PDAMeta: Meta-Learning Framework with Progressive Data Augmentation for Few-Shot Text Classification

Xurui Li[1], **Kaisong Song**[3,1†], **Tianqianjin Lin**[2], **Yangyang Kang**[1†]
Fubang Zhao[1], **Changlong Sun**[1], **Xiaozhong Liu**[4†]

[1] Institute for Intelligent Computing, Alibaba Group, China,
[2] Zhejiang University, China, [3] Northeastern University, China, [4] Worcester Polytechnic Institute, USA
xurui.lee@msn.com, lintqj@zju.edu.cn, changlong.scl@taobao.com, xliu14@wpi.edu
{kaisong.sks, yangyang.kangy, fubang.zfb}@alibaba-inc.com,

Abstract

Recently, we have witnessed the breakthroughs of meta-learning for few-shot learning scenario. Data augmentation is essential for meta-learning, particularly in situations where data is extremely scarce. However, existing text data augmentation methods can not ensure the diversity and quality of the generated data, which leads to sub-optimal performance. Inspired by the recent success of large language models (LLMs) which demonstrate improved language comprehension abilities, we propose a **Meta**-learning framework with **P**rogressive **D**ata **A**ugmentation (**PDAMeta**) for few-shot text classification, which contains a two-stage data augmentation strategy. First, the prompt-based data augmentation enriches the diversity of the training instances from a global perspective. Second, the attention-based data augmentation further improves the data quality from a local perspective. Last, we propose a dual-stream contrastive meta-learning strategy to learn discriminative text representations from both original and augmented instances. Extensive experiments conducted on four public few-shot text classification datasets show that PDAMeta significantly outperforms several state-of-the-art models and shows better robustness.

Keywords: Meta-learning, Data Augmentation, Large Language Model, Few-shot Learning

1. INTRODUCTION

The effectiveness of natural language processing (NLP) models heavily relies on the quality and quantity of the training data. The challenge of training data insufficiency is especially prominent in few-shot learning (FSL) scenarios, where the model trained on the source domain data is expected to generalize from only a few examples to the target domain. Existing FSL methods have shown promising results by overcoming this challenge in various tasks, which mainly focus on improving the learning and generalization capability of the model via meta-learning (Yin, 2020; Lee et al., 2022a) or prompt-based methods (Lester et al., 2021; Han et al., 2022; Wang et al., 2022). However, the performance of all these methods is still intrinsically limited by the data quality and quantity in both the source and target domains.

To mitigate the challenge of training data insufficiency, various text data augmentation methods are widely used and work well together with other FSL methods in NLP (Wei and Zou, 2019; Kumar et al., 2019; Sun et al., 2021). Traditional text data augmentation methods are usually model-agnostic and rely on direct operations on the training samples, such as *synonym replacement*, *random deletion*, and *random insertion* (Feng et al.). However, all these methods only make local or word-level changes in the samples, and they can not generate sufficient and effective training data. Given the exemplar sentence $s_i =$ "*Argentina won the world cup, thanks to the great football player Messi.*" in Fig. 1(a), *random deletion* masks the useless stop-word "*the*" but such operation makes no semantic changes, *synonym replacement* is another commonly used data augmentation technique but it heavily relies on the size and quality of hand-crafted synonym lexicons. In comparison, we prefer to make local changes based on automatic selection of the prominent words. Thus, an attention-based data augmentation method is proposed, which achieves better text representation via carefully designed attention mechanisms.

More recent studies explore the possibilities of language models, and tray to generate reliable training samples for more effective data augmentation, including back translation (Xie et al., 2020), seq2seq generation (Yoo et al., 2020; Zhang et al., 2020) and word vector interpolation in the latent space (Jindal et al., 2020; Bayer et al., 2023). All these methods make changes in the training data from the global or sentence-level perspective, however they are still limited in the accuracy and diversity of the generated training data. Recently, the advent of large language model (LLM) such as the GPT family (Yoo et al., 2021; Dai et al., 2023) bring new opportunities for generating text samples based on carefully designed prompting and significantly alleviate the burden of human annotators. In Fig.1 (b), *back translation* can generate simi-

† Corresponding authors.

LREC-COLING 2024, pages 12668–12678

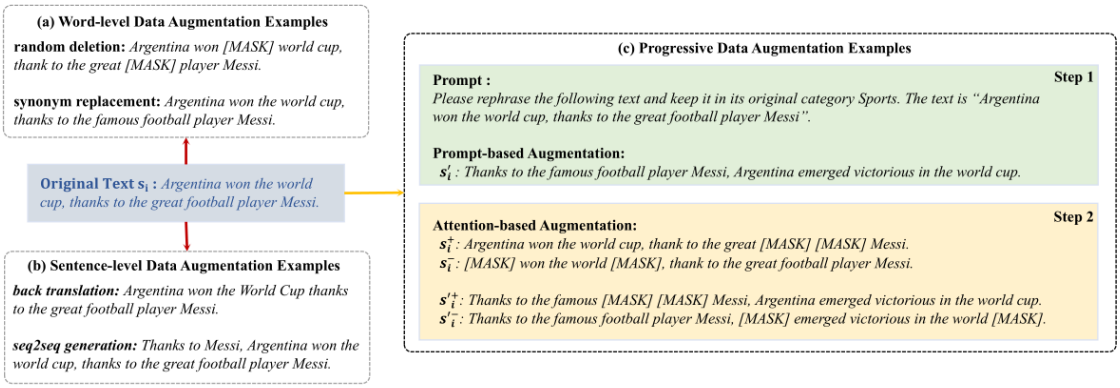

Figure 1: Comparison among different data augmentation strategies. (a) illustrates the word-level data augmentation from local perspective. (b) illustrates the sentence-level data augmentation from global perspective. (c) illustrates our progressive data augmentation.

lar samples but still lack of syntax and vocabulary changes, while *seq2seq generation* easily results in syntax redundancy or errors. In comparison, prompt-based data augmentation has been validated effectiveness in alleviating the issue of data scarcity, which can capture the semantics from global perspective and generate sufficient similar samples with well-designed prompting.

Intuitively, neither word-level nor sentence-level data augmentation method alone can overcome such data challenges in the few-shot learning scenario well. Different from previous work, we propose a novel progressive data augmentation strategy, which contains a two-stage process. First, the prompt-based data augmentation method generates sufficient training samples with prompting from a global perspective. Second, an attention-based data augmentation method further changes the training samples from a local perspective. We have demonstrated our strategy in Fig. 1(c), where the quantity and quality of training samples (i.e., positive and negative sample) can be enhanced step by step. The generated instances will finally be used for learning discriminative text representation.

In this paper, we propose a novel Meta-learning framework with Progressive Data Augmentation (PDAMeta) for few-shot text classification, which overcomes the challenge of data scarcity by a progressive data augmentation method, and learns better discriminative representations via a dual-stream contrastive meta-learning method. In summary, we make the contributions in three-folds:

- We propose a progressive data augmentation strategy for meta-learning, which first uses a prompt-based data augmentation method to generate sufficient training samples from a global perspective, and then uses a well-designed attention-based data augmentation method to improve the data quality from a local perspective.
- We propose a novel dual-stream contrastive meta-learning method, which can learn better discriminative text representations from both the original samples and the augmented samples by supervised and unsupervised contrastive learning techniques.
- Extensive experiments conducted on four public few-shot text classification datasets validate that our PDAMeta meta-learning framework outperforms comparative methods significantly and achieves better robustness. *All the resources will be publicly available.*

2. METHODOLOGY

2.1. Problem Formulation

The meta-learning paradigm of few-shot text classification aims to learn a prior meta-knowledge over hypothesis from a sample of meta-training tasks for fast adaptation on meta-testing tasks. Formally, let $\mathcal{C}_{train}$ and $\mathcal{C}_{test}$ denote the disjoint set of training classes and test classes, i.e., they have no overlapping classes. In the meta-training process, we sample N classes (i.e., N-way) from $\mathcal{C}_{train}$. Then, we randomly sample K examples (i.e., K-shot) for each class as the support set $\mathcal{S} = \{s_i\}_{i=1}^{N\times K}$ and L examples as the query set $\mathcal{Q} = \{q_j\}_{j=1}^{N\times L}$, where $\mathcal{S} \cap \mathcal{Q} = \emptyset$. All the examples from the remaining classes are called source pool (Bao et al., 2020) and denoted as $\mathcal{O}$. Notation $s_i = (x_i, y_i)$ denotes the i_{th} instance in $\mathcal{S}$, $q_j = (x_j, y_j)$ denotes the j_{th} instance in $\mathcal{Q}$, x and y represent text and label, respectively. In the meta-training procedure, we use $\mathcal{S}$ to learn a base-learner, then we use $\mathcal{Q}$ to evaluate and update the base-learner by optimizing a meta-learner $F(\Phi)$, which is parameterized by a

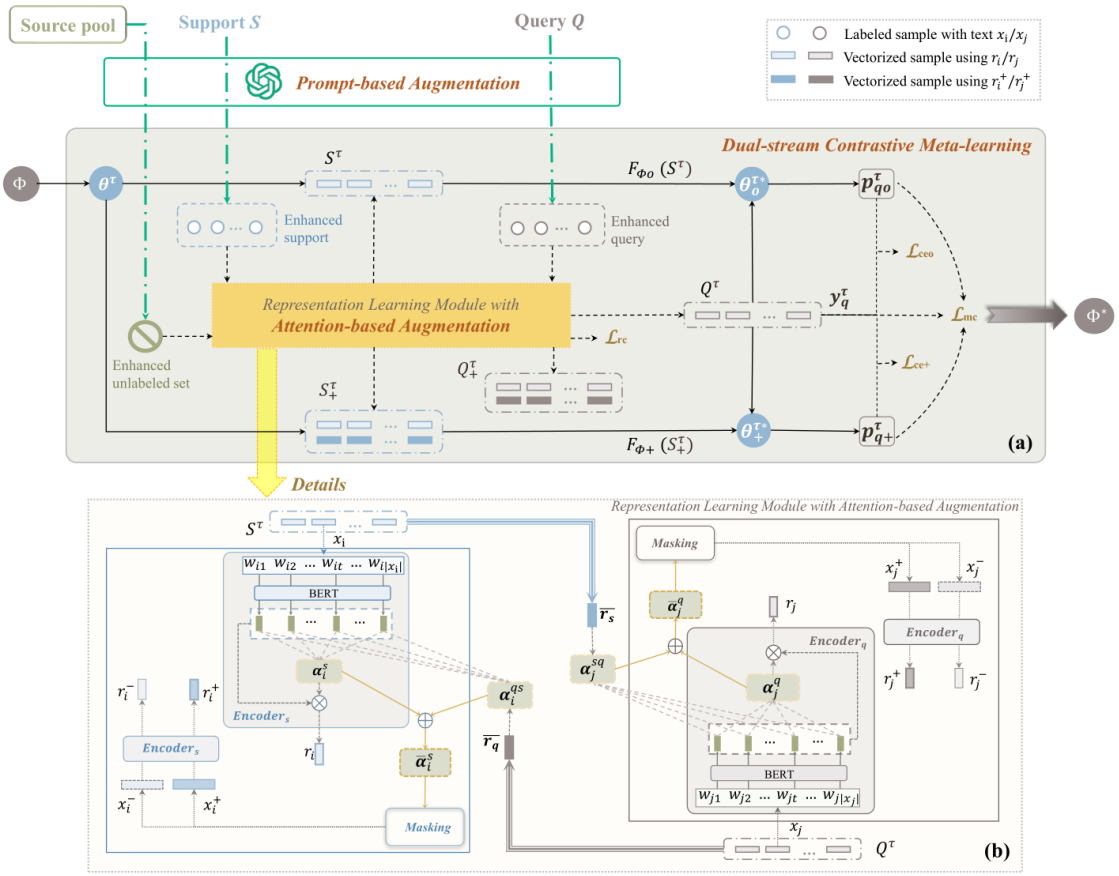

Figure 2: The Framework of PDAMeta. (a) The overall meta-learning skeleton integrated with progressive data augmentation and dual-stream contrastive meta-learning. $\mathbf{p}_{qo}^{\tau}$ represents the predictions of query set $\mathcal{Q}^{\tau}$ from vanilla base-learner with optimal parameters $\theta_o^{\tau*}$, and $\mathbf{p}_{q+}^{\tau}$ represents those from augmented base-learner with optimal parameters $\theta_+^{\tau*}$. (b) The details for the attention-based data augmentation.

global parameter set Φ. The procedure above is also called a training episode or task τ, which will be repeated multiple times (Han et al., 2021). The optimal global meta-parameters Φ^* can be finally used for fast adaption to the test classes $\mathcal{C}_{test}$ (Lee et al., 2022b).

2.2. Prompt-based Data Augmentation

In this work, we first use ChatGPT (Ouyang et al., 2022), a popular LLM tool to conduct data augmentation[1]. Compared to previous data augmentation methods, ChatGPT is more suitable for data augmentation for three reasons (Dai et al., 2023):

- ChatGPT is pre-trained on large-scale corpora, so it has a broader semantic expression space for generating more diverse data.
- The supervised fine-tuning stage of ChatGPT introduces numerous manual annotations, the language generated by ChatGPT is more in line with human expression habits.
- Reinforcement learning stage of ChatGPT can compare the advantages and disadvantages of different expressions and ensure it to generate more informative and impartial data.

Our prompt-based data augmentation is used in two ways. On one hand, we augment the support set and query set for discriminative text representation with supervised learning. On the other hand, to prevent training overfitting, we augment unlabeled data in the source pool for learning more robust text representations with unsupervised learning.

Specifically, we use ChatGPT to augment both the labeled and unlabeled data, and we design the prompts in a single-turn dialogue (Dai et al., 2023). For augmenting the labeled data, the exemplar prompt template is "*Please rephrase the following text and keep it in its original category* $\{category\}$. *The text is* $\{text\}$". The slots $\{text\}$ and $\{category\}$ can be instantiated with specific text and corresponding category. For augmenting the unlabeled data, the prompt can be "*Please rephrase the following text. The text is* $\{text\}$". More exemplar prompts have been displayed in

[1]Our method can be easily adaptive to other LLMs.

Fig. 1(c). In this work, we combine the original data and the ChatGPT augmented data into the enhanced support and query sets, which are then fed into our text representation learning module.

In practical scenario, direct application of ChatGPT or other LLMs for data augmentation is still problematic, because it highly depends on the selection of specific LLMs and the design of prompts. In addition, the generated instances may be very similar. For better generalization, we will consider making automatic changes of local words with the help of attention mechanisms.

2.3. Attention-based Data Augmentation

Compared with previous data augmentation methods (Sun et al., 2021; Ni et al., 2021), our attention-based data augmentation contributes in two ways: (1) The attention-based data augmentation is the successor to the first-step prompt-based augmentation, and focuses on selection of prominent words for masking. (2) Our data augmentation operation is applied on both support set and query set, meanwhile reduces their distribution difference via interactive attention mechanisms.

Text Encoder. In the support stage, the input $x_i = [w_{i1}, ...w_{it}..., w_{i|x_i|}]$ is a text and $|x_i|$ is the text length. We transform each x_i into a sequence of hidden states $\{\mathbf{h}_{it}\}$ via a pre-trained language model BERT (Devlin et al., 2019). Afterwards, we use attention mechanism to select most important words to obtain informative text representation. We denote the text encoder in Fig. 2 (b) as $Encoder_s$. The attention weight $\alpha_{it}^s \in (0,1)$ of the word w_{it} in any text x_i can be formulated as below:

$$\begin{aligned} \alpha_{it}^s &= \frac{\exp(\mathbf{u}_{it}^\top \mathbf{U}_w)}{\sum_{k=1}^{|x_i|} \exp(\mathbf{u}_{ik}^\top \mathbf{U}_w)} \\ \mathbf{u}_{it} &= \tanh(\mathbf{W}_w \mathbf{h}_{it} + \mathbf{b}_w) \end{aligned} \tag{1}$$

where $\mathbf{W}_w$, $\mathbf{b}_w$ and $\mathbf{U}_w$ are learnable model parameters and $\exp(\cdot)$ is an exponential function. The final text representation $\boldsymbol{r_i}$ is the weighted summation of all the hidden states $\{\mathbf{h}_{it}\}$ can be formulated as: $\boldsymbol{r_i} = \sum_{t\in[1,|x_i|]} \alpha_{it}^s \mathbf{h}_{it}$.

Considering the symmetrical characteristics of the augmentation during support and query phases, we mainly elaborate the operations for support phase here. In the query stage, we replace the subscript i with j for the corresponding calculations in support stage. We obtain the attention weight $\alpha_{jt}^q \in (0,1)$ for each word in the text x_j. Then, we can obtain the sentence representation $\boldsymbol{r_j}$ of the input x_j from text encoder $Encoder_q$.

Attention From Query to Support. As shown in Fig. 1(b), random deletion is problematic for data augmentation. Intuitively, the prominent words should be paid more attention because of higher attention values, while the non-prominent words usually have lower attention values (Moon et al., 2021). The attention weight α_{it}^s in Formula 1 obviously provides prior guidance for word selection.

In addition, we aim to mitigate the distribution difference between the support and query instances, which contributes to meta-knowledge transferring. Specifically, we use the query representation $\bar{\boldsymbol{r}}_q$ to guide the selection of prominent words in the support set. The attention weight α_i^{qs} can be formulated as below:

$$\begin{aligned} \alpha_{it}^{qs} &= \frac{\exp(\bar{\mathbf{u}}_{it}^\top \bar{\boldsymbol{r}}_q)}{\sum_{k=1}^{|x_i|} \exp(\bar{\mathbf{u}}_{ik}^\top \bar{\boldsymbol{r}}_q)} \\ \bar{\mathbf{u}}_{it} &= \tanh(\bar{\mathbf{W}}_w \mathbf{h}_{it} + \bar{\mathbf{b}}_w) \\ \bar{\boldsymbol{r}}_{\boldsymbol{q}} &= AvgPooling(\{\boldsymbol{r}_j\}) \end{aligned} \tag{2}$$

where $\bar{\mathbf{W}}_w$ and $\bar{\mathbf{b}}_w$ are learnable model parameters. $\bar{\boldsymbol{r}}_q$ is the averaged representation of all the samples in the query set $\mathcal{Q}$.

Finally, we combine the influence of both attention weight α_{it}^s and α_{it}^{qs} and obtain the final attention weight $\bar{\alpha}_{it}^s = \frac{1}{2}(\alpha_{it}^s + \alpha_{it}^{qs})$ for guiding word masking. Specifically, we aim to generate positive instance x_i^+ and negative instance x_i^- with the guidance of attention weight $\bar{\alpha}_{it}^s$, and build triplets (x_i, x_i^+, x_i^-). Negative instance x_i^- is generated by masking the words with the highest attention weights. Positive instance x_i^+ is generated by masking the words with top-k lowest attention weight. However, directly sampling words from a discrete distribution is a non-differentiable operation and makes the standard back-propagation invalid. Thus, we use *Gumbel-Softmax* (Gu et al., 2018; Paulus et al., 2020) trick to approximate the discrete operation. The probability distribution for the w_{it} being selected in a single Gumbel-Softmax sampling step is:

$$m_{it} = \frac{\exp(log(\bar{\alpha}_{it}^s) + g_{it})/\kappa_g}{\sum_{j=1}^{|x_i|} \exp(log(\bar{\alpha}_{ij}^s) + g_{ij})/\kappa_g} \tag{3}$$

where gumbel noise $g_{it} = -log(-log(u_{it}))$ and $u_{it} \sim Uniform(0,1)$, κ_g is a temperature hyper-parameter. In the backward pass we simply use the continuous value, thus the error signal is able to back-propagate. In the forward propagation, we discretize the continuous probability vector sampled from the *Gumbel-Softmax* distribution into a one-hot vector $\mathbf{z}_i$, where

$$z_{it} = \begin{cases} 1 & m_{it} = argmax_j m_{ij}, \\ 0 & otherwise. \end{cases} \tag{4}$$

If $z_{it} = 1$, the word w_{it} will be masked by a special token $[\mathrm{MASK}]$. We repeat the operation until p percent of words in one instance are masked, which produces the negative instance $\mathbf{x}_i^-$. To generate

positive instance $\mathbf{x}_i^+$, we mask the words with lowest attention values via a revered attention weight $(1-\bar{\alpha}_{it}^s)/\sum_{t=1}^{|x_i|}(1-\bar{\alpha}_{it}^s)$.

Attention From Support to Query. Symmetrically, we use the representation $\bar{\boldsymbol{r}}_s$ of the support set to select prominent words in the query set. The attention weight α_j^{sq} for each word w_j can be obtained. $\bar{\boldsymbol{r}}_s$ is the averaged representation of all the samples in the support set. We combine the influence of both attention weights α_{jt}^q and α_{jt}^{sq}, and obtain the attention weight $\bar{\alpha}_{jt}^q = \frac{1}{2}(\alpha_{jt}^q + \alpha_{jt}^{sq})$ for guiding word masking.

Contrastive Text Representation. In order to learn discriminative text representations between similar classes, we implement two types of contrastive constraints. Specifically, there are $N_b^s = N \times K$ for original support set and $N_b^q = N \times L$ for original query set. $\boldsymbol{r}_i$ is the representation vector of instance x_i, $\boldsymbol{r}_i^+$ and $\boldsymbol{r}_i^-$ are representation vectors of the samples $\boldsymbol{x}_i^+$ and $\boldsymbol{x}_i^-$ from the attention-based augmentation.

The first contrastive learning is to ensure the effectiveness of the masking strategy, by forcing the masked positive sample more close to original sample then masked negative sample. By combing all the triplets of $\{(\boldsymbol{r}_i, \boldsymbol{r}_i^+, \boldsymbol{r}_i^-)\}$ from the progressive data augmentation into one task batch with $N_b = 2N \times (K+L)$ samples, we can obtain the following margin-based ranking loss:

$$\mathcal{L}_b = \frac{1}{N_b}\sum_{i=1}^{N_b}\max(0, \Delta + d(\boldsymbol{r}_i, \boldsymbol{r}_i^+) - d(\boldsymbol{r}_i, \boldsymbol{r}_i^-)) \quad (5)$$

where Δ is a margin value, $d(\boldsymbol{r_i}, \boldsymbol{r_j}) = \|\boldsymbol{r_i} - \boldsymbol{r_j}\|_2$ denotes the distance between the representations.

The second contrastive learning is to build a more comprehensive constraint across different samples in each batch for better discriminative text representation learning. Different from the simple dropout augmentation in SIMCSE (Gao et al., 2021), we use the progressive data augmentation to form the positive counterpart. We group the pairs of $(\boldsymbol{r_i}, \boldsymbol{r_i^+})$ from the progressive data augmentation for both support and query with that of the N_u unlabeled data to form a batch with $N_{bu} = N_b + N_u$ samples. The representation $\boldsymbol{r_i}$ for unlabeled data is encoded by BERT, and the positive counterpart $\boldsymbol{r_i^+}$ is ChatGPT-augmented $\boldsymbol{r_i}$. Finally, a normalized temperature-scaled loss is used as below:

$$\mathcal{L}_u = \frac{-1}{N_{bu}}\sum_{i=1}^{N_{bu}} log \frac{\exp(\frac{\delta(\boldsymbol{r_i}, \boldsymbol{r_i^+})}{\kappa_c})}{\sum_{j=1}^{N_{bu}}(\exp(\frac{\delta(\boldsymbol{r_i}, \boldsymbol{r_j})}{\kappa_c}) + \exp(\frac{\delta(\boldsymbol{r_i}, \boldsymbol{r_j^+})}{\kappa_c}))} \quad (6)$$

where notation κ_c is a temperature hyperparameter, $\delta(\boldsymbol{r_i}, \boldsymbol{r_j})$ is the cosine similarity score between the representation $\boldsymbol{r_i}$ and $\boldsymbol{r_j}$.

Finally, we combine all the contrastive objectives for samples as the merged local constraints for learning a better representation:

$$\mathcal{L}_{rc} = \mathcal{L}_b + \mathcal{L}_u \quad (7)$$

2.4. Dual-stream Contrastive Meta-learning

Meta-learning aims to learn prior meta-knowledge across tasks to achieve fast adaptation to the specific task. We propose an innovative dual-stream contrastive learning as a global constraint, which force the meta-learner to capture better latent meta-knowledge. The vanilla meta-learner updates the original base-learner (trained on original support set), while the augmented meta-learner updates another base-learner (trained on augmented support set) based on the augmented query set in the outer-loop optimization. Specifically, we conduct the contrastive procedure between the outputs from original and augmented meta-learner for the query set respectively.

As shown in Fig. 2, we define S^τ and Q^τ are original support and query samples of task τ, while the samples for S_+^τ and Q_+^τ are the positive instances augmented by our proposed progressive data augmentation. In our PDAMeta, we train two different meta-functions $F_{\Phi o}$ and $F_{\Phi +}$, where $F_{\Phi o}(S^\tau)$ represents that the inner-loop update is based on the original support set S^τ and $F_{\Phi +}(S_+^\tau)$ represents the one based on the progressively augmented support set S_+^τ. Then, we use Q^τ and Q_+^τ to evaluate the corresponding optimal task-specific parameters $\theta_o^{\tau *}$ and $\theta_+^{\tau *}$, respectively.

Our objective is to train a classifier capable of acquiring meta-knowledge from the support set. This enables the classifier to rapidly assimilate knowledge from a small number of annotations when tasked with classifying previously unseen classes. Inspired by previous optimization-based methods, we also adopt the ridge regression to fit the labeled support set (Bertinetto et al., 2018; Bao et al., 2020; Han et al., 2021). It admits a closed-form solution that enables end-to-end differentiation through the model with proper regularization. During query training phase, suppose $\boldsymbol{y_q^\tau}$ is the ground-truth labels of Q^τ. Finally, we can utilize the *cross-entropy* as loss function for the outer-loop optimization objective of N-way classification. Let the prediction vectors for Q^τ from $\theta_o^{\tau *}$ denoted as $\boldsymbol{p_{qo}^\tau}$, and that from $\theta_+^{\tau *}$ denoted as $\boldsymbol{p_{q+}^\tau}$. The *cross-entropy* loss objective calculated from $\boldsymbol{p_{qo}^\tau}$ can be written as:

$$\mathcal{L}_{ceo} = -\frac{1}{N_b^q}\sum_{i=1}^{N_b^q}\sum_{j=1}^{N} y_{ij}^\tau \log p_{ijo}^\tau \quad (8)$$

where y_{ij}^τ is the j_{th} label of y_i^τ. The subscript j in p_{ijo}^τ represents the probability that the outputs p_{qi}^τ

DataSet	train/val/test classes	examples	vocab size
HuffPost	20/5/16	36,900	8218
Amazon	10/5/9	24,000	17062
Reuters	15/5/11	620	2234
20News	8/5/7	18,820	32137

Table 1: Statistics of the four benchmark datasets.

been predicted to be the j_{th} label. The definition of the classification loss objective $\mathcal{L}_{ce+}$ calculated from $\boldsymbol{p}^{\tau}_{q+}$ resembles that of $\mathcal{L}_{ceo}$ by replacing the tail subscript o into $+$. The combined *cross-entropy* loss objective is $\mathcal{L}_{ce} = \mathcal{L}_{ceo} + \mathcal{L}_{ce+}$.

Then we implement the meta-aspect contrastive loss objective and force the $\boldsymbol{p}^{\tau}_{q+}$ more closely to $\boldsymbol{y}^{\tau}_{q}$ than $\boldsymbol{p}^{\tau}_{qo}$:

$$\mathcal{L}_{mc} = \frac{1}{N^q_b} \sum_{i=1}^{N^q_b} \max(0, \Delta + d(\boldsymbol{y}^{\tau}_{q}, \boldsymbol{p}^{\tau}_{q+}) - d(\boldsymbol{y}^{\tau}_{q}, \boldsymbol{p}^{\tau}_{qo})) \tag{9}$$

Finally, we combine all the loss objective from different parts into the final objective function:

$$\mathcal{L}_{all} = \mathcal{L}_{ce} + \beta \mathcal{L}_{rc} + \gamma \mathcal{L}_{mc} \tag{10}$$

where $\beta \in (0, 1)$ and $\gamma \in (0, 1)$ are hyper-parameters for the weights of loss $\mathcal{L}_{rc}$ and $\mathcal{L}_{mc}$ respectively. Since the dual-stream contrastive meta-learning applies additional constraint during learning the meta-knowledge across the support and query training phases, a better robustness can be achieved from the progressive augmentations.

3. EXPERIMENTS AND RESULTS

3.1. Datasets

To evaluate our PDAMeta for few-shot text classification, we conduct experiments on four public datasets, including HuffPost, Amazon, Reuters and 20News, which are widely used in FSL tasks (Bao et al., 2020; Han et al., 2021; Hou et al., 2022). The statistics of the datasets are given in Tab. 1.

3.2. Experimental Settings

We evaluate our models based on typical 5-way 5-shot and 5-way 10-shot text classification settings with $L = K$. We perform 100 episodes in each meta-training epoch with the early stopping at a maximum tolerance of 20 epochs without performance growth. We evaluate the performance on 1,000 episodes for meta-test. We run each experimental setting for 5 times over different random seeds and report the averaged performance. We use commonly used pre-trained language model *bert-base-uncased* as the basic text encoder for a fair comparison with the previous state-of-the-art method (Hou et al., 2022). However, our method is extensible and bert-base-uncased can be easily replaced with stronger pre-trained language models. All hyper-parameters are selected using greedy search on the validation set. The temperature factor κ_c for contrastive learning is set 0.1, and the temperature factor κ_g for *Gumbel-Softmax* is initialized to 10 and decreases during training using loss annealing strategy. The mask percent p is set to 15. The loss weights β and γ are all set to 0.2. Margin value Δ is set to 1. Unlabeled data number N_u is 2 times that of the labeled set. Parameters are optimized using Adam at the learning rate of 1e-5. Our programs are implemented by Pytorch and run on a server configured with a Tesla A100 GPU. We use the OpenAI's GPT-3.5-Turbo API to implement the ChatGPT experiments.

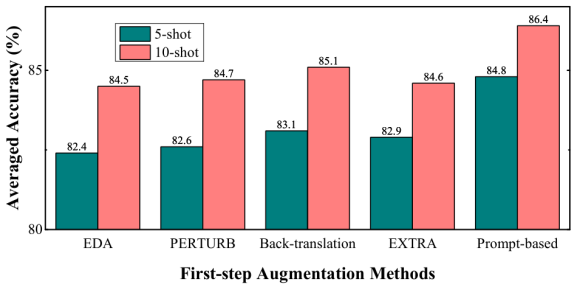

Figure 3: Accuracy (%) performance comparisons of models with different augmentation methods for the first-step of the progressive data augmentation.

3.3. Performance on Few-Shot Learning

We compare our PDAMeta with several few-shot text classification models: 1) **MAML** is a typical optimization-based method which learns easily adaptable model parameters through gradient descent (Finn et al., 2017); 2) **PN** is a typical metric-based method striving to learn effective distance measurements (Snell et al., 2017); 3) **R2D2** explores the feasibility of incorporating fast solvers with closed-form solutions as the base learning component for meta-learning (Bertinetto et al., 2018); 4) **HATT** extends the prototypical networks by incorporating a hybrid attention mechanism (Gao et al., 2019); 5) **Induction Networks** introduces dynamic routing algorithm to learn class-aspect representation (Geng et al., 2019); 6) **Relation Networks** simultaneously learns an embedding and a deep non-linear distance metric (Sung et al., 2018); 7) **DS-FSL** aims to extract more transferable features by mapping distribution signatures to attention scores (Bao et al., 2020); 8) **MLADA** introduces an adversarial domain adaptation network in meta-training episodes to extract domain invariant features and improve the adaptability of meta-learner in new tasks (Han et al., 2021); 9) **ContrastNet** tries to learn discriminative text represen-

Model	HuffPost		Amazon		Reuters		20News		Average	
	shot-5	shot-10	shot-5	shot-10	shot-5	shot-10	shot-5	shot-10	shot-5	shot-10
MAML	49.3	50.8	47.1	49.5	62.9	64.1	43.7	46.0	50.8	52.6
PN	41.3	42.9	52.1	54.4	66.9	68.7	45.3	46.9	51.4	53.2
R2D2	41.4	43.5	53.1	54.9	65.6	66.8	78.6	79.7	59.7	61.2
HATT	56.3	58.5	66.0	67.4	56.2	58.4	55.1	56.9	58.4	60.3
Induction Networks	49.1	50.4	41.3	42.1	67.9	69.2	33.3	34.7	47.9	49.1
Relation Networks	53.8	55.2	64.5	66.1	63.2	65.0	45.7	46.8	56.8	58.3
DS-FSL	63.5	64.8	81.1	82.9	96.0	97.2	68.3	69.8	77.2	78.7
MLADA	64.9	68.2	86.0	86.1	96.7	97.7	77.8	83.7	81.4	84.0
ContrastNet	65.3	70.7	85.2	85.7	95.3	96.6	81.6	83.9	81.9	84.2
ChatGPT	71.5	72.5	81.5	83.4	94.3	97.1	77.3	81.1	81.2	83.5
P-Tuning	65.8	68.4	79.1	82.6	96.7	97.8	71.5	76.2	78.3	81.4
MetaPrompting	**76.3**	**78.3**	85.5	87.3	97.2	97.7	76.6	78.2	83.9	85.4
PDAMeta	70.3	72.8	**87.2**	**89.4**	**98.5**	**98.9**	**83.1**	**84.5**	**84.8**	**86.4**

Table 2: Comparison among different few-shot text classification models. We run all experiments for 5 times and achieve the t-test result with $t \leq 0.005$.

tation by contrast the mean embeddings across different tasks (Chen et al., 2022); 10) **ChatGPT** is directly used for few-shot classification, with the prompt designed using in context learning template; 11) **P-Tuning** is a prompt-based method that uses masked language model to convert target tasks into cloze problems (Liu et al., 2021); 12) **MetaPrompting** employs optimization based meta-learning algorithm to find adaptive initialization for soft-prompt methods with pre-trained language models (Hou et al., 2022).

From Tab. 2, we can find that the averaged accuracy of our PDAMeta outperforms all baselines. Compared to methods such as R2D2, DS-FSL and MLADA with relatively weak text encoder, the BERT encoder in PDAMeta can learn better semantic representation. Compared with P-Tuning and ChatGPT using prompt-based methods for direct text classification, the carefully designed meta-learning framework and task-specific fine-tuning of PDAMeta improves the performance. Our method also outperforms ContrastNet owing to the novel progressive data augmentation overcomes the data challenge from both local and global perspectives with the help of LLM, as well as the novel dual-stream contrastive strategy. The multi-aspect contrastive constraints enhanced by prompt-based augmentation also help learning a better representation and meta-knowledge, and make our method outperforms MetaPrompting.

3.4. First-step Augmentation Comparisons

To further demonstrate the effectiveness of the first-step prompt-based augmentation, we compared the performances for models which replace the first-step prompt-based augmentation into other augmentation methods. In Fig. 3 we compared four typical data augmentation methods with ChatGPT: 1) EDA is easy data augmentation which adopts random deletion (Wei and Zou, 2019); 2) PERTURB applies both additive and multiplicative perturbation in feature space for data augmentation (Kumar et al., 2019); 3) Use Back-translation method; 4) EXTRA trains an auto-encoder to increase the variability for instances, resulting in a more robust model (DeVries and Taylor, 2017); The averaged performance across four datasets show that the prompt-based augmentation outperforms all other augmentation methods for the first-step of the progressive augmentation methods.

3.5. Robustness Analysis

To validate the robustness of the models, we intentionally modify the data distribution of support and query set in meta-learning. We first cluster each class into 2 subcategories for the original N classes using the K-means method. It is supposed that although these two subcategories belong to the same major class, there are still distinct differences in semantic distribution between them. While preparing an episode for robustness validation, the K instances of support and L instances of query are sampled from different subcategories for each major class.

The performances of datasets with the sampling strategy with intentionally designed shift for different models are show in Shift columns at Tab. 3. Comparing with the Original columns where we use randomly sampling from all N classes for support and query, we can see that all models decrease for Shift columns. It is because that the distribution for support and query are more different, it is more hard for model to learn transfer knowledge and cause the decrease. However, our PDAMeta model has the least decrease percentage, and outperforms state-of-the-art models by a large margin. Note that we run the experiments 5 times with different clustering and sampling and report the averaged performances, and achieve the t-test result with $t \leq 0.005$.

Model	Original		Shift		Decrease	
	shot-5	shot-10	shot-5	shot-10	shot-5	shot-10
MLADA	81.4	84.0	78.0	80.1	$\downarrow 4.2$	$\downarrow 4.6$
ContrastNet	81.9	84.2	78.9	81.2	$\downarrow 3.7$	$\downarrow 3.6$
MetaPrompting	83.9	85.4	80.7	82.7	$\downarrow 3.8$	$\downarrow 3.2$
PDAMeta	84.8	86.4	**83.5**	**84.9**	$\downarrow$ **1.5**	$\downarrow$ **1.7**

Table 3: Robustness study among different models. The "*Decrease*" columns represent the mean decrease percentage (%) of averaged accuracy for models under the distribution shift settings compared to that under the original settings.

Model	HuffPost		Amazon		Reuters		20News		Average	
	shot-5	shot-10	shot-5	shot-10	shot-5	shot-10	shot-5	shot-10	shot-5	shot-10
w/o prompt-based augmentation	65.6	68.6	86.4	86.6	96.9	98.0	77.9	83.9	81.7	84.3
w/o distribution alignment attention	69.3	71.6	86.9	87.6	97.5	98.0	81.9	83.9	83.9	85.3
w/o $\mathcal{L}_{rc}$&$\mathcal{L}_{mc}$	63.6	66.6	84.2	84.1	94.5	95.6	75.4	81.2	79.5	81.9
w/o $\mathcal{L}_u$	68.8	71.4	85.8	87.5	96.2	97.4	80.3	83.2	82.8	84.9
w/o $\mathcal{L}_{rc}$	66.4	68.4	84.9	86.2	95.6	95.9	79.8	81.7	81.7	83.1
w/o $\mathcal{L}_{mc}$	67.3	69.6	86.4	87.6	97.5	98.0	80.9	83.9	83.0	84.8
PDAMeta	**70.3**	**72.8**	**87.2**	**89.4**	**98.5**	**98.9**	**83.1**	**84.5**	**84.8**	**86.4**

Table 4: Ablation Study for the PDAMeta model. Here w/o means without the component.

3.6. Ablation Study

We implement several variants for ablation study for PDAMeta. The results are shown in Tab. 4. The first line labeled as "*w/o prompt-based augmentation* " means that we use a simple dropout instead of the ChatGPT for the first-step augmentation. The second line labeled as "*w/o distribution alignment attention*" means that we remove the attention α_i^{qs} and α_j^{sq}. For the following lines, we validate the effect for different contrastive constraints by removing them in turn. Finally, we add all the aforementioned components for the entire PDAMeta model. It can be found that the progressive augmentation and dual-stream contrast plays a positive role for the performance respectively. The key component distribution alignment also improves the performance during the augmentation.

3.7. Case Study

To better understand the usefulness of our progressive data augmentation, we show two cases in Fig. 4. Case 1 shows a positive effect for the distribution alignment which use the support set distribution to guide the prominence detection for a query instance. It can be seen that without using distribution alignment, words "*singing*" and "*dancing*" have the highest attention weights. It eventually led to the mistaken prediction from the ground-truth "*Parents*" into "*Entertainment*". By considering the distribution shift, the word "toddler" become the prominent one and help making a correct prediction. Case 2 also demonstrates an effective attention modification by using distribution alignment, but in an opposite direction from query to support. The category-independent words "*movie*", "*poster*" and "*dramatic*" for "*Parents*" are modified into "*toddler*", "*Dad's*" and "*bedtime*", which obviously pull the model's prediction to the right direction.

4. RELATED WORK

Meta-learning pursues the goal of fast adapting to new classes/domains/tasks given the experience of multiple learning episodes with rich annotations. Existing work on meta-learning can be divided into three categories: optimization-based, metric-based and model-based (Hospedales et al., 2022). MAML is a typical optimization-based method based on learning easily adaptable model parameters through gradient descent (Finn et al., 2017). Prototypical Networks (Snell et al., 2017) and R2D2 (Bertinetto et al., 2018) are typical metric-based methods and they strive to learn effective distance measurements. Model-based methods such as MANN (Chavan et al., 2022) aim at establishing models for fast learning, which is either archived by model's internal design or with the help of another meta-model. Nevertheless, meta-learning still suffers from low robustness problem due to the limited data in few-shot settings.

Robust representation learning is significant for the application of meta-learning. A typical solution strategy for robust representation learning is data augmentation. Contrastive learning is a recently popular method for robust representation learning (Falcon and Cho, 2020). Self-supervised contrastive learning methods such as SIMCSE (Gao

Alignment	Attention distribution	Prediction
Case 1: This singing and dancing toddler is what we all need right now. ***Label: Parents***		
None	This singing and dancing toddler is what we all need right now.	✗ *Entertainment*
Support->query	This singing and dancing toddler is what we all need right now.	✓ *Parents*
Support->query	**CHATGPT Augmented:** Right now, we all could benefit from this adorable toddler who sings and dances.	✓ *Parents*
Case 2: Dad's fake movie poster gives bedtime with a toddler the dramatic treatment it deserves. ***Label: Parents***		
None	Dad's fake movie poster gives bedtime with a toddler the dramatic treatment it deserves.	✗ *Entertainment*
Query->support	Dad's fake movie poster gives bedtime with a toddler the dramatic treatment it deserves.	✓ *Parents*
Query->support	**CHATGPT Augmented:** Dad creatively dramatizes bedtime with a toddler through a fake movie poster, emphasizing its deserving significance.	✓ *Parents*

Figure 4: Case study for the progressive data augmentation in Huffpost dataset. The darker the color, the more important the word is. We consider the influence of query (support) on support (query).

et al., 2021) make a global comparison among the input instance, augmented positive instance and in-batch negatives for improving the uniformity of the representation space. ContrastNet (Chen et al., 2022) is a representative work in meta-learning by contrasting the averaged representation across different tasks, and achieves good results in FSL tasks. However, all these works resort to simple and inadequate augmentation methods.

5. Conclusion

In this paper, we propose a PDAMeta meta-learning framework for the few-shot text classification. We first propose a progressive data augmentation method which overcomes the challenge of data from both local and global perspectives. In addition, we propose a novel dual-stream contrastive meta-learning method to learn better discriminative representations. In the future, we will apply PDAMeta to more few-shot tasks.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (62106039).

7. Bibliographical References

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with distributional signatures. In *Proceedings of the ICLR*.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2023. A survey on data augmentation for text classification. *ACM Comput. Surv.*, 55(7):146:1–146:39.

Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. 2018. Meta-learning with differentiable closed-form solvers. In *Proceedings of the ICLR (Poster)*.

Arnav Chavan, Rishabh Tiwari, Udbhav Bamba, and Deepak K. Gupta. 2022. Dynamic kernel selection for improved generalization and memory efficiency in meta-learning. In *Proceedings of the CVPR*, pages 9851–9860.

Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. 2022. Contrastnet: A contrastive learning framework for few-shot text classification. In *Proceedings of the AAAI*, pages 10492–10500.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, et al. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

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 NAACL-HLT*, pages 4171–4186.

Terrance DeVries and Graham W. Taylor. 2017. Dataset augmentation in feature space. In *Proceedings of the ICLR Workshop*.

William Falcon and Kyunghyun Cho. 2020. A framework for contrastive self-supervised learning and designing a new approach. *CoRR*.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. A survey of data augmentation approaches for NLP. In *ACL Findings*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the ICML*, pages 1126–1135.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the AAAI*, pages 6407–6414.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the EMNLP*, pages 6894–6910.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text classification. In *Proceedings of the EMNLP-IJCNLP*, pages 3902–3911.

Jiatao Gu, Daniel Jiwoong Im, and Victor O. K. Li. 2018. Neural machine translation with gumbel-greedy decoding. In *Proceedings of the AAAI*, pages 5125–5132.

Chengcheng Han, Zeqiu Fan, Dongxiang Zhang, Minghui Qiu, Ming Gao, and Aoying Zhou. 2021. Meta-learning adversarial domain adaptation network for few-shot text classification. In *ACL Findings*, pages 1664–1673.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022. PTR: prompt tuning with rules for text classification. *AI Open*, 3:182–192.

Timothy M. Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. 2022. Meta-learning in neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):5149–5169.

Yutai Hou, Hongyuan Dong, Xinghao Wang, Bohan Li, and Wanxiang Che. 2022. Metaprompting: Learning to learn better prompts. In *Proceedings of the COLING*, pages 3251–3262.

Amit Jindal, Arijit Ghosh Chowdhury, Aniket Didolkar, Di Jin, Ramit Sawhney, and Rajiv Ratn Shah. 2020. Augmenting NLP models using latent feature interpolations. In *Proceedings of the COLING*, pages 6931–6936.

Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and William Campbell. 2019. A closer look at feature space data augmentation for few-shot intent classification. In *Proceedings of the EMNLP-IJCNLP Workshop*, pages 1–10.

Hung-yi Lee, Shang-Wen Li, and Thang Vu. 2022a. Meta learning for natural language processing: A survey. In *Proceedings of the 2022 NAACL*, pages 666–684.

Hung-yi Lee, Shang-Wen Li, and Thang Vu. 2022b. Meta learning for natural language processing: A survey. In *Proceedings of the NAACL*, pages 666–684.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the EMNLP, EMNLP 2021*, pages 3045–3059.

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

Seung Jun Moon, Sangwoo Mo, Kimin Lee, Jaeho Lee, and Jinwoo Shin. 2021. MASKER: masked keyword regularization for reliable text classification. In *Proceedings of the AAAI*, pages 13578–13586.

Renkun Ni, Micah Goldblum, Amr Sharaf, Kezhi Kong, and Tom Goldstein. 2021. Data augmentation for meta-learning. In *Proceedings of the ICML*, volume 139, pages 8152–8161.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. volume 35, pages 27730–27744.

Max B. Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, and Chris J. Maddison. 2020. Gradient estimation with stochastic softmax tricks. In *Proceedings of the NeurIPS*.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the NeurIPS*.

Pengfei Sun, Yawen Ouyang, Wenming Zhang, and Xinyu Dai. 2021. MEDA: meta-learning with data augmentation for few-shot text classification. In *Proceedings of the IJCAI*, pages 3929–3935.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the CVPR*, pages 1199–1208.

Jianing Wang, Chengyu Wang, Fuli Luo, Chuanqi Tan, Minghui Qiu, Fei Yang, Qiuhui Shi, Songfang Huang, and Ming Gao. 2022. Towards unified prompt tuning for few-shot text classification. In *EMNLP Findings*, pages 524–536.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the EMNLP-IJCNLP*, pages 6381–6387.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Wenpeng Yin. 2020. Meta-learning for few-shot natural language processing: A survey. *CoRR*, abs/2007.09604.

Kang Min Yoo, Hanbit Lee, Franck Dernoncourt, Trung Bui, Walter Chang, and Sang-goo Lee. 2020. Variational hierarchical dialog autoencoder for dialog state tracking data augmentation. In *Proceedings of the EMNLP*, pages 3406–3425.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. In *EMNLP findings*, pages 2225–2239.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. In *Proceedings of the ACL*, pages 3221–3228.

8. Language Resource References

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with distributional signatures. In *Proceedings of the ICLR*.

Chengcheng Han, Zeqiu Fan, Dongxiang Zhang, Minghui Qiu, Ming Gao, and Aoying Zhou. 2021. Meta-learning adversarial domain adaptation network for few-shot text classification. In *ACL Findings*, pages 1664–1673.

Yutai Hou, Hongyuan Dong, Xinghao Wang, Bohan Li, and Wanxiang Che. 2022. Metaprompting: Learning to learn better prompts. In *Proceedings of the COLING*, pages 3251–3262.