Semantic Transfer from Head to Tail: Enlarging Tail Margin for Long-Tailed Visual Recognition (Supplementary Materials)

Shan Zhang [*,1], Yao Ni[*,1], Jinhao Du[2], Yanxia Liu[†,3], Piotr Koniusz[†,4,1]
[1]Australian National University, [2]Peking University, [3]Beijing Union University, [4]Data61♥CSIRO
[1]firstname.lastname@anu.edu.au, [2]dujinhao02@gmail.com, [3]yanxia.liu@163.com

A. Sketch Proofs in Section 3

A.1. Proof of Assumption 1

Assumption 1. *Assume that after augmentation, the feature $\tilde{\boldsymbol{a}}_i^t$ passed to the classifier, which comes from a tail sample $\boldsymbol{x}_i^t$, can be approximately represented by a distribution $\tilde{\boldsymbol{a}}_i^t \sim \mathcal{N}(\boldsymbol{a}_i^t, \Delta\boldsymbol{\Sigma}_{th}^i)$. Here, $\boldsymbol{a}_i^t$ is the feature obtained without augmentation, and $\Delta\boldsymbol{\Sigma}_{th}^i$ is a positive definite covariance matrix.*

Proof: For a tail class k_t, each unaugmented feature $\boldsymbol{a}_i^t$ comes from $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, where $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ represent the mean and covariance for class k_t. The augmentation operation in Eq. 6 transforms the $\boldsymbol{a}^t$ into $\tilde{\boldsymbol{a}}^t$, which follows $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_{th})$, with $\boldsymbol{\Sigma}_{th}$ being an approximation to the covariance matrix $\boldsymbol{\Sigma}_h$ (Referring to §A.4, the covariance of augmented tail samples will be close to that of semantically similar head samples under the optimal transformation matrix.). For each i, there is always a suitable matrix $\Delta\boldsymbol{\Sigma}_{th}^i$ that ensures augmented features for class k_t align closely with $\boldsymbol{\Sigma}_{th}$ or $\boldsymbol{\Sigma}_h$. This alignment arises from the synergy of the augmentation method and variations during training, effectively shaping $\boldsymbol{a}_i^t$ into $\tilde{\boldsymbol{a}}_i^t \sim \mathcal{N}(\boldsymbol{a}_i^t, \Delta\boldsymbol{\Sigma}_{th}^i)$.

Remark: Assumption 1 posits that under mild assumptions the augmentation naturally shifts $\boldsymbol{a}_i^t$ to $\tilde{\boldsymbol{a}}_i^t \sim \mathcal{N}(\boldsymbol{a}_i^t, \Delta\boldsymbol{\Sigma}_{th}^i)$.

A.2. Proof of Lemma 2

Lemma 2. *Given the negative log softmax function, the loss L_k for samples of class k without augmentation can be derived as:*

$$\begin{aligned} L_k &= \frac{1}{n_k}\sum_{i=1}^{n_k} -\log \frac{e^{\boldsymbol{w}_k^T\boldsymbol{a}_i + b_k}}{\sum_{j=1}^{K} e^{\boldsymbol{w}_j^T\boldsymbol{a}_i + b_j}} \\ &= \frac{1}{n_k}\sum_{i=1}^{n_k} \log\left(1 + \sum_{j\neq k} e^{d_i\|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2 \cdot \mathrm{sign}(\cos\theta_{i,jk})}\right) \end{aligned} \tag{12}$$

Drawing upon [37], the decision boundary between class j and class k can be formulated as: $(\boldsymbol{w}_j - \boldsymbol{w}_k)^T\boldsymbol{a} + (b_j - b_k) = 0$. d_i is the distance from point $\boldsymbol{a}_i$ to the decision boundary, $\theta_{i,jk}$ denotes the angle between $\boldsymbol{w}_j - \boldsymbol{w}_k$ and $\boldsymbol{a}_i$.

---
[*]Equal contribution. [†]Corresponding author.

Proof:

$$
\begin{aligned}
L_k &= \frac{1}{n_k}\sum_{i=1}^{n_k} -\log \frac{e^{\boldsymbol{w}_k^T \boldsymbol{a}_i + b_k}}{\sum_{j=1}^{K} e^{\boldsymbol{w}_j^T \boldsymbol{a}_i + b_j}} \\
&= \frac{1}{n_k}\sum_{i=1}^{n_k} \log \frac{\sum_{j=1}^{K} e^{\boldsymbol{w}_j^T \boldsymbol{a}_i + b_j}}{e^{\boldsymbol{w}_k^T \boldsymbol{a}_i + b_k}} \\
&= \frac{1}{n_k}\sum_{i=1}^{n_k} \log(1 + \sum_{j \neq k} e^{(\boldsymbol{w}_j - \boldsymbol{w}_k)^T \boldsymbol{a}_i + (b_j - b_k)}) \\
&= \frac{1}{n_k}\sum_{i=1}^{n_k} \log(1 + \sum_{j \neq k} e^{d_i \|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2 \cdot \mathrm{sign}(\cos\theta_{i,jk})})
\end{aligned} \tag{13}
$$

To derive Eq. 13, let us inspect the geometric representation. Focus on the distance d_i from $\boldsymbol{a}_i$ to the decision boundary $(\boldsymbol{w}_j - \boldsymbol{w}_k)^T \boldsymbol{a} + (b_j - b_k) = 0$:

$$
d_i = \frac{|(\boldsymbol{w}_j - \boldsymbol{w}_k)^T \boldsymbol{a}_i + (b_j - b_k)|}{\|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2} \tag{14}
$$

For orientation, when $\boldsymbol{a}_i$ and $\boldsymbol{w}_k$ lie on the same side of the decision boundary, the sign of cosine of the angle between $(\boldsymbol{w}_j - \boldsymbol{w}_k)$ and $\boldsymbol{a}_i$ is given by $\mathrm{sign}(\cos\theta_{i,jk}) = -1$. In the opposite scenario, the sign would be positive. Integrating this insight with Eq. 14, we deduce:

$$
(\boldsymbol{w}_j - \boldsymbol{w}_k)^T \boldsymbol{a}_i + (b_j - b_k) = d_i \|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2 \cdot \mathrm{sign}(\cos\theta_{i,jk}) \tag{15}
$$

This directly gives rise to Eq. 13.

A.3. Proof of Theorem 1

Theorem 1. *Assume Assumption 1 holds when using our augmentation. The loss function L_k^t for tail class k is:*

$$
\begin{aligned}
L_k^t &= \frac{1}{n_k}\sum_{i=1}^{n_k} \mathbb{E}_{\tilde{\boldsymbol{a}}_i^t}\Big[-\log \frac{e^{\boldsymbol{w}_k^T \tilde{\boldsymbol{a}}_i^t + b_k}}{\sum_{j=1}^{K} e^{\boldsymbol{w}_j^T \tilde{\boldsymbol{a}}_i^t + b_j}} \Big] \\
&\leq \frac{1}{n_k}\sum_{i=1}^{n_k} \log\left(1 + \sum_{j \neq k} \beta_{jk}^i e^{d_i \|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2 \cdot \mathrm{sign}(\cos\theta_{i,jk})}\right)
\end{aligned} \tag{16}
$$

where $\beta_{jk}^i = e^{\frac{1}{2}(\boldsymbol{w}_j - \boldsymbol{w}_k)^T \Delta\boldsymbol{\Sigma}_{th}^i (\boldsymbol{w}_j - \boldsymbol{w}_k)}$. Furthermore,

$$
\beta_{jk}^i = \exp\left(\frac{1}{2} {\boldsymbol{v}_{jk}^i}^T \boldsymbol{\Lambda}^i \boldsymbol{v}_{jk}^i\right) > 1 \tag{17}
$$

where $\boldsymbol{V}^i \boldsymbol{\Lambda}^i {\boldsymbol{V}^i}^T = \Delta\boldsymbol{\Sigma}_{th}^i$ and $\boldsymbol{v}_{jk}^i = {\boldsymbol{V}^i}^T (\boldsymbol{w}_j - \boldsymbol{w}_k)$.

Proof of Eq. 16:

$$
\begin{aligned}
L_k^t &= \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbb{E}_{\tilde{\boldsymbol{a}}_i^t}[-\log\frac{e^{\boldsymbol{w}_k^T\tilde{\boldsymbol{a}}_i^t+b_k}}{\sum_{j=1}^{K}e^{\boldsymbol{w}_j^T\tilde{\boldsymbol{a}}_i^t+b_j}}] \\
&= \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbb{E}_{\tilde{\boldsymbol{a}}_i^t}[\log(1+\sum_{j\neq k}e^{(\boldsymbol{w}_j-\boldsymbol{w}_k)^T\tilde{\boldsymbol{a}}_i^t+(b_j-b_k)})] \\
&\leq \frac{1}{n_k}\sum_{i=1}^{n_k}\log(1+\sum_{j\neq k}\mathbb{E}_{\tilde{\boldsymbol{a}}_i^t}[e^{(\boldsymbol{w}_j-\boldsymbol{w}_k)^T\tilde{\boldsymbol{a}}_i^t+(b_j-b_k)}]) &(18)\\
&= \frac{1}{n_k}\sum_{i=1}^{n_k}\log\Big(1+\sum_{j\neq k}e^{(\boldsymbol{w}_j-\boldsymbol{w}_k)^T\boldsymbol{a}_i^t+(b_j-b_k)+\frac{1}{2}(\boldsymbol{w}_j-\boldsymbol{w}_k)^T\Delta\boldsymbol{\Sigma}_{th}(\boldsymbol{w}_j-\boldsymbol{w}_k)}\Big) &(19)\\
&= \frac{1}{n_k}\sum_{i=1}^{n_k}\log\Big(1+\sum_{j\neq k}e^{\frac{1}{2}(\boldsymbol{w}_j-\boldsymbol{w}_k)^T\Delta\boldsymbol{\Sigma}_{th}^i(\boldsymbol{w}_j-\boldsymbol{w}_k)}\cdot e^{(\boldsymbol{w}_j-\boldsymbol{w}_k)^T\boldsymbol{a}_i^t+(b_j-b_k)}\Big) \\
&= \frac{1}{n_k}\sum_{i=1}^{n_k}\log\big(1+\sum_{j\neq k}\beta_{jk}e^{(\boldsymbol{w}_j-\boldsymbol{w}_k)^T\boldsymbol{a}_i^t+(b_j-b_k)}\big) \\
&= \frac{1}{n_k}\sum_{i=1}^{n_k}\log\big(1+\sum_{j\neq k}\beta_{jk}e^{d_i\|\boldsymbol{w}_j-\boldsymbol{w}_k\|_2\cdot\mathrm{sign}(\cos\theta_{i,kj})}\big). &(20)
\end{aligned}
$$

In the above derivation, the inequality Eq. 18 is a direct consequence of Jensen's inequality $\mathbb{E}[\log X]\leq\log\mathbb{E}[X]$. Eq. 19 is obtained by leveraging the moment-generating function $\mathbb{E}[e^{tX}]=e^{t\mu+\frac{1}{2}\sigma^2t^2}$ where $X\sim\mathcal{N}(\mu,\sigma^2)$, and the fact that $(\boldsymbol{w}_j-\boldsymbol{w}_k)^T\tilde{\boldsymbol{a}}_i^t+(b_j-b_k)$ is a Gaussian random variable drawn from $\mathcal{N}\Big((\boldsymbol{w}_j-\boldsymbol{w}_k)^T\boldsymbol{a}_i^t+(b_j-b_k),(\boldsymbol{w}_j-\boldsymbol{w}_k)^T\Delta\boldsymbol{\Sigma}_{th}^i(\boldsymbol{w}_j-\boldsymbol{w}_k)\Big)$. Lastly, Eq. 20 is derived by incorporating Eq. 15.

Proof of Eq. 17: Performing SVD on the positive definite symmetric covariance matrix $\Delta\boldsymbol{\Sigma}_{th}^i$, we obtain $\Delta\boldsymbol{\Sigma}_{th}^i=\boldsymbol{V}^i\boldsymbol{\Lambda}^i{\boldsymbol{V}^i}^T$, where $\boldsymbol{V}^i$ represents the eigenvectors and $\boldsymbol{\Lambda}^i$ is the diagonal matrix of eigenvalues. By incorporating $\boldsymbol{w}_j-\boldsymbol{w}_k$ and $\boldsymbol{V}^i$ into a single term, we define $\boldsymbol{v}_{jk}^i={\boldsymbol{V}^i}^T(\boldsymbol{w}_j-\boldsymbol{w}_k)$. On deriving β_{jk}^i, we get:

$$
\begin{aligned}
\beta_{jk}^i &= e^{\frac{1}{2}(\boldsymbol{w}_j-\boldsymbol{w}_k)^T\Delta\boldsymbol{\Sigma}_{th}^i(\boldsymbol{w}_j-\boldsymbol{w}_k)} \\
&= e^{\frac{1}{2}(\boldsymbol{w}_j-\boldsymbol{w}_k)^T\boldsymbol{V}^i\boldsymbol{\Lambda}^i{\boldsymbol{V}^i}^T(\boldsymbol{w}_j-\boldsymbol{w}_k)} \\
&= e^{\frac{1}{2}({\boldsymbol{V}^i}^T(\boldsymbol{w}_j-\boldsymbol{w}_k))^T\boldsymbol{\Lambda}^i\big({\boldsymbol{V}^i}^T(\boldsymbol{w}_j-\boldsymbol{w}_k)\big)} \\
&= e^{\frac{1}{2}{\boldsymbol{v}_{jk}^i}^T\boldsymbol{\Lambda}^i\boldsymbol{v}_{jk}^i}
\end{aligned}
$$

Let v_c be the c^{th} element of $\boldsymbol{v}_{jk}^i$ and $\lambda_c\geq 0$ be the c^{th} $\mathrm{diag}(\boldsymbol{\Lambda}^i)$, we have:

$$\beta_{jk}^i=e^{\frac{1}{2}\sum_c v_c^2\lambda_c}>e^0.$$

For any non-zero vector $\boldsymbol{w}$ and a positive definite matrix $\boldsymbol{A}$, the result $\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}>0$ implies $\beta_{jk}^i>1$ due to the non-zero vector $\boldsymbol{w}_j-\boldsymbol{w}_k$ and positive definiteness of $\Delta\boldsymbol{\Sigma}_{th}^i$. Furthermore, the relationship $\sum_c\lambda_c=\mathrm{trace}(\boldsymbol{\Lambda}^i)=\mathrm{trace}(\Delta\boldsymbol{\Sigma}_{th}^i)$ indicates the larger the semantic similarities between the head and tail samples, and the more diverse the head class, the greater the value of β_{jk}^i.

A.4. Design of transformation matrix

We aim to design a transformation matrix such that the covariance of transformed tail feature $\boldsymbol{F}_t$, aligns closely with the covariance of head samples. The objective is formulated as:

$$\tilde{\boldsymbol{F}}_t^*=\arg\min_{\tilde{\boldsymbol{F}}_t}\|\tilde{\boldsymbol{F}}_t^T\tilde{\boldsymbol{F}}_t-\boldsymbol{F}_h^T\boldsymbol{F}_h\|_F^2 \tag{21}$$

$$s.t.\quad \tilde{\boldsymbol{F}}_t=\boldsymbol{T}\boldsymbol{F}_t. \tag{22}$$

Substituting the constraint from Eq. 22 into Eq. 21, we find optimality at:

$$\boldsymbol{F}_t^T \boldsymbol{T}^T \boldsymbol{T} \boldsymbol{F}_t = \boldsymbol{F}_h^T \boldsymbol{F}_h. \tag{23}$$

Upon applying singular value decomposition (SVD) to $\boldsymbol{F}_t$ and $\boldsymbol{F}_h$, yielding $\boldsymbol{V}_t \boldsymbol{\Sigma}_t \boldsymbol{V}_t^T$ and $\boldsymbol{V}_h \boldsymbol{\Sigma}_h \boldsymbol{V}_h^T$ and insert them into Eq. 23, a solution set emerges:

$$\boldsymbol{T} = (\boldsymbol{V}_h \boldsymbol{\Sigma}_h^{\frac{1}{2}} \boldsymbol{V}_h^T) \boldsymbol{U} (\boldsymbol{V}_t \boldsymbol{\Sigma}_t^{-\frac{1}{2}} \boldsymbol{V}_t^T)^T, \tag{24}$$

where $\boldsymbol{U} \in \mathbb{R}^{C \times C}$ is a orthogonal group. Eq. 24 indicates that the transformation matrix, $\boldsymbol{T}$, is influenced by the covariance matrices of both tail and head classes. Empirical validation of this design is presented in Sec. 4.3.