Integrating Model-Agnostic Meta-Learning with Advanced Language Embeddings for Few-Shot Intent Classification

Ali Rahimi
College of Interdisciplinary Science and Technologies
University of Tehran
Tehran, Iran
ali.rahimi97@ut.ac.ir

Hadi Veisi
College of Interdisciplinary Science and Technologies
University of Tehran
Tehran, Iran
h.veisi@ut.ac.ir

2024 32nd International Conference on Electrical Engineering (ICEE) | 979-8-3503-7634-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICEE63041.2024.10667921

***Abstract*—Addressing the challenge of few-shot learning in intent classification tasks within Natural Language Processing (NLP), this study introduces a novel approach that harnesses the robust adaptation capabilities of Model-Agnostic Meta-Learning (MAML) combined with sophisticated language embeddings, namely BERT, LaBSE, and text-embedding-ada-002. The need for models to understand and classify intents with minimal training data is imperative to progress in creating versatile, responsive AI systems. We propose a methodology that leverages the generalizability of MAML and the deeply contextualized representations offered by state-of-the-art embeddings, allowing for significant improvements in Accuracy and data efficiency. We evaluate our approach using the CLINC150 dataset across a series of N-way & K-shot configurations, demonstrating the efficacy of the proposed model with varying numbers of intent classes and examples. Our findings reveal that the text-embedding-ada-002 embeddings consistently provide superior performance in both 1-shot and 5-shot settings across all class configurations tested, indicating their potent synergy with meta-learning strategies. Specifically, text-embedding-ada-002 achieved an accuracy of 97.07% in the 5-Way & 1-Shot setting and 99.1% in the 5-Way & 5-Shot setting. The outcomes of our experimental evaluation suggest that our approach also illuminates the potential of harmonious integration of cutting-edge language embeddings with meta-learning frameworks. This work provides a solid foundation for further exploration in optimizing few-shot intent classification, paving the way for creating AI systems proficient in understanding user intents with minimal exemplars. This research lays the groundwork for future advancements in few-shot intent classification, enabling the development of AI systems that require minimal training data to interpret user intent accurately.**

***Index Terms*—Few-shot learning, Model-Agnostic Meta-Learning (MAML), Intent classification, Natural Language Processing (NLP), BERT, LaBSE, and Ada.**

I. Introduction

In the swiftly evolving domain of natural language processing (NLP), the challenge of intent classification has garnered significant attention, particularly in the context of developing conversational agents and intelligent systems capable of understanding human intent with minimal exemplars. The conventional paradigms for training such systems require extensive datasets covering a comprehensive spectrum of potential intents. However, the need for labeled data in many real-world scenarios poses a daunting barrier, increasing interest in few-shot learning approaches.

Few-shot learning aims to construct learning algorithms that gain proficiency with a minimal number of training examples. This pursuit aligns closely with the inherent learning efficiency exhibited by human learners. This methodology is vital in intent classification tasks where acquiring and annotating vast amounts of data for each possible intent can be impractical or infeasible. MAML [1], a pioneering algorithm in the meta-learning landscape, has demonstrated a remarkable capacity to prepare models for quick adaptation to new tasks with limited data. By training a model on many learning tasks, MAML effectively instills learning versatility, enabling rapid convergence to optimal performance on new tasks with only a few gradient updates. Moreover, the emergence of sophisticated language representations rooted in transformer-based architectures, such as BERT (Bidirectional Encoder Representations from Transformers) [2], LaBSE (Language-agnostic BERT Sentence Embedding) [3], and text-embedding-ada-002[1] embeddings, has revolutionized our ability to capture the semantic essence of language. These embeddings have pushed the frontiers of NLP, offering deep, contextualized representations that are transferable across various tasks, including intent classification. In this work, we seek to bridge the gap between the adaptability of MAML and the profound representational power of state-of-the-art embedding techniques. By amalgamating MAML with embeddings such as BERT, LaBSE, and text-embedding-ada-002, we hypothesize that we can facilitate a meta-learning framework that not only adapts rapidly to new intents but also leverages the nuanced understanding of language inherent in these embeddings. This

This work was partially supported by ITRC (IRAN Telecommunication Research Center)

979-8-3503-7634-0/24/$31.00 ©2024 IEEE

[1]https://platform.openai.com/docs/guides/embeddings/what-are-embeddings

integration promises to be incredibly potent in the context of few-shot learning, potentially setting a new benchmark for intent classification performance in low-resource settings.

Here, we detail our exploration into this synergy, with a robust examination of MAML and its application to few-shot learning in NLP. We then delve into the intricacies of various embedding techniques, analyzing their advantages and compatibility with meta-learning paradigms. After establishing the theoretical underpinnings, our paper presents an empirical investigation where we implement and evaluate our combined approach across diverse intent classification tasks. The results aim not only to shed light on the viability of this hybrid approach but also to pave the way for future research trajectories and practical applications that require efficient, scalable, and intelligent systems adept at understanding human language with minimal supervision.

Thus, this work intends to contribute to the broader research discourse by demonstrating how the confluence of MAML and advanced embedding strategies can address a critical limitation in current NLP systems—achieving high levels of Accuracy in intent classification with minimal available data.

In the ensuing discourse, we first delve into the Related Works in section II, surveying the landscape of few-shot learning within NLP and examining the strides in meta-learning frameworks and embedding techniques that have informed our research direction. Following this, section III articulates the intricate details of our experimental design, including dataset configuration, model architecture, and the nuanced process of embedding that we employed, as well as the few-shot learning tasks, training details, and the optimization procedures we adopted. Section IV then recounts the actual application of our methodology, where we benchmark against baseline models, delineate the few-shot learning configurations, and expound upon the specific training details before presenting our results. In section V, we critically probe into our findings, dissecting the implications and significance within the context of the broader research field, culminating the paper encapsulates the main takeaways, reassessing the contributions made, and canvassing the potential avenues for future research based on the work at hand.

II. Related Works

The challenge of intent classification in natural language processing is multi-faceted, requiring the accurate deciphering of user intents from textual inputs—a critical component in developing intelligent dialogue systems. Significant strides have been made with the advent of deep learning [4], but the dependency on large annotated datasets remains a critical bottleneck [5]. As such, a rich corpus of literature focuses on overcoming data scarcity.

Few-shot Learning: Few-shot learning has been a compelling answer to the data dilemma, where models are designed to learn information from a few examples. Earlier works primarily focused on metric-based approaches such as Siamese Networks [6], Matching Networks [7], Prototypical Networks [8], and Relation Networks [9]. However, our work is more closely aligned with the meta-learning paradigm, which has been the center of attention in recent studies. Specifically, MAML [1] has shown promising results in enabling models to rapidly adapt to new tasks, with its application in NLP being explored in studies like "Domain generalization via model-agnostic learning of semantic features" [10].

Meta-Learning: The genre of meta-learning, conceptualized as 'learning to learn,' has found a spectrum of uses in NLP, yet its assimilation with intent classification remains partially untapped. Meta-learning algorithms, including MAML as well as developments like Simple Neural Attentive Meta-Learner (SNAIL) [11], AVID [12], and Meta-SGD [13], establish a protocol permitting models to assimilate and extrapolate from limited samples of unseen tasks. This functionality is advantageous in domains constrained by scant labeled data [8] [9].

Embeddings in NLP: Embedding techniques have experienced a paradigm shift by introducing transformer-based models like BERT [2], which offer deep contextualized word representations. These embeddings have set new benchmarks across various NLP tasks. LaBSE [3] extends this notion by providing language-agnostic sentence embeddings facilitating cross-lingual tasks. Furthermore, the text-embedding-ada-002 embedding paradigm introduces efficiencies and improvements in adaptability within language models, aiding in the fine-tuning process for specific tasks.

Combining Meta-Learning with Embeddings: While MAML effectively deals with few-shot learning, combining it with robust embeddings from transformer models presents a novel exploration. Studies have begun integrating these ideas, like [14], who showed the efficacy of utilizing BERT, ELMo [15], and GloVe [16] within a meta-learning framework for few-shot text classification.

Our work contributes to this emerging intersection by integrating the robust adaptability of MAML with the rich, context-aware representations of BERT, LaBSE, and text-embedding-ada-002. By doing so, we aim to create a meta-learning model that learns quickly from new intents and encapsulates a deeper understanding of semantics and transferability across languages. Such endeavor has the potential to benefit the research community and industry practitioners dealing with varied and sparse data environments, extending the reach of automated dialog systems in multiple linguistic setups.

In summary, the literature reviewed highlights the importance of few-shot learning and meta-learning, particularly MAML, in addressing data scarcity in intent classification. Despite significant advances made by large pre-trained embeddings, the exploration of their combination with meta-learning techniques like MAML for intent classification remains nascent. Our work seeks to fill this gap, offering a comprehensive study that will add a valuable perspective to the field.

III. Methodology

This section provides a detailed description of our methodology, including the preprocessing of our dataset, the nuanced

architecture of our model, and the criteria employed for evaluation.

A. Dataset Configuration

The CLINC150 dataset [17], recognized for its diverse intents, serves as the benchmark for our study. Comprising 150 distinct classes, this dataset was bifurcated into two subsets: 120 classes were dedicated to meta-training (training the model on various learning tasks). In comparison, the remaining 30 classes were reserved for meta-testing (evaluating the model's ability to generalize to new tasks).

B. Model Architecture and Embedding Process

Our model's architecture is spearheaded by transforming sentences into high-dimensional embeddings, which function as the subsequent Multi-Layer Perceptron (MLP) input. The MLP consists of an input layer, a single hidden layer, and an output layer that conducts N-way classification. The embedding dimension directly corresponds to the input size of the MLP, ensuring a harmonious transition from the embedding to the classification process. "Fig. 1", indicates our model architecture.

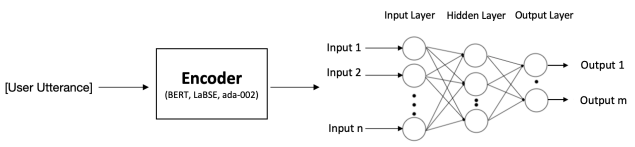

Fig. 1. Model Architecture

Across various experiments, the hidden layer architecture was assessed through the lens of different activation functions: linear, sigmoid, and ReLU. The linear activation function emerged as superior in Accuracy, making it our final model design choice.

C. Optimization Procedure

Our approach utilized the meta-learning framework inherent in MAML, which adopts a hierarchical learning rate structure to facilitate a two-tiered optimization process. The optimization was delineated into two distinct loops, each with its strategic learning function.

In the outer loop, the learning mechanism was designed to capture overarching patterns across multiple tasks. This meta-optimization step was critical in updating the model's initial parameters, ensuring that the model acquired a generalized prior that could effectively bootstrap learning for new tasks.

Conversely, the inner loop targeted rapid learning adaptation, tuning the model to absorb task-specific details. Here, the model engaged in several gradient descent steps, ensuring it could swiftly acclimate to the particulars of each task. This process was meticulously balanced to embody an efficient computational demand while allowing for the nuanced convergence required by the intricacies of the tasks.

We simplified the model architecture to a single hidden layer within the MLP to facilitate a swifter optimization process. This choice was instrumental in preventing the vanishing gradient problem, particularly in deep networks when dealing with the sparse data regimes characteristic of few-shot learning environments. This simplification heightened the optimization efficiency and supported the model's ability to learn quickly from limited data, a prominent aspect of few-shot learning scenarios.

D. Query Set and Evaluation Metrics

The intent classification was subjected to rigorous evaluation, using a query set consisting of 25 samples for each intent to calculate the model's Accuracy. The measurement of Accuracy—the proportion of correctly predicted samples in the query set—directly reflected the model's few-shot classification capabilities.

IV. Experiments and Evaluations

Our experiments aimed to evaluate our model's performance on the task of few-shot intent classification using different embedding techniques: BERT, LaBSE, and text-embedding-ada-002. We focused on various "N-way and K-shot" scenarios, where "N" represents the number of classes and "K" is the number of examples from each class during training.

A. Baseline Model

As a fundamental point of comparison, our experiments utilize a Multi-layer Perceptron (MLP) with a single hidden layer and linear activation function, excluding the application of MAML techniques. This baseline MLP model was selected for its simplicity and transparency in illustrating the raw learning capability in few-shot learning scenarios. Tailored to accommodate input feature vectors of varying sizes aligned with the embedding dimensions of the preprocessed data (768 for BERT and LaBSE, 1536 for text-embedding-ada-002), the network architecture facilitates direct Feature-to-Class mapping. It comprises an input layer that takes the specified embedding size, followed by a hidden layer whose number of neurons was methodically matched to the complexity of respective few-shot tasks, and an output layer that uses a softmax activation to yield a probability distribution over N classes.

This single-hidden-layer MLP is designed to establish baseline performance without the benefits of complex transformations or meta-learning enhancements. The choice of a linear activation function is deliberate, striving to lay bare the models' inherent discriminative power absent of non-linearities. The output through the softmax function ensures the generation of a probabilistic class membership prediction for each input. By imposing such minimality, the baseline model's outcomes ground the discussion about the advantages introduced by more sophisticated models and meta-learning techniques later in our evaluations.

B. Few-Shot Learning Configurations

We examined the adaptability and efficiency of our model across an array of few-shot learning configurations, encompassing 1-shot and 5-shot scenarios combined with 5-way, 10-way, and 20-way classification challenges. This variety ensures a comprehensive analysis of the model's performance under varying levels of task difficulty.

C. Training Details

In our experimentation, we meticulously crafted a training regimen aimed at optimizing model performance for few-shot intent classification tasks. Surprisingly, we observed that a model with a single hidden layer outperformed its multi-layered counterparts, sparking a point of significant interest. This insight prompted us to delve into the training process that led to this pivotal discovery.

Initially, our hypothesis presumed that deeper networks would deliver superior performance due to their augmented feature representation capacity. To our surprise, experimental results consistently favored a single hidden layer MLP model over those with multiple hidden layers. This peculiar finding strongly suggests that model simplicity correlates positively with effective generalization and performance within few-shot learning and intent classification.

Our MLP culminates with a softmax activation function, aligning with the standard approach for multi-class classification problems. This layer ensures the representation of a probability distribution over the 'N-way' classes, producing a probability distribution between 0 and 1 for each input sample, summing to 1 across the class labels. As a result, the highest probability within this distribution serves as the model's predicted class.

Throughout our experiments, we explored hidden layer sizes of 64, 128, 256, 384, 512, and 768 to understand their impact on the model's performance.

The MLP model, using MAML, was optimized with a range of learning rates for adaptability (outer loop learning rates: 0.001, 0.004, 0.005, 0.007; inner loop learning rates: 0.01 and 0.03). We used 50 gradient steps for adaptation during the inner loop of MAML. For activations in the hidden layer, the linear function yielded the best Accuracy, which was noteworthy given the nonlinear nature of the embeddings.

D. Optimization Configuration

In configuring our optimization, we honed the MLP model with an array of learning rates to enhance its flexibility, selecting outer loop rates at 0.001, 0.004, 0.005, and 0.007, along with 0.01 and 0.03 for the inner loop. The adaptation phase within MAML's inner loop was conducted over 50 gradient adjustments. Interestingly, a linear function for hidden layer activations delivered the highest accuracy despite the typically nonlinear characteristics of the embeddings involved.

Moreover, we adopted a streamlined network design featuring just one hidden layer in the MLP to strike a balance between rapid computation and thorough training. This simplified architecture hastened the optimization phase and helped avoid the vanishing gradient problem—a significant hurdle in few-shot learning scenarios with sparse data.

E. Results

To interpret the performance of different embeddings in few-shot learning tasks, we measured the Accuracy of both baseline models and our proposed model across various task configurations. The results are methodically organized in two tables, delineating outcomes with and without the MAML approach, providing a comparative view and insight into the efficacy of meta-learning on model performance.

Table I encapsulates the Accuracy achieved by baseline models utilizing BERT, LaBSE, and text-embedding-ada-002 embeddings in the absence of the MAML approach. The models were assessed under 5-way 1-shot, 5-way 5-shot, 10-way 1-shot, 10-way 5-shot, 20-way 1-shot, and 20-way 5-shot configurations, where a clear pattern emerges, indicating that the text-embedding-ada-002 embedding tends to yield superior performance in more constrained (1-shot) scenarios.

TABLE I
ACCURACY IN TASKS WITHOUT MAML APPROACH

Task Configuration	BERT	LaBSE	text-embedding-ada-002
5-Way 1-Shot	80.37%	85.27%	96.7%
5-Way 5-Shot	95.17%	97.95%	98.9%
10-Way 1-Shot	69.04%	76.8%	87.16%
10-Way 5-Shot	92.3%	89.75%	93.4%
20-Way 1-Shot	60.5%	50.73%	52.34%
20-Way 5-Shot	87.16%	57.2%	55.0%
Average	80.76%	76.28%	80.58%

Conversely, Table II reports the accuracy improvements invoked by applying the MAML approach to the same models and task configurations. The enhancements are particularly noteworthy in 1-shot scenarios, suggesting that MAML significantly leverages minimal data to foster better generalization capabilities in the models. The optimization technique has a profound impact, with each embedding type exhibiting a marked improvement.

TABLE II
ACCURACY IN TASKS WITH MAML APPROACH

Task Configuration	BERT	LaBSE	text-embedding-ada-002
5-Way 1-Shot	89.75%	92.6%	97.07%
5-Way 5-Shot	95.8%	98.0%	99.1%
10-Way 1-Shot	83.4%	88.0%	94.04%
10-Way 5-Shot	94.97%	96.6%	98.54%
20-Way 1-Shot	78.0%	81.5%	90.67%
20-Way 5-Shot	92.43%	91.5%	97.27%
Average	88.92%	91.12%	95.92%

V. CONCLUSION AND DISCUSSION

Throughout all evaluated task configurations, the models utilizing text-embedding-ada-002 embeddings have displayed superior Accuracy to those employing BERT and LaBSE. This consistent outperformance underscores the prowess of text-embedding-ada-002 embeddings in more effectively capturing the nuances of sentence semantics crucial for few-shot

learning. A distinct observation is the remarkable accuracy improvement in the 5-shot scenarios, which demonstrates the models' ability to capitalize on the increment in example data, translating it into noteworthy enhancements in outcomes. We can see the comparison of models in "Fig. 2".

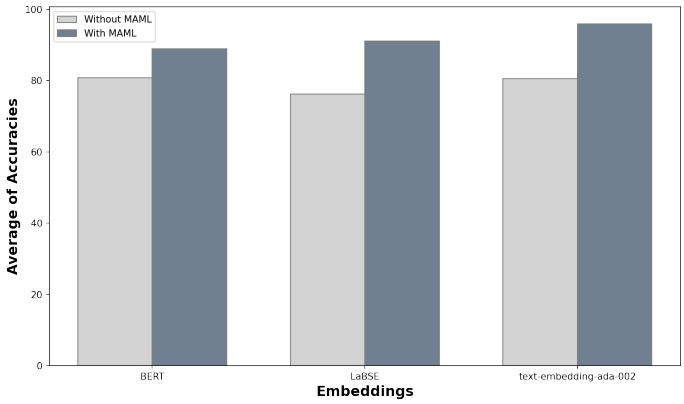

Fig. 2. Comparison of Average Embeddings Scores: Without MAML vs With MAML

Further analysis reveals that with meticulous tuning, the MLP, coupled with the MAML framework, shows an impressive capacity to acclimate swiftly to novel tasks within a few-shot learning paradigm. This adaptability is especially pronounced with text-embedding-ada-002 embedding integration. Intriguingly, the unexpectedly high performance of the linear activation function used within the MLP architecture suggests a synergy with the embedding space that merits deeper exploration. The implications of such results indicate that the relationship between simplicity in activation and the complexity of embeddings can profoundly affect the model's learning dynamics in few-shot settings.

Through these experiments, we have substantiated that combining MAML with cutting-edge embeddings substantially advances the field of few-shot intent classification. The text-embedding-ada-002 embeddings, when paired with MAML, present a significant improvement in model adaptability and performance across various few-shot learning scenarios, illuminating a clear path for future research and practical implementations.

REFERENCES

[1] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 1126–1135, PMLR, 2017. ISSN: 2640-3498.

[2] 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, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), pp. 4171–4186, Association for Computational Linguistics, 2019.

[3] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (S. Muresan, P. Nakov, and A. Villavicencio, eds.), pp. 878–891, Association for Computational Linguistics, 2022.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," vol. 521, no. 7553, pp. 436–444, 2015.

[5] J. Schmidhuber, "Deep learning in neural networks: An overview," 2014.

[6] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, Lille, 2015. Issue: 1.

[7] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," vol. 29, 2016.

[8] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[9] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1199–1208, 2018.

[10] Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," vol. 32, 2019.

[11] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," 2018.

[12] A. Javaheri, S. R. Kheradpisheh, H. Farahani, A. G. Khoee, and M. Ganjtabesh, "Avid: A variational inference deliberation for meta-learning," in *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 268–273, 2022.

[13] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-SGD: Learning to learn quickly for few-shot learning," 2017.

[14] J. Krone, Y. Zhang, and M. Diab, "Learning to classify intents and slot labels given a handful of examples," in *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI* (T.-H. Wen, A. Celikyilmaz, Z. Yu, A. Papangelis, M. Eric, A. Kumar, I. Casanueva, and R. Shah, eds.), pp. 96–108, Association for Computational Linguistics, 2020.

[15] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (M. Walker, H. Ji, and A. Stent, eds.), pp. 2227–2237, Association for Computational Linguistics, 2018.

[16] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Association for Computational Linguistics, 2014.

[17] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang, and J. Mars, "An evaluation dataset for intent classification and out-of-scope prediction," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), pp. 1311–1316, Association for Computational Linguistics, 2019.