Visual Prompt Tuning for Few-Shot Text Classification

Jingyuan Wen[1,2], Yutian Luo[1,2], Nanyi Fei[1,2], Guoxing Yang[1,2], Zhiwu Lu[1,2,*]
Hao Jiang[3], Jie Jiang[3], Zhao Cao[3]
[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods
[3]Huawei Poisson Lab, Hangzhou, Zhejiang
{wenjingyuan,luzhiwu}@ruc.edu.cn

Abstract

Deploying large-scale pre-trained models in the prompt-tuning paradigm has demonstrated promising performance in few-shot learning. Particularly, vision-language pre-training models (VL-PTMs) have been intensively explored in various few-shot downstream tasks. However, most existing works only apply VL-PTMs to visual tasks like image classification, with few attempts being made on language tasks like text classification. In few-shot text classification, a feasible paradigm for deploying VL-PTMs is to align the input samples and their category names via the text encoders. However, it leads to the waste of visual information learned by the image encoders of VL-PTMs. To overcome this drawback, we propose a novel method named Visual Prompt Tuning (VPT). To our best knowledge, this method is the first attempt to deploy VL-PTM in few-shot text classification task. The main idea is to generate the image embeddings w.r.t. category names as visual prompt and then add them to the aligning process. Extensive experiments show that our VPT can achieve significant improvements under both zero-shot and few-shot settings. Importantly, our VPT even outperforms the most recent prompt-tuning methods on five public text classification datasets.

1 Introduction

Pre-training models have achieved great success across a variety of tasks in recent years. Pre-training language models (PLMs) like BERT (Devlin et al., 2019), GPT (Radford et al., 2018) and their variants (Liu et al., 2019; Raffel et al., 2020; Yang et al., 2019; Lewis et al., 2020) firstly appeared as the milestones in the AI field. They brought huge boost to natural language processing (NLP) tasks, such as text classification (Devlin et al., 2019), named entity recognition (NER) (Jia et al., 2020), and text generation (Chan and Fan, 2019). In computer vision, large-scale pre-training models (e.g., BiT (Kolesnikov et al., 2020) and ViT (Dosovitskiy et al., 2021)) became popular as in NLP. With convolutional neural networks or Transformers (Vaswani et al., 2017) as the backbones, they were shown to be effective on a wide range of visual downstream tasks (e.g., image classification, object detection, and semantic segmentation). More recently, inspired by these pre-training models in NLP and computer vision, vision-language pre-training models (VL-PTMs) have been intensively explored (Su et al., 2020; Li et al., 2020; Huo et al., 2021; Lu et al., 2022; Fei et al., 2022). They achieve excellent performance in cross-modal tasks like image-text retrieval, visual question answering (VQA), and image caption. Besides, they also show great potential in single-modal tasks (Lin et al., 2021; Yuan et al., 2021). These achievements clearly declare the power of large-scale pre-training models.

With GPT-3 (Brown et al., 2020) demonstrating astonishing zero-shot and few-shot ability, researchers are encouraged to explore the potential of large-scale pre-training models in few-shot learning. Recently, prompt-tuning has been widely used in few-shot tasks as a paradigm for deploying pre-training models. Compared with prompt-tuning, the performance of the traditional fine-tuning paradigm has apparent drawbacks when only few training samples are available (Schick and Schütze, 2021). PLM based prompt-tuning methods (e.g. Prefix-tuning (Li and Liang, 2021), P-tuning (Liu et al., 2021b), ADAPET (Tam et al., 2021)) have shown their effectiveness and robustness on NLP tasks. Meanwhile, VL-PTM based prompt-tuning methods like CoOp (Zhou et al., 2021), Clip-Adapter (Gao et al., 2021a) and CPT (Yao et al., 2021) apply VL-PTMs to few-shot visual tasks including few-shot image classification and visual ground-

*Zhiwu Lu is the corresponding author.

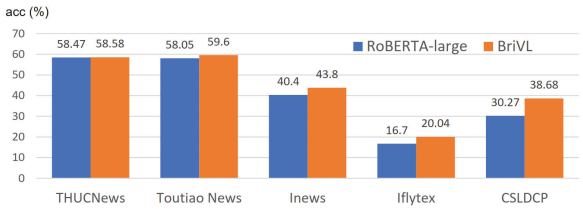


Figure 1: Zero-shot results on five public text datasets. The cross-modal model BriVL (Fei et al., 2022) is shown to outperform the comparably-sized single-modal model RoBERTa-large, indicating that the visual information may bring benefits to textual tasks.

ing. These successes reveal that textual information is beneficial for visual tasks. However, there still lacks a method to utilize VL-PTMs in few-shot NLP tasks like few-shot text classification. Importantly, we notice that the cross-modal model BriVL (Fei et al., 2022) achieves better zero-shot results than the comparably-sized single-modal model RoBERTa-large on five public text classification datasets (see Figure 1), indicating that the visual information may bring benefits to textual tasks. This thus motivates us to introduce VL-PTM into few-shot text classification.

In this work, we propose a novel method named Visual Prompt Tuning (VPT) for few-shot text classification. It is a prompt-tuning method designed to apply VL-PTM in few-shot text classification. To make use of the visual understanding ability of VL-PTM, we design a visual prompt generation module based on model inversion (see Figure 2), which can obtain sound visual representations of the categories offline as visual prompts. In the classification process, we still adopt the standard prompt-tuning pipeline, using the text encoder of VL-PTM as backbone. Specifically, we append learnable soft prompt to the front of category names' text embeddings. We then add visual prompts to the obtained embeddings of the corresponding categories and conduct the text classification by computing the cosine similarity scores between the input embeddings and the summed (both visual and textual) category embeddings. In addition, to make extensive evaluation, we collect five public available datasets for Chinese text classification, which cover a diverse set of data domains including news, emotions, types of app, and specialized subjects.

Our main contributions are three-fold: (1) To the best of our knowledge, this is the first work on introducing VL-PTM into few-shot NLP tasks. Particularly, the importance of VL-PTM has been successfully shown in few-shot text classification. (2) We devise a novel VPT method for VL-PTM, which can utilize the visual information to boost the text classification performance. (3) Extensive experiments are conducted on five benchmark datasets to show that our proposed VPT outperforms the state-of-the-art approaches.

2 Related Work

2.1 Vision-Language Pre-Training Models

We first deliver an overview of Vision-Language Pre-Training Models (VL-PTMs). Note that existing VL-PTMs can be broadly divided into two groups according to their network architectures: single-tower models (Su et al., 2020; Li et al., 2020) and two-tower ones (Radford et al., 2021; Jia et al., 2021; Yuan et al., 2021).

Single-tower models appear as the pioneers of VL-PTMs, using a joint network (mostly multi-layer Transformers) to encode the image and text pair. (Su et al., 2020) employs BERT-like objectives to learn cross-modal representations from a concatenated sequence of visual region features and language token embeddings. (Li et al., 2020) makes use of object tags as anchor points for aligning elements in two modalities. This method is motivated by the observation that the salient objects in an image can be accurately detected, and are often mentioned in the paired text. Single-tower models have strong ability to fuse visual and linguistic information. However, they still have a lot of limitations due to the model structure, such as limited understanding ability of high-level semantics, long inference time, etc.

As the successors of VL-PTMs, two-tower models demonstrate greater potential in cross-modal pre-training. They adopt separate image and text encoders, typically taking image-text retrieval as the pre-training task. (Radford et al., 2021; Jia et al., 2021) introduces contrastive learning with SimCLR-based loss for visual-language pre-training. The training goal is to learn powerful encoders that can embed image and paired text samples into the same latent space for effective image-text retrieval. With acceptable inference time and ideal comprehension skill in both cross-modal and single-modal tasks, two-tower models are deployed in various of application scenarios. To expand the learned representations to more visual tasks, (Yuan et al., 2021) concurrently uses

self-attention and cross-attention in their network, enhancing the understanding ability in both single-modal and cross-modal tasks.

In this work, we devise our VPT based on the latest BriVL (Fei et al., 2022), which is a two-tower large-scale Chinese VL-PTM (see Sec. 3.1 for more details). Theoretically, our VPT can be extended to other VL-PTMs in the same way.

2.2 Prompt-Tuning of VL-PTMs

With the recent rapid development of VL-PTMs, there is a growing interest in prompt-tuning with VL-PTMs for various downstream tasks. As a representative model of VL-PTMs, CLIP (Radford et al., 2021) employs prompt template like "A photo of a {label}." in image classification task without further training, which shows the competitive performance against linear probe on ResNet50 (He et al., 2019) (a fully supervised baseline). This success declares the potential of the combination of prompt-tuning and VL-PTM. To alleviate the instability and manpower cost of manual hard prompt, (Zhou et al., 2021) introduces learnable prompt for few-shot image classification. In addition to automated prompt engineering, (Gao et al., 2021a) proposes to insert lightweight learnable module named adapter into VL-PTM, which is a simpler alternative than soft prompt.

Beside image classification, recent works have applied visual-language pre-training to more downstream tasks with prompt-tuning. For example, (Yao et al., 2021) reformulates the visual grounding task into a fill-in-the-blank problem. This recent work creatively uses the RGB value of different colors to build the CLIP-like image sub-prompt. (Tsimpoukelli et al., 2021) designs a unified framework for multi-modal conditional text generation. The proposed pipeline is compatible with seven cross-modal tasks including Referring Expression Comprehension (REC) and Visual Commonsense Reasoning (VCR).

Note that VL-PTMs are generally deployed for visual or cross-modal tasks in the previous works mentioned above. Differently, our proposed VPT extends the application scenarios of VL-PTMs, and forms the first prompt-tuning method that induces VL-PTM into few-shot text classification.

2.3 Prompt-Tuning Methods in NLP Tasks

"Pre-train, prompt, and predict" paradigm is a sea change in NLP (Liu et al., 2021a). Instead of adapting PLMs to downstream tasks through objective engineering, the downstream problems are reformed with the use of a textual prompt to seem more like those solved during the original PLM training. For instance, (Schick and Schütze, 2021) maps each class into a masked token and inserts it into cloze-style phrases, then predicting it using the pre-trained masked language model. This method ensembles multiple models trained with several manual prompts. To get better prompt templates, (Gao et al., 2021b) adopts a T5 model to automatically generate prompts in cloze-style. Then another PLM like RoBERTA is deployed to conduct the label name generation process. (Liu et al., 2021b) introduces trainable continuous prompt embeddings as a better choice than manual prompts, which significantly improves the understanding ability of generative PLMs like GPT. Unlike cloze question-based methods, (Devlin et al., 2019) reformulates the classification task into textual entailment task. This setting can be used as a unified approach to modelling different kinds of classification tasks. Different from the above-mentioned PLM based methods, we introduce VL-PTM to few-shot text classification and propose a novel Visual Prompt Tuning (VPT) method, which blazes a new trail for few-shot NLP tasks.

3 Methodology

In this section, we give the details of the proposed VPT. The overall architecture is shown in Figure 2. Our main idea is to improve the performance of text classification by utilizing visual information in help of VL-PTM. Specifically, VPT deploys model inversion of VL-PTM to generate visual representations of category names and then add them to the text embeddings of the corresponding category names. We first describe the overall framework of VPT in Sec. 3.1, followed by a detailed description of visual prompt generation in Sec. 3.2. Finally, the training objective is presented in Sec. 3.3.

3.1 Overall Framework

We employ BriVL as the backbone of VPT. To learn better cross-modal representations, the representative contrastive learning algorithm MoCo (He et al., 2020) is adopted in the pre-training stage of BriVL. The text encoder of BriVL consists of RoBERTa-large (Cui et al., 2020) and a successive self-attention block. The self-attention

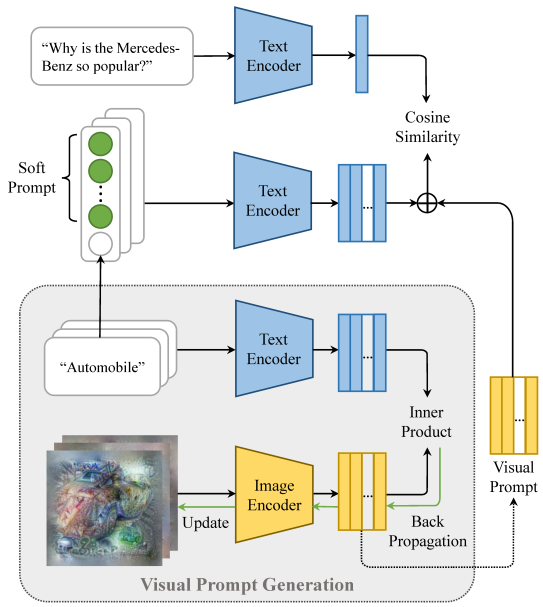


Figure 2: A schematic illustration of the proposed VPT model for few-shot text classification. The bottom panel presents the visual prompt generation module, and the top panel presents the prompt-tuning module for few-shot text classification.

block with four layers of Transformers is designed for keep RoBERTa from catastrophic forgetting. To perform few-shot text classification with BriVL, we adopt similar paradigm used in few-shot image classification. In particular, given a text classification dataset with M categories, we have natural language expressions $\{C_1, \cdots, C_M\}$ for all categories. For the m-th class C_m, we append learnable soft prompts to the front of its word embedding. That is, the total input embedding sequence $\mathbf{t}_m$ for C_m is designed as:

$$\mathbf{t}_m = [CLS][V]_1 \cdots [V]_N[CLASS]_m[SEP], \tag{1}$$

where $[V]_n$ $(n = 1, \cdots, N)$ denotes a learnable vector with the same dimension as the word embedding of BriVL, N is the hyperparameter specifying the number of learnable tokens, $[CLASS]_m$ is the word embedding of the m-th class name C_m, and $[CLS]$ (or $[SEP]$) is the word embedding of the special token CLS (or SEP). Note that we adopt class-specific soft prompts and use word embeddings of sampled tokens from the vocabulary as the initialization of $[V]_n$.

With the text encoder of BriVL fixed during the training stage, our training goal is to optimize the soft prompts. We first get the visual prompts of all categories offline from the visual prompt generation module. We then encode the tokenized input sequence $\mathbf{x}_i$ and class prompt $\mathbf{t}_m$ into text embeddings $\mathbf{r}_i$ and $\mathbf{r}_m^C$ via the text encoder of BriVL, respectively. For cross-modal information fusion, we thus adopt a simple operation-based method: adding visual prompt to its corresponding class name embedding $\mathbf{r}_m^C$ with the weight of α. Classification is finally conducted by computing cosine similarity scores between the input embedding $\mathbf{r}_i$ and fused class embeddings.

Algorithm 1 Pseudocode of Visual Prompt Generation in a PyTorch-like style.

```
# text_list: list of all class names
# Shape: shape of pseudo image ([C, H, W])
# VP: list of visual prompts corresponding
    to all class names
# f_image, f_text: image encoder and text
    encoder of the adopted VL-PTM

VP = []
for text in text_list:
  pseudo_image = random_tensor(*Shape)
  pseudo_image.requires_grad_(True)
  for i in range(max_iteration):
    imageFea = f_image.forward(pseudo_image)
    textFea = f_text.forward(text)

    # Eqn.(2)
    loss = -mm(imageFea, textFea.t()).mean()

    # Adam update: pseudo_image
    loss.backward()
    update(pseudo_image.params)
  VP.append(imageFea)
save(VP)
```

Notations: `mm` – matrix multiplication.

3.2 Visual Prompt Generation

The bottom panel of Figure 2 illustrates the pipeline of our visual prompt generation. Given a text classification task, we choose to generate a series of pseudo images according to the class names. In this work, we take the embeddings of pseudo images as the visual representations of the class names, namely visual prompts.

For each class name C_m, we can obtain its text embedding $\mathbf{r}_m^C$ through the text encoder of BriVL. Then we randomly initialize a noisy image and also compute its image embedding $\mathbf{r}_m^I$ through the image encoder of BriVL. Since there should be a one-to-one correspondence between pseudo images and class names, we compute the inner product similarity score between the image embedding and the class embedding and maximize it for model inversion. The loss function for the m-th class can be written as follows:

$$\mathcal{L}_{inversion} = - < \mathbf{r}_m^C, \mathbf{r}_m^I >, \tag{2}$$

where $< \cdot, \cdot >$ is the dot product of two vectors.

Algorithm 2 Pseudocode of VPT in a PyTorch-like style.

```
# VP: output from visual prompt generation
# prompt_tokens: N tokens from vocabulary
# alpha: weight of visual prompt
# f_text: text encoder of the adopted VL-PTM
# f_text.emb: word embedding layer of f_text
# class_names: tokenized class names in the
    following form: [CLS][CLASS][SEP]

load(VP)
C = f_text.emb(class_names)
# initialize soft prompt
soft_prompt = f_text.emb(prompt_tokens)
soft_prompt.requires_grad_(True)
for input_text in loader: # load a minibatch
  # Eqn.(1)
  t = cat([C[:,0,:], soft_prompt, C[:,1:,:])

  inputFea = f_text.forward(input_text)
  labelFea = f_text.forward(input_embs=t)
  labelFea += alpha * VP.ToTensor() # Eqn.(3)

  logits = bmm(inputFea, labelFea)
  # Eqn.(5)
  loss = CrossEntropyLoss(logits, labels)

  # Adam update: soft prompt
  loss.backward()
  update(soft_prompt.params)
```

Notations: `bmm` – batch matrix multiplication; `cat` – concatenation.

Because both encoders of BriVL are frozen during pseudo image generation, only the noisy image is set to be learnable, i.e., it can be updated through back propagation. After a number of iterations, we obtain the pseudo image that depicts a picture of what BriVL knows about the category.

We take the image embedding $\bar{\mathbf{r}}_m^C$ of the pseudo image as our "visual prompt", which can supplement the insufficient information in $\mathbf{r}_m^C$. Because the generation process is done offline, there is no extra time for classification, ensuring VPT's efficiency. The details of the visual prompt generation pipeline are presented in Algorithm 1.

3.3 Training Objective

In this subsection, we describe our training objective and explain the role of visual prompt. The similarity between the i-th input text $\mathbf{x}_i$ and the m-th class name C_m is calculated as follows:

$$s_{im} = < \mathbf{r}_i, \mathbf{r}_m^C + \alpha \bar{\mathbf{r}}_m^C >, \tag{3}$$

where $\mathbf{r}_i$ and $\mathbf{r}_m^C$ are respectively the text embedding of $\mathbf{x}_i$ and class prompt $\mathbf{t}_m$, $\bar{\mathbf{r}}_m^C$ is the visual prompt of C_m, and α is the weight of visual prompt. A softmax function is then used to define the probability value:

$$P(y_i = m|\mathbf{x}_i) = \frac{\exp(s_{im}/\tau)}{\sum_{j=1}^{M} \exp(s_{ij}/\tau)}, \tag{4}$$

where $P(y_i = m|\mathbf{x}_i)$ means the chance of the i-th input text x_i belonging to the m-th class (y_i is the predicted label), and τ is the temperature. Given k shots per class, model training is performed by minimizing the cross-entropy loss:

$$\mathcal{L}_C = \frac{-1}{k * M} \sum_i \sum_m^M y_{im} \log P(y_i = m|\mathbf{x}_i), \tag{5}$$

where $y_{im} = 1$ if $y_i = m$, otherwise $y_{im} = 0$.

Note that Equation (3) indicates the core idea of our proposed VPT for few-shot text classification. That is, each input sentence is forced to be matched with not only textual but also visual semantics of class names. From this perspective, visual prompts act as augmentations of text embeddings of class names. They provide extra information when searching the nearest class name in the latent space for the input sentence. The full VPT algorithm for few-shot text classification is outlined in Algorithm 2.

Table 1: Statistics of five text classification datasets.

Dataset	Classes	Train	Val	Test
THUCNews	14	661,785	83,000	83,000
Toutiao News	15	306,688	38,000	38,000
Inews	3	3,356	1,000	1,000
Iflytex	119	6,935	2,599	2,599
CSLDCP	67	536	536	1,784

4 Experiments

4.1 Datasets

We collect five public available datasets for text classification in Chinese: THUCNews (Li et al., 2006), Toutiao News[1], Inews[2], Iflytex (Xu et al., 2020) and CSLDCP (Xu et al., 2021). Diverse textual tasks are covered, including classification on news (THUCNews, Toutiao News), emotions (Inews), types of app (Iflytex), and specialized subjects (CSLDCP). We follow the original dataset split from public benchmarks (Inews, CSLDCP) and randomly split the others (THUCNews, Toutiao News) with the train/validation/test ratio 8:1:1. Particularly, for THUCNews, we only use titles of the news in our experiments. Since the test set of Iflytex is not labeled, we use the public validation set as the test set, and split the public training set into the training and validation sets. The details of datasets are shown in Table 1.

[1]https://github.com/aceimnorstuvwxz/toutiao-text-classfication-dataset

[2]https://github.com/ChineseGLUE/ChineseGLUE

Table 2: Comparative results for few-shot text classification on five public datasets. We report the mean (and standard deviation) performance over 5 repeated trials. The best performance and the second best performance are denoted in bold and underlined fonts, separately.

Method	THUCNews	Toutiao News	Inews	Iflytex	CSLDCP
Soft Prompt	64.70 (3.64)	71.98 (1.15)	51.76 (1.90)	28.92 (1.67)	37.24 (1.40)
PET	66.33 (1.70)	75.75 (3.31)	62.10 (0.96)	<u>33.49</u> (2.44)	41.87 (0.95)
LM-BFF	71.56 (0.99)	<u>76.67</u> (1.24)	63.72 (2.25)	29.70 (1.68)	38.23 (2.71)
EFL	70.17 (2.12)	71.03 (2.89)	60.20 (5.83)	22.80 (5.06)	42.80 (1.45)
P-tuning	<u>73.46</u> (2.29)	76.56 (1.02)	<u>65.96</u> (2.18)	32.36 (2.63)	<u>44.03</u> (1.59)
VPT (ours)	**74.73** (0.90)	**79.24** (1.44)	**67.20** (2.85)	**34.24** (1.35)	**47.03** (0.84)

In our few-shot experiments, we set $k = 16$ shots per class for THUCNews, Toutiao News, and Inews, but only $k = 1$ shot per class for Iflytex and CSLDCP (given their large number of classes).

4.2 Implementation Details

For the visual prompt generation process, the total iteration number is set to 2,000. The noisy image is optimized by Adam with a learning rate of 0.02. The size of generated pseudo image is set to 600*600, and the dimension of VP is 2,560 according to the image encoder of BriVL.

For the classification process, the weight of visual prompt is set as $\alpha = 1$, and the prompt length is set as $N = [15, 20, 40, 10, 5]$ for THUCNews, Toutiao News, Inews, Iflytex, and CSLDCP, respectively. We optimize the soft prompts using Adam with the learning rate 1e-5.

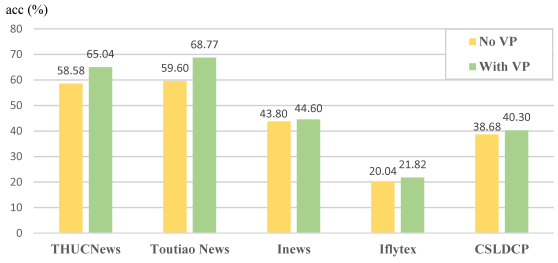
4.3 Evaluation and Training Protocol

It is commonly accepted that fine-tuning on small datasets can suffer from instability and results may change dramatically given a new split of data. To obtain a robust measure of the model performance, we follow the setting of existing works (Gao et al., 2021b; Devlin et al., 2019). Concretely, on each dataset, we randomly sample $k * M$ labeled training samples from the training set and $k*M$ labeled validation samples from the validation set for few-shot fine-tuning, which is repeated five times. The average performance across five repeated trials is reported. Note that the number of validation samples is set the same as the number of training samples during few-shot fine-tuning on each dataset. Although few-shot fine-tuning using a larger set of validation samples leads to significant improvements (Gao et al., 2021b), its initial goal of learning from limited data is subverted.

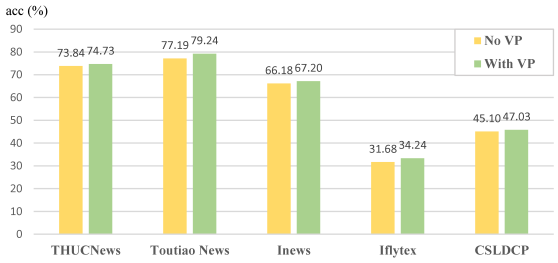
4.4 Main Results

We compare our proposed VPT with a series of few-shot learning methods based on PLMs, including PET (Schick and Schütze, 2021), LM-BFF (Gao et al., 2021b), P-tuning (Liu et al., 2021b) and EFL (Devlin et al., 2019). The widely-used RoBERTa-large is adopted as the backbone for these methods. In addition, we apply our framework without visual prompt generation to the original RoBERTa-large as a baseline, denoted as "soft prompt". The comparative results are shown in Table 2. We compare methods based on VL-PTM and PLM, with two backbones (text encoders) of similar sizes: BriVL and RoBERTa-large, respectively. Note that the four PLM-based baselines typically adopt the pre-trained masked language modeling (MLM) head for prompt learning and thus we cannot apply BriVL (without the MLM head) as their backbone. Moreover, since the text embedding of RoBERTa-large is not aligned with the visual prompt, it is unreasonable to apply RoBERTa-large as the backbone to our proposed VPT model.

We can clearly observe from Table 2 that our proposed VPT consistently outperforms the recent state-of-the-art methods for few-shot text classification on all five datasets. Particularly, our proposed VPT yields more than 2.5% improvements over the second best on Toutiao News and CSLDCP. These observations indeed identify the important role of our proposed VPT as a better approach to few-shot text classification. Although two different text backbones (i.e., BriVL and RoBERTa-large) of similar sizes have been employed, these observations are still remarkable since the pre-training data of BriVL does not bring benefits as expected (see Table 3). Moreover, PLM-based prompt methods demonstrate unstable performance across text classification datasets



(a) 'Zero-Shot'



(b) 'Few-Shot'

Figure 3: Ablation study results of the proposed VPT with the large-scale pre-training model BriVL as the backbone. Average accuracy (%) on 5 repeated trials is reported. VP stands for visual prompt.

with different data distributions, while our proposed VPT demonstrates great robustness in few-shot text classification.

4.5 Ablation Study Results

We conduct ablation studies to show the contribution of the visual prompt. We run experiments with and without visual prompt in both zero-shot and few-shot scenarios, using the large-scale pre-training model BriVL as the backbone. The ablation results are shown in Figure 3. We have two main observations. Firstly, after adding visual prompt into few-shot prompt-tuning (see the comparison VPT vs. VPT w/o VP in Figure 3(b)), the few-shot performance increases by 1.7% in average, as compared with that using standard prompt-tuning alone (i.e., VPT w/o VP). Secondly, by directly adding visual prompt into zero-shot classification, the zero-shot performance of BriVL can be improved by 4.0% in average (see the comparison Zero-shot w/ VP vs. Zero-shot w/o VP in Figure 3(a)). These evidences clearly show that the visual prompt is indeed beneficial for deploying BriVL in few-shot text classification.

Note that the usage of visual prompt at test time is clearly shown in the experiments in Figure 3(a). Our visual prompt brings considerable boost in zero-shot scenario, and only the inference process is included. That is, we conduct the classification by directly computing the cosine similarity scores between the embeddings of input sentences and the embeddings of category names. After adding visual prompts to the embeddings of category names, each input sentence is forced to be matched with not only textual but also visual semantics of class names. Therefore, visual prompts serve as augmentations of text embeddings of class names at test time.

Table 3: Results obtained by base models using different pre-training data. The average accuracy (%) over all five datasets is reported for each model.

Model	Zero-shot	Soft Prompt
RoBERTa-base	30.39	51.48
RoBERTa-base (finetune)	28.88	51.92
BriVL w/ RoBERTa-base	**38.85**	**55.45**

Furthermore, we notice that the pre-training data of BriVL is different from that of RoBERTa, which may cause a bit of unfairness in the comparison of our experiments in Table 2. Therefore, we make comparison among the following three models: (1) RoBERTa-base; (2) RoBERTa-base (finetune): we finetune the pre-trained RoBERTa-base on the text data of 22 million image-text pairs, which is the same pre-training dataset of BriVL w/ RoBERTa-base; (3) BriVL w/ RoBERTa-base: it is a smaller version (using RoBERTa-base as backbone instead) of the standard BriVL, which is pre-trained with the aforementioned 22 million image-text pairs. Due to the limited GPU resource, only the base models are considered here. The ablation results in Table 3 show that RoBERTa-base (finetune) yields only slight improvements (or even performance drops) over RoBERTa-base, while BriVL w/ RoBERTa-base outperforms RoBERTa-base by large margins. This is mainly due to that the text data from large-scale image-text pairs has not been filtered (without any reprocessing), and PLMs like RoBERTa can hardly benefit from this noisy data. Therefore, the observations/conclusions from Table 2 can still be drawn, even if RoBERTa-large is first finetuned for the baselines with the pre-training data of BriVL.

Finally, it would be easier to demonstrate the advantages of our VPT by examining the impact of visual prompt on accuracy when alternative image representations are used for generating visual prompt. Concretely, we consider three im-

Table 4: Results obtained by using different image representations for generating visual prompt. Results of few-shot text classification are reported only on Iflytex.

Image Representation	Accuracy
Random Noise	7.81
Visual Prompt (1K iterations)	30.94
Visual Prompt (2K iterations)	**34.24**

age representations for generating visual prompt: (1) Random Noise: visual prompt initialized by random noise (without optimization); (2) Visual Prompt (1K iterations): low-quality visual prompt obtained only with 1K iterations of optimization; (3) Visual Prompt (2K iterations): standard visual prompt obtained with 2K iterations of optimization. Note that the results of few-shot text classification are reported only on the small dataset Iflytex for quick evaluation. Two observations can be drawn from the ablation results in Table 4. Firstly, the visual prompt initialized by random noise (without optimization) causes serious damage to the performance of our VPT for few-shot text classification. Secondly, the visual prompt obtained with 2K iterations of optimization leads to 4.7% improvements over that obtained with 1K iterations of optimization, showing that the visual prompt of higher quality brings larger benefits to our VPT for few-shot text classification.

4.6 Influence of Hyperparameters

In this subsection, we discuss the influence of two crucial hyperparameters on the performance of VPT: prompt length – N, and weight of VP – α. The detailed results are provided in Figure 4. Only two small datasets Iflytex and CSLDCP are considered for quick evaluation.

We first conduct experiments by varying the prompt length in {1; 5; 20; 40; 60; 80; 100}, while fixing the rest hyperparameters. Figure 4(a) and Figure 4(c) show that increasing prompt length is beneficial for better performance when the prompt length is modest. When the prompt length is further increased (e.g., more than 5), the performance tends to gradually deteriorate. Notably, VPT has a strong performance even with a single prompt token. This indicates that VPT inherits the advantages of high efficiency of prompt-tuning. Increasing the token count above 20 leads to marginal gains (or even drops). Going above 60 tokens appears to be consistently damaging on

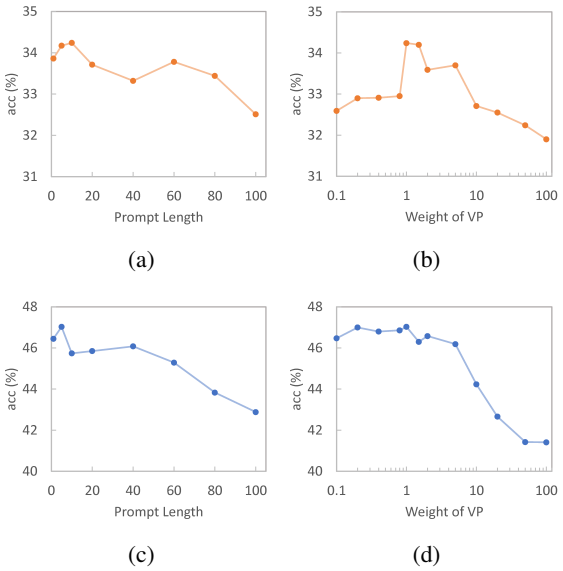


Figure 4: Effect of hyperparameters on the performance of VPT. Results on two small datasets (Iflytex, CSLDCP) are reported. The orange line refers to Iflytex ((a), (b)) and the blue one refers to CSLDCP ((c), (d)). (Left) On prompt length: employing more tokens for VPT leads to improvements when the number of prompt tokens is small. (Right) On weight of VP: VPT performs the best on both datasets when $\alpha = 1$.

both datasets. (Lester et al., 2021) discovered a similar pattern of declining performance beyond a particular prompt length. In addition, in Figure 4(a) and Figure 4(c), the minimum prompt length is 1. If we set the prompt length to 0 (i.e., the soft prompt is not used), the performance drops significantly. This means that the soft prompt does lead to significant improvements in text classification.

Further, the influence of weight of VP is explored in the range from 0.1 to 100, while the other hyperparameters are fixed. As shown in Figure 4(b) and Figure 4(d), increasing the weight α yields performance improvements when $\alpha < 1$ and causes negative effects when $\alpha > 1$. Notably, when $\alpha >> 1$, the performance of VPT still outperforms the baselines like Soft Prompt. This observation demonstrates that the generated VP indeed contains rich semantic information inherited from the class name. However, VP can only play an auxiliary role, i.e., it cannot take the place of the textual class name.

4.7 Visualization Results

To directly figure out the effect of visual prompt, we visualize the obtained pseudo images in the visual prompt generation process in Figure 5. Note

(a) "automobile" (b) "furniture" (c) "camera" (d) "machinery" (e) "gardening"

(f) "culture" (g) "art" (h) "military" (i) "technology" (j) "positive emotion"

Figure 5: Visualizations of generated images for different class names.

that the input texts of class names are originally in Chinese and translated into English for illustration purpose. For text descriptions with concrete meanings, the generated visualizations provide intuitive pictures (e.g., "automobile": a car; "furniture": table, chair and vase; "camera": a telecamera; "machinery": some gears and pipes; "gardening": morning glories). For text descriptions with abstract meanings, the generated visualizations are able to show concrete embodiment of these concepts (e.g., "culture": a figure in traditional Chinese clothing with Chinese "culture" in the upper right corner; "art": geometric figures and abstract portraits; "military": soldiers, a military aircraft and a armored vehicle; "technology": monitors and consoles; "positive emotion": a smiling face in the bottom left). Overall, these visualization results clearly demonstrate that our visual prompts have actually be learned to well represent the semantic content of the corresponding class names.

5 Conclusion

We propose a novel prompt-based method termed Visual Prompt Tuning (VPT) for deploying VL-PTM like BriVL in few-shot text classification. The main component of our proposed VPT is a visual prompt generation module based on model inversion of VL-PTM. Extensive experimental results on five benchmark datasets demonstrate that our proposed VPT achieves the new state-of-the-art in few-shot text classification. The ablation study and visualization results further show the effectiveness of our proposed VPT. In our ongoing research, we will apply our proposed VPT to other few-shot NLP tasks.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. This work was supported in part by National Natural Science Foundation of China (61976220 and 61832017), and Beijing Outstanding Young Scientist Program (BJJWZYJH012019100020098).

References

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent bert-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 657–668.

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, Volume 1 (Long and Short Papers), pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13:3094.

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. *arXiv preprint arXiv: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 (Volume 1: Long Papers)*, pages 3816–3830.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. 2019. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 558–567.

Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. 2021. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 4904–4916. PMLR.

Chen Jia, Yuefeng Shi, Qinrong Yang, and Yue Zhang. 2020. Entity enhanced bert pre-training for chinese ner. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6384–6396.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Big transfer (bit): General visual representation learning. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 491–507.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3045–3059.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Jingyang Li, Maosong Sun, and Xian Zhang. 2006. A comparison and semi-quantitative analysis of words and character-bigrams as features in chinese text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 545–552.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 121–137.

Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. 2021. M6: A chinese multimodal pretrainer. *arXiv preprint arXiv:2103.00823*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

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*.

Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. 2022. COTS: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15692–15701.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI Blog*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, pages 140:1–140:67.

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*, pages 255–269.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Derek Tam, Rakesh R Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4980–4991.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Liang Xu, Xiaojing Lu, Chenyang Yuan, Xuanwei Zhang, Huilin Xu, Hu Yuan, Guoao Wei, Xiang Pan, Xin Tian, Libo Qin, et al. 2021. Fewclue: A chinese few-shot learning evaluation benchmark. *arXiv preprint arXiv:2107.07498*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2021. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2021. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*.