Distinguishability Calibration to In-Context Learning

Hongjing Li[1]*, Hanqi Yan[1]*, Yanran Li, Li Qian[2], Yulan He[1,3,4], and Lin Gui[3]

[1]Department of Computer Science, University of Warwick, UK
[2]Xiaomi AI Lab, China
[3]Department of informatics, King's College London, UK
[4]The Alan Turing Institute, UK
{Hongjing.Li,Hanqi.Yan}@warwick.ac.uk, yanranli.summer@gmail.com,
qianli@xiaomi.com, {yulan.he,lin.1.gui}@kcl.ac.uk

Abstract

Recent years have witnessed increasing interests in prompt-based learning in which models can be trained on only a few annotated instances, making them suitable in low-resource settings. When using prompt-based learning for text classification, the goal is to use a pre-trained language model (PLM) to predict a missing token in a pre-defined template given an input text, which can be mapped to a class label. However, PLMs built on the transformer architecture tend to generate similar output embeddings, making it difficult to discriminate between different class labels. The problem is further exacerbated when dealing with classification tasks involving many fine-grained class labels. In this work, we alleviate this *information diffusion* issue, i.e., different tokens share a large proportion of similar information after going through stacked multiple self-attention layers in a transformer, by proposing a calibration method built on feature transformations through rotation and scaling to map a PLM-encoded embedding into a new metric space to guarantee the distinguishability of the resulting embeddings. Furthermore, we take the advantage of hyperbolic embeddings to capture the hierarchical relations among fine-grained class-associated token embedding by a coarse-to-fine metric learning strategy to enhance the distinguishability of the learned output embeddings. Extensive experiments on the three datasets under various settings demonstrate the effectiveness of our approach. [1]

1 Introduction

Large pre-trained language models (PLMs) (Devlin et al., 2019; Lan et al., 2020; Liu et al., 2019) have been achieved state-of-the-art performance in many Natural Language Processing (NLP) downstream tasks. More recently, the PLMs with prompt learning demonstrate surprising capabilities in numerous

---

*Equal contribution.
[1]Our code can be found at https://github.com/dontt al/TARA

tasks both in NLP and computer vision, even outperforming their fine-tuned counterparts (Brown et al., 2020; Liu et al., 2021; Lester et al., 2021; Zhou et al., 2022b; Gao et al., 2021a).

Train#1:	Gotta protect'em! It was [MASK].
Train#2:	That's why it's only 20$. It was [MASK].
Test:	On a boat trip to Denmark. It was [MASK].

Table 1: The prompt templates for emotion classification. The samples are from *GoEmotion* (Demszky et al., 2020) dataset.

In an emotion classification task shown in Table 1, an input sentence X, followed by a prompt, *"It was [MASK]"*, is fed to a PLM to predict the missing token at the position of *[MASK]*. The predicted word can be used to identify the emotion label of the input sentence. Such few-shot learning generates a probability distribution over the *[MASK]* conditioning on the given prompt/context, which is considered as in-context learning of language models.

However, as in-context learning does not require updating PLM parameters, there arises the problem of distribution mismatch between the data used for LM pre-training and the test samples used in in-context learning, which hinders the full exploitation of the knowledge encoded in PLMs (Xie et al., 2022; Zhao et al., 2021; Ge et al., 2022; Shin et al., 2022). To alleviate the context shift, existing methods rely on prior knowledge to increase the overlapping between the two distributions. For example, *PTR* (Han et al., 2021) appends domain-agnostic tokens to prompts to discriminate the domains, such as *"sports"*, *"politics"*. Another line of studies designs sophisticated handcrafted verbalizers to map the test samples onto the label word space derived from PLMs (Schick and Schütze, 2021; Gao et al., 2021b). Although the gradient-optimized verbalizers (Hu et al., 2022) are proposed to ease the human effort and can be adapted to different downstream tasks via training, it is still consid-

ered inferior to the manual verbalizers, especially in both the few-shot and zero-shot settings where training data are scarce.

In this paper, we first show that PLMs have an inherent *information diffusion* issue in their generated output token embeddings, which share a large proportion of similar information after going through a stack of transformer layers (Gao et al., 2019; Yan et al., 2022). Such token embeddings occupy a narrow cone, leading to largely overlapped output distributions when applied to in-context learning. Next, we elaborate that the overlapped output distributions would violate the distinguishability condition (Xie et al., 2022) under in-context learning. To this end, we propose to flatten the singular value distributions of the output embeddings generated from PLMs to shape the space spanned by the singular values to a desirable manifold. On the one hand, we apply an orthogonal and a scaling constraints to the weight matrix applied to the output embeddings, which can avoid exploding and vanishing values in the feature matrix (Saxe et al., 2014), leading to better discriminative features when trained with limited labelled data. On the other hand, we leverage hyperbolic embeddings to capture the hierarchical relations among fine-grained class labels of training examples to further enhance the distinguishability of output embeddings.

Our proposed framework has been implemented on top of existing prompt-based few-shot learning methods and it demonstrates an average 5.86% performance improvement of F1-measure on three classification tasks under 100-shot learning. We also verify that the improvement stems from a more balanced singular value distribution for the output features and the learnt hierarchical feature space.

In summary, our contributions include:

- We propose a transformation-based constraint to output embeddings by rotation and ratio balancing which is able to guarantee the distinguishability of learned embeddings.

- The proposed hyperbolic embedding-based metric learning strategy not only improves the performance of prompt learning but also measures the relation between different categories.

- The experimental results outperform many strong baselines and the visualisation illustrates that the proposed method is able to project the embedding to a less overlapping

distribution and improve the interpretability and distinguishability of output. Specifically, across three evaluated datasets, our method surpasses the state-of-the-art by 9.60%, 5.11% and 2.87%, respectively, in the 100-shot setting.

2 Related works

***Information diffusion* in PLMs.** In a typical L-layer transformer-based PLM, assuming the prompt is a concatenation of a few training examples and a test input X_{test}, consisting of m tokens in total, the goal of in-context learning is to predict the output distribution over the masked token at the t-th position, *[MASK]*. It is formally defined by the following equation:

$$p(\mathcal{O}_t|X_{\text{test}}) = \mathbb{E}_{h\sim p_{\text{prompt}}(h|X_{\text{test}})}[p(\mathcal{O}_t|X_{\text{test}}, h, \theta)],$$

where h denotes the last-layer hidden state corresponding to the token of X_{test}, θ is the parameters in prompt-based learning.

Although we have limited knowledge of the output distribution $p(\mathcal{O}_t|X_{\text{test}})$ over token *[MASK]*, many existing studies analyzed the geometry properties of the last layer feature h^L, and examined its effects in downstream tasks (Goyal et al., 2020; Zhou and Srikumar, 2022). Due to the softmax bottleneck (Yang et al., 2018) and the likelihood loss in language generation tasks (Gao et al., 2019), the output feature distribution in PLMs tends to be anisotropic and rank-deficient, which limits the expressiveness of the generated representations. Goyal et al. (2020) discussed the information diffusion issue among tokens within a sentence that feeding the tokens in different positions for classification only resulted in a 1.2% variance in classification accuracy. Gao et al. (2019) explored the information diffusion among different sentences via singular value decomposing and they found that the singular value distributions are skewed especially in deeper PLM layers, i.e., larger singular values become more predominant compared to the smaller ones.

Context shift in in-context learning. Many researchers studied the distribution shift (aka. domain shift) between the pretraining corpora and test samples and proposed solutions to decrease the performance variance in prompt-based few-shot learning (Xie et al., 2022; Zhao et al., 2021; Hu et al., 2022; Zhou et al., 2022b; Shin et al., 2022).
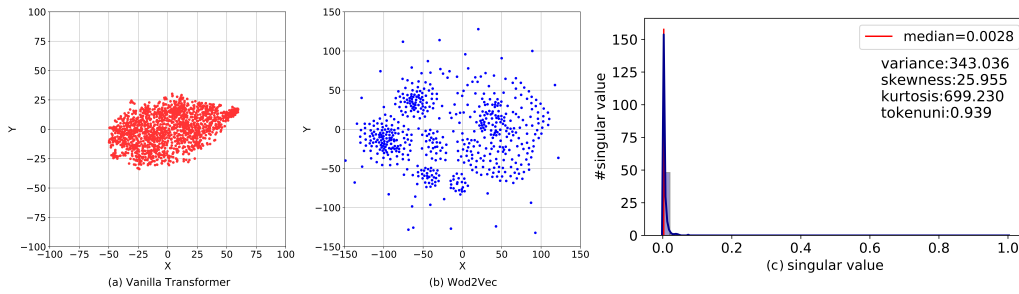
Figure 1: **(a)**: The mapping results of 1,500 *[MASK]* tokens randomly sampled from the *GoEmotions* dataset. Each red dot is the output representations derived from prompt-based learning for the *[MASK]* token of an input example, which will be used to predict the masked token in the corresponding position. **(b)**: Each blue dot is the static word representation of the corresponding predicted token with the largest probability on *[MASK]* for one of the 1,500 samples in (a) from the *GoEmotions* dataset. **(c)**: Singular value distribution (after normalisation) of the output representations of the randomly selected 1,500 *[MASK]*s. It is clear that the representations are dominated by very few singular values.

On the one hand, some in-context learning methods incorporated domain-specific words or learnable tokens in the prompt to discriminate different context. Ben-David et al. (2022) proposed to first generate the name of the domain and then generate domain-related features (DRFs) conditioned on the domain in a supervised manner. Both the generated domain name and DRFs were used as the prompt fed to the model. On the other hand, the sophisticated verbalizers contributed to minimising the distance between the two distributions (Schick et al., 2020; Schick and Schütze, 2021; Gao et al., 2021b; Hu et al., 2022). To broaden the coverage of single-choice verbalizer, *Knowledge Prompt Tuning (KPT)* (Hu et al., 2022) used the knowledge graph to extract more topic-related words as label words and then refine the label word candidates. To incorporate prior knowledge to calibrate the context shift, Xie et al. (2022) simplified a language model as the Hidden Markov Model, where the observed tokens are sampled from a family of concepts and proposed the *distinguishability condition* to measure context shift as the Kullback–Leibler (KL) divergence.

3 Contextual Calibration for Output Distribution

Many existing methods calibrate the probabilities of the generated tokens in a language model in order to improve the generation quality. In prompt-based learning, we want to find out if the output distribution $p(\mathcal{O}_t|X_{\text{test}})$ or the output feature $h^{[\text{mask}]}$, which is a part of the hidden representation from the last layer of a PLM, h^ℓ, suffers from the *information diffusion* issue and occupies a narrow

cone. We take RoBERTa-based prompt learning as an example and derive the value of $h^{[\text{mask}]}$ from 1,500 randomly selected test samples from an emotion classification dataset, *GoEmotions* (Demszky et al., 2020), and visualise the results in a 2D plane in Figure 1(a). For comparison, we select the predicted token with the largest probability on each *[MASK]* and map their corresponding vectors from Word2Vec (Mikolov et al., 2013) to a 2D plane in 1(b). It is clear that the word embeddings learned from Word2Vec has a more uniform distribution around the origin. In contrast, the representations derived by RoBERTa degenerate into a narrow cone, which implies limited expressiveness. Inspired by the approach proposed in (Yan et al., 2022), we display the singular value distribution of $h^{[\text{mask}]}$ and calculate the distribution statistics, i.e., the matrix moment and the average cosine similarity between every *[MASK]* pair in Figure 1(c). From the empirical results, we can see that the value of the hidden representation for *[MASK]* in different samples share much similar information with the token uniformity value (Yan et al., 2022) (*tokenuni* in Figure 1(c)) of 0.939. This shows that most $h^{[\text{mask]}}$ concentrates at very few singular values, which implies a severe information diffusion issue.

3.1 Uniform Ratio-based Distinguishability

Although many calibration methods have been proposed, few of them focuses on explicitly addressing the information diffusion issue in the prompt-based learning framework. One main challenge in this task is that the unlabelled data used in language model pre-training is significantly larger than the labelled samples used for prompt tuning. Hence,

the optimised distribution in prompt-based few-shot learning can be very different from the true distribution. To avoid inheriting the *information issue* caused in the pre-training phase, we propose a calibration method to reduce the skewness of the output token distributions, such that the output representations are evenly distributed in the embedding space. The idea is to rotate the original embedding space to an isotropic metric space by an inner product-based operator on a learnable basis. For each dimension of the basis, we use the inner product to measure its relevance with a given input. The dimension-dependent relevance scores are sent to a Multi-layer Perceptron (MLP) decoder to generate the calibrated output embedding for final prediction.

The framework of the proposed calibration method is shown in Figure 2. In practice, due to the small size of training samples in prompt learning, the relevance scores might be dominated by very few dimensions. Therefore, inspired by Zhou et al. (2022a), who proposed a ratio estimator to balance the distribution from different label categories, we design a scaling matrix in our isotropic distribution scenario. That is, for both labelled and unlabelled data, the multi-class ratio between different dimensions should be similar.
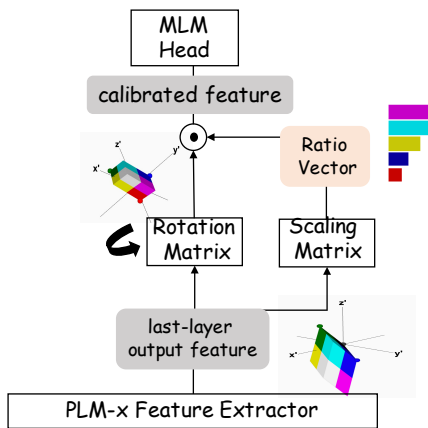

Figure 2: Our proposed calibration method is applied to the output embeddings from the last layer of a PLM. After being transformed with a rotation matrix through a Multi-layer Perception (MLP), the resulting output feature is assumed to have a more balanced singular value distribution in different basis directions. Moreover, as the vector norm on each projected direction would change in the new base, we derive a ratio vector to balance the distribution along the rotated directions.

Concretely, assuming we have N labelled data $\{y_j, x_j\}_{j=1}^N$ and M unlabelled data from pre-training $\{x_j\}_{j=N+1}^{N+M}$, where x_j is the input sample,

y_j is the true label, and $M \gg N$. To simplify the notation, in the rest of this paper, we use x_j to represent the feature of the last embedding layer and h_j to represent the output of our calibrated feature. Then, for the representation of a masked token, x_j, we assume there are K isotropic directions in the metric space and the corresponding inner product based relevance score is:

$$\mathcal{H}_k(x_j) = \sigma(\langle x_j, W_k \rangle),\ (1 \le k \le K), \quad (1)$$

where $\sigma(\cdot)$ is the softmax activation function. Here, we can define a **rotation matrix** based on W_k since Eq. (1) projects an input embedding onto a new metric space by rotation. To guarantee the orthogonality of the basis in the new metric space, we use the following regulariser during training:

$$\mathcal{L}_{orth} = \left\| W^\top W - \boldsymbol{I} \right\|_2^2, \quad (2)$$

where W is the stacking of $\{W_k\}_{k=1}^K$. Correspondingly, for each dimension k, we can define a ratio score which aims to better separate them to avoid the skewed distribution by minimising the following loss:

$$\mathcal{L}_t = \frac{1}{N+M} \Sigma_{k=1}^K \Sigma_{j=1}^{N+M} ||\mathcal{R}_k(x_j) - \frac{1}{K}||^2, \quad (3)$$

where $\mathcal{R}_k(x_j)$ is an MLP-based estimator with a softmax activation:

$$\mathcal{R}_k(x_j) = \sigma(S_k \cdot x_j + \beta). \quad (4)$$

By minimising $\mathcal{L}_t$, even if one input sample x_j is similar to a basis vector along a popular dimension k, there will still be a probability to assign it a low ratio score $\mathcal{R}_k(x_j)$ if there are other samples which are more closer to the basis vector in dimension k. In this way, we can balance the distribution after rotation. We define the stacking of S_k as a **scaling matrix** which aims to distribute x_j uniformly into K clusters in the metric space.[2]

However, it is difficult to optimise the loss defined in Eq. (3) since the size of the unlabelled data for pre-training is much larger than the labelled data and the unlabelled data is usually unseen to the downstream tasks. We instead define an alternative optimisation objective. First, according to Eq. (3), we need to ensure that for any two dimensions k and t, we have $\frac{1}{N+M}e^{S_k \cdot x_j} = \frac{1}{N+M}e^{S_t \cdot x_j}$.

---

[2]We measured the impact of different weight initialisations on S_k in Appendix A.2.

By the Jensen's inequality, we have the following lower bound: $e^{\frac{1}{N+M}S_k \cdot x_j} \leq \frac{1}{N+M}e^{S_k \cdot x_j}$, in which we can achieve the lower bound for any two independent dimensions by taking $\frac{1}{N+M}S_k \cdot x_j = \frac{1}{N+M}S_t \cdot x_j$. It means that for any two dimensions, the sum of their ratio scores should be similar. As such, Eq. (3) can be approximated by:

$$\mathcal{L}_t \sim \Sigma_{k=1}^K (||S_k||^2 - 1)^2. \qquad (5)$$

Accordingly, we can define the distinguishability loss in a more general form by both the relevance score and the ratio score without the need of sampling from unlabelled data:

$$\mathcal{L}_{dis} = \mathcal{L}_{orth} + \mathcal{L}_t. \qquad (6)$$

From our findings in Section 3, much information encoded by the output representations generated by the last layer of a PLM only occupies a space spanned by very few singular value directions. This leads to the information diffusion issue. Therefore, our solution here is to re-project the output features into a new hyperplane, in which the information is more evenly distributed in different directions, and at the same time we can derive a ratio vector by aggregating the rotated components.

3.2 Supervised Prompt Learning

By our proposed distinguishability loss-based learning in Section 3.1, an input embedding has been separated into vectors along K independent dimensions. Then, for the labelled data $\{x_j\}_{j=1}^N$, we propose to use k independent decoders to produce the final prediction. The decoding result is based on the relevance score and ratio score on each independent dimension:

$$\boldsymbol{h}_j = \Sigma_{i=k}^K \text{Decoder}_k(\mathcal{H}_k(x_j) \cdot \mathcal{R}_k(x_j)), \quad (7)$$

where the Decoder_k is a decoder for the k-th dimension. Then the representation of $\boldsymbol{h}_j$ can be used in the verbalizer $p_{verbalizer}(\hat{\mathcal{O}}|\boldsymbol{h}_j)$, where $\hat{\mathcal{O}}$ is the predicted masked token. Finally, the cross-entropy loss H is defined by the predicted $\hat{\mathcal{O}}$ and the true label y_j:

$$\mathcal{L}_{cls}(x_j) = H(y_j, p_{verbalizer}(\hat{\mathcal{O}}|\boldsymbol{h}_j)). \quad (8)$$

By combining the uniform ratio-based distinguishability loss of $\mathcal{L}_{dis}$ and the prompt-based classification loss $\mathcal{L}_{cls}$, we propose our first model, named as **T**ransformation based **A**daptation for

Ratio b**A**lanced (**TARA**) prompt learning, which aims to minimise $\mathcal{L}_{\textbf{TARA}} = \mathcal{L}_{cls}(x_j) + \mathcal{L}_{dis}$. Note that $\mathcal{L}_{cls}(x_j)$ is the default loss term in all the baselines and our proposed methods.

3.3 Dimension Rotation by Hyperbolic Embeddings

In Section 3.1, we project the input mask embedding into a K dimensional metric space to avoid skewed distributions. However, we ignore the potential class relations between the dimensions. For example, in emotion classification, both the emotions of 'gratitude' and 'approval' belong to the *coarse* positive class, but they are associated with different *fine-grained* labels in the GoEmotions dataset (Demszky et al., 2020). Hence, in this section, we only consider those positive pairs under the same coarse category to achieve a better class disambiguation by a proxy based metric learning (Movshovitz-Attias et al., 2017; Yang et al., 2022), which uses an anchor vector to represent a category for metric loss optimisation and capture the hierarchical structure between coarse- and fine-grained labels in the hyperbolic space.

Strategies for Constructing Sample Pairs. Inspired by the hierarchical structure of coarse-to-fine emotion categories, we assume that a fine-grained emotion should be close to the coarse-grained emotion it belongs to. To implement this idea, we construct sample-anchor pairs $(\boldsymbol{h}_j, z_i^+)$ for training, where $\boldsymbol{h}_j$ is the representation for prompt prediction and $z_i^+ \in \mathbb{R}^d$ is a learnable anchor representation for each coarse class.

Metric Learning in a Hyperbolic Space. To maximise the similarity in sample-anchor positive pairs, where the sample and the anchor share the same coarse-grained label, while minimising the similarity in negative pairs, we adopt the following metric learning objective:

$$\mathcal{L}_{metric}(\boldsymbol{h}_j) = -log \frac{e^{-d(\boldsymbol{h}_j, z_{pj}^+)}}{\sum_{i=1}^C e^{-d(\boldsymbol{h}_j, z_i^+)}}, \quad (9)$$

where $\{(\boldsymbol{h}_j, z_i^+)\}_{i=1}^C$ represents a set of sample-anchor pairs that we constructed for each sample i, C denotes the number of anchors, z_{pj}^+ is the representation of positive pairing anchor of j-th sample, and $d(\cdot)$ is the hyperbolic distance metric defined by the Poincaré ball model of the hyperbolic space (Nickel and Kiela, 2017). In a n-dimensional hyperbolic space, all points will fall into a unit open

interval: $\mathcal{I}^n = \{x \in \mathbf{R}^n | \|x\| < 1\}$, where $\|\cdot\|$ donates the Euclidean norm. The distance $d(\cdot)$ between two points $u, v \in \mathcal{I}^n$ can be formulated as:

$$d(u, v) = \text{arcosh}(1 + 2\frac{\|u - v\|^2}{(1 - \|u\|^2)(1 - \|v\|^2)}). \tag{10}$$

The motivation of using $\mathcal{L}_{metric}(\boldsymbol{h}_j)$ is to push similar categories together in the metric space. Hence, we can obtain our final learningn objective by adding the loss of tree-structured metric learning $\mathcal{L}_{metric}(\boldsymbol{h}_j)$ to **TARA** as:

$$\mathcal{L}_{final} = \mathcal{L}_{cls}(x_j) + \mathcal{L}_{metric}(\boldsymbol{h}_j) + \mathcal{L}_{dis}. \tag{11}$$

For a comparison, we propose a variant called **TML** by keeping the learning architectue and simply adding $\mathcal{L}_{metric}(\boldsymbol{h}_j)$ to the classification loss of $\mathcal{L}_{cls}(x_j)$, but without the ratio balancing term of $\mathcal{L}_{dis}$, that is, $\mathcal{L}_{\textbf{TML}} = \mathcal{L}_{cls}(x_j) + \mathcal{L}_{metric}(\boldsymbol{h}_j)$.

4 Experiments

Datasets We evaluate our proposed approach on three multi-class text classification datasets, the *Emotion*[3] (Saravia et al., 2018) dataset, an academic paper classification dataset, *WOS* (Kowsari et al., 2017), and a fine-grained emotion classification dataset, *GoEmotions*[4] (Demszky et al., 2020). All of these datasets have hierarchical label structures. The datasets statistics are shown in Table 2.

Name	#Classes	#Train	#Dev	#Test
Emotion	6	16,000	2,000	2,000
WOS	11	5,736	1,147	1,147
GoEmotions	28	23,485	2,956	2,984

Table 2: Dataset statistics.

For all datasets, we remove punctuation, digits, and special characters that do not have specific semantic meanings. For the *Emotion* dataset which consists of tweet, we also remove user mentions.

Baselines We implement our proposed framework on top of the commonly used prompt-based learning methods and compare it with existing approaches including those which can be used for learning more discriminative representations:

- *Prompt-baselines.* Three commonly used prompt-based methods are selected including

Soft Prompts (Brown et al., 2020), *Prompt-Tuning* (Lester et al., 2021) and *PTR* (Han et al., 2021). The best-performing methods is used as the default prompt-based training method for the following three comparison models, and denoted as *Prompt-baseline*.[5]

- *KPT* (Hu et al., 2022). It uses a knowledge graph to incorporate topic-related label words to increase the coverage of the verbaliser.

- *Context Calibration* (Zhao et al., 2021). This method calibrates the output representations by one-layer linear transformation, whose weight matrix is optimised to be diagonal.

- *Proxy-NCA* (Movshovitz-Attias et al., 2017). It creates a proxy for each class and uses the Neighbourhood Component Analysis (NCA) loss to pull samples closer to their assigned proxies while pushing negative samples away.

Prompt Settings As the performance of prompt-based methods heavily relies on prompt templates and verbalisers, we use the same template and verbaliser for all models for fair comparison. The prompt templates are shown in Table 3. The original class labels are used as label words in the verbaliser as in (Schick and Schütze, 2021).

Datasets	Prompt template
Emotion	*<X>It's [MASK].*
WOS	*<X>The domain of the text is [MASK].*
GoEmotions	*<X>The emotional aspect of this text is [MASK].*

Table 3: Prompt templates used in three datasets.

4.1 Few-shot Learning on Three Datasets

We randomly select k different training samples for few-shot learning and show the results across the three datasets in Table 4.

For metric-learning, *Proxy-NCA* with contrastive loss leads to performance degradation compared to the *Prompt-baseline*, with more significant performance drops on the *GoEmotions* dataset, which has the largest label categories. By contrast, **TML** gives better results over the *Prompt-baseline* and *Proxy-NCA*, showing its efficiency in encoding hierarchical relations between the coarse- and fine-grained labels. It can be further demonstrated in Figure 3, which shows the similarity matrix

	Emotion				WOS				GoEmotions			
K-shot	5	10	50	100	5	10	50	100	5	10	50	100
Prompt-baseline	0.336	0.363	0.431	0.625	0.236	0.252	0.359	0.435	0.161	0.173	0.281	0.310
Proxy-NCA	0.333	0.384	0.412	0.637	0.214	0.246	0.295	0.383	0.149	0.166	0.208	0.233
Context Calibration	0.337	0.352	0.531	0.706	0.212	0.361	0.687	0.707	0.164	0.224	0.355	0.420
TML	0.339	0.387	0.466	0.699	0.229	0.277	0.372	0.529	0.158	0.227	0.309	0.355
TARA	0.348	0.401	0.697	0.783	0.245	0.418	0.705	0.728	0.172	0.249	0.364	0.442
Ours full model	**0.355**	**0.441**	**0.713**	**0.802**	**0.278**	**0.439**	**0.719**	**0.757**	**0.206**	**0.255**	**0.384**	**0.448**

Table 4: Weighted F1 scores on three Datasets. The proposed **TML** is better than *Proxy-NCA*. Our full method (**TML+TARA**) achieves the best performance among all the settings.

(28×28) of the 28 fine-grained emotion labels from 3 high-level categories, i.e., *"anger"*,*"joy"* and*"sad"*. The results of *Proxy-NCA* in (c) are similar to the *Prompt-baseline* as shown in (b). Our proposed **TML** in (d) can capture the hierarchical relations among the 28 labels, where the correlations among labels belonging to the same high-level emotion category are similar. By comparison, we replace the hyperbolic distance in **TML** with the Euclidean distance and show the results in (c). It can be observed that the resulting label embeddings fail to exhibit different patterns within and across different high-level emotion categories.

For the calibration methods, *Context Calibration* and **TARA** are overall better than the *Prompt-baseline*. This shows that the simple linear transformation of the output representations can greatly improve the performance of prompt-based learning. The superior performance of **TARA** over *Context Calibration* demonstrates the benefit of using our proposed rotation and scaling transformations. Combining **TML** with **TARA**, our full model achieves the best performance and the improvements are more predominant when K is larger. In the 100-shot setting, our method surpasses the state-of-the-art method, *Context Calibration*, by 9.6% on Emotion, 5.1% on WOS, and 2.9% on GoEmotions, respectively, verifying its superiority in the few-shot text classification task.

4.2 Information Diffusion Alleviation

In addition to the classification results, we also examine the characteristics of the generated output representations to check whether the information diffusion issue has been addressed. Figure 4 shows the PCA projection results of all the *[MASK]* representations, i.e., $h^{[\text{MASK}]}$ in the test samples, which are colour-coded according to their assigned class labels by the model. It is clear that our method can generate more widely distributed [MASK] representations, therefore better reducing the overlaps
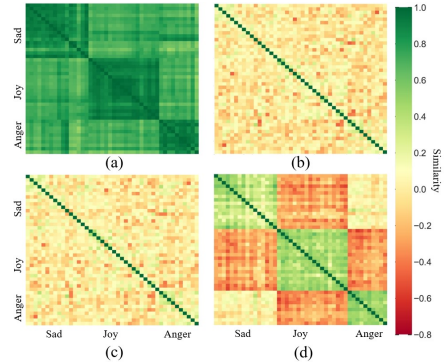

Figure 3: Heatmap for the pair-wised cosine similarity of fine-grained classes on *GoEmotion*. (a) Label representations from PLM without fine-tuning. (b) Fine-tuned label representations by classification module only. (c) Fine-tuned label representations with proposed constraint but based on Euclidean distance, i.e., *Proxy-NCA*. (d) Fine-tuned label representations by **TML**.

of the features from different class labels. For example, in the *Emotion* dataset, the output features from the baseline model mostly reside along the horizontal direction, while ours distribute more evenly across different directions.[6]

We also calculate the summary statistics of the singular value distribution of the output features, as well as the average similarity between every two *[MASK]* pairs. The results are shown in Table 5. The average cosine similarity (*CosSim*) between every token pair is used as a proxy measure of the degree of information diffusion. We can observe that the *CosSim* value calculated on the output representations generated by our model is significantly lower compared to the other baselines. We also observe an increase in the median and the decrease in variance of the singular value distribution from our model outputs in comparison to the prompt learning baseline. The results show that our model produces the output representations which have a more balanced singular value distribution. The

---

[6]The T-SNE results and singular value distribution of the output representations in *Emotion* and *GoEmotions* are shown in Figure A1 and Figure A2.

(a) *Emotion.*
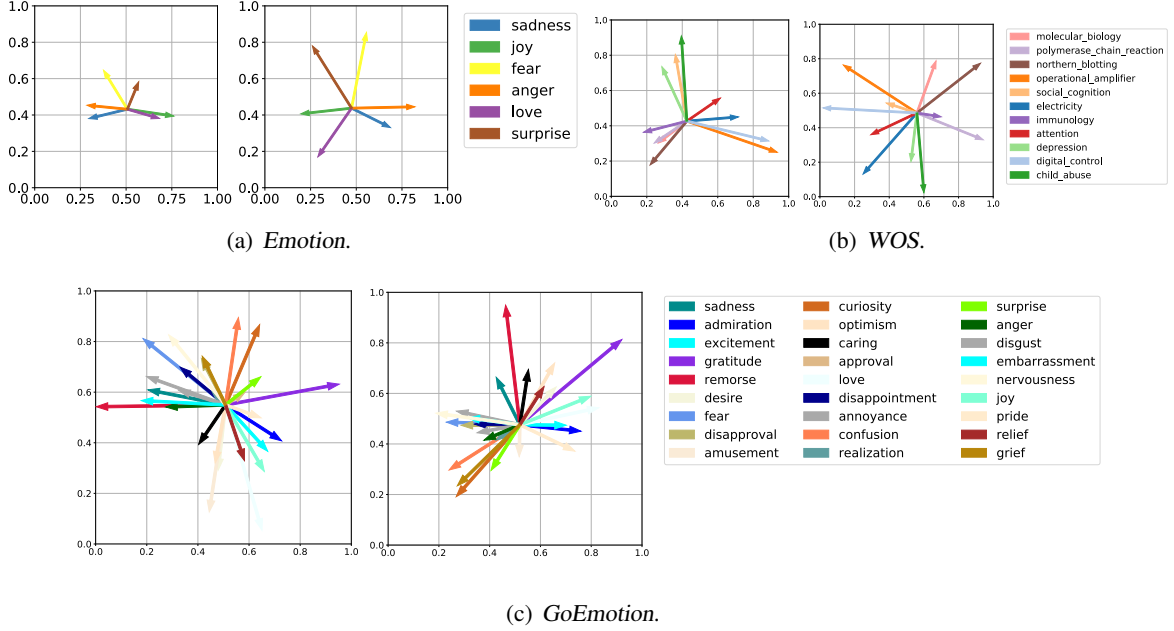
(b) *WOS.*

(c) *GoEmotion.*

Figure 4: The PCA projection of the output representations belonging to different classes. In each sub-figure, the **left figure is the prompt-baseline**, while **the right figure is our method**. It is clear that our method distributes the output representations more evenly in the embedding space, while the output representations from the baseline appear to be more concentrated.

	Median	Variance	Skewness	CosSim
Emotion-prompt	0.0028	371.9	24.57	0.898
Emotion-Ours	0.0145	5.211	8.960	0.256
WOS-prompt	0.0036	235.8	22.06	0.817
WOS-Ours	0.0117	5.681	9.088	0.191
GoEmotions-prompt	0.0028	822.1	24.64	0.899
GoEmotions-Ours	0.0268	11.20	7.728	0.243

Table 5: The statistics of the singular value distribution of the output features, as well as the average cosine similarity of all *[MASK]* token pairs.

	Ours	**w/o $\mathcal{L}_{orth}$**	**w/o $\mathcal{L}_t$**	**w/o l_2**	**w/o all**
Emotion	0.802	0.725	0.719	0.723	0.724
WOS	0.757	0.728	0.687	0.741	0.699
GoEmotions	0.448	0.422	0.415	0.427	0.412

Table 6: Ablation study of various loss terms in the learning objective for the distinguishability loss.

smaller skewness value further verifies that our proposed model can generate isotropic representations where the embedding dimensions are uncorrelated.

4.3 Ablation Study

To study the effect of different components of our proposed distinguishability loss, i.e., the constraints applied to the transformation operation for ratio balancing, we remove one of them and compare the performance changes in Table 6. Here, $\mathcal{L}_{orth}$ is applied on W in Eq.2, $\mathcal{L}_t$ is applied on

S_k (from Eq.4 and Eq.5), and l_2 is the weight for the L_2 regularisation term on all the other learnable parameters. The $\mathcal{L}_{orth}$ and L_2 constraints have similar effects on the overall performance, as they both act as axis transformations, while the constraint L_t applied on S_k plays a more important role, whose removal leads to a larger performance drop among all the settings. It partly demonstrates the importance of the balancing ratio vector after the rotation transformation.

5 Conclusion

In this paper, to address the information diffusion issue in prompt-based few-shot learning, we propose a calibration method based on featuretransformation which first rotates output embeddings into a new metric space, and then scales the ratio of each dimension to a uniform distribution to guarantee the distinguishability of the transformed embeddings. On the other hand, we utilise hyperbolic embeddings to capture the hierarchical relations between class labels to guide the metric learning strategy to enhance the interpretability of the learned output embeddings. Extensive experiments on the three multi-class classification tasks under various settings demonstrate the effectiveness of our approach with an average 5.9% performance improvement on the F1-measure.

Limitations

In this work, we only focus on the multi-class classification task with hierarchical class labels. Future work could explore extending our idea to other tasks, such as controllable text generation, which has the similar information diffusion issue. Another potential direction in future work is to learn a prior distribution rather than simply using the uniform distribution in ratio balancing. Since the uniform distribution-based ratio balancing is a strong assumption, it might not be suitable for some tasks in real-world applications. One could use VAE or VQ-VAE to learn a distribution which could be subsequently used to regularise the optimisation of feature transformation.

Acknowledgements

References

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2022. Pada: Example-based prompt learning for on-the-fly adaptation to unseen domains. *Transactions of the Association for Computational Linguistics*, 10:414–433.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4040–4054. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021a. Clip-adapter: Better vision-language models with feature adapters. *CoRR*, abs/2110.04544.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021b. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830. Association for Computational Linguistics.

Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. 2022. Domain adaptation via prompt learning. *CoRR*, abs/2202.06687.

Saurabh Goyal, Anamitra Roy Choudhury, Saurabh Raje, Venkatesan T. Chakaravarthy, Yogish Sabharwal, and Ashish Verma. 2020. Power-bert: Accelerating BERT inference via progressive word-vector elimination. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3690–3699. PMLR.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: word-level adversarial reprogramming. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4921–4933. Association for Computational Linguistics.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021. Ptr: Prompt tuning with rules for text classification. *arXiv preprint arXiv:2105.11259*.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong

Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2225–2240. Association for Computational Linguistics.

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, , Matthew S Gerber, and Laura E Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*. IEEE.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3045–3059. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Yair Movshovitz-Attias, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, and Saurabh Singh. 2017. No fuss distance metric learning using proxies. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 360–368. IEEE Computer Society.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *CoRR*, abs/1705.08039.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. 2014. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woo-Myoung Park, Jung-Woo Ha, and Nako Sung. 2022. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5168–5186. Association for Computational Linguistics.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hanqi Yan, Lin Gui, Wenjie Li, and Yulan He. 2022. Addressing token uniformity in transformers via singular value transformation. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 2181–2191. PMLR.

Zhibo Yang, Muhammet Bastan, Xinliang Zhu, Doug Gray, and Dimitris Samaras. 2022. Hierarchical proxy-based loss for deep metric learning. In *IEEE/CVF Winter Conference on Applications of*

Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022, pages 449–458. IEEE.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. 2018. Breaking the softmax bottleneck: A high-rank RNN language model. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Doudou Zhou, Molei Liu, Mengyan Li, and Tianxi Cai. 2022a. Doubly robust augmented model accuracy transfer inference with high dimensional features.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Yichu Zhou and Vivek Srikumar. 2022. A closer look at how fine-tuning changes BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1046–1061. Association for Computational Linguistics.

A Appendix

A.1 Model Selection

Following previous research (Gao et al., 2021b; Hambardzumyan et al., 2021; Lester et al., 2021), BERT (Devlin et al., 2019), Roberta and ALBERT (Lan et al., 2020) were used when using the *cloze* prompts. The *cloze* is to fill in the blanks in the prompt template by the model itself.

Model	Zero-shot		
	Accuracy	**Macro-f1**	**Weighted-f1**
BERT (fine-tuning)	0.0342	0.0196	0.0329
BERT	0.0716	0.0165	0.0384
RoBERTa	0.1094	0.0465	0.0994
ALBERT	0.0538	0.0217	0.0459

Table A1: Classification results on GoEmotion dataset of different baseline models.

To select the baseline model used as the backbones of our proposed method, we evaluate the baseline models on *GoEmotions* dataset for zeroshot learning, the results are shown in Table A1. We compare the effects of the same model using prompt learning and fine-tuning, respectively (difference in effects between BERT (Devlin et al., 2019) and BERT (fine-tuning). After comparison, we chose RoBERTa (Liu et al., 2019) as it shows the overall best performance.

Based on the large pretrained language model backbone, we compare different prompt-based training methods and select the best as our *Prompt-Baseline*. The details are shown in Table A3.

A.2 Weight Initialisation

The optimisation of W_k and S_k can be affected by different weight initialisations. As such, we experiment with different initialisation strategies and show the results of 100-shot learning in Table A2 (We use the same initialisation for W_k and S_k.). The Gaussian distribution initialisation performs the best overall. Therefore we use the Gaussian distribution initialisation in all the experiments reported in the paper.

B Visualisation results

To better compare the results of Baseline methods and ours, we visualize the output of different labels by mapping them into 2D plane via T-SNE (Figure A1). It is clear that our model separates the data points of different labels ((b) and (d)) rather than

Initialisation	Emotion	WOS	GoEmotions
Gaussian	0.802	0.757	0.448
Xavier	0.817	0.749	0.420
Eye	0.798	0.747	0.387
Orthogonal	0.801	0.757	0.431

Table A2: Initialisation of different distributions on weight matrix.

mixing them up (shown in (a) and (c)). To explore the corresponding effects of singular values distribution, we visualise the normalized singular value distribution of the output embeddings in Figure A2. We observe a more balanced distribution after applying our transformation and metric learning.

K-shot	Emotion			WOS			GoEmotions		
	Soft	P-Tuning	PTR	Soft	P-Tuning	PTR	Soft	P-Tuning	PTR
5	0.295	**0.336**	0.330	0.165	**0.236**	0.213	0.072	0.135	**0.161**
10	0.312	**0.363**	0.351	0.180	**0.252**	0.230	0.151	0.151	**0.173**
50	0.363	**0.431**	0.409	0.328	**0.359**	0.319	0.230	0.245	**0.281**
100	0.423	0.625	**0.631**	0.412	**0.435**	0.391	0.331	**0.336**	0.310

Table A3: Weighted F1 of few-shot for different prompt-based training methods.

(a) P-Tuning on Emotion   (b) Our Model on Emotion   (c) PTR on GoEmotion   (d) Our Model on GoEmotion
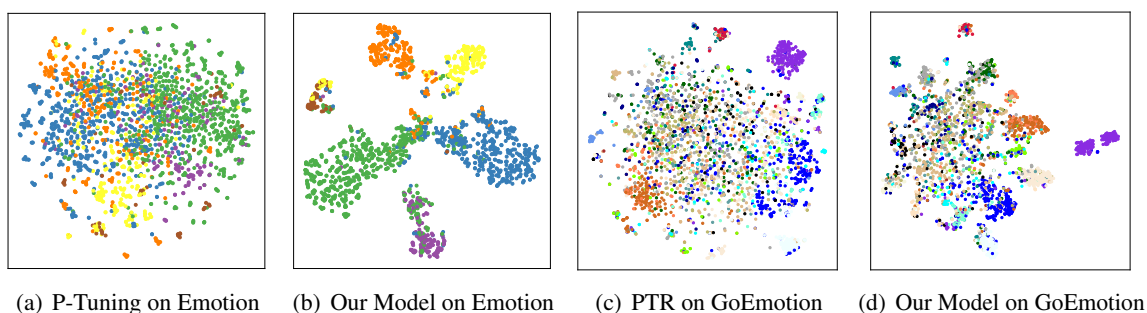
Figure A1: The T-SNE results of test samples in Emotion/GoEmotions Dataset under 100-shot.

(a) P-Tuning on Emotion dataset

(b) Our Model on Emotion dataset

(c) PTR on GoEmotion dataset

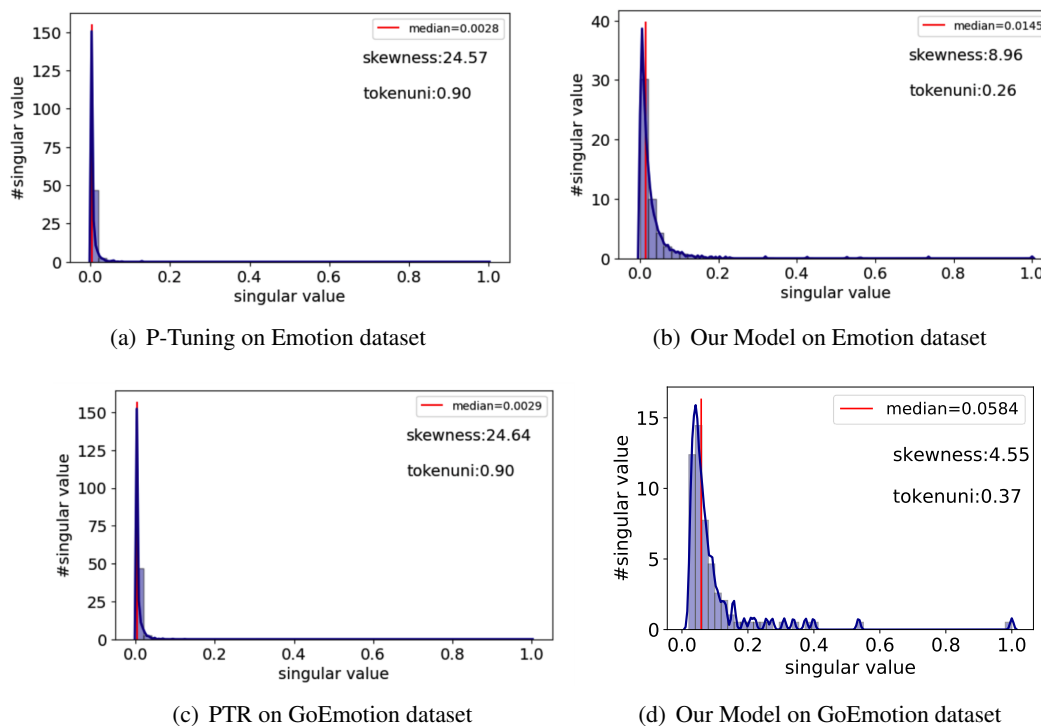(d) Our Model on GoEmotion dataset

Figure A2: The singular value distribution of test samples under 100-shot. Our methods greatly balance the singular distribution, i.e., decrease the skewness, and alleviate the *information diffusion* issue, i.e., decrease the token similarity (*tokenuni*).