Supplementary Material for "Stealing Knowledge from Pre-trained Language Models for Federated Classifier Debiasing"

Meilu Zhu, Qiushi Yang, Zhifan Gao, Jun Liu, and Yixuan Yuan

In this supplementary material, we show the detailed derivation of the upper bound $\overline{\mathcal{L}}_{align}^{\infty}$ of $\mathcal{L}_{align}^{\infty}$ and the prompts of concepts of OCT-C8 and Kvasir-v2 datasets using in this paper.

1 Derivation of the Upper Bound $\overline{\mathcal{L}}_{align}^{\infty}$

$$
\begin{aligned}
\mathcal{L}_{align}^{\infty} &= \frac{1}{N_c}\sum_{i=1}^{N_c}\mathbb{E}_{\boldsymbol{e}^{(\mathbf{y}_i)}\sim\mathcal{N}^{(\mathbf{y}_i)}}\left(-\log\frac{e^{\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{e}^{(\mathbf{y}_i)}}}{e^{\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{e}^{(\mathbf{y}_i)}}+\sum_{k\neq\mathbf{y}_i}^{K}\mathbb{E}_{\boldsymbol{e}^{(k)}\sim\mathcal{N}^{(k)}}e^{\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{e}^{(k)}}}\right)\\
&= \frac{1}{N_c}\sum_{i=1}^{N_c}\mathbb{E}_{\boldsymbol{e}^{(\mathbf{y}_i)}\sim\mathcal{N}^{(\mathbf{y}_i)}}\left(\log(e^{\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{e}^{(\mathbf{y}_i)}}+\sum_{k\neq\mathbf{y}_i}^{K}\mathbb{E}_{\boldsymbol{e}^{(k)}\sim\mathcal{N}^{(k)}}e^{\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{e}^{(k)}})-\log(e^{\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{e}^{(\mathbf{y}_i)}})\right)\\
&\text{//using the Jensen's inequality: } \mathbb{E}[\log(X)]\leq\log(\mathbb{E}X)\\
&\leq \frac{1}{N_c}\sum_{i=1}^{N_c}\left[\log(\mathbb{E}_{\boldsymbol{e}^{(\mathbf{y}_i)}\sim\mathcal{N}^{(\mathbf{y}_i)}}\left(e^{\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{e}^{(\mathbf{y}_i)}}+\sum_{k\neq\mathbf{y}_i}^{K}\mathbb{E}_{\boldsymbol{e}^{(k)}\sim\mathcal{N}^{(k)}}e^{\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{e}^{(k)}}\right)-\mathbb{E}_{\boldsymbol{e}^{(\mathbf{y}_i)}\sim\mathcal{N}^{(\mathbf{y}_i)}}\log(e^{\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{e}^{(\mathbf{y}_i)}})\right]\\
&= \frac{1}{N_c}\sum_{i=1}^{N_c}\left[\log(\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{e}^{(k)}\sim\mathcal{N}^{(k)}}e^{\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{e}^{(k)}})-\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\mu}_{(\mathbf{y}_i)}\right]\\
&\text{//using the moment generation function for Gaussian variable } X:\mathbb{E}[e^{\mathbf{h}^{\mathrm{T}}X}]=e^{\mathbf{h}^{\mathrm{T}}\boldsymbol{\mu}+\frac{1}{2}\mathbf{h}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{h}}\\
&= \frac{1}{N_c}\sum_{i=1}^{N_c}\left[\log(\sum_{k=1}^{K}e^{\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\mu}_k+\frac{1}{2}\tau^2\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\Sigma}_k\mathbf{h}_i})-\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\mu}_{(\mathbf{y}_i)}\right]\\
&\text{//Let } \mathcal{F}(\mathbf{h},\mathbf{y})=\tau\mathbf{h}^{\mathrm{T}}\boldsymbol{\mu}+\frac{1}{2}\tau^2\mathbf{h}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{h}\\
&= \frac{1}{N_c}\sum_{i=1}^{N_c}\left[\log(\sum_{k=1}^{K}e^{\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\mu}_k+\frac{1}{2}\tau^2\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\Sigma}_k\mathbf{h}_i})-\mathcal{F}(\mathbf{h}_i,\mathbf{y}_i)+\mathcal{F}(\mathbf{h}_i,\mathbf{y}_i)-\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\mu}_{(\mathbf{y}_i)}\right]\\
&= \frac{1}{N_c}\sum_{i=1}^{N_c}\left[-\log\frac{e^{\mathcal{F}(\mathbf{h}_i,\mathbf{y}_i)}}{\sum_{k=1}^{K}e^{\mathcal{F}(\mathbf{h}_i,k)}}+\mathcal{F}(\mathbf{h}_i,\mathbf{y}_i)-\tau\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\mu}_{(\mathbf{y}_i)}\right]\\
&= \frac{1}{N_c}\sum_{i=1}^{N_c}\left[-\log\frac{e^{\mathcal{F}(\mathbf{h}_i,\mathbf{y}_i)}}{\sum_{k=1}^{K}e^{\mathcal{F}(\mathbf{h}_i,k)}}+\frac{1}{2}\tau^2\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{(\mathbf{y}_i)}\mathbf{h}_i\right]\\
&= \overline{\mathcal{L}}_{align}^{\infty}
\end{aligned}
$$

2 Prompts of OCT-C8 and Kvasir-v2 Datasets

Table 1. The prompts of OCT-C8 dataset

Prompts
“an image of { }.”,
“a photo of { }.”,
“an OCT scan of { }.”,
“this is a photo of { }.”,
“this is an image of { }.”,
“this is an OCT scan of { }.”,
“this is an OCT photo of { }.”,
“this is an OCT image of { }.”,
“{ } presented in photo.”,
“{ } presented in image.”,
“{ } presented in OCT scan.”,
“{ } presented in OCT image.”,
“{ } presented in OCT photo.”,
“the image shows { }.”,
“the photo shows { }.”,
“the OCT scan shows { }.”,
“the OCT image shows { }.”,
“the OCT photo shows { }.”,
“the image shows the presence of { }.”,
“the photo shows the presence of { }.”,
“the OCT scan shows the presence of { }.”,
“the OCT image shows the presence of { }.”,
“the OCT photo shows the presence of { }.”,
“the presence of { } in image.”,
“the presence of { } in photo.”,
“the presence of { } in OCT image.”,
“the presence of { } in OCT photo.”,
“the presence of { } in OCT scan.”.

Table 2. The prompts of Kvasir-v2 dataset

Prompts
“an image of { }.”,
“a photo of { }.”,
“a endoscopic image of { }.”,
“a endoscopic photo of { }.”,
“this is a photo of { }.”,
“this is an image of { }.”,
“this is a endoscopic photo of { }.”,
“this is a endoscopic image of { }.”,
“{ } presented in photo.”,
“{ } presented in image.”,
“{ } presented in endoscopic photo.”,
“{ } presented in endoscopic image.”,
“the image shows { }.”,
“the photo shows { }.”,
“the endoscopic image shows { }.”,
“the endoscopic photo shows { }.”,
“the image shows the presence of { }.”,
“the photo shows the presence of { }.”,
“the endoscopic image shows the presence of { }.”,
“the endoscopic photo shows the presence of { }.”,
“the presence of { } in image.”,
“the presence of { } in photo.”,
“the presence of { } in endoscopic image.”,
“the presence of { } in endoscopic photo.”.