International Journal of Knowledge and Innovation Studies
https://www.acadlore.com/journals/IJKIS

AMBERT-DWPM: An Adaptive Masking and Dynamic Prototype Learning Framework for Few-Shot Text Classification

Junyu Li[1], Jialin Ma[1*], Ashim Khadka[2]

[1] Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, 223003 Huaian, China
[2] Nepal College of Information Technology, Pokhara University, 44700 Lalitpur, Nepal

[*] Correspondence: Jialin Ma (majl@hyit.edu.cn)

Received: 02-11-2025 **Revised:** 03-14-2025 **Accepted:** 03-21-2025

Citation: J. Y. Li, J. L. Ma and A. Khadka, "AMBERT-DWPM: An adaptive masking and dynamic prototype learning framework for few-shot text classification," *Int J. Knowl. Innov Stud.*, vol. 3, no. 1, pp. 37–49, 2025. https://doi.org/10.56578/ijkis030104.

Abstract: Transformer-based language models have demonstrated remarkable success in few-shot text classification; however, their effectiveness is often constrained by challenges such as high intraclass diversity and interclass similarity, which hinder the extraction of discriminative features. To address these limitations, a novel framework, Adaptive Masking Bidirectional Encoder Representations from Transformers with Dynamic Weighted Prototype Module (AMBERT-DWPM), is introduced, incorporating adaptive masking and dynamic weighted prototypical learning to enhance feature representation and classification performance. The standard BERT architecture is refined by integrating an adaptive masking mechanism based on Layered Integrated Gradients (LIG), enabling the model to dynamically emphasize salient text segments and improve feature discrimination. Additionally, a DWPM is designed to assign adaptive weights to support samples, mitigating inaccuracies in prototype construction caused by intraclass variability. Extensive evaluations conducted on six publicly available benchmark datasets demonstrate the superiority of AMBERT-DWPM over existing few-shot classification approaches. Notably, under the 5-shot setting on the DBpedia14 dataset, an accuracy of 0.978±0.004 is achieved, highlighting significant advancements in feature discrimination and generalization capabilities. These findings suggest that AMBERT-DWPM provides an efficient and robust solution for few-shot text classification, particularly in scenarios characterized by limited and complex textual data.

Keywords: Few-shot text classification; Dynamic Weighted Prototype Module (DWPM); Adaptive Masking Bidirectional Encoder Representations from Transformers (AMBERT); Contrastive learning; Feature discrimination

1 Introduction

In the field of Natural Language Processing (NLP) and others, models based on the Transformer architecture [1] have become mainstream and have achieved significant performance in various tasks. However, their dependence on large amounts of labeled data [2–4] limits their effectiveness in few-shot learning (FSL) environments where resources are scarce. This has given rise to the research need for few-shot text classification. Few-shot text classification requires models to learn new categories from a limited number of labeled instances, which poses challenges to the models' generalization ability and feature discrimination capabilities.

Few-shot text classification [5, 6] has experienced several significant developmental stages. Early research primarily depended on transfer learning and meta-learning strategies [7, 8], acquiring general knowledge representations by pretraining models on large-scale datasets. Subsequently, the introduction of prototypical networks and other metric learning methods enabled models to better capture the similarities between categories. Prototypical networks mainly construct decision boundaries by calculating the mean vectors of categories. However, when there are variations in the expressions of the same category, such as the "logistics service" category containing expressions with different semantic focuses like "fast delivery speed" and "beautiful packaging" the mean prototype can be diluted by diverse features, leading to insufficient intra-class consistency.

The emergence of pre-trained language models has further propelled the development of this field [9]. Models such as BERT [10] have acquired robust language comprehension capabilities through self-supervised learning, providing a better foundation for feature representation in few-shot scenarios. Recent studies have shown that although pre-trained

language models can offer rich semantic representations, they still have limitations when dealing with complex scenarios. Some studies have attempted to enhance the few-shot ability of models through prompt learning [11–13] and prompt tuning. However, their performance heavily relies on the domain adaptability of manually designed prompts. For instance, in medical text classification, if the prompt template does not cover professional terms such as "diagnosis results" and instead uses general expressions like "medical conclusions", the model may completely overlook key semantic clues. Other studies have explored contrastive learning and adaptive feature extraction to improve the discriminability of features. However, these methods are limited in scenarios with high inter-class similarity. For example, in legal text classification tasks, the semantic boundary between "breach of contract" and "negligence in contract formation" is so blurred that it leads to significant increases in negative sampling errors in contrastive learning, further confusing the feature space.

Traditional FSL methods, despite their successes across various scenarios, often struggle to extract discriminative features from limited samples when confronted with the aforementioned challenges. The inherent complexity of natural language, compounded by the scarcity of labeled samples, makes it particularly difficult to establish clear inter-class decision boundaries while preserving intra-class expression diversity. For instance, distinguishing between "anxiety" and "depression" in psychological state classification tasks requires capturing subtle differences in semantic intensity, which traditional methods fail to model due to their over-reliance on surface-level lexical statistical features.

To address these challenges, this paper proposes a novel framework called AMBERT-DWPM, which integrates an adaptive masking strategy [14] and DWPM [15]. The adaptive masking strategy is used to fine-tune the BERT model [8, 16] by selectively masking irrelevant information, guiding the model to focus on discriminative text segments in the classification task, thereby enhancing feature extraction. The DWPM module dynamically weights sample features through cross-attention, effectively addressing the issue of inaccurate category distribution caused by high intraclass diversity and high interclass similarity [17].

The AMBERT-DWPM framework integrates an adaptive masking strategy and a DWPM to enhance feature extraction and discrimination in few-shot text classification. This study not only improves the feature extraction capabilities but also optimizes the model's generalization performance through a hybrid loss function combining contrastive learning and prototypical learning. Experimental results demonstrate that the AMBERT-DWPM framework achieves excellent performance across multiple datasets [18]. Specifically, on the DBpedia14 dataset, the accuracy reaches 0.978±0.004 under the 5-shot setting, approaching the performance levels of fully supervised learning. Additionally, the framework exhibits strong robustness and generalization on other challenging datasets, such as AG News and Symptoms. These findings validate the effectiveness of the adaptive masking strategy and DWPM in addressing high intraclass diversity and high interclass similarity [17]. The proposed framework offers an efficient and reliable solution for few-shot text classification tasks [19], showcasing significant potential for practical applications.

2 Related Work

2.1 Few-shot Text Classification

Few-shot text classification methods can generally be categorized into five types: non-parametric methods, data augmentation methods, prompt-based methods, meta-learning methods, and fine-tuning-based methods. The model we propose belongs to the fine-tuning-based methods [11].

Among non-parametric methods, compressor-based text classification has achieved considerable success. This approach leverages compressor techniques to estimate entropy or approximate Kolmogorov complexity and information distance for text classification. Jiang et al. [20] proposed a method for few-shot text classification that combines lossless compressors with compressor-based distance metrics and employs the k-nearest neighbors (KNN) classifier for text classification [11]. However, these non-parametric methods rely on information-theoretic distances (such as compression rates or entropy) to measure the similarity between texts. The semantic similarity of texts cannot be fully characterized by compression rates, which severely limits their application value in complex classification scenarios.

Data augmentation methods address the few-shot problem by expanding existing datasets to generate additional data. For example, Aug GPT [21] uses ChatGPT to rewrite texts, thereby enhancing the performance of few-shot classification tasks. Piedboeuf and Langlais [22] proposed that directly leveraging large language models to generate texts aligned with corresponding categories can effectively improve the performance of few-shot classification. However, data augmentation methods face challenges in ensuring that the synthesized data maintains the same distribution as the original dataset [11]. If the distribution of the generated data does not match that of the real task data, it may lead to model bias and even affect classification performance.

Prompt-based methods have made significant progress in FSL in recent years. These methods construct task-specific prompts that enable models to better leverage the knowledge of pre-trained language models [23]. For example, token-level prompting relies on masked language models (MLMs), with PET [24] being a method of this type; sentence-level prompting relies on the next sentence prediction task (NSP), such as the NSP-BERT sentence-level prompting method proposed by Sun et al. [25]. Although prompt-based methods can fully utilize pre-trained language models, they are highly dependent on prompt templates, which need to be adjusted specifically for different tasks.

Moreover, models are highly sensitive to prompts, and different prompts can lead to significantly different classification results, affecting the stability of the model.

The core idea of meta-learning methods is to enhance a model's adaptability to new tasks by learning from multiple few-shot tasks. The main strategies include metric-based meta-learning and optimization-based meta-learning. Metric-based meta-learning methods, such as Siamese neural networks [26], prototypical networks [27], and matching networks [28], construct a metric space across tasks to enhance class separability, enabling models to classify by computing similarities between samples without large-scale parameter updates. In contrast, optimization-based meta-learning [9] (such as MAML) learns an optimal model initialization that allows the model to quickly adapt to new tasks under few-shot conditions, thereby improving training efficiency and generalization ability. Müller et al. [29] proposed a model that combines Siamese networks with label fine-tuning, which achieved good results in few-shot text classification tasks. However, meta-learning methods typically rely on a large number of auxiliary tasks for training, and their generalization ability may significantly decrease if the target task distribution differs greatly from the training tasks. Moreover, metric-based meta-learning has limited discriminability in scenarios with high inter-class similarity, leading to potential misclassification issues when dealing with complex semantic categories. Therefore, despite the advantages of meta-learning in FSL, its requirement for consistency in task distribution is high, making it less applicable to real-world scenarios with significant task differences.

Fine-tuning-based methods achieve efficient few-shot classification by replacing the output layer on the basis of pre-trained language models and fine-tuning on a small number of samples. Compared with data augmentation and prompt-based methods, this approach is more concise and efficient, as it does not require additional data generation or complex prompt design, and it inherits the powerful feature representation capabilities of pre-trained language models [9], which gives it better adaptability in few-shot tasks. However, traditional fine-tuning methods struggle to learn discriminative features from limited samples when confronted with tasks that have high intra-class diversity and high inter-class similarity, thus restricting their classification performance.

To address this issue, we propose adding an adaptive masking operation to BERT and combining it with a DWPM for fine-tuning. Since BERT is pre-trained on open-domain datasets, which have different data distributions compared to the target domain of specific tasks, further pre-training BERT on target-domain data can also help with domain adaptation. However, in the FSL setting with extremely limited samples, it is challenging to train the BERT parameters. In contrast, our proposed model only adds an adaptive masking operation to BERT and combines it with the DWPM for further fine-tuning. It directly inherits the parameters of the pre-trained model without the need for additional expensive pre-training.

2.2 Prototype Learning

Prototype learning, as an important method of metric learning, is based on the idea of representing the overall feature distribution of each category by calculating class prototypes and classifying samples based on the distance between samples and prototypes. In FSL scenarios, prototype learning has attracted widespread attention due to its simplicity and effectiveness.

Basic prototypical networks construct class prototypes by simply averaging the features of support set samples and perform classification predictions by calculating the Euclidean distance between query samples and class prototypes. They also introduce cosine similarity to improve the distance metric. These methods are characterized by their simple structural design, ease of implementation and extension, and low computational overhead. However, since the prototypes are constructed using simple averaging, they fail to fully account for the heterogeneity of sample features. As a result, they perform poorly when dealing with data that have uneven intraclass distributions. This is especially true when the number of samples is extremely small, as the constructed prototypes tend to lack representativeness.

In recent years, researchers have proposed various methods to enhance the performance of prototype learning. Ragno et al. [30] explored prototype learning based on Graph Neural Networks (GNNs). By modeling the relationships between samples, these methods optimize prototype representations, enabling class prototypes to better capture the structural information between samples. These approaches significantly enhance feature-representation capabilities by deeply exploring the structural information between samples. However, they also bring higher computational complexity, which notably reduces training efficiency on large-scale datasets. Moreover, when the number of samples is extremely limited, the stability of graph structure construction is poor, which can easily affect the overall performance of the model.

Gogoi et al. [31] proposed the Adaptive Prototypical Network, which introduces an attention mechanism to dynamically adjust the contributions of different support samples in the prototype construction process. This effectively enhances the model's ability to adapt to sample heterogeneity. TPN [32] alleviates the problem of distribution differences of NOTA in cross-domain tasks by dynamically adjusting the representation of NOTA prototypes. However, it still has shortcomings in dealing with sample heterogeneity. Especially when the number of samples is small, it is difficult to fully capture the complex relationships between samples, which limits the representativeness of prototypes. RAPL [33] further optimizes the prototype learning method. It improves the model's performance in

cross-domain tasks through instance-level prototype construction and relation-weighted contrastive learning. But its computational complexity is still high, and the sensitivity to differences in data distribution across domains has not been fundamentally solved. Additionally, Wu et al. [15] introduced a dynamic prototype update mechanism, which adaptively adjusts the prototype update strategy based on the distribution characteristics of samples during training, thereby improving the representation capability of prototypes. However, the integration with triplet loss is relatively loose, and it fails to effectively utilize the relationship information between other sample pairs within the batch [9].

To address the above issues, we propose the AMBERT-DWPM architecture, a prototype learning method centered on the DWPM. By combining adaptive feature weighting and an improved contrastive learning strategy, it effectively enhances the model's ability to model complex text distributions while maintaining computational efficiency. This approach optimizes the feature space structure by mining multi-granular relationship information between sample pairs within a batch. On the basis of retaining the original advantages of triplet loss, it introduces intra-class and inter-class contrastive constraints, thereby significantly enhancing the model's discriminative power.

3 Methods

We propose the overall architecture of AMBERT-DWPM for few-shot classification, as shown in Figure 1, which consists of two main components: the mask-guided BERT model and the DWPM.

Figure 1 illustrates the proposed AMBERT-DWPM framework. The BERT model extracts features from both support and query data. When computing classification prototypes, the DWPM module dynamically assigns weights to the limited support samples. The model is optimized using a hybrid loss function.

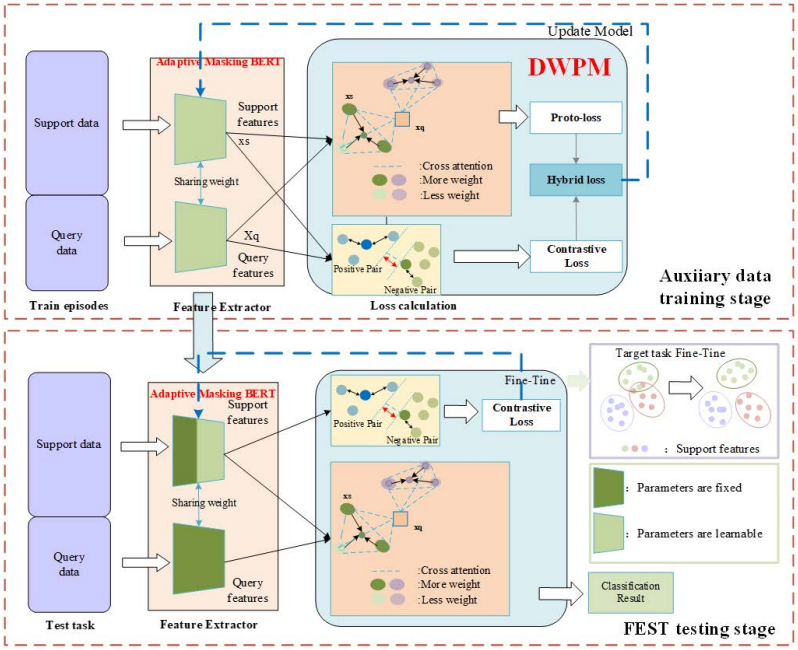

Figure 1. Overall architecture

3.1 Feature Extraction Based on Mask Guidance

Traditional Mask-BERT enhances the model's focus on key features and improves feature extraction by applying predefined masking operations to the input text. However, its masking strategy lacks dynamic adjustment capabilities, as mask generation often relies on fixed rules and cannot adaptively adjust according to the specific characteristics of each input sample. Therefore, when dealing with few-shot tasks that have high intraclass diversity or high interclass similarity, the model's feature discrimination is often insufficient, making it difficult to fully leverage the advantages of its masking mechanism.

The adaptive masking strategy is based on the LIG method, which calculates the importance of each token to the final classification output, thereby determining whether to mask the token. Let the input text sequence be $X = \{x_1, x_2, \ldots, x_n\}$, with the corresponding token embedding representation $E(X) = \{e_1, e_2, \ldots, e_n\}$. A reference input X' is selected, with the token embedding representation $\mathrm{E}(\mathrm{X}') = \{\mathrm{e}', \mathrm{e}'_2, \ldots, \mathrm{e}''\}$, where the reference input can be chosen as a zero vector to measure the contribution of each token. The contribution of tokens at different

levels is calculated using the integrated gradients method:

$$\mathrm{IG}\left(\mathrm{X}, \mathrm{X}'\right) = \left(\mathrm{E}(\mathrm{X}) - \mathrm{E}\left(\mathrm{X}'\right)\right) \times \frac{\partial f\left(\mathrm{E}\left(\mathrm{X}'\right)\right) + \alpha\left(\mathrm{E}(\mathrm{X}) - \mathrm{E}\left(\mathrm{X}'\right)\right)}{\partial \mathrm{E}(\mathrm{X})} \mathrm{d}\alpha$$

where, f denotes the output of BERT after passing through the classification layer, and α represents the integration step size [34]. To calculate the contribution of tokens at different levels, the cumulative sum of LIG is introduced:

$$S\left(\chi_i\right) = \sum_{\mathrm{l}=1}^{\mathrm{L}} \mathrm{IG}_{\mathrm{l}}\left(\chi_i, \chi_i'\right)$$

where, L denotes the number of Transformer layers, and $S(xi)$ represents the comprehensive importance score of token x_i across different layers. A higher $S(xi)$ ultimately indicates that the token plays a more significant role in the classification task. The computation of LIG involves calculating the integrated gradients for each token, with a computational complexity of:

$$O(n \cdot L \cdot F)$$

where, n represents the length of the token sequence, L represents the number of Transformer layers, and F represents the integration step size.

Based on the calculated token importance scores $S(xi)$ the adaptive masking strategy dynamically selects the mask positions as follows: First, set a mask proportion threshold τ, calculate the importance scores $S(xi)$ for all tokens, and then sort the tokens. Select the tokens in the bottom τ proportion of the ranking for masking to filter out redundant or low-discriminative tokens. Finally, replace the selected tokens with [MASK] using a hard masking method to ensure that the model focuses on more discriminative features. The mask selection involves token importance ranking, and its computational complexity is:

$$O(n \cdot \log \cdot n)$$

After the adaptive masking process is completed, a new input sequence X_{masked} is obtained. This sequence is processed by BERT and then fed into the classification layer for training:

$$P\left(\mathrm{y} \mid X_{\text{masked}}\right) = \mathrm{softmax}\left(W h_{\text{BERT}}\left(X_{\text{masked}}\right)\right)$$

where, W denotes the classification weight matrix, and $h_{\mathrm{BERT}}\left(X_{\text{masked}}\right)$ represents the feature vector after BERT processing. During training, the loss function employs the cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{\mathrm{N}} y_i \log \mathrm{P}\left(y_i \mid \mathrm{x}_{\text{masked}}\right)$$

where, N represents the number of training samples, and y_i denotes the true class label. BERT training involves the forward and backward propagation of the Transformer, with its computational complexity mainly determined by the self-attention mechanism:

$$O\left(n^2 d\right)$$

where, d is the dimension of the hidden layer. The overall computational complexity of the adaptive masking strategy can be expressed as:

$$O\left(T\left(n \cdot L \cdot F + n\log n + n^2 d\right)\right)$$

where, T represents the number of training epochs.

3.2 Structure of the DWPM Module

Prototypical networks have demonstrated remarkable capabilities in few-shot classification tasks, thanks to their simple model structure and computational efficiency. However, they have limitations when dealing with high intraclass diversity and high interclass similarity [17]. They struggle to fully capture the varying contributions of different samples to prototype generation, which affects the accuracy of the class prototypes.

To address this issue, this study proposes a DWPM. This module aims to dynamically assign weights to each support sample to accurately construct class prototypes. Combined with a contrastive learning module to further optimize query samples, the DWPM enhances intraclass consistency and interclass discriminability through a hybrid loss function that integrates contrastive and prototypical losses. This approach improves classification performance.

As shown in Figure 2, the DWPM module maps the feature representations of samples through a projection head, projecting them from the original feature space to an attention space. It uses an n-layer cross-attention block based on the Transformer structure to facilitate information interaction between support samples and query samples.

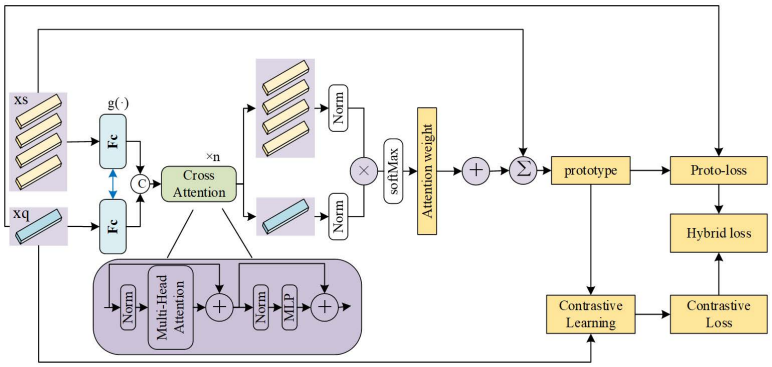

Figure 2. Diagram of the DWPM

In the DWPM module, a 4-layer Transformer structure is adopted. Each layer consists of Multi-Head Attention, Feed-Forward Network (FFN), and Residual Connection to enhance the ability of information aggregation. The hidden layer dimension of the Transformer layer is set to 512, and the number of attention heads is set to 8, to balance computational efficiency and feature representation capability. Moreover, in the cross-attention mechanism, by calculating the correlations between query samples and support samples of each category, the classification prototype assigns higher weights to more discriminative samples. The strength of the attention weights is controlled by the hyperparameter λ. When λ is set to 0, the prototype calculation uses average weighting, similar to traditional prototype learning. As the value of λ increases, the effect of attention weights becomes stronger. The formula for calculating the category prototype c_k is as follows:

$$c_k = \sum_{(x_i, y_i) \in S_k} \left(\lambda \cdot w_{attn}^i + (1 - \lambda) \cdot w_{avg}^i \right) f_\theta \left(x_i \right)$$

where, S_k denotes the set of support samples for a class, $f_\theta(x)$ represents the features extracted by the feature extractor,and w_{attn}^i and w_{avg}^i represent the attention weights and average weights, respectively. After obtaining the prototype representation, the distance between the query sample and the prototype is measured using Euclidean distance. The probability that the query sample x belongs to class k is calculated using the softmax function, as follows [35]:

$$\mathrm{p}_\theta(y = k \mid x) = \frac{\exp\left(-d\left(f_\theta(x), c_k\right)\right)}{\sum_{k'} \exp\left(-d\left(f_\theta(x), c_{k'}\right)\right)}$$

where, $d\left(f_\theta(x), c_k\right)$ represents the Euclidean distance between the query sample feature $f_\theta(x)$ and the class prototype c_k [36].

In few-shot text classification tasks, distributional shifts are inevitable due to intraclass variations and interclass similarities. To simultaneously reduce the impact of intraclass diversity on feature distribution and enhance interclass discriminability, we introduce a hybrid loss function that combines contrastive learning loss and prototypical loss to fine-tune the model, thereby improving the discriminative power between samples. During fine-tuning, only the shared weights of the BERT model are updated, enabling the model to better adapt to the feature distribution of the target task. Combined with the DWPM, the model can leverage limited samples to mitigate distributional shifts caused by intraclass diversity and interclass similarity, ultimately enhancing the accuracy and robustness of few-shot text classification tasks [37].

The hybrid loss function we employ combines contrastive loss and prototypical classification loss. The formula for calculating the prototypical classification loss is as follows:

$$L_{cls} = -\frac{1}{|Q|} \sum_{(x_i, y_i) \in Q} \log P_\emptyset \left(y = y_i \mid x_i \right)$$

where, |Q| denotes the number of query samples during the training process. (x_i, y_i) represents the query sample x_i and its true class y_i, while $P_\emptyset \left(y = y_j \mid x_i \right)$ denotes the predicted probability of the query sample x_i being classified into class y_i.

For each query sample x_q in the batch, the contrastive learning loss is calculated based on the distances between the query sample and its positive class prototype c_+ and negative class prototype c_-. The formula for calculating the contrastive learning loss is as follows:

$$L_{contrastive} = \frac{1}{|\mathrm{B}|} \sum_{(x_q, c)} \left[y \cdot d\left(x_q, c_+\right) + (1 - y) \cdot \max\left(0, m - d\left(x_q, c_-\right)\right) \right]$$

where, $|B|$ denotes the total number of query samples in the batch, and y is a binary indicator variable. When the query sample x_q belongs to the positive class, y=0; when the query sample x_q belongs to the negative class, y=1. m represents the margin threshold in contrastive learning, which is used to adjust the distance of negative samples, ensuring that the feature separation between negative samples and the query sample is at least m to enhance discriminability.

The hybrid loss is the weighted sum of the prototypical classification loss and the contrastive learning loss:

$$L_{hybrid} = \alpha \cdot L_{proto} + (1 - \alpha) \cdot L_{contrastive}$$

where, α is the weighting coefficient that controls the contributions of the two types of losses. When α is large, the model primarily relies on the prototype classification loss for optimization; when α is small, the model emphasizes enhancing class discriminability through the contrastive loss.

This hybrid loss fine-tunes the BERT model, bringing significant advantages to few-shot text classification tasks. By minimizing the distance between query samples and their correct class prototypes [38], the prototypical classification loss enhances intraclass consistency, making samples from the same class more clustered in the feature space and effectively reducing the impact of intraclass variations. Meanwhile, the contrastive learning loss, by pushing query samples away from incorrect class prototypes, improves interclass discriminability and avoids the overlap of features from different classes. During fine-tuning, the hybrid loss further optimizes the shared weights of the BERT model, making its feature extraction capabilities more adaptive to the feature distribution in few-shot tasks. This allows the model to quickly adapt to new classes and improve classification accuracy. Additionally, by adjusting the weighted ratio between the prototypical and contrastive losses, the hybrid loss achieves an effective balance between intraclass consistency and interclass discriminability, enabling the model to exhibit stronger generalization and robustness across various few-shot scenarios.

4 Experiments

4.1 Dataset

The experiments used six publicly available datasets, including three open-domain datasets and three medical-domain datasets. Each dataset was divided into base classes and novel classes according to their categories, forming the base dataset and the novel dataset, respectively [14]. Table 1 summarizes the average length of samples in each dataset, as well as the number of classes in the base and novel datasets.

-AG news: A text classification dataset containing 4 news categories, including World, Sports, Business, and Tech.

-DBpedia14: A text classification dataset with 14 categories.

-Snippets: A text dataset containing web snippets from Google search.

-Symptoms: This dataset, available on Kaggle, contains over 8 hours of audio data describing common medical symptoms. We used the text transcriptions of the audio data and removed duplicates. After preprocessing, the dataset contains 231 samples divided into 7 symptom categories.

-PubMed20k: A sequence classification dataset based on PubMed.

-NICTA-PIBOSO: A dataset based on the ALTA 2012 Shared Task, used for classifying sentences from biomedical abstracts into predefined categories [14].

Table 1. Statistics of sample length and class numbers in each dataset

Dataset	Avg. Length	$\lvert\mathbf{Y_b}\rvert$ / $\lvert\mathbf{Y_n}\rvert$	Domain
AG news	39	2/2	Open-domain
DBpedia14	50	8/6	Open-domain
Snippets	18	4/4	Open-domain
Symptom	11	4/3	Medical-domain
PubMed20k	26	3/2	Medical-domain
NICTA-PIBOSO	24	3/2	Medical-domain

4.2 Experimental Settings

In the experiments, all models were initialized with the pre-trained BERT-base-cased model. Our method employed a hybrid loss function that combines prototypical loss and contrastive learning loss for optimization. During training, the batch size for the base dataset was set to 64, while the batch size for the novel dataset was set to 32. The learning rates were set to 2×10^{-5} and 4×10^{-5}, respectively, and the AdamW optimizer was employed. During the optimization process of the hybrid loss function, the weight parameter α is selected through grid search within the range {0.2,0.4,0.6,0.8}, and is ultimately set to 0.6. In the Adaptive Masking strategy, the masking ratio threshold τ is

set to 0.15, the number of Transformer layers L is set to 12, and the integration step size F is set to 50, to ensure that the model can effectively learn important feature information. DWPM module employs a 4-layer Transformer structure, with each layer comprising 8 attention heads and a FFN hidden dimension of 512. In the experiments, the batch size for all comparison models is uniformly set to 32, the learning rate is set to 4×10^{-5}, and the number of training epochs is set to 10, to ensure a fair comparison with the method proposed in this paper. Each set of experiments is repeated five times, and the average accuracy and standard deviation are reported to ensure the stability and fairness of the results.

4.3 Experimental Results and Analysis

To validate the effectiveness of our proposed method, we compared it with several few-shot classification models on six publicly available datasets. The comparison methods included standard BERT, further pre-trained BERT (FPT-BERT), sentence prompt-based NSP-BERT, contrastive learning-based CPFT, and some mainstream few-shot classification models such as Prototypical Networks (SN-FT) and Siamese Networks (SN-FT). Experiments were conducted under both 5-shot and 8-shot settings to comprehensively evaluate the performance of each method under different sample sizes.

The main results of the performance comparison are listed in Table 2 and Table 3. Overall, our method outperformed the comparison methods on both open-domain and medical-domain datasets [14]. On open-domain datasets (such as AG News and DBpedia14), our method achieved performance comparable to the state-of-the-art, demonstrating strong generalizability. On medical-domain datasets (such as Symptoms and PubMed20k), our method showed significant advantages, proving its effectiveness in handling few-shot classification problems in more challenging scenarios (e.g., domain-specific tasks).

Table 2. Comparison results under the 5-shot setting

K-short	Models	AG News	DBpedial4	Snippets	Symptoms	PubMed20k	NICTA-PIBOSO
5-short	BERT	0.752±0.098	0.947±0.028	0.854±0.025	0.782±0.040	0.845±0.027	0.696±0.054
	FPT-BERT	0.779±0.035	0.963±0.026	0.888±0.028	0.800±0.112	0.864±0.030	0.727±0.042
	Re-init-BERT	0.762±0.024	0.941±0.029	0.799±0.061	0.830±0.047	0.862±0.019	0.702±0.063
	CPFT	0.768±0.047	0.965±0.013	0.853±0.029	0.876±0.044	0.848±0.028	0.723±0.068
	SN-FT	0.768±0.055	0.971±0.003	0.867±0.024	0.858±0.059	0.782±0.031	0.736±0.041
	NSP-BERT	0.820±0.016	0.973±0.002	0.885±0.016	0.870±0.043	0.883±0.043	0.732±0.047
	our	0.824±0.022	0.978±0.004	0.891±0.013	0.884±0.025	0.881±0.015	0.752±0.033

Table 3. Comparison results under the 8-shot setting

K-short	Models	AG News	DBpedial4	Snippets	Symptoms	PubMed20k	NICTA-PIBOSO
8-short	BERT	0.790±0.063	0.979±0.004	0.888±0.021	0.855±0.066	0.875±0.024	0.714±0.038
	FPT-BERT	0.801±0.035	0.978±0.012	0.903±0.023	0.848±0.094	0.870±0.014	0.726±0.048
	Re-init-BERT	0.789±0.026	0.964±0.011	0.863±0.028	0.882±0.046	0.880±0.008	0.738±0.034
	CPFT	0.816±0.021	0.978±0.002	0.888±0.017	0.936±0.029	0.871±0.022	0.776±0.046
	SN-FT	0.817±0.024	0.986±0.002	0.881±0.020	0.876±0.081	0.796±0.044	0.757±0.033
	NSP-BERT	0.833±0.009	0.986±0.003	0.898±0.018	0.897±0.031	0.761±0.051	0.749±0.040
	our	0.834±0.026	0.987±0.001	0.908±0.011	0.938±0.047	0.899±0.005	0.779±0.034

Under the 5-shot setting, our method achieves an accuracy of 0.891±0.013 on the Snippets dataset and 0.884±0.025 on the Symptoms dataset, showing significant improvements compared to methods such as NSP-BERT and CPFT. The performance improvement is primarily attributed to the introduction of the Adaptive Masking strategy. This strategy dynamically and selectively masks irrelevant information, enabling the model to focus on the key feature segments that are crucial for the classification task, thereby significantly enhancing the discriminability of feature extraction. This advantage is particularly evident on the Snippets dataset, which comprises short texts of various categories. Within this dataset, the expression styles of texts within the same category are highly diverse, while texts from different categories may overlap in certain keywords or expressions. For instance, for the "movie review" category, sentences may cover multiple different focuses such as actors, plot, and ratings. Similarly, texts in the "news

report" category may also involve related social events, making it difficult for traditional methods to distinguish between them. Without Adaptive Masking, the model may be distracted by redundant information and struggle to extract features with high inter-class discriminability. Adaptive Masking, however, identifies and masks the most discriminative tokens through gradient integration, allowing the model to concentrate on the information in the text that truly aids classification (for example, words like "box office" and "director" are more representative of movie reviews, while "government" and "policy" are more indicative of news reports). This approach enhances intra-class consistency, reduces inter-class confusion, and ultimately improves classification performance.

Data in the medical field typically exhibits high terminological density and semantic complexity. For instance, in the symptoms dataset, symptom descriptions of certain categories such as "headache" and "migraine," or "influenza" and "cold," are highly similar. Moreover, cases within the same category may be described using different medical terminologies. This makes traditional text classification methods susceptible to noise and challenges in effectively distinguishing between similar categories. In contrast, our method, through the dynamic weighted prototype learning mechanism of the DWPM module, assigns dynamic weights to support samples, enabling the category prototypes to more accurately describe the characteristics of each category. Unlike fixed-weight methods, DWPM adjusts the weights of category prototypes through cross-attention mechanisms, so that features with stronger discriminative power for certain categories are assigned higher weights, thereby optimizing the construction of category prototypes. Specifically, in the symptoms dataset, DWPM assigns higher attention weights to key medical terms such as "fever" and "cough," while reducing the weights of potentially cross-category irrelevant information. This enables the model to more precisely capture subtle differences between categories in the face of high inter-class similarity tasks, thereby enhancing classification performance.

At the same time, under the 8-shot setting, as the number of support samples increases, the method proposed in this paper maintains a high accuracy rate on all datasets, demonstrating strong stability. On the DBpedia14 dataset, the accuracy of the proposed method under the 8-shot setting reaches 0.987±0.001, close to the optimal performance. This performance improvement is mainly attributed to the design of the hybrid loss function. This loss function combines contrastive loss and prototype loss, thereby performing well in optimizing the structure of the feature space. In traditional few-shot classification tasks, the feature representations of categories are easily influenced by the limited number of samples, leading to unstable prototype distributions. Particularly when category boundaries are ambiguous, the model's discriminative ability may decline. However, the hybrid loss function, by incorporating contrastive learning to bring similar samples closer and push dissimilar samples farther apart, while also combining prototype loss to enhance intra-class consistency, ultimately optimizes the geometric structure of the feature space. For example, on the DBpedia14 dataset, under the 8-shot setting, the model is able to better leverage the additional samples, causing samples of the same category to cluster more tightly in the feature space, while increasing the feature distances between different categories. This improves classification accuracy and reduces category confusion. Moreover, the hybrid loss function further balances intra-class consistency and inter-class discriminability, enabling the model to exhibit excellent performance across different datasets and task settings.

4.4 Ablation Study

In this section, we conducted experiments on the AG News dataset under the 5-shot setting to investigate the specific impacts of the adaptive encoding strategy (A) and contrastive learning (B) within the DWPM on the performance of DWPM. This study aims to address the challenges of high intraclass diversity and high interclass similarity in few-shot text classification tasks [17].

The starting point of the ablation study was a baseline BERT model integrated with the DWPM without contrastive learning (B). At this stage, DWPM focused solely on optimizing the class center feature representations through prototypical loss, without leveraging contrastive learning to enhance feature discrimination. The accuracy achieved by this baseline model was 0.745±0.063. In the second stage, when only the adaptive encoding strategy (A) was introduced to the baseline model, the accuracy improved to 0.769±0.048, an increase of 1.79 percentage points compared to the baseline model.g This improvement can be attributed to the strategy's ability to optimize the BERT model by selectively masking text information irrelevant to the classification task, thereby prompting the model to focus more on features critical to classification.

In the third stage, contrastive learning (B) was introduced into the DWPM module based on the baseline model to combine with the prototypical loss, forming a new hybrid loss function. This step aims to evaluate how contrastive learning works in synergy with the DWPM module and its specific impact on model performance. The experimental accuracy further increased to 0.794±0.057.

Ultimately, when both the adaptive encoding strategy (A) and the contrastive learning mechanism (B) were integrated into the baseline model, the precision reached its highest level at 0.824±0.022, demonstrating the model's robust performance across various evaluation metrics. The main results are summarized in Table 4. These findings fully substantiate the advantages of our proposed model in optimizing interclass distances, enhancing intraclass consistency, and improving feature discriminability. The adaptive masking strategy significantly enhances the

extraction of key features, while the DWPM module further optimizes model performance by dynamically adjusting the feature distribution of samples. The hybrid loss function ensures a balance between interclass separation and intraclass aggregation. The synergistic action of these modules markedly improves the model's performance in few-shot text classification tasks, providing a solid foundation for classification and relationship mining tasks. Through this structured experimental design, we are able to clearly demonstrate the specific contributions of each module to the performance of DWPM and how they collectively enhance the overall performance in few-shot text classification tasks.

Table 4. Ablation study results

Methods			Dataset
BERT	A	B	AG news
✓			0.745 ± 0.063
✓	✓		0.769 ± 0.048
✓		✓	0.794 ± 0.057
✓	✓	✓	0.824 ± 0.022

In the previous ablation studies, we analyzed the contributions of the adaptive encoding strategy (A) and the contrastive learning mechanism (B) to model performance, demonstrating the advantages of these two modules in optimizing intraclass consistency and interclass distances.

To more intuitively illustrate the effects of the model at different stages, we mapped the 5-shot training iteration results on the Snippets dataset to two dimensions using UMAP and colored the results according to class labels, as shown in Figure 3. In Snippets, the label IDs for the base dataset range from 0 to 7, while those for the novel dataset range from 8 to 13. For comparison, we also visualized the training processes of BERT and AMBERT-DWPM. The visualization results indicate that although both BERT and our module can effectively separate samples from different classes, our module achieves a more uniform class distribution, avoiding the clustering of samples observed in BERT during training. This characteristic is crucial for the effectiveness of FSL.

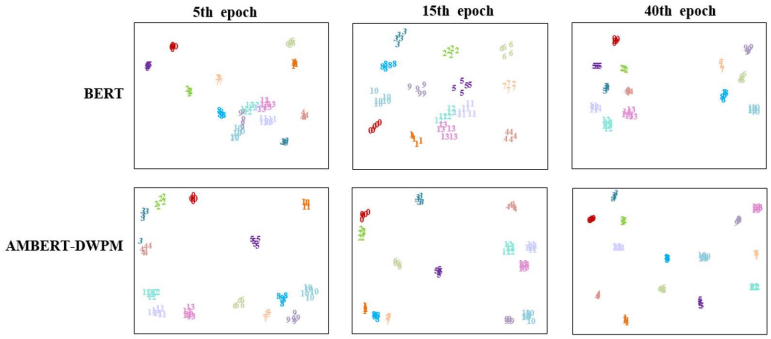

Figure 3. Visualization of training iterations on the snippets dataset

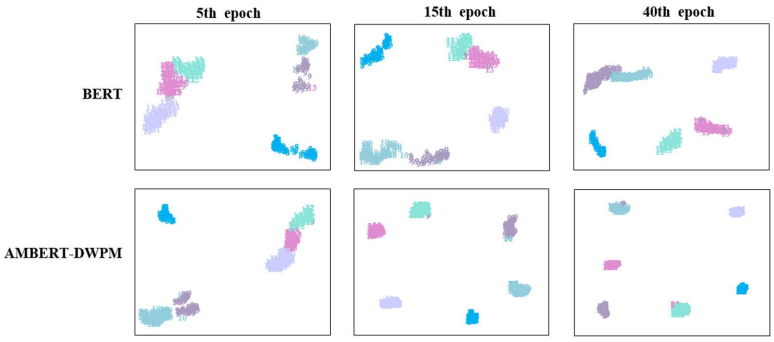

Figure 4. Visualization of the test set on snippets

During the UMAP dimensionality reduction process, we set n_neighbors=15, min_dist=0.1, n_components=2, and random_state=42 to ensure the stability of the experimental settings. First, we extracted the hidden states of the last Transformer layer from the BERT and AMBERT-DWPM models as feature vectors and performed dimensionality

reduction using UMAP to map them onto a two-dimensional plane. Subsequently, to further evaluate the feature distribution of the models on the test set, we randomly selected 200 test samples from the novel dataset and visualized them using the same method, ensuring a balanced number of samples from each class to avoid the impact of uneven data distribution on the visualization results. As shown in Figure 4, the results reveal that BERT struggles to separate certain classes, especially those that are similar. In contrast, AMBERT-DWPM effectively addresses this issue and enhances the compactness of samples within the same class, thereby improving the model's discriminative power.

5 Conclusion

This paper proposes AMBERT-DWPM, a novel few-shot text classification framework that effectively improves classification accuracy and stability by integrating an adaptive masking strategy and a DWPM. The adaptive masking strategy dynamically guides the feature extraction process of BERT, enabling the model to focus on discriminative features critical to classification tasks. Simultaneously, the DWPM module dynamically weights the contributions of support samples, effectively addressing challenges associated with high intraclass diversity and high interclass similarity. Comprehensive experimental results demonstrate that AMBERT-DWPM significantly outperforms existing baseline methods, achieving notable improvements across multiple open-domain and domain-specific datasets. Nevertheless, the model may face limitations when encountering extremely imbalanced or noisy datasets. Future research directions include further exploration of adaptive masking strategies tailored for noisy data, enhancement of DWPM's robustness to data imbalance, and applying the proposed method to cross-domain and unsupervised FSL scenarios.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflict of Interests

The authors declare that they have no conflicts of interest.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. https://doi.org/10.48550/arXiv.2010.11929

[3] A. Parnami and M. Lee, "Learning from few examples: A summary of approaches to few-shot learning," *arXiv preprint arXiv:2203.04291*, 2022. https://doi.org/10.48550/arXiv.2203.04291

[4] I. D. Mienye and T. G. Swart, "A comprehensive review of deep learning: Architectures, recent advances, and applications," *Information*, vol. 15, no. 12, p. 755, 2024. https://doi.org/10.3390/info15120755

[5] J. Wei, C. Huang, S. Vosoughi, Y. Cheng, and S. Xu, "Few-shot text classification with triplet networks, data augmentation, and curriculum learning," *arXiv preprint arXiv:2103.07552*, 2021. https://doi.org/10.48550/arXiv.2103.07552

[6] S. Wang, H. Fang, M. Khabsa, H. Mao, and H. Ma, "Entailment as few-shot learner," *arXiv preprint arXiv:2104.14690*, 2021. https://doi.org/10.48550/arXiv.2104.14690

[7] W. Yin, "Meta-learning for few-shot natural language processing: A survey," *arXiv preprint arXiv:2007.09604*, 2020. https://doi.org/10.48550/arXiv.2007.09604

[8] H. Y. Lee, S. W. Li, and N. T. Vu, "Meta learning for natural language processing: A survey," *arXiv preprint arXiv:2205.01500*, 2022. https://doi.org/10.48550/arXiv.2205.01500

[9] X. D. Luo, Z. Q. Deng, B. X. Yang, and M. Y. Luo, "Pre-trained language models in medicine: A survey," *Artificial Intelligence in Medicine*, vol. 154, p. 102904, 2024. https://doi.org/10.1016/j.artmed.2024.102904

[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "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*, Minneapolis, Minnesota, 2019, pp. 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

[11] X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun, "PTR: Prompt tuning with rules for text classification," *AI Open*, vol. 3, pp. 182–192, 2022. https://doi.org/10.1016/j.aiopen.2022.11.003

[12] S. Sun, X. Pan, T. Yang, and J. Gao, "STID-Prompt: Prompt learning for sentiment-topic-importance detection in financial news," *Knowl.-Based Syst.*, vol. 284, p. 111347, 2024. https://doi.org/10.1016/j.knosys.2023.111347

[13] J. N. Wang, C. Y. Wang, F. L. Luo, C. Q. Tan, M. H. Qiu, F. Yang, Q. H. Shi, S. F. Huang, and M. Gao, "Towards unified prompt tuning for few-shot text classification," *arXiv preprint arXiv:2205.05313*, 2022. https://doi.org/10.48550/arXiv.2205.05313

[14] W. X. Liao, Z. L. Liu, H. X. Dai, Z. H. Wu, and et al., "Mask-guided BERT for few-shot text classification," *Neurocomputing*, vol. 610, p. 128576, 2024. https://doi.org/10.1016/j.neucom.2024.128576

[15] C. W. Wu, J. Y. Yang, Y. Z. Shang, J. F. Pei, and et al., "Dynamically weighted prototypical learning method for few-shot SAR ATR," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024. https://doi.org/10.1109/LGRS.2024.3365147

[16] Z. Ren, "Enhancing Seq2Seq models for role-oriented dialogue summary generation through adaptive feature weighting and dynamic statistical conditioninge," in *2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE)*, Guangzhou, China, 2024, pp. 497–501. https://doi.org/10.1109/CISCE62493.2024.10653360

[17] Y. B. Zhao, J. J. Liu, J. L. Yang, and Z. B. Wu, "EMSCNet: Efficient multisample contrastive network for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023. https://doi.org/10.1109/TGRS.2023.3262840

[18] Y. H. Jung and A. Bellot, "Efficient policy evaluation across multiple different experimental datasets," in *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024, pp. 136 361–136 392.

[19] W. Liu, J. Pang, N. Li, F. Yue, and G. Liu, "Few-shot short-text classification with language representations and centroid similarity," *Appl. Intell.*, vol. 53, no. 7, pp. 8061–8072, 2023. https://doi.org/10.1007/s10489-022-03880-y

[20] Z. J. Jiang, M. Yang, M. Tsirlin, R. Tang, Y. Dai, and J. Lin, ""Low-resource" text classification: A parameter-free classification method with compressors," in *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, 2023, pp. 6810–6828. https://doi.org/10.18653/v1/2023.findings-acl.426

[21] H. Dai, Z. Liu, W. Liao, X. Huang, and et al., "AugGPT: Leveraging ChatGPT for text data augmentation," *IEEE Trans. Big Data*, vol. 1, no. 1, pp. 1–12, 2025. https://doi.org/10.1109/TBDATA.2025.3536934

[22] F. Piedboeuf and P. Langlais, "Data augmentation is dead, long live data augmentation," *arXiv preprints arXiv:2402.14895*, 2024. https://doi.org/10.48550/arXiv.2402.14895

[23] M. Chakraborty, "Analysis of textual-based reviews with minimal supervision," Ph.D. dissertation, Iowa State University, Ames, Iowa, 2024.

[24] T. Schick and H. Schütze, "Exploiting cloze questions for few shot text classification and natural language inference," *arXiv preprint arXiv:2001.07676*, 2020. https://doi.org/10.48550/arXiv.2001.07676

[25] Y. Sun, Y. Zheng, C. Hao, and H. Qiu, "NSP-BERT: A prompt-based few-shot learner through an original pre-training task–next sentence prediction," *arXiv preprint arXiv:2109.03564*, 2021. https://doi.org/10.48550/arXiv.2109.03564

[26] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, 2015.

[27] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017. https://doi.org/10.48550/arXiv.1703.05175

[28] R. Geng, B. Li, Y. Li, and et al., "Induction networks for few-shot text classification," *arXiv preprint arXiv:1902.10482*, 2019. https://doi.org/10.48550/arXiv.1902.10482

[29] T. Müller, G. Pérez-Torró, and M. Franco-Salvador, "Few-shot learning with siamese networks and label tuning," *arXiv preprint arXiv:2203.14655*, 2022. https://doi.org/10.48550/arXiv.2203.14655

[30] A. Ragno, B. La Rosa, and R. Capobianco, "Prototype-based interpretable graph neural networks," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 4, pp. 1486–1495, 2022. https://doi.org/10.1109/TAI.2022.3222618

[31] M. Gogoi, S. Tiwari, and S. Verma, "Adaptive prototypical networks," *arXiv preprint arXiv:2211.12479*, 2022. https://doi.org/10.48550/arXiv.2211.12479

[32] Y. Zhang and Z. Kang, "TPN: Transferable Proto-Learning Network towards few-shot document-level relation extraction," in *2024 International Joint Conference on Neural Networks (IJCNN)*, Yokohama, Japan, 2024, pp. 1–9. https://doi.org/10.1109/IJCNN60899.2024.10650913

[33] T. K. Tran, H. P. Tran, T. L. Le, and et al., "FedNTProto: A prototype-based approach for personalized federated learning," in *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, Da Nang, Vietnam, 2024, pp. 1–6. https://doi.org/10.1109/MAPR63514.2024.10660707

[34] Q. Wang, Y. He, S. Dong, X. Gao, S. Wang, and Y. Gong, "Non-exemplar domain incremental learning via cross-domain concept integration," in *European Conference on Computer Vision*. Cham: Springer Nature

Switzerland, 2024, pp. 144–162. https://doi.org/10.1007/978-3-031-72967-6_9

[35] G. Duan, Y. Song, Z. Liu, S. Ling, and J. Tan, "Cross-domain few-shot defect recognition for metal surfaces," *Meas. Sci. Technol.*, vol. 34, no. 1, p. 015202, 2022. https://doi.org/10.1088/1361-6501/ac90de

[36] H. Zhang, H. Liu, L. Liang, W. Ma, and D. Liu, "BiLSTM-TANet: An adaptive diverse scenes model with context embeddings for few-shot learning," *Appl. Intell.*, vol. 54, no. 6, pp. 5097–5116, 2024. https://doi.org/10.1007/s10489-024-05440-y

[37] T. Hu, Z. Chen, J. Ge, Z. Yang, and J. Xu, "A Chinese few-shot text classification method utilizing improved prompt learning and unlabeled data," *Appl. Sci.*, vol. 13, no. 5, p. 3334, 2023. https://doi.org/10.3390/app13053334

[38] Z. Sun, W. Zheng, and M. Wang, "SLTRN: Sample-level transformer-based relation network for few-shot classification," *Neural Netw.*, vol. 176, p. 106344, 2024. https://doi.org/10.1016/j.neunet.2024.106344