

**University of Science**

**Vietnam National University,**

**Ho Chi Minh City**

**Student information:**

21127175 – Le Anh Thu

21127191 – Nguyen Nhat Truyen

|  |  |
| --- | --- |
|  | On Transferability of Prompt Tuning for Natural Language Processing |
|  |  |
|  | Class: 21CNTT  Theoretical Instructor: Dr. Nguyen Hong Buu Long  Practical Instructor: Dr. Le Thanh Tung, Dr. Luong An Vinh |
|  |  |

21127592 – Nguyen Minh Dat

**Table of Contents**

**[1.](#_Toc155560733)****[Introduction](#_Toc155560733)** [1](#_Toc155560733)

**[2.](#_Toc155560734)****[Related Work](#_Toc155560734)** [2](#_Toc155560734)

**[3.](#_Toc155560735)****[Proposed method](#_Toc155560735)** [3](#_Toc155560735)

[3.1 Objective 3](#_Toc155560736)

[3.2 Experiment 3](#_Toc155560737)

[3.2.1 Cross-Task Transfer 3](#_Toc155560738)

[3.2.2 Cross-Model Transfer 5](#_Toc155560739)

[3.2.3 Transferability Indicator 6](#_Toc155560740)

**[4.](#_Toc155560741)****[Describe the corpus](#_Toc155560741)** [7](#_Toc155560741)

**[5.](#_Toc155560742)****[Results](#_Toc155560742)** [11](#_Toc155560742)

[5.1 Cross-task transfer 11](#_Toc155560743)

[5.2 Cross-model transfer 12](#_Toc155560744)

[5.3 Prompt Similarity Metrics 13](#_Toc155560745)

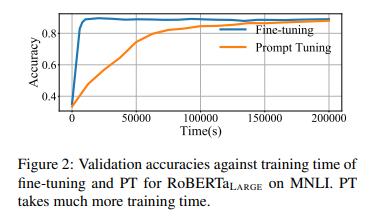
**[6.](#_Toc155560746)****[Conclusion](#_Toc155560746)** [15](#_Toc155560746)

**[References](#_Toc155560747)** [16](#_Toc155560747)

2. **Introduction**

The current landscape of pre-trained language models (PLMs) boasts impressive capabilities but is hindered by the computational demands and time-intensive nature of Fine Tuning due to their vast number of parameters. Addressing this challenge, a novel approach known as **Prompt Tuning** has emerged, focusing on adjusting a minimal number of soft prompts within very large PLMs to achieve performance comparable to full parameter Fine Tuning.

Despite its promising potential, Prompt Tuning encounters a drawback in its *extended learning time* (*Figure 1*). To enhance its efficiency, this study explores the feasibility of leveraging **knowledge transfer** as a means of expediting Prompt Tuning. The primary aim is to investigate whether prompt transfer can be a key factor in optimizing the efficiency of Prompt Tuning.

**

*Figure 1. Validation accuracies against training time of*

*Fine tuning and Prompt tuning.*

The experimental investigation delves into the *transferability* of learned soft prompts across *different downstream tasks* and *PLMs*.

The findings reveal noteworthy insights:

* In a zero-shot setting, soft prompts acquired from one task not only transfer effectively to similar tasks within the same PLM but also demonstrate effectiveness when transferred to other PLMs as cross-model projectors for analogous tasks.
* The study demonstrates that utilizing learned soft prompts from similar tasks and transferred prompts from different PLMs as initializations significantly accelerates learning and enhances the overall performance of Prompt Tuning.

Through these experiments, they aim to contribute valuable insights into the potential of **Prompt Transfer** as a tool for improving the efficiency of Prompt Tuning, thereby advancing the landscape of large-scale language model applications.

1. **Related Work**

**Prompt Tuning**, demonstrated by *Brown et al.*, *2020* on GPT-3, excels in few-shot NLP tasks using textual prompts. Various efforts, including *hard prompts* and *soft prompts* have aimed to enhance NLP tasks.

*Lester et al., 2021* show Prompt Tuning's competitive performance with full-parameter Fine Tuning in large PLMs, but with longer training times. This study focuses on mitigating this challenge through prompt transfer for knowledge transfer, exploring prompt transferability across tasks and PLMs.

In broader **Knowledge Transfer** strategies for NLP, existing approaches often involve intermediate task Fine Tuning. *Lester et al., 2021* highlight Prompt Tuning's superior cross-domain transferability compared to Fine Tuning.

Contrastingly, *Vu et al., 2021* enhance Prompt Tuning effectiveness through cross-task transfer with prompt initialization and similarity metrics. Their work emphasizes improving efficiency, examining transferability indicators like the overlapping rate of activated neurons and exploring cross-model transfer inspired by methodologies such as Net2Net (*Chen et al., 2016*), knowledge distillation (*Hinton et al., 2015*), and knowledge inheritance (*Qin et al., 2021*).

1. **Proposed method**

**3.1 Objective**

* Analyze the transferability of prompts across different tasks and Pre-trained Language Models (PLMs).
* Investigate the reasons behind the transferability of soft prompts across tasks and identify the factors that determine the extent of transferability between them.

**3.2 Experiment**

**3.2.1 Cross-Task Transfer**

**Dataset: 17** NLP tasks in **6** categories

**- Sentiment Analysis**: 6 datasets, including:

+ IMDB

+ SST-2

+ laptop

+ restaurant

+ Movie Rationales

+ TweetEval.

- **NLI**: 3 datasets, including

+ MNLI

+ QNLI

+ SNLI.

- **Ethical Judgment**: 2 datasets, including

+ deontology

+ justice.

- **Paraphrase Identification**: 2 datasets, including

+ QQP

+ MRPC.

- **QA**: 2 datasets, including

+ SQuAD

+ NQ-Open.

- **Summarization**: 2 datasets, including

+ Multi-News

+ SAMSum.

**Implementation**: Zero shot setting

- Conduct Prompt Tuning on a source task

- Reuse the trained prompts on other target tasks

- Assess the model's performance on the target tasks

**Evaluation**:

- For the tasks within the same type, transferring soft prompts between them can generally perform well and may even outperform original Prompt Tuning on the target task, especially when the source task has more data.

- For the tasks of different types, the transferability of soft prompts among them is generally poor, and transferring soft prompts often achieve similar performance to randomly initialized prompts.

### **3.2.2 Cross-Model Transfer**

**Models**: RoBERTa, T5

- **RoBERTa-large**:

+ A transformer-based pre-trained language model (PLM) for masked language modeling.

+ Limitation: predict only a single token or a fixed length of tokens.

- **T5-XXL**:

+ A transformer-based pre-trained language model (PLM) for sequence-to-sequence pre-training.

+ Advantage: suitable for tasks that involve generating sequences of text, such as QA and summarization.

**Prompt Projector**:

- **Why do we need a Prompt Projector?** To map the learned soft prompts of a PLM onto the semantic space of another PLM.

- Two possible **learning objectives** to train a Prompt Projector:

+ **Distance Minimizing**: Minimize the distance between Project(PS) and Pt, where PS is the prompt trained on source model and Pt is the prompt trained on target model.

+ **Task Tuning**: Directly tune the projected prompts Project(PS) and back propagate the supervision signals to train the weights of the projector.

**Implementation**: Zero shot setting

- Train soft prompts on the source PLM.

- Project the trained soft prompts to the embedding space of the target PLM using the Prompt Projector.

- Use the projected soft prompts on the target models.

**Evaluation**:

- Distance Minimizing is effective in transferring prompts of the projector-training tasks, but yields random performance on other unseen tasks, making it impractical for practical use.

- Task Tuning demonstrates better performance and effectively generalizes to unseen tasks of the same type as the projector-training tasks, proving the feasibility of practical cross-model prompt transfer.

- The projectors trained with Task Tuning still cannot work for tasks of a different type.

**3.2.3 Transferability Indicator**

**Prompt Similarity Metrics**:

- **Embedding Similarity**: compute the similarity between trained soft prompts in the embedding space using Euclidean distance and cosine similarity. Given two groups of trained soft prompts, each having tokens, there are two ways to compute the similarity score:

+ **Concatenate**: concatenate all tokens to get the embeddings representation for each group, and then compute similarity between two concatenation embeddings.

+ **Average**: calculate the similarity for each pair of tokens and use the averaged results as the final similarity score.

- **Model Stimulation Similarity**: examine the similarities in the responses of Pre-trained Language Models (PLMs) to the two soft prompts using the overlapping rate of activated neurons.

1. **Describe the corpus**

17 NLP datasets divided into 6 categories

**Category 1: Sentiment Analysis**

**Task description**: Sentiment analysis is the task of determining the sentiment or emotional tone expressed in a piece of text. The goal is to classify the sentiment as positive, negative, or neutral. It is often used to analyze opinions, reviews, or social media content.

**Datasets**:

* **IMDB**: A dataset for sentiment analysis based on movie reviews from the Internet Movie Database, where the goal is to classify reviews as positive or negative.
* **SST-2**: Binary sentiment classification dataset derived from movie reviews with sentences annotated for sentiment.
* **laptop**: This dataset is designed for aspect-based sentiment analysis on laptop reviews, aiming to discern sentiments related to specific aspects like performance and design.
* **restaurant**: Tailored for aspect-based sentiment analysis on restaurant reviews, this dataset focuses on sentiments associated with various aspects such as food quality and service.
* **Movie Rationales**: Sentiment analysis dataset with associated rationales, providing explanations for model predictions.
* **TweetEval**: Diverse sentiment analysis tasks on Twitter data, covering various domains in short texts.

**Category 2: NLI (Natural Language Inference)**

**Task Description**: Natural Language Inference involves determining the logical relationship between two given text snippets. The three common categories are "entailment" (one statement logically follows from another), "contradiction" (the statements cannot both be true), and "neutral" (there is no clear logical relationship).

**Datasets**:

* **MNLI (Multi-Genre Natural Language Inference)**: Multi-genre natural language inference dataset where models predict the relationship (entailment, contradiction, or neutral) between sentence pairs.
* **QNLI (Question-Answering Natural Language Inference)**: Question-answering based NLI dataset that transforms sentences into questions to assess the understanding of entailment.
* **SNLI (Stanford Natural Language Inference)**: Dataset for predicting the logical relationship between pairs of sentences (entailment, contradiction, or neutral).

**Category 3: Ethical Judgment**

**Task Description**: Ethical judgment tasks involve assessing a model's ability to make ethical decisions or judgments. This can include tasks related to ethical principles, moral reasoning, or evaluating the ethical implications of certain actions or situations.

**Datasets**:

* **deontology**: This dataset focuses on ethical judgment tasks rooted in deontology principles, evaluating models' ability to make ethical decisions based on adherence to rules and moral principles.
* **justice**: Centered on ethical judgment tasks emphasizing principles of justice, this dataset evaluates models' capacity to make ethical decisions in scenarios involving considerations of fairness and equity.

**Category 4: Paraphrase Identification**

**Task Description**: Paraphrase identification focuses on determining whether two given pieces of text convey the same or similar meaning. It is a binary classification task where the model determines if the provided text pairs are paraphrases of each other.

**Datasets**:

* **QQP**: Binary classification dataset determining if pairs of questions are paraphrases.
* **MRPC**: Dataset for paraphrase identification, where models decide if two sentences are paraphrases of each other.

**Category 5: QA**

**Task Description**: Question Answering is the task of developing models that can understand a question posed in natural language and provide accurate and relevant answers.

**Datasets**:

* **SQuAD**: Reading comprehension dataset where models answer questions based on a given passage.
* **NQ-Open**: Open-domain question answering dataset, extending the question-answering task to a broader range of topics.

**Category 6: Summarization**

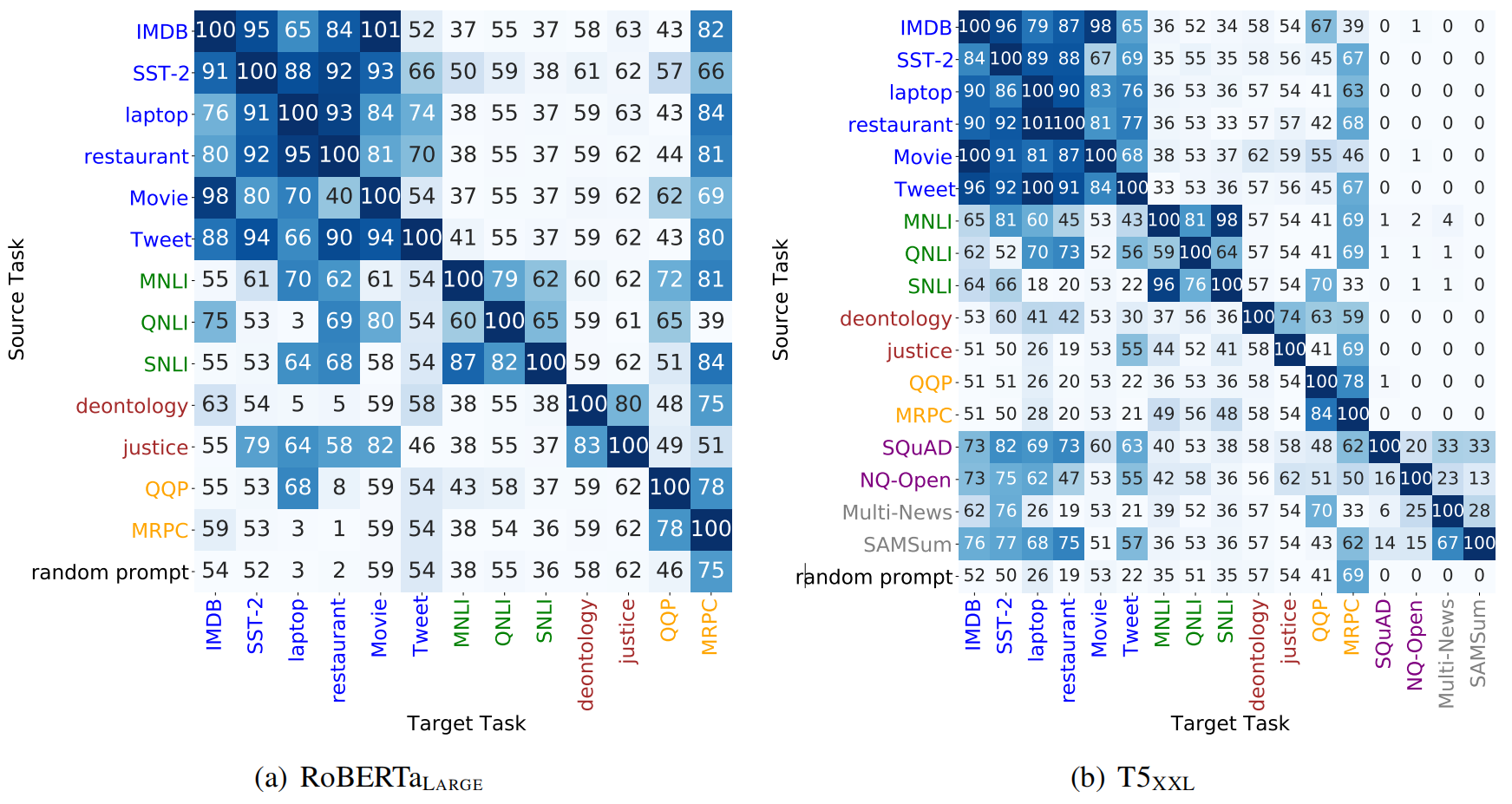
**Task Description**: Summarization involves generating concise and coherent summaries of longer pieces of text. There are two main types: extractive summarization, which involves selecting and rearranging existing sentences, and abstractive summarization, where the model generates new sentences to convey the main ideas of the original text.

**Datasets**:

* **Multi-News**: Summarization dataset focusing on news articles, requiring models to generate concise summaries.
* **SAMSum**: Abstractive summarization of dialogues, challenging models to summarize conversational content in a concise manner.

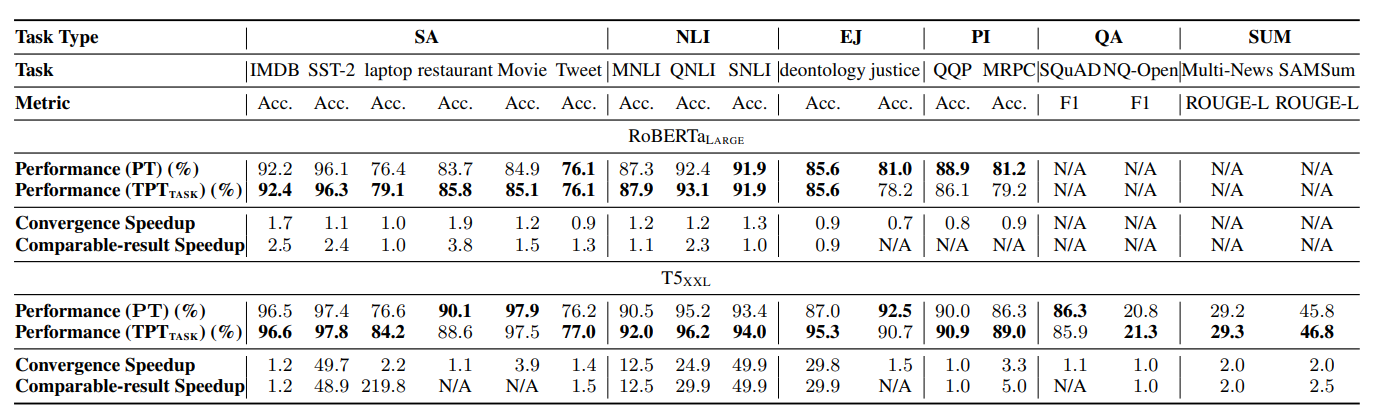
1. **Results**

**5.1 Cross-task transfer**



*Figure 2. Heat map presenting the ratio of zero-shot transfer performance over original Prompt Tuning performance, with columns representing target tasks, rows representing source tasks, and tasks within the same category share the same color.*

Further comparison is demonstrated in the table below, where the cross-task transferable prompt tuning (TPTTASK) method is compared with the original Prompt Tuning method on 2 criteria: performance and convergence speed. Note that TPTTASK initialize PT with the soft prompt performing best in zero-shot transfer.

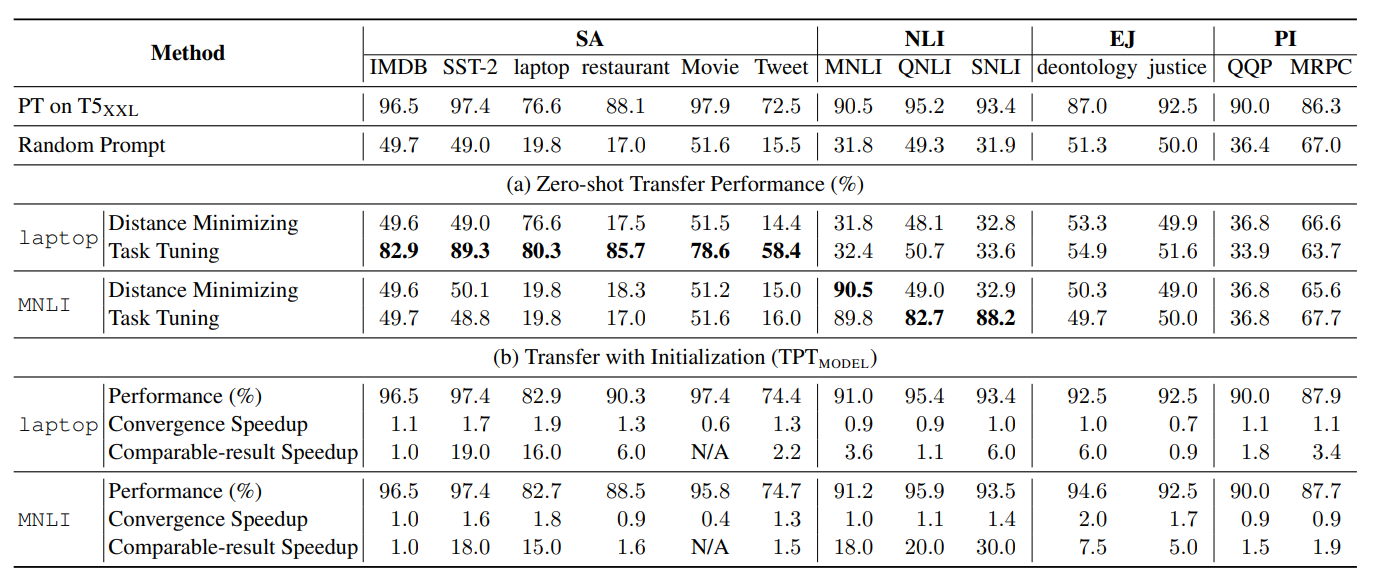


*Figure 3. Table presenting the comparison between TPTTASK and vanilla PT.*

**Observation:**

* TPTTASK can mostly achieve better or comparable performance to vanilla Prompt Tuning starting from random initialization.
* TPTTASK typically takes less training time to converge.

## **5.2 Cross-model transfer**

****

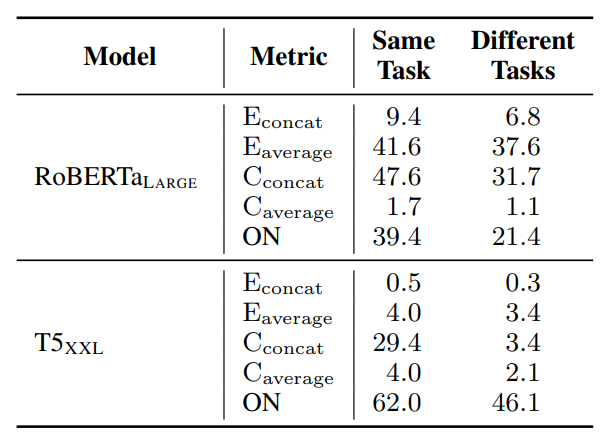
*Figure 4. Cross-model prompt transfer (RoBERTaLARGE to T5XXL) results.*

Above is the cross-model transferable prompt tuning TPTMODEL results, which uses the Task Tuning projectors to map soft prompt trained on source model to the embedding space of target model and then initialize prompt tuning with the projected soft prompt.

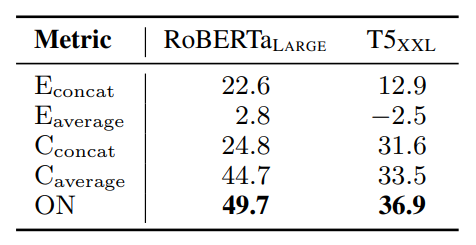
Observation: For tasks of the same kind as the projector-training task, TPTMODEL generally achieves similar or better performance compared to original Prompt Tuning, and it does so with significantly less training time.

→ The promising potential of practical cross-model prompt transfer in enhancing the efficiency and effectiveness of PT.

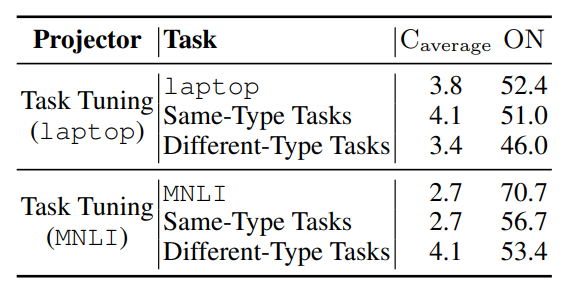
## **5.3 Prompt Similarity Metrics**

****

*Figure 5. Table presenting the average values (%) of the 5 similarity metrics for prompt pairs of the same task and different tasks, with “E” stands for Euclidean, “C” stands for cosine and “ON” stands for Overlapping Neurons.*

****

*Figure 6. The Spearman's rank correlation scores (%) between different similarity metrics and the zero-shot transfer performance of soft prompts across tasks, with “E” stands for Euclidean, “C” stands for cosine and “ON” stands for Overlapping Neurons.*



*Figure 7. The similarities (%) between the prompts projected*

*with Task Tuning projector and the original prompts trained on T5XXL.*

**Observation:**

* All metrics effectively differentiate between prompts within the same task as well as across different tasks.

→ Soft prompts from diverse tasks create distinguishable clusters in the embedding space and stimulate different abilities within the PLM.

* The overlapping rate of activated neurons (ON) metric outperforms all embedding similarity measures.

→ The effectiveness of prompt transferability is more dependent on model stimulation than on the distances in the embeddings.

* ON performs significantly less effectively on T5XXL (11 billion parameters) compared to RoBERTaLARGE (330 million parameters).

1. **Conclusion**

Key conclusions from the paper:

* **Cross-Task Transferability**: Soft prompts exhibit transferability to similar tasks without additional training in the cross-task setting.
* **Cross-Model Transferability**: Successful projection of prompts into the space of other Pre-trained Language Models (PLMs) in the cross-model setting.
* **Prompt Initialization for Accelerated Training**: Trained soft prompts from other tasks or PLMs can be used as initialization, resulting in significantly accelerated training and improved effectiveness.
* **Importance of Model Stimulation**: Explore prompt transferability indicators and highlight how the prompts stimulate PLMs is important in transferability.
* **Facilitating Further Research**: The paper aims to provide insights and ideas, through analyses and the concept of model stimulation, to facilitate further research on transferable and efficient Prompt Tuning.

# **References**

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. 2021. Spot: Better frozen model adaptation through soft prompt transfer. ArXiv preprint, abs/2110.07904.

Tianqi Chen, Ian J. Goodfellow, and Jonathon Shlens. 2016. Net2net: Accelerating learning via knowledge transfer. In 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. ArXiv preprint, abs/1503.02531.

Yujia Qin, Yankai Lin, Jing Yi, Jiajie Zhang, Xu Han, Zhengyan Zhang, Yusheng Su, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2021. Knowledge inheritance for pre-trained language models. ArXiv preprint, abs/2105.13880.

*- The end -*