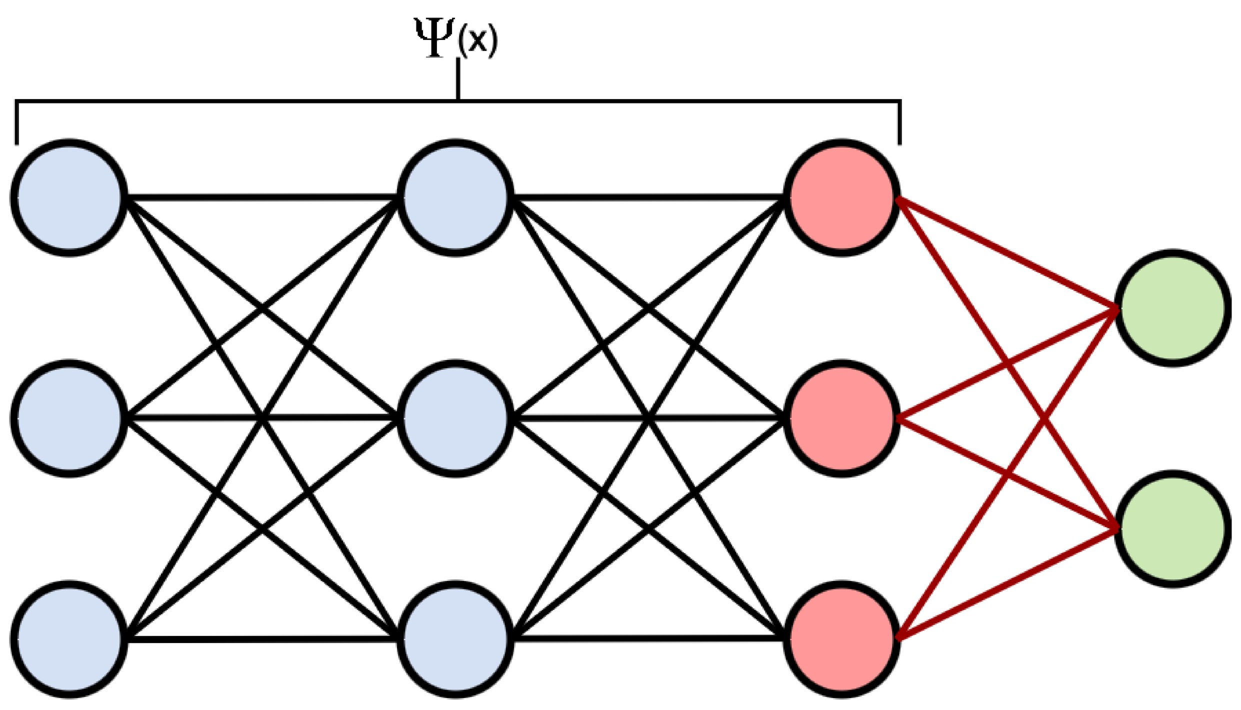
$$\tilde{\mathrm{p}}^*(x, y) \approx \mathrm{p}^*(x, y)$$

Linear Probing

Fix the pre-trained NN up to the penultimate layer

Obtain the function $\Psi(x)$: activations of the penultimate layer

Train a linear classifier using $\Psi(x)$ on the downstream tasks data



Minimax Risk Classifiers

Feature mapping $\Phi : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}^m$ with $\Phi(x, y) = \boldsymbol{e}_y \otimes \Psi(x)$

Uncertainty set $\mathcal{U} = \{\, \mathrm{p} \in \Delta(\mathcal{X} \times \mathcal{Y}) : |\mathbb{E}_{\mathrm{p}}[\Phi(x, y)] - \boldsymbol{\tau}| \preceq \boldsymbol{\lambda} \text{ and } \mathrm{p}(x) = \mathrm{p}^*(x)\}$

Mean vector $\boldsymbol{\tau} = \frac{1}{n}\sum_{i=1}^{n} \Phi(x_i, y_i)$

Confidence vector $\boldsymbol{\lambda} = \lambda_0 \sqrt{\frac{\boldsymbol{Var}}{n}}$

Minimax risk classifier (MRC) is the solution of the minimax problem:

$$\mathcal{P}_{MRC} : \min_{h \in T(\mathcal{X}, \mathcal{Y})} \max_{\mathrm{p} \in \mathcal{U}} R(h, p)$$

where:

- $R(h, \mathrm{p})$ is the risk of classifier h over distribution p with $0-1$ loss
- $T(\mathcal{X}, \mathcal{Y})$ is the set of all possible classifiers

Theorem 1. *If $\mathcal{U}$ is non-empty and satisfies standard regularity conditions, then:*

$$h^{\mathcal{U}}_{MRC}(y|x) = \left(\Phi(x, y)^t \boldsymbol{\mu}^* - \varphi(x, \boldsymbol{\mu}^*)\right)_+$$

Where:

- *$\boldsymbol{\mu}^*$ is solution of the convex non-smooth optimization problem*

$$\min_{\boldsymbol{\mu}} 1 - \tau^t \boldsymbol{\mu} + \lambda^t |\boldsymbol{\mu}| + \mathbb{E}_{\mathrm{p}^*(x)}[\varphi(x, \boldsymbol{\mu})] \tag{1}$$

- $$\varphi(x, \boldsymbol{\mu}) = \max_{\mathcal{C} \subseteq Y} \frac{\Sigma_{y \in \mathcal{C}} \Phi(x, y)^t \boldsymbol{\mu} - 1}{|\mathcal{C}|}$$

Since the optimization problem (1) is the Lagrange dual of the minimax problem and strong duality holds, the minimax risk of $\mathcal{U}$ is $R(\mathcal{U}) = 1 - \tau^t \boldsymbol{\mu}^* + \lambda^t |\boldsymbol{\mu}^*| + \mathbb{E}_{\mathrm{p}^*(x)}[\varphi(x, \boldsymbol{\mu}^*)]$

Since p^* is unkown, also $\mathbb{E}_{\mathrm{p}^*(x)}[\varphi(x, \boldsymbol{\mu}^*)]$ is unknow. Hence, the optimization problem (1) is solved with the stochastic subgradient method

Theorem 2. *Bound on stochastic subgradient method's excess population risk, with probability at least $1 - \delta$:*

$$\epsilon_{SSM} \leq O\left(\frac{\log n \log(n/\delta)}{\sqrt{n}}\right)$$

Generalization bounds

Theorem 3. *If $\mathcal{U}$ is non-empty and satisfies standard regularity conditions, then:*

$$R(h^{\mathcal{U}}_{MRC}, \mathrm{p}^*) \leq R(\mathcal{U}) + (|\mathbb{E}_{\mathrm{p}^*}[\Phi(x, y)] - \boldsymbol{\tau}| - \boldsymbol{\lambda})^t |\boldsymbol{\mu}^*| + \epsilon_{SSM}$$

Where:

- $R(\mathcal{U}) = \min_{h \in T(X, Y)} \max_{\mathrm{p} \in \mathcal{U}} R(h, p)$
- *ϵ_{SSM} upper bound on Stochastic Subgradient Method*

In particular, if $\mathbb{P}\left(|\mathbb{E}_{\mathrm{p}^*}[\Phi] - \boldsymbol{\tau}| \leq \boldsymbol{\lambda}\right) \geq 1 - \delta$, then w.p. at least $1 - \delta$:

$$R(h^{\mathcal{U}}_{MRC}, \mathrm{p}^*) \leq R(\mathcal{U})$$

The inequality holds also for $\boldsymbol{\lambda} \approx |\mathbb{E}_{\mathrm{p}^*}[\Phi(x, y)] - \boldsymbol{\tau}|$

Theorem 4. *For any classifier $h \in T(\mathcal{X}, \mathcal{Y})$, given*

$$\overline{R}(\mathcal{U}, h) = \min_{\boldsymbol{\mu}} 1 - \boldsymbol{\tau}^t \boldsymbol{\mu} + \boldsymbol{\lambda}^t |\boldsymbol{\mu}| + \mathbb{E}_{\mathrm{p}^*(x)}\left[\Phi(x, y)^t \boldsymbol{\mu} - h(y|x)\right]$$

$$\underline{R}(\mathcal{U}, h) = \max_{\boldsymbol{\mu}} 1 - \boldsymbol{\tau}^t \boldsymbol{\mu} - \boldsymbol{\lambda}^t |\boldsymbol{\mu}| + \mathbb{E}_{\mathrm{p}^*(x)}\left[\Phi(x, y)^t \boldsymbol{\mu} - h(y|x)\right]$$

then, for any $p \in \mathcal{U}$, these bounds hold:

$$\underline{R}(\mathcal{U}, h) \leq R(h, p) \leq \overline{R}(\mathcal{U}, h)$$

Model selection

When dealing with few samples, k-folds Cross-Validation:

- High variance, loses reliability
- Training k times for every model
- No performance estimate if not nested (even more computations needed for nested)

Our model selection method using MRC consists of choosing the pre-trained model that gives the lowest upper bound $R(\mathcal{U})$, this means:

- More robustness
- Only 1 training per model
- Tight performance guarantees

Experimental results

Metric	Acc	UB	Mean CV	Std CV
MRC with UB	0.834	0.108	0.810	0.058
MRC with 5-folds CV	0.818	0.130	0.870	0.087
SVM with 5-folds CV	0.832	No	0.880	0.087
LogReg with 5-folds CV	0.842	No	0.880	0.087

- Model selection over 10 different pretrained NNs for computer vision
- Dataset consisting of 100 samples of two classes of Fashion-MNIST dataset without class imbalance
- Model selection with upper bound (UB) is 5 times faster than CV on MRC

References

[1] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

[2] Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C Mozer. Head2toe: Utilizing intermediate representations for better transfer learning. In *International Conference on Machine Learning*, pages 6009–6033. PMLR, 2022.

[3] Santiago Mazuelas, Andrea Zanoni, and Aritz Pérez. Minimax classification with 0-1 loss and performance guarantees. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 302–312. Curran Associates, Inc., 2020.

Acknowledgments

The authors acknowledge the technical and human support provided by the DIPC Supercomputing Center