2021 International Conference on Artificial Intelligence and Electromechanical Automation (AIEA) | 978-1-6654-3265-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/AIEA53260.2021.00069

A Survey on Few-Shot Learning in Natural Language Processing

Mengde Yang*
School of Information Engineering
China University of Geosciences, Beijing
Beijing, China
*1004181116@cugb.edu.cn

***Abstract*—The annotated dataset is the foundation for Supervised Natural Language Processing. However, the cost of obtaining dataset is high. In recent years, the Few-Shot Learning has gradually attracted the attention of researchers. From the definition, in this paper, we conclude the difference in Few-Shot Learning between Natural Language Processing and Computer Vision. On that basis, the current Few-Shot Learning on Natural Language Processing is summarized, including Transfer Learning, Meta Learning and Knowledge Distillation. Furthermore, we conclude the solutions to Few-Shot Learning in Natural Language Processing, such as the method based on Distant Supervision, Meta Learning and Knowledge Distillation. Finally, we present the challenges facing Few-Shot Learning in Natural Language Processing.**

Keywords- Few-Shot Learning; Natural Language Processing; Transfer Learning; Meta Learning; Distant Supervision

I. INTRODUCTION

Humans are good at identifying a new object through a very small number of samples, such as children who only need some pictures in a book to know what a zebra is and what a rhinoceros is. When we apply the machine learning method of practical situations, there are often problems with fewer samples. Inspired by the rapid learning ability of human beings, researchers hope that after learning a large amount of data from a certain category, the machine learning model can learn quickly with only a small number of samples for the new category, which is the goal of the Few-Shot Learning solve [1,2].

A. What Is the Few-Shot Learning?

Suppose a P is used to evaluate the performance of a computer program on a task class T. If a program improves performance on T task by using experience E, then we say that the program learns about the T and the program.

Few-Shot Learning is defined as A type of machine learning problems (specified by E, T and P) where E contains a little supervised information for the target T. Then Few-Shot Learning as a part of machine learning[4], the difference between them is the size of the experience E. As the name implies, a small sample size leads to less information from a small number of samples, which leads to less experience E.

One goal of Few-Shot Learning is to solve the problem of few data, and it can also solve the long-tail problem. A small number of categories occupy the majority of samples. A number of categories have only a small number of samples.

B. Few Shot Learning on Natural Language Processing

The early research of Few-Shot Learning algorithms focuses on image, and Few-Shot Learning models can be divided into Mode Based, Metric Based, and Optimization Based. At present, there are many algorithms for Few-Shot Learning which focused on small sample image recognition.

Few-Shot Learning data sets and models have also begun to appear in the field of Natural Language Processing in recent years. Compared with images, text semantics contains more changes and noise. The current research status of Few-shot Learning algorithms focusing on natural language tasks is that there are some data sets such as FewRel, ARSC, and ODIC[3].

C. Motivation

Few-Shot learning is very effective in solving the problem of Few-Shot learning, so Few-Shot Learning development is very important for Natural Language Processing. However, there are still many problems, such as the core problem of Few-Shot Learning is the unreliable empirical risk minimization mechanism, making it difficult for Few-Shot Learning to learn[4]. Therefore, we investigate the existing work, show the current progress in this field, report on representative work, and discuss the key problems and challenges in this field.

II. BACKGROUND

Relevant knowledge background includes the development of technology related to Few-Shot learning in Natural Language Processing aspects. They provide technical support for Few-Shot learning. The following three main technologies are introduced :(1) Transfer Learning; (2)Meta-Learning; (3) Knowledge Distillation.

A. Transfer Learning

1) What is Transfer Learning?

Transfer Learning is also called inductive transfer, domain adaptation. Its goal is to apply the knowledge or patterns learned in a field or task to different but related fields or problems [5]. For example, training embedded layers and fine-tuning models.

The main idea of Transfer Learning is to transfer annotated data or knowledge structure from relevant auxiliary fields, and to improve the learning effect of target areas or tasks.

There are many solutions of Transfer Learning in Few-Shot Learning on Natural Language Processing, for example, Multi-task Transfer Learning. The main idea of Multi-task Transfer Learning is to use data samples of source tasks in Transfer

978-1-6654-3265-8/21/$31.00 ©2021 IEEE
DOI 10.1109/AIEA53260.2021.00069

Learning, but to consider all characters together in multi-task. The goal is to learn effectively from a few target samples under reasonable assumptions about the source hypothesis. In addition, there are nonparametric sparse topic model migration algorithms, integrated learning instance migration algorithms and so on [6].

2) Transferability of deep learning

In the deep learning model, the first few layers learn general features; as the network level deepens, the latter network focuses on learning task-specific features so that the general features can be transferred to other fields.

- **The simplest deep network migration: Finetune (Finetuning, fine-tuning).** Finetune is to use the network that others have trained to fix the parameters of the previous layers, only for our task, fine-tune the latter layers. We don't usually train a neural network from scratch for a new task because in practical applications. Such an operation is clearly very time-consuming. Especially, our training data cannot be as large as ImageNet, which can train deep neural networks with strong generalization ability. Even with so much training data, the cost of training from scratch is unbearable.

  In the current task in the field of computer vision, the proposed method has generally used the strategy of deep migration for pre-training. The depth CNN model is trained using large-scale image data sets, for example, ImageNet, because the number of samples and parameters is very large, even using GPU acceleration will take a long training time. But another advantage of depth CNN architecture is that the pre-trained network model can separate the network structure from the parameter information. So long as the network structure is consistent, the network can be constructed and initialized by using the trained weight parameters. Greatly saves the network training time.

3) The Limits of Transfer Learning

It includes the overfitting and underfitting problems of classical machine learning, as well as the underfitting and negative transfer problems of Transfer Learning.

- **Negative migration.** Auxiliary domain tasks have a negative effect on target domain tasks. At present, the main idea of studying negative migration from the perspective of algorithm design is to reduce the knowledge structure of interdomain migration, such as sharing the prior probability of the model only between domains, but not the model parameters or likelihood functions.
- **Under adaptation.** The problem of probability distribution adaptation across domains cannot be fully corrected.
- **Underfitting.** The learning model fails to fully characterize the important structure of probability distribution.
- **Overfitting.** Learning model overfitting independent information of sample distribution.

B. Meta Learning

At present, there are many methods for meta-learning[7], mainly divided into the following categories:

- Based on the memory approach. The basic idea is learning from experience, based on the memory approach by adding Memory to neural networks.
- Based on prediction gradient approach. The purpose of Meta Learning is to realize fast learning, and the key point to realize fast learning is that the gradient descent of neural network should be accurate and fast, so that neural network can use previous tasks to learn how to predict the gradient. In the face of new tasks, learning will be fast [8].

C. Knowledge Distillation

Hinton proposes the concept [9] of Knowledge Distillation to achieve knowledge transfer by introducing soft targets associated with teacher networks (teacher network： complex but superior inferential performance) as part of the total loss to induce training in student networks (student network： streamlined, low complexity).

1) Typical Method of Knowledge Distillation and Application

At present, Knowledge Distillation has become an independent research direction[10], various new technologies emerge endlessly. However, if summed up roughly, the mainstream Knowledge Distillation technology has two main development lines: Logits method and characteristic distillation method.

Common recommendation systems generally have three cascading processes: recall, coarse row, and fine row. The recall link quickly selects some items that may be of interest to some users from the mass goods library and passes them to the coarse row module. The coarse row link usually uses a simple sorting model with a small number of features to sort the recalled materials. Among them, the coarse row link according to a specific application can choose not to choose.

The performance and effect of the existing recommendation system can be optimized by using the Knowledge Distillation technology in the fine row, the coarse row and the model recall service and the good quality of the recommendation.

III. FEW-SHOT LEARNING METHODS IN NATURAL LANGUAGE PROCESSING

There are many related methods in solving Natural Language Processing problems with Few-Shot Learning. In the following, we mainly introduce the three methods that are now more mainstream: Distant Supervision, Few-Shot Learning based on Meta Learning and Knowledge Distillation.

A. Few-Shot Learning based on Distant Supervision

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

Assumption. If two entities have a relationship in a known knowledge base, then all sentences that mention these two entities will express that relationship in some way [11].

Distant Supervision for Few-Shot Learning consists of two parts:

(1) Distant Supervision way to build data set. Create a large candidate sentence set web encyclopedia as a corpus through Distant Supervision and use network data as a knowledge base[12]. Delete the relationship where the number of instances is less than a certain value and leave the instance randomly for the rest of the relationships. A candidate set containing relationships and instances is obtained.

If the sentence is incomplete or the reference is incorrectly linked to the entity, the annotator is asked to mark an instance as negative[12,13]. To ensure the quality of the annotation, each instance is marked by at least two annotators. If two annotators disagree on this instance, it will be assigned to the third annotator. So, each instance has at least two identical annotations, which will be the final decision.

(2) Learning from Distant Supervision dataset. Since the Distant Supervision method can bring a large amount of noise data, most methods based on distant Supervision use attention mechanism to reduce noise [14].

B. Few-Shot Learning based on Meta Learning

The Meta-Learning can be achieved well in Few-Shot Learning, so researchers focus on Meta-Learning. There are three ways to achieve Few-Shot Learning by Meta-Learning: Metric-based, model-based, and based on optimization. Next, we take model-based Meta-Learning as an example to introduce how to implement Few-Shot Learning by Meta-Learning.

Firstly, we can use the distributed features of words and a Meta-Learning text classification model with fewer samples.

Then we can use an attention generator and a ridge regressor.

In order to create a training segment [15], a model-based method first randomly takes a category from each category, and then takes a training sample from each category as the training set and a sample as the test set. So, there is a training sample and a test sample. Generally speaking, we refer to this training sample as a support set (support set) and the test sample as an inquiry set (query set). In the Meta-Learning test phase, for each test segment, the model-based method randomly takes a class from it, then takes the support set and the query set from this class, and then verifies the model effect in the query set of all segments[16].

Attention generator. This module generates class-based attention size by combining source pool distributed features and support sets, and then the generated attention is used in the ridge regression to correct the deviation of word importance

Ridge Tractor. For each segment, the module accepts the attention force value and constructs a lexical representation, which is then predicated on the query set

C. Few-Shot Learning based on Knowledge Distillation

The mainstream knowledge distillation[17] technology has two main development lines: Logits method and characteristic distillation method. We mainly take the logits scheme as an example to introduce knowledge distillation to general classification problems, such as inputting a text. After DNN various nonlinear transformations of the network, near the last layer of the network, the size of the text belongs to each category Zt. the larger the Zt value of a category, the more likely the model thinks that input text belongs to this category. Logits is the summary score z of each category after summarizing the various information within the network. i represents the i category, Zt represents the possibility of falling into category i. Since Logits is not a probability value, Softmax function is generally used to transform the Logits value, and the probability value is taken as the probability of the final classification result[18]. Softmax on the one hand, the probability of Logits value is normalized between different categories; on the other hand, it amplifies the difference between Logits values and polarizes the Logits score. The probability of obtaining high Logits is larger, while that of lower Logits is smaller. On the basis of understanding what is Logits, the following is a Logits distillation method. Suppose we have a Teacher network, a Student network, input the same data to these two networks. Teacher will get a Logits vector, representing the possibility that the input data belongs to each category. Student also has a Logits vector. Represent the possibility that the input data belongs to each category. The simplest and earliest work of knowledge distillation is to allow Student Logits to fit the Teacher. That is, the Student loss function is:

$$L_{student} = \|Z_t - Z_s\|^2 \quad (1)$$

The Zt is Teacher Logits, Zs is Student Logits. The Logits of Teacher here is the dark knowledge passed on to Student.

Hinton put forward an improved method called Softmax Temperature. He was the first man who formally put forward the "knowledge distillation" and introduced Temperature T. If we set the T to 1, A standard Softmax function, that is, the extremely polarized version. If the T is big, Then the Logits value after Softmax, the probability score gap between categories will narrow. Conversely, it will increase the polarization of the probability between categories. Hinton version of knowledge distillation is that let Student fit the Teacher after T influence [19]. Student loss function consists of two terms, a subitem is Ground Truth, let student fit the training data. The other is distillation loss, let Student fit the Logits of the Teacher:

$$L_{student} = H(y, f(x)) + \lambda * H\big(ST(Z_t), ST((Z_S))\big) \quad (2)$$

H is a cross - entropy loss function. f (x) is the mapping function of the Student model. y is Ground Truth Label. Zt is Teacher Logits. Zs is Student Logits. ST() are Softmax Temperature functions. λ used to regulate the degree of influence of distillation Loss.

Typically, temperature T should be set to a value greater than 1, which reduces the polarization of different categories of attribution probability. Because in Logits method, the additional information Teacher can provide to the Student is included in the Logits value.

IV. The Challenges Of Few-Shot Learning In Natural Language Processing

Existing Few-Shot Learning methods often used prior knowledge from a single mode (e.g., image, text, or video). However, although there are some examples of modes currently used, there may be another mode with a large number of supervised samples. The study of extinct animals is an example.

Recently, there are some efforts to borrow techniques from Zero-Shot Learning methods to solve Few-Shot Learning problems. However, using a small number of samples for fine-tuning may lead to overfitting.

The realization of language intelligence is very difficult. Although robots frequently surpass human beings in e-sports and go, the current system cannot achieve the language and understanding ability of children aged three or four after Few-Shot Learning..

V. Conclusions

Few-Shot Learning aims to bridge the gap between artificial intelligence and human learning. By integrating prior knowledge, it can learn new tasks that contain only a few examples and supervised information. In this study, we provide a comprehensive and systematic review of Few-Shot Learning. A definition of the Few-Shot Learning is presented, and the circumstances that are consistent with it are discussed. The difference between Zero-Shot Learning and Few-Shot Learning is finally explained. Analysis of why Few-Shot Learning should be studied and introduce the progress of current Few-Shot Learning in Natural Language Processing. This paper also introduces the current knowledge background from three aspects: Transfer Learning, Meta-learning, and Knowledge Distillation. The third part introduces three related.

References

[1] Wang, Yaqing, et al. "Generalizing from a Few Examples: A Survey on Few-shot Learning." ACM Computing Surveys 53.3(2020):1-34.

[2] Blaes, Sebastian, and T. Burwick. "Few-Shot learning in deep networks through global prototyping." Neural networks: the official journal of the International Neural Network Society 94(2017):159-172.

[3] Han, Xu, et al. "FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation." (2018).

[4] Li, Shuai, et al. "Few-Shot Learning for Monocular Depth Estimation Based on Local Object Relationship." 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) IEEE, 2020.

[5] Pan, Sinno Jialin, and Q. Yang. "A Survey on Transfer Learning." IEEE Transactions on Knowledge and Data Engineering 22.10(2010):1345-1359.

[6] Taylor, Matthew Edmund, and P. H. Stone. "Transfer Learning for Reinforcement Learning Domains: A Survey." The Journal of Machine Learning Research 10.10(2009):1633-1685.

[7] Ye, Han Jia, X. R. Sheng, and D. C. Zhan. "Few-shot learning with adaptively initialized task optimizer: a practical meta-learning approach." Machine Learning 109.3(2020):643-664.

[8] Alexandros, Kalousis, and H. Melanie. "Meta-learning." International Journal on Artificial Intelligence Tools 10.04(2001):525-554.

[9] Lu, Liang, M. Guo, and S. Renals. "Knowledge distillation for small-footprint highway networks." (2017).

[10] Fukuda, Takashi, et al. "Efficient Knowledge Distillation from an Ensemble of Teachers." Interspeech 2017 2017.

[11] Purver, Matthew, and S. Battersby. "Experimenting with Distant Supervision for Emotion Classification." Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL) 2012.

[12] Woods, D. D., and L. G. Shattuck. "Distant Supervision–Local Action Given the Potential for Surprise." Cognition Technology & Work 2.4(2000):242-245.

[13] Mintz, Mike, et al. "Distant supervision for relation extraction without labeled data." International Joint Conference on Acl Association for Computational Linguistics, 2009.

[14] Gao, Tianyu, et al. "Hybrid Attention-Based Prototypical Networks for Noisy Few-Shot Relation Classification." Proceedings of the AAAI Conference on Artificial Intelligence 33(2019):6407-6414.

[15] Bettis, Richard, and M. A. Hitt. "Dynamic Core Competences through Meta-Learning and Strategic Context." Journal of Management 22.4(1996):549-569.

[16] Doya, Kenji. "Metalearning and neuromodulation." Neural Networks 15.4-6(2002):495-506.

[17] Yim, Junho, et al. "A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, 2017.

[18] Nakashole, Ndapandula, and R. Flauger. "Knowledge Distillation for Bilingual Dictionary Induction." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing 2017.

[19] Jankowski, Norbert, and K. Grabczewski. "Gained Knowledge Exchange and Analysis for Meta-Learning." International Conference on Machine Learning & Cybernetics IEEE, 2007.