NTT DATA
Trusted Global Innovator

Few-shot Learning in Natural Language Processing

Trends and the Future

NTT DATA Corporation
Toyosu Center Building, 3-3, Toyosu 3-chome, Koto-ku, Tokyo
Phone: +81-3-5546-8202
https://www.nttdata.com/global/en/

table of contents

CHAPTER.1

Background

Besides Machine Learning (ML) algorithms, training data is also the backbone of most Natural Language Processing (NLP) empowered products. Especially, since the era of Deep Neural Network (DNN) models become dominant in nearly all AI areas including Computer Vision (CV) and Automatic Speech Recognition (ASR), etc., a greater abundance of training data was required to fit a multi-layer Neural Network (NN) model containing a significantly larger number of parameters. Because a model trained for a certain purpose barely has versatility for other tasks, it always has been a pain to prepare a large amount of training data for a new task.

Figure 1: Traditional model training

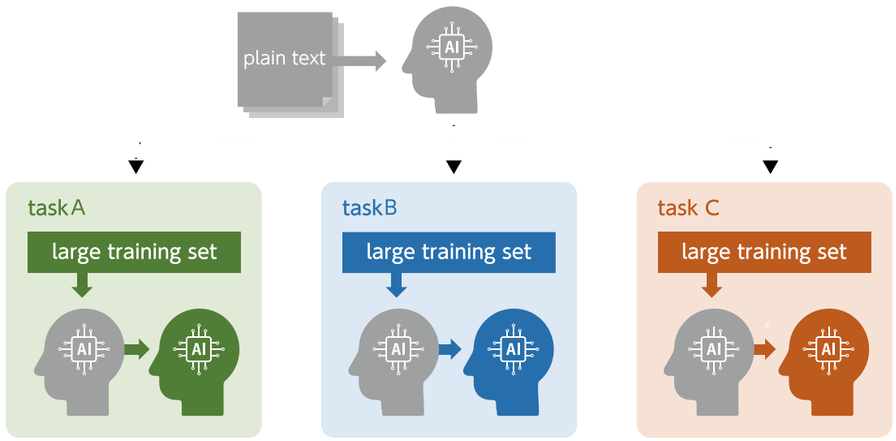


Figure 2: Pre-train and Fine-tune

In late 2018, the proposal of Bidirectional Transformers NN structure (e.g., BERT) made it possible to pre-train a large-scale language model for general purpose using only easy-to-acquire plain text data (Wikipedia text, books, etc.). For each word in an input text, this type of architecture can take into consideration of the full context by looking at words come both before and after it, which makes it suitable for various types of NLP tasks. Using the large pre-trained language model (PLM) as a good start point, NLP tasks could be then solved using a smaller amount of task-specific training data. To put it in detail, the parameters of the PLM are then fine-tuned using task-specific training data to be optimal for the task in question. This type of Pre-train & Fine-tune paradigm achieved state-of-the-art performance in most of the benchmark tasks even with a limited amount of training data and began to be dominant in NLP recently. Figure 1 and Figure2 demonstrate the difference between traditional model training and the Pre-train & Fine-tune paradigm. From a human perspective, it is not difficult to understand that becoming a specialist in different areas after learning the required language is more efficient than starting from specialized knowledge without knowing the language itself. However, current technology still roughly requires at least over a thousand training examples to fine-tune a qualified model.

Creating training data can involve different manual efforts according to different tasks. Crafting training data for general-purpose tasks such as sentiment analysis or topic classification normally can be accomplished by annotators with basic common sense. On the other hand, domain-specific tasks such as symptom name extraction or legal document classification usually require a certain level of in-depth expertise in a particular area. The above-mentioned problems have become one of the biggest bottlenecks for real-world businesses. Because there is hardly a versatile NLP model for all purposes, it always takes a large portion of the budget to create training data for every use case. Worst cases happen when clients are the only persons who have the expertise to create appropriate training data. Therefore, NLP providers tend to be in a predicament of cooking without gradients.

Approaches such as automatic data augmentation were proposed to alleviate the lack of training data problems. Data augmentation mainly focuses on automatically generating different variations of the original training data to increase the training data size. Like data augmentation techniques used in CV where images are either rotated, cropped, or processed by different types of filters, NLP data augmentation follows the same basic idea to apply modifications to different components of the original text data. However, text data is sometimes more sensitive to modifications. There are many cases where even a single word replacement changes the meaning of the original text to the complete opposite.

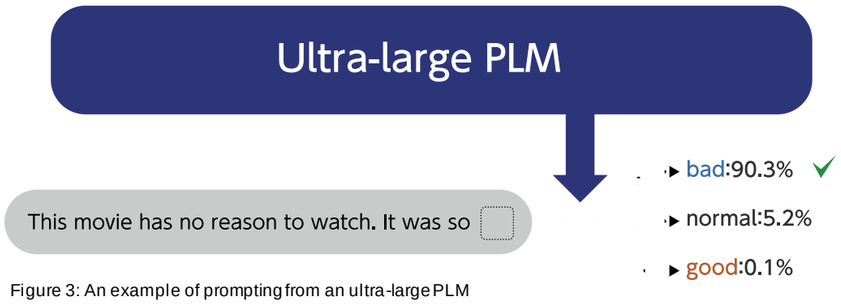


Figure 3: An example of prompting from an ultra-large PLM

Ultra-large PLM such as GPT-3 have shown the ability to solve a new task using natural language prompts and few labeled samples as demonstrations without task-specific fine-tuning. As shown in Figure3, with the high capability of context-based word prediction, information such as sentiment of movie reviews can be naturally induced from a large PLM. Although this type of approach becomes a milestone in PLMs applications and achieved competitive results compare to state-of-the-art pre-train & fine-tune paradigm, the imperative number of parameters up to 175B makes it difficult to apply domain-specific fine-tuning for further improvement.

CHAPTER.2

Survey of few-shot technology for NLP

To get a good grasp of the latest few-shot learning techniques for NLP, we investigated PLM-based few-shot applications provided in real-world services. In the theoretical perspective, NLP few-shot approaches in fundamental research were also taken into consideration.

2.1. Use cases in real-world business

We first investigated more than 100 businesses and start-ups, and their use of PLMs -- GPT-3 in particular. GPT-3 is available as a service from OpenAI, so it is quite easy to build a successful business in very little time, as the requests to GPT-3 are handled through a simple API. We catalogued the use cases on a variety of multiple factors: the task they are tackling and how they are solving it, the level of "maturity" and novelty of their application, the type of data needed and overall performance.

Here we present some other concrete examples of successful and mature applications of GPT-3 in business from smaller start-ups. They utilize GPT-3 in its most basic task of text generation with very little human input. Copysmith.ai offers a copywriting assistant based on different templates, from content enhancement and blog writing. JiraPT-3 can generate JIRA tickets based on task description written by users. In addition, Figma deals with more complex generation tasks. Figma is a code generator application that can handle multiple coding languages such as HTML and SQL.

Other than generation tasks, other applications such as Algolia provides a helpdesk service that returns the best match of a given question from a list of documents such as manuals. Viable is a tool that analyzes user feedback using embedding information from GPT-3 model. Rare applications such as Algomo utilizes GPT-3 to perform few-shot name entity recognition (NER) task.

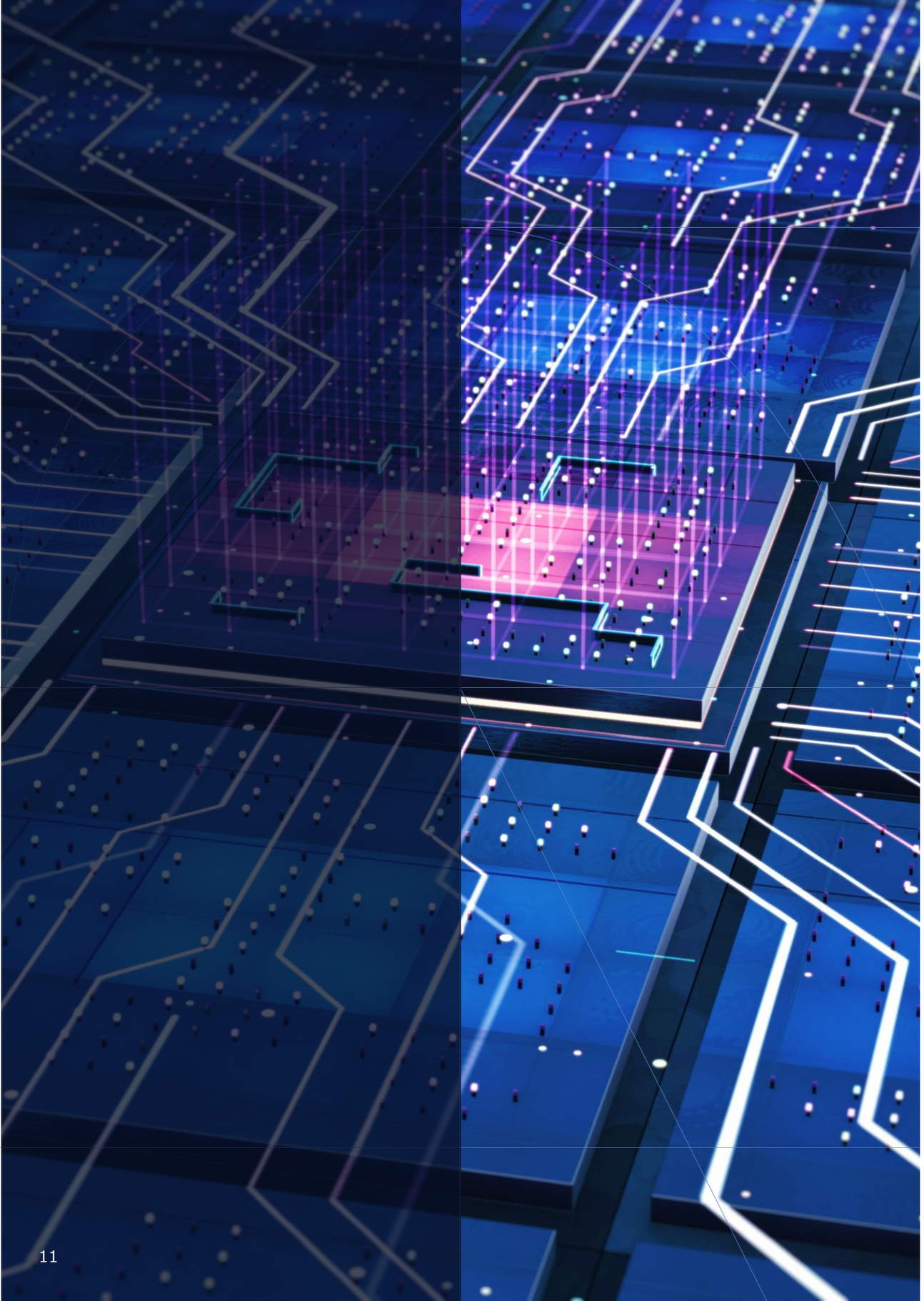
In conclusion, among these businesses, most of them utilize GPT-3 as a tool to generate content, to summarize, or to translate texts. Most of these services are offered by tech start-ups which focus on one of those specific tasks. Moreover, most solutions require as little human input as possible, while, since GPT-3 can output very human-like texts, the generated content is directly targeted to and read by humans.

2.2. Fundamental Research

We conducted a brief survey of the latest papers published in the top conferences including ACL 2022 and EMNLP 2021. We narrowed down all the titles to those contain keywords such as “few-shot” or “zero-shot” and picked up around 90 papers as our main investigation targets.

2.2.1 Types of few-shot learning approaches

Since fine-tuning PLMs usually requires a target dataset of 1K ~ 10K of examples to achieve good performance, many works tried to integrate domain knowledge from downstream tasks into the PLMs to reduce the need of training data. For example, unlabeled data from the target domain were used to fine-tune the PLMS[1][2]. Automatically generated synthetic data were



used to train an intermediate task which is closely related and beneficial to the target task[3]. Text information from labels was also exploited when integrate domain knowledge into pre-trained models[4].

Other research emphasized more on combining different types of learning strategies to improve performance. For example, given a set of unlabeled data along with another set of auto-augmented data using the same unlabeled dataset, a well-trained model should have consistent prediction distribution for both datasets. Based on this idea, consistency training was proposed[5]. Other learning strategies such as meta learning was also widely used to apply to new tasks with less training data. Self-training approaches[6][7] was proposed to use an iteration-based strategy to repeat the processes of acquiring new labeled data produced by a base model and training a new model using the extended dataset.

Inspired by human behaviors when learning from very few examples, contrastive learning[8] was introduced whose learning objective can push the examples from the same class close and the examples from different classes apart.

Recently, prompt-based approaches[9] became a trend among few-shot learning works. This type of approach leveraged the generative capability of the PLMs to naturally produce specific tokens representing the label to be predicted. Other than relying on PLM like GPT-3, recent works[10] made it feasible to utilize smaller PLMs such as BERT as well. On the other hand, prompting involved a certain level of manual effort such as engineering proper prompting examples and manual mapping between generated tokens to pre-defined labels. There were also approaches[11] proposed to reduce these types of manual effort.

[1] Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. Phang et al., 2019
[2] Label Semantic Aware Pre-training for Few-shot Text Classification. Mueller et al., 2022
[3] Unsupervised Data Augmentation for Consistency Training. Xie et al., 2019
[4] Self-training Improves Pre-training for Natural Language Understanding. Du et al., 2021
[5] STraTA: Self-Training with Task Augmentation for Better Few-shot Learning. Vu et al., 2021
[6] Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. Gunel et al., 2022
[7] Language Models are Few-Shot Learners. Brown et al., 2020
[8] Making Pre-trained Language Models Better Few-shot Learners. Gao et al., 2021
[9] LM-BFF-MS: Improving Few-Shot Fine-tuning of Language Models based on Multiple Soft Demonstration Memory. Park et al., 2022
[10] Making Pre-trained Language Models Better Few-shot Learners. Gao et al., 2021
[11] LM-BFF-MS: Improving Few-Shot Fine-tuning of Language Models based on Multiple Soft Demonstration Memory. Park et al., 2022

2.2.2 Performance of few-shot learning techniques

We explored the latest experimental results for 23 commonly used NLP benchmark tasks (Table 1) under different few-shot settings. Even for the same benchmark task, the amount of training data was used could vary in different previous works. As a result, commonly used few-shot settings such as zero-shot, 8-shot, 16-shot, and 32-shot are chosen in our survey. The best scores among all the approaches we investigated were recorded to represent the state-of-the-art results for each few-shot setting. Note that within this scale of investigation, it is still intractable to cover every experimental result for each benchmark task. Experimental results using full-size training data are essential to be set as an upper bound for each benchmark task but were sometimes missing in these papers. Therefore, we also investigated the leaderboards in “Papers with Code”, a free resource to find the state-of-the-art machine learning related papers and source code, as supplementary materials.

task name	labels descriptions	input type	task type
SST-2	positive, negative	one sentence	sentiment
MR	positive, negative	one sentence	sentiment
CR	positive, negative	one sentence	sentiment
Subj	subjective, objective	one sentence	subjectivity
OS	hate speech, benign twitter	one sentence	topic classification
MNLI	contradiction, entailment, neutral	sentence pair	entailment
IMDB	positive, negative	one sentence	sentiment
TREC	abbr., entity, description, human, loc, num	one sentence	question classification
AG News	world, sports, business, science	one sentence	topic classification
QQP	equivalent, not_equivalent	sentence pair	paraphrase
MRPC	equivalent, not_equivalent	sentence pair	paraphrase
QNLI	entailment, not_entailment	sentence pair	NLI
SNLI	contradiction, entailment, neutral	sentence pair	NLI
RTE	entailment, not_entailment	sentence pair	NLI
STS-B	[0, 5] continuous score	sentence pair	sentence similarity
BoolQ	yes, no	sentence pair	QA
SciTail	yes, no	sentence pair	NLI
SICK-E	contradiction, entailment, neutral	sentence pair	NLI
CoLA	grammatical, not grammatical	one sentence	grammtical acceptability
MPQA	positive, negative	one sentence	opinioan polarity
SICK-R	[0, 5] continuous score	sentence pair	similarity
SST-5	negative, somewhat negative, neutral, somewhat positive or positive	one sentence	sentiment
Yelp	5-scale user sentiments	one sentence	sentiment

Table 1: list of benchmarktasks

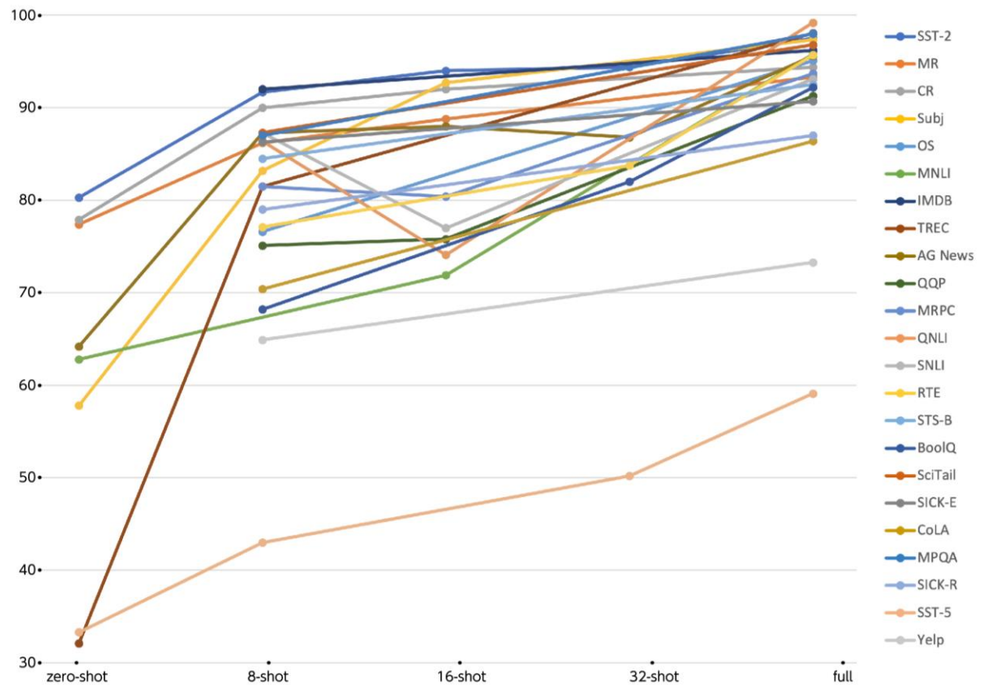


Figure 4: best scores under different few-shot settings

Thanks to the rapid evolution in techniques utilizing large-scale pre-trained language models, we found that most of the common benchmark tasks have achieved the overall full-dataset fine-tuning accuracy over 90% despite very few cases due to data-specific characteristics. Most tasks were able to maintain a score over 80% even when given only 16 samples per class for training, except some tasks take two sentences as input such as “QQP” and “MNLI” etc.

To compare few-shot with full-dataset fine-tuning for each task, we calculated the score drop rate from the score using full dataset for each few-shot setting. Figure 5 shows the results for each benchmark task. Few-shot settings that were not found within investigated works were simply ignored.

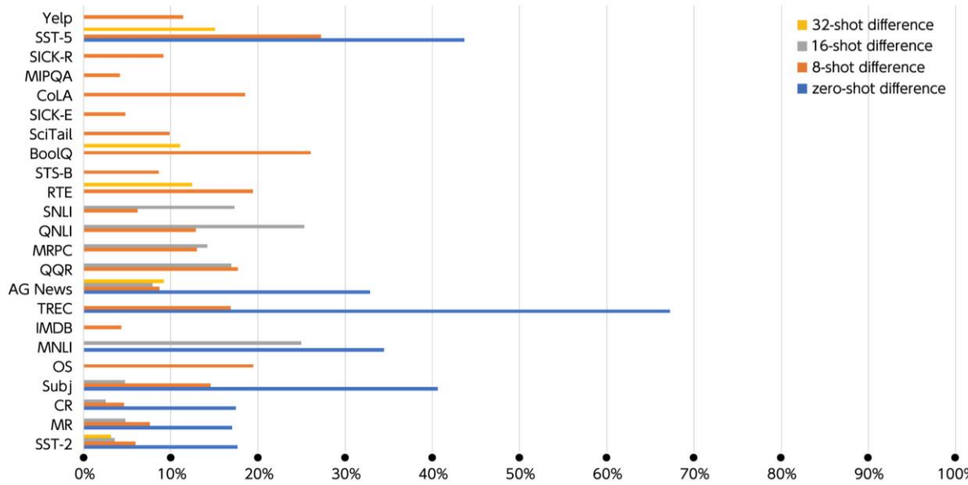


Figure 5: score drop rate using benchmarks data

Putting zero-shot results aside, we found sentiment analysis tasks such as “SST-2”, “MR”, “CR”, and topic classification tasks such as “AG News” achieved an overall score drop rate within 10% under multiple few-shot settings. We regarded these types of tasks as well-performed tasks. The abovementioned sentiment analysis tasks include reviews of movies, forum posts, blogs, and products which normally do not require much domain-specific knowledge to make correct predictions thus achieved relatively better performance even without any labeled training data.

On the other hand, even though tasks such as “Yelp” and “SST-5” are also sentiment analysis related, we found them suffering from higher drop rate, especially for “SST-5” which dropped at least over 10%. In fact, these two tasks adopt more fine-grained sentiment labels (negative, somewhat negative, neutral, somewhat positive, positive). Most of the works we investigated treated them as independent labels and simply tried to solve this problem as a multi-class classification problem without considering the quantitative relations within these labels. We assumed this fact mainly caused the difference in

few-shot performance between these two types of sentiment analysis tasks.

Compared to “AG News”, another topic classification task “OS” dropped nearly 20% when using 8 training samples. Since “OS” is mainly about detecting hate speech yet hate speech is still a difficult phenomenon to define even for human, we considered this performance drop to be reasonable.

Focusing on the tasks with drop rates over 10%, we considered tasks such as “QQP”, “MRPC”, “QNLI”, “RTE”, and “BoolQ” as relatively poorly performed. These tasks are closely related to NLP problems including recognizing textual entailment (RTE), natural language inference (NLI), question answering (QA), and paraphrasing. Comparing to sentiment analysis and topic classification where keyword-level or syntactic-level information are the major factor for correct prediction, these types of tasks highly rely on semantic-level information which is relatively more difficult to catch within an extreme small amount of training data. For example, “BoolQ” dataset mainly includes non-factoid questions which often require high-level comprehension of given passage to infer the correct answers. It is also worth noting that most of the well-performed tasks take a single sentence as input while poorly performed ones were mainly dealing with sentence pairs.

2.2.3 Experiments using domain-specific data

Most of the benchmark tasks deal with general-purpose data such as product reviews, news articles, Quora questions, etc. Many of the data were generated by crowd-sourced annotators that basically anyone with common sense could be qualified to become. Especially general purpose one sentence classification tasks achieved relatively promising few-shot results. The domain of data used to pre-train the large language models is consistent with the downstream benchmark at a certain level which makes it possible to dig out desired information with only a few training examples.

Real-world use cases normally deal with data related to more specialized areas such as finance, medicine, or other industry categories. Therefore, it is also crucial to test the capability of the latest few-shot approaches for solving tasks that require more target domain-related literacy. We utilized an open dataset related to patent document classification for experiments. In this dataset, each patent document is labeled using one of the 9 predefined classes. Considering the effect caused by the number of classes, besides the original dataset we also composed another version using only three most common classes

("Physics", "Electricity", and "Human Necessities"). We named these two versions "patent-9" and "patent-3" accordingly. For experiments, we employed the above-mentioned LM-BFF-MS approach which achieved promising few-shot results on general-purpose datasets.

Similarly, we evaluated the accuracy under zero-shot, 8-shot, 16-shot, and 32-shot settings. Figure 6 shows the score drop rate when using these two datasets along with results of several similar benchmark datasets. Note that all these tasks are one-sentence classification problems in which each input document/sentence is assigned with a single label. While the benchmark datasets maintain a drop rate within 10%, both "patent-3" and "patent-9" drop over 10 points compared to full-dataset fine-tuning. "Patent-9" dataset has a drop rate over 20% when using only 8 and 16 training samples for each class. The results get worse when applying zero-shot on both datasets. Since patent classification dataset involves various types of patent terminologies that are quite different from the way people describe things in daily life, it is still challenging to acquire correct answers from a general-purpose pre-trained language model using only a limited amount of training samples.

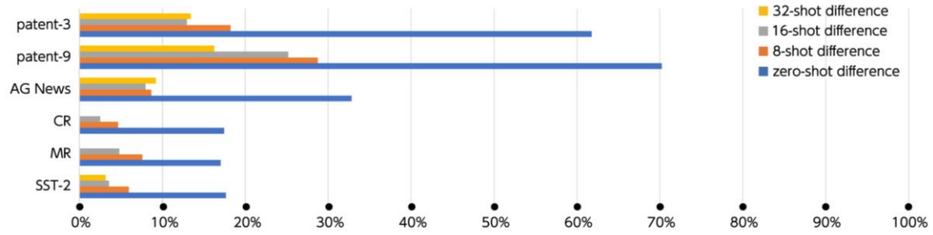


Figure 6: score drop rate using domain-specific open data

CHAPTER.3

Future perspective

3.1 Improvements of few-shot learning algorithms

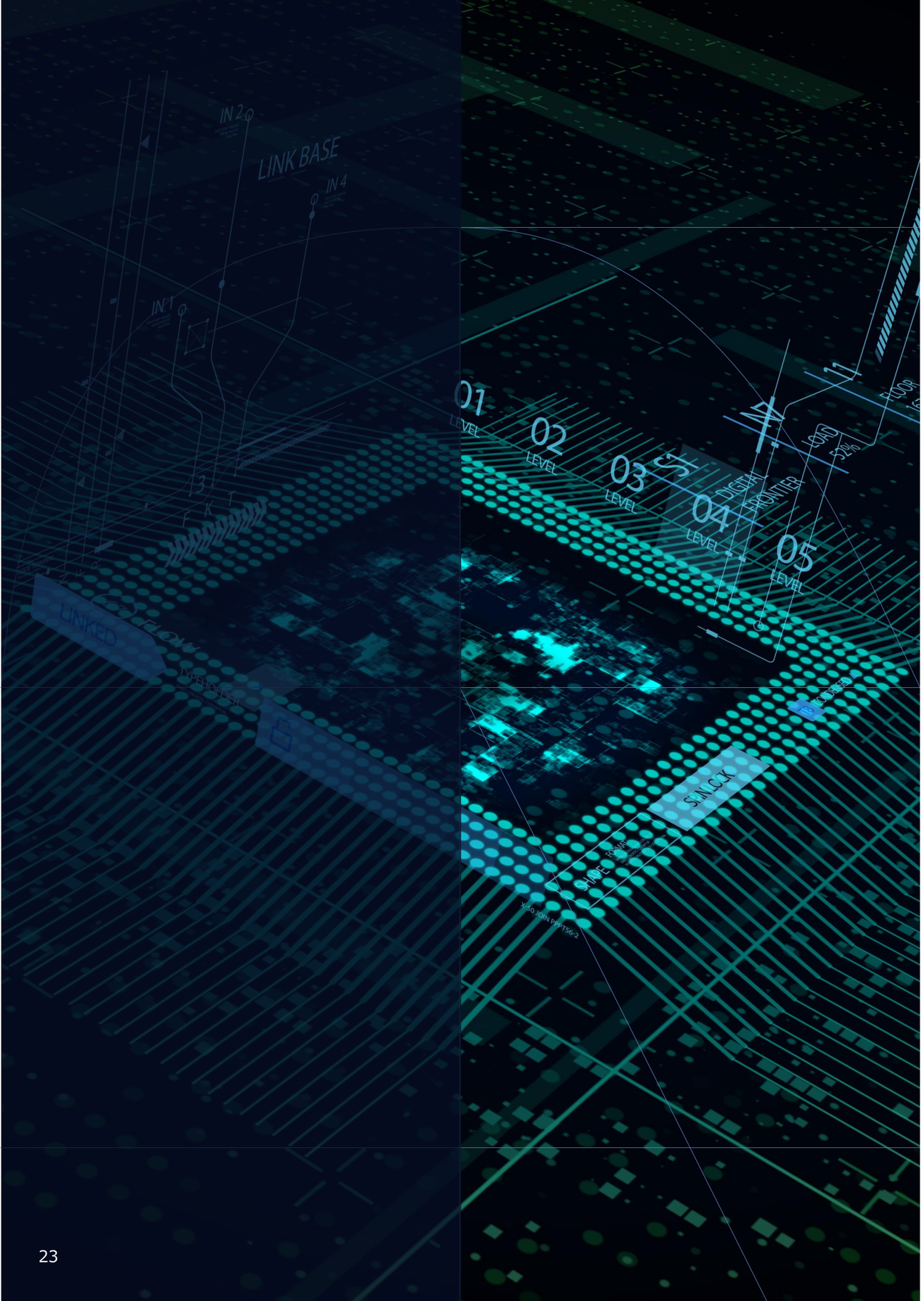
Transformer-based language models have changed the landscape of NLP in the past few years and enlarging the model size has been the primary way where improvements have been carried out, especially for few-shot learning. Nonetheless, research is still going on to introduce new advancements to refine the current frameworks. Recently, Meta proposed a technique that used a sparse all-MLP (Multi-Layer Perceptron) and achieved comparable results to Transformer architectures while surpassing them on several zero-shot downstream tasks. This fact indicates a possible path for few-shot technique improvements by refining the traditional architectures such as MLP.

Other improvements might come in the form of prompt engineering. The ability to provide the right prompts have been shown to substantially improve zero and few-shot accuracy. Even though current research introduced auto-generated prompts for general-purpose tasks, manual effort is still required in a certain level for domain-specific tasks. It is worth expecting new technical breakthroughs that make prompt engineering more automatic and flexible to different domain-specific tasks.

Within current architectures, developing novel learning strategies may also benefit few-shot learning performance. In particular, meta learning algorithms exploit other related tasks to optimize model selection and algorithm hyper-parameter tuning for the target task. Evolution of this type of learning strategy could potentially make it possible to optimize the NN architecture itself to better adapt to undergoing tasks.

3.2 The bigger the better?

Recently Google has announced PaLM, a 540B parameter model, which further advanced the state-of-the-art in many benchmarks. In particular, Google focused on few-shot settings, and PaLM even outperformed the average human performances in many tasks. Without doubt, increasing model size has been proven to be the fundamental to the improvements in language understanding and few-shot capabilities. It will not be surprising that ultra-large language models composed by further increased parameters be published for greater few-shot learning improvement, in the not-so-distant future. To meet the requirements for different use cases in real-world applications, training domain-specific large-scale language models using documents collected from different fields is also an essential task. Besides enlarging the size of general-purpose language models, composing and providing large language models for specific areas such as finance, manufacturing industry, medicine, etc., could also be the future



trends in the NLP community.

On the other hand, further considerations should also be given to environmental protection and energy efficiency. It has been said that "Training a single AI model can emit as much carbon as five cars in their lifetimes". It is severe enough to reconsider if it is the best solution to simply expand the language model size for better downstream task few-shot performance. Regarding this point, DeepMind showed that a model with 70B parameters (Chinchilla), which trained on more training data, outperformed GPT-3 and Microsoft along with NVIDIA's 540B Megatron-Tuning natural language generation model (MT-NLG). Moreover, Google's dialog model Lambda which is slightly smaller than GPT-3 achieved better results in one of the OpenAI's aims -- safety and truthfulness, by using more and better annotated data. Therefore, instead of simply aiming at building high-performance large language models, environment-friendly will be a prominent factor in future AI developing.

3.3 Using information other than language

Human perception of the world is always based on several different modalities. To build a generalist AI who can see, hear, speak, understand, reason and so forth, the research progress in multi-modal learning has grown rapidly especially in the areas of CV and NLP. Inspired by the success of pre-training large language models in NLP, multimodal models pre-training which share the same concept began to be applied in multi-modal tasks among which vision-language models are most representative. Thanks to the huge amount of multi-modal data existing in webpages, EC sites, SNS, or YouTube, where images and text data are essentially aligned, it does not require massive manual labor to collect training data to pre-train large multi-modal models for tasks such as Visual Question Answering (VQA) and Image Captioning. However, reducing data noise and data bias from such training data remains a challenging task.

From NLP perspective, assigning extra visual information to words, phrases, or sentences can undoubtedly enrich the feature representation of text information, hence making it easier to disambiguate abstract linguistic concepts. To give an intuitive analogy, picture teaching human children a new concept. Instead of barely giving stacks of verbal descriptions, showing a single image is more efficient to help understanding the concept. Similarly, learning to classify human sentiment utilizing additional audio information such as intonation will possibly much faster than using text data alone. This coincident with the concept of few-shot learning and may better explain the well-known symbol grounding problem for Strong AI. With the multi-modal nature of human perception, human beings normally only require a handful of examples to learn a new concept. It is not hard to imagine that more modalities including text, images, animations, phonemes, narrations, etc., would be better utilized in future evolution of multi-modal models to achieve more human-like AI models which are able to learn from a very limited amount of training data.