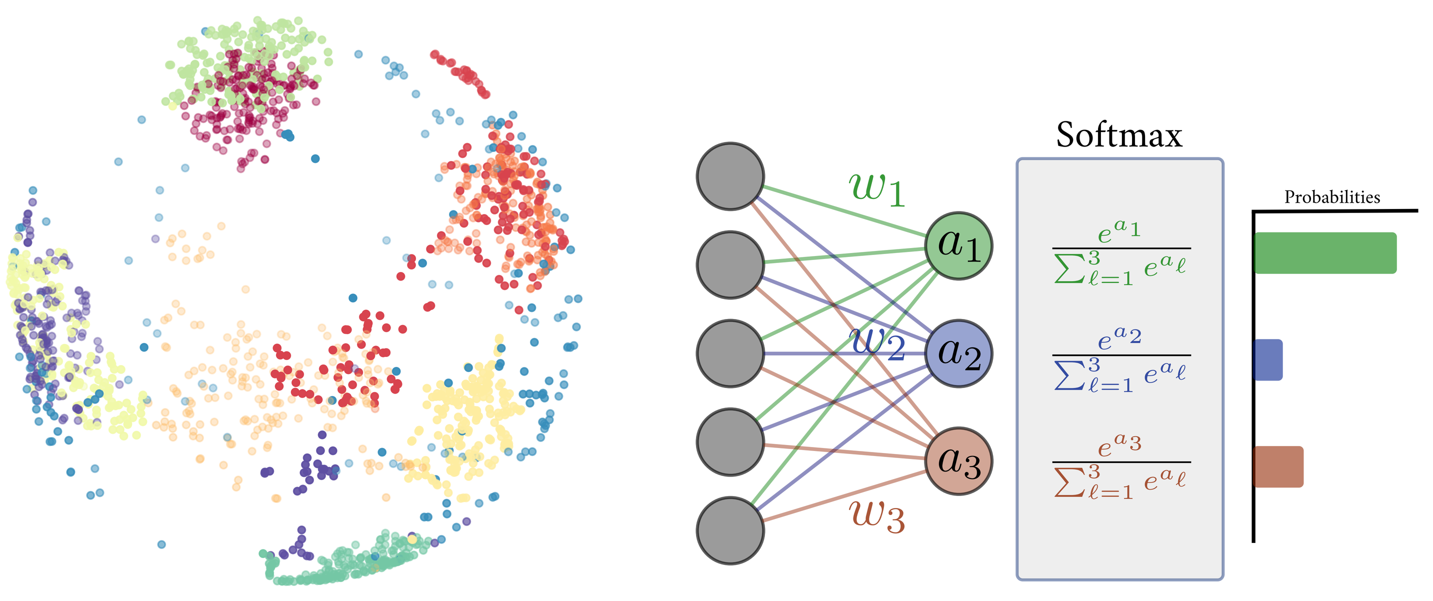
Neural Networks Classify through the Class-wise Means of their Representations

Mohamed El Amine Seddik [1] Mohamed Tamaazousti [2]

[1]Huawei Paris Research Center, France [2]University Paris-Saclay, CEA List, France

Abstract



$x_1, \ldots, x_n \in \mathbb{R}^p$

- **Theoretically:** Asymptotic analysis of the Softmax classifier suggests that weights correlate with the class-wise means of the input features.
- **Empirically:** The theoretical observation extends to the Softmax layer of feed-forward neural networks trained for classification tasks.
- Consequences for transfer learning: a simple initialization procedure of the Softmax weights is proposed based on the theoretical findings.

Settings

We consider n data points $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n$ with their corresponding labels $\ell_1, \ldots, \ell_n$ distributed in k different classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$. Denote

$$\boldsymbol{x} = \varphi \circ \phi(\boldsymbol{d}; \Theta) \in \mathbb{R}^p$$

where

- ϕ is implemented by a deep CNN model parameterized by Θ.
- φ is the activation function at the representation layer.

The final class prediction is given by a classifier function $\psi : \mathbb{R}^p \to \mathbb{R}^k$ as

$$\arg\max \psi(\boldsymbol{x}) \quad \text{with} \quad \psi(\boldsymbol{x}) = softmax(\boldsymbol{W}^\top \boldsymbol{x}).$$

Suppose that the following statistics exist and well defined

$$\boldsymbol{m}_\ell = \mathbb{E}_{\boldsymbol{d} \in \mathcal{C}_\ell}[\boldsymbol{x}] \quad \boldsymbol{C}_\ell = \mathbb{E}_{\boldsymbol{d} \in \mathcal{C}_\ell}[(\boldsymbol{x} - \boldsymbol{m}_\ell)(\boldsymbol{x} - \boldsymbol{m}_\ell)^\top]$$

Softmax Classifier

Minimize:

$$\mathcal{L}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k) = -\frac{1}{n} \sum_{i=1}^n \sum_{\ell=1}^k y_{i\ell} \log p_{i\ell}$$

$$p_{i\ell} = \frac{\exp(\boldsymbol{w}_\ell^\intercal \boldsymbol{x}_i)}{\sum_{j=1}^k \exp(\boldsymbol{w}_j^\intercal \boldsymbol{x}_i)}, \quad y_\ell^{(i)} = \alpha_{c(i)} \frac{|\delta_{\ell,c(i)} - \varepsilon|}{1 + (k-2)\varepsilon}$$

where $\alpha_{c(i)}$ and $\varepsilon > 0$ are hyper-parameters, $c(i)$ returns the class index of the i-th datum and $\delta_{i,j}$ stands for the Kronecker delta.

The classical labels are recovered by setting $\alpha_{c(i)} = 1$ and $\varepsilon = 0$.

Assumption on statistical model

For $\boldsymbol{x}_i \in \mathcal{C}_\ell$, assume $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{m}_\ell, \boldsymbol{C}_\ell)$.

Denote $\pi_\ell = \lim_n \frac{|\mathcal{C}_\ell|}{n}$ the proportion of class $\mathcal{C}_\ell$.

Expression of the gradients of the Softmax class-weight vectors

$$\nabla_{\boldsymbol{w}_\ell} \mathcal{L} = \frac{1}{n} \sum_{i=1}^n \alpha_{c(i)} \left(\frac{e^{\boldsymbol{w}_\ell^\intercal \boldsymbol{x}_i}}{\sum_{j=1}^k e^{\boldsymbol{w}_j^\intercal \boldsymbol{x}_i}} - \frac{|\delta_{\ell,c(i)} - \varepsilon|}{1 + (k-2)\varepsilon} \right) \boldsymbol{x}_i \equiv \frac{1}{n} \sum_{i=1}^n \alpha_{c(i)} f_{\ell,i}(\boldsymbol{w}_\ell^\intercal \boldsymbol{x}_i) \boldsymbol{x}_i.$$

Asymptotic Softmax class-weight vectors

Let $\bar{\boldsymbol{w}}_1, \ldots, \bar{\boldsymbol{w}}_k$ be the deterministic vectors satisfying $\mathbb{E}[\nabla_{\bar{\boldsymbol{w}}_\ell} \mathcal{L}] = 0$. Thus, each $\bar{\boldsymbol{w}}_\ell$ satisfies the implicit equation

$$\bar{\boldsymbol{w}}_\ell = \left(\sum_{j=1}^k \alpha_j \pi_j \mathbb{E}_j[f'_{\ell,i}(\bar{\boldsymbol{w}}_\ell^\intercal \boldsymbol{x}_i)] \boldsymbol{C}_j \right)^{-1} \left(\sum_{j=1}^k \alpha_j \pi_j \mathbb{E}_j[f_{\ell,i}(\bar{\boldsymbol{w}}_\ell^\intercal \boldsymbol{x}_i)] \boldsymbol{m}_j \right)$$

where the notation $\mathbb{E}_j[g(\boldsymbol{x}_i)] \equiv \mathbb{E}[g(\boldsymbol{x}_i) \,|\, \boldsymbol{x}_i \in \mathcal{C}_j]$ for some $g : \mathbb{R} \to \mathbb{R}$.

Setting the parameters $\alpha_j = (k\pi_j)^{-1}$ results in the class-weight vectors becoming independent of the proportion π_j:

$$\bar{\boldsymbol{w}}_\ell = \left(\sum_{j=1}^k \mathbb{E}_j[f'_{\ell,i}(\bar{\boldsymbol{w}}_\ell^\intercal \boldsymbol{x}_i)] \boldsymbol{C}_j \right)^{-1} \left(\sum_{j=1}^k \mathbb{E}_j[f_{\ell,i}(\bar{\boldsymbol{w}}_\ell^\intercal \boldsymbol{x}_i)] \boldsymbol{m}_j \right)$$

Near optimal representations (NOR)

Let $\epsilon > 0$, assume

- $\boldsymbol{m}_i^\intercal \boldsymbol{m}_j = \delta_{i,j}\, \mu_1 + (1 - \delta_{i,j})\, \mu_2$ with $\mu_1 = \mathcal{O}(1)$ and $\mu_2 = \mathcal{O}(p^{-\epsilon})$.
- $\boldsymbol{C}_j = \sigma_{1,j}^2 \boldsymbol{I}_p + \sigma_{2,j}^2 \left(\boldsymbol{1}_p \boldsymbol{1}_p^\intercal - \boldsymbol{I}_p \right)$ with $\sigma_{1,j}^2 = \mathcal{O}(p^{-\epsilon})$ and $\sigma_{2,j}^2 = \mathcal{O}(p^{-1-3\epsilon})$.

Notably, these conditions ensure (as $p \to \infty$) that the between-class means are asymptotically orthogonal (maximize the between-class variance) and the within class covariances asymptotically isotropic (independent features).

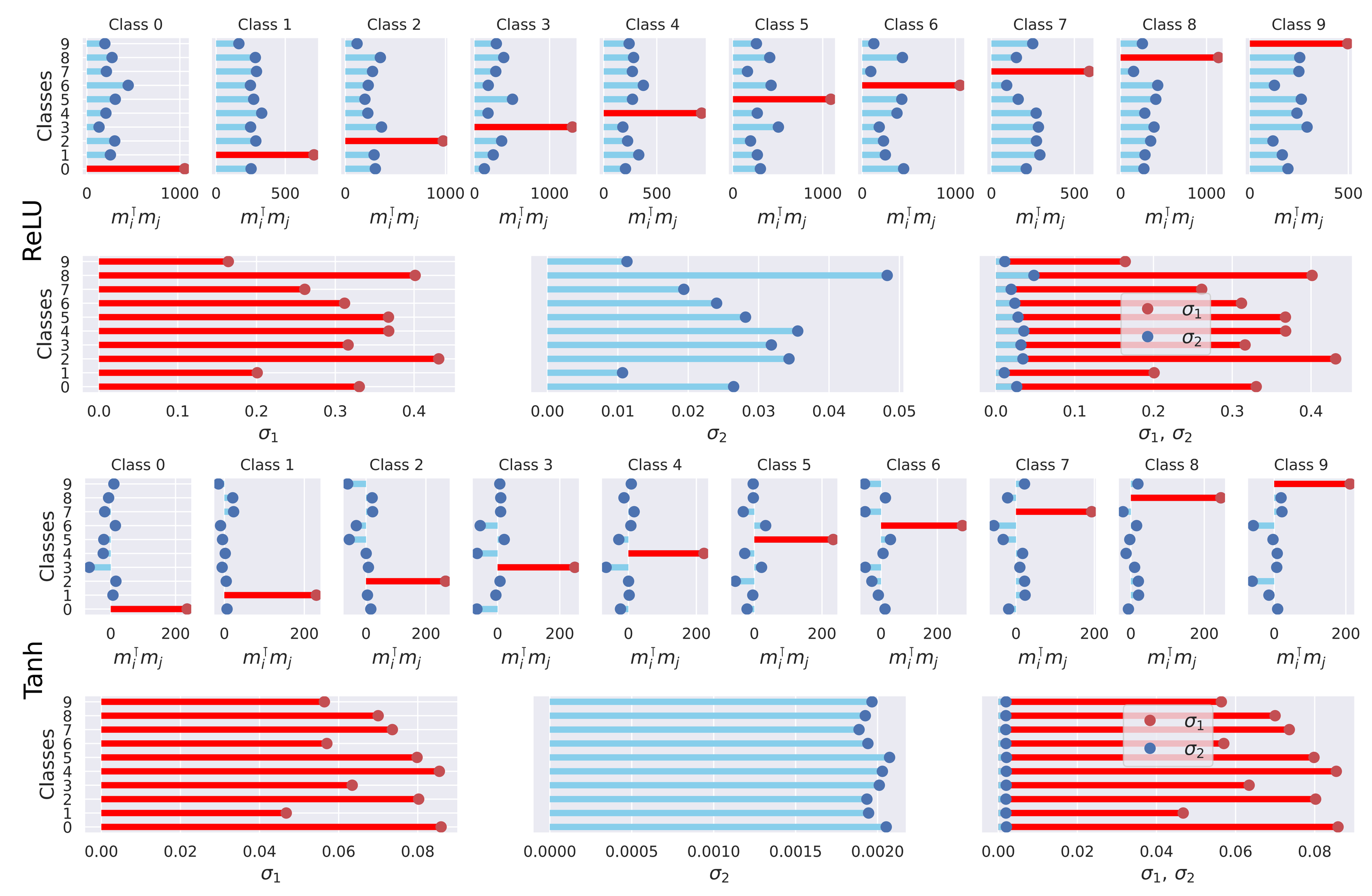
Asymptotic Softmax weight vectors under NOR assumption

For near optimal representations, for sufficiently large p and letting $\varepsilon \to 0$ in the expression of the generalized labels $y_\ell^{(i)}$, the class-weight vectors are asymptotically proportional to the centred class-wise means as

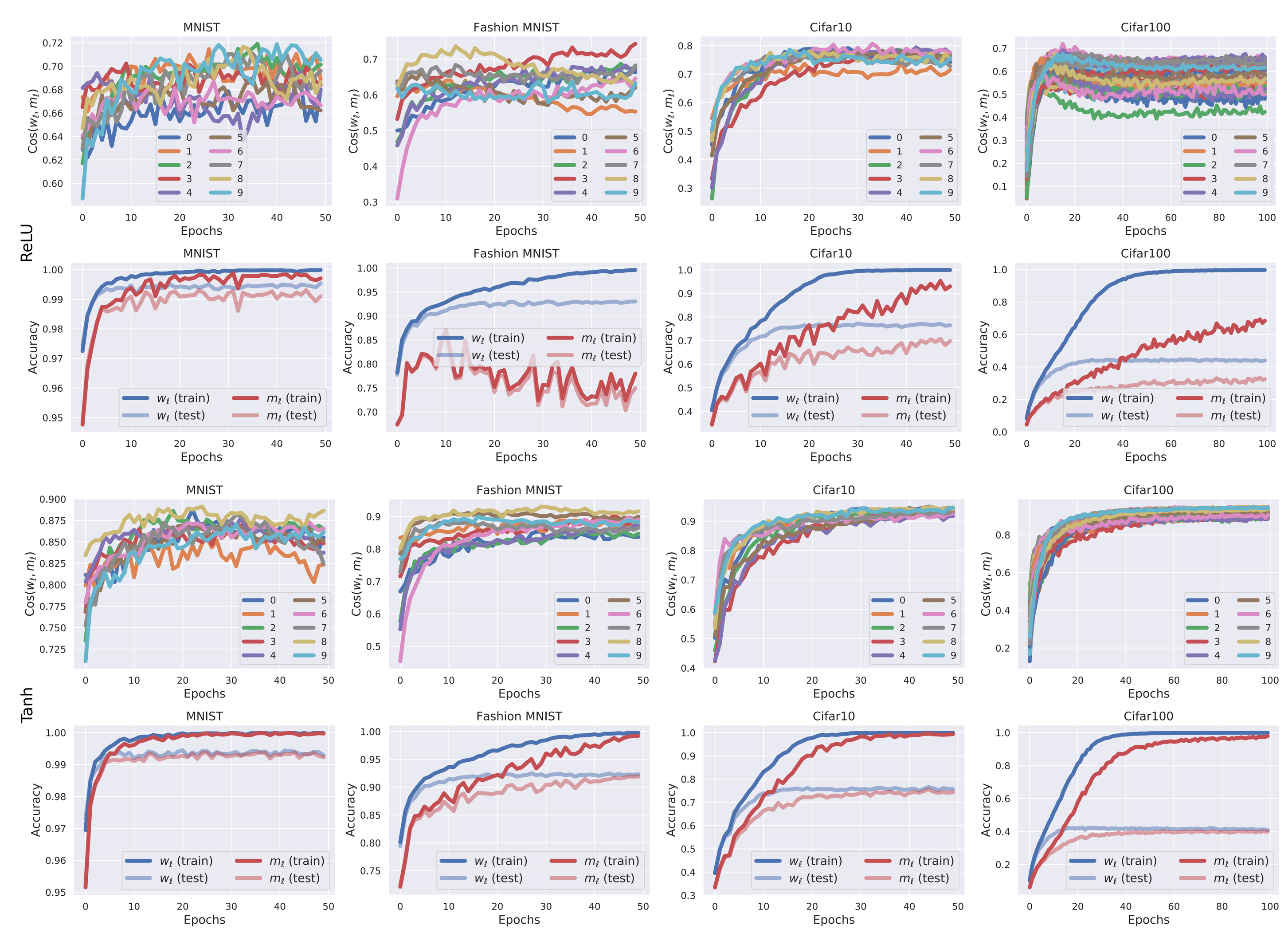
$$\bar{\boldsymbol{w}}_\ell = \frac{\gamma_\ell\, k\, e^{-\kappa \mu_1}}{1 + (k-1)\, e^{-\kappa \mu_1}} \left(\boldsymbol{m}_\ell - \frac{1}{k} \sum_{j=1}^k \boldsymbol{m}_j \right) + \mathcal{O}_{\|\cdot\|}(1) \qquad (1)$$

for some constant $\kappa > 0$ and $\gamma_\ell = \left(\sum_{j=1}^k \sigma_{1,j}^2 \mathbb{E}_j[-f'_{\ell,i}(\bar{\boldsymbol{w}}_\ell^\intercal \boldsymbol{x}_i)] \right)^{-1} \geq \frac{4}{k\, \sigma_{1,\max}^2}$.

NOR in practice



Correlation between $\bar{w}_\ell$ and $\bar{m}_\ell$



Conclusion

Our findings suggest three main procedures for efficient transfer learning: **(i)** use of symmetric representation activations to ensure the near-optimal representations assumption; **(ii)** source model selection without training the Softmax layer; **(iii)** initialization procedure which accelerates the training of the Softmax layer as the target domain gets closer to the source domain.

AAAI 2022 - **Contact:** melaseddik@gmail.com