Recent Work on Prompt-Based Fine-Tuning

Ercong Nie

Center for Information and Language Processing (CIS),
Ludwig Maximilians University of Munich (LMU)

April 2, 2024

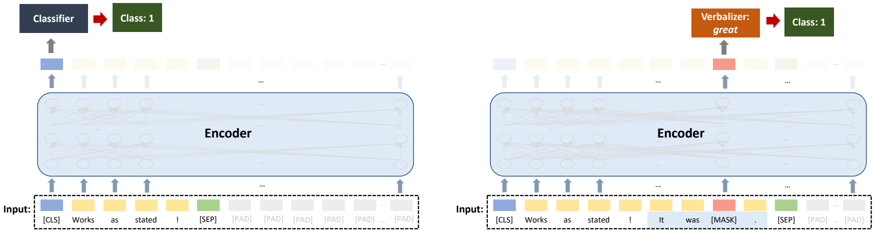
Outline

About me

Ercong Nie

- 2nd-year PhD student at CIS.
- **Master**: Computational Linguistics + Informatics at CIS, LMU.
- **Bachelor**: German + Finance at Shanghai Jiao Tong University, China.
- **Research interest**: multilingual NLP, low-resource NLP, etc.

Fine-Tuning: Prompt-based vs. Vanilla



(a) Vanilla finetuning

(b) Prompt-based finetuning

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Overview

- Recent work on prompt-based fine-tuning:
  - **Multilingual Adaptation** (Ma et al., 2023)
  - **Sequence Labeling**: Generalization to Complex Tasks (Ma et al., 2024)
  - **Parameter-Efficient Method**: Integration with Graph Neural Networks (GNNs) (Yuan et al., 2024)

Outline

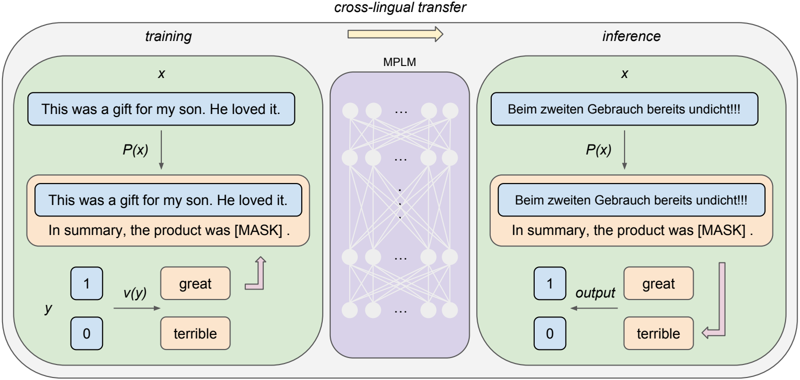
Multilingual Adaptation

We applied prompt-based fine-tuning to zero-shot cross-lingual transfer learning.

- **Prior work:** Zhao and Schütze (2021) implemented prompt-based fine-tuning in multilingual natural language inference tasks, (**XNLI**, Conneau et al., 2018).
- We (Ma et al., 2023) further conducted an extensive comparative analysis of the cross-lingual transfer capabilities of prompt-based fine-tuning compared to vanilla fine-tuning.

Prompt-Based Fine-Tuning: Multilinugal Setting

- **Training on English data**: prompt pattern, verbalizer, fine-tuning by mask token prediction.
- **Inference in the cross-lingual setting**:
  - input given in target languages
  - no changes in prompt pattern, verbalizer



Datasets and Models

- **Datasets**
  - Amazon Reviews Dataset:
    Multi-class sentiment analysis task in **6** languages (Keung et al., 2020)
  - PAWS-X:
    Binary paraphrase identification task in **7** languages (Yang et al., 2019)
  - XNLI:
    Multi-class natural language inference task in **15** languages (Conneau et al., 2018)
- **Multilingual Models**
  - Multilingual BERT model (M) (Devlin et al., 2019)
  - XLM-R model (X) (Conneau et al., 2020)

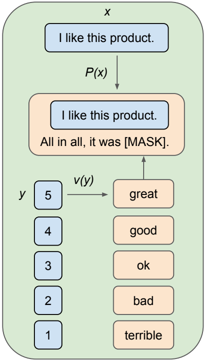


Figure 3: A prompt example for Amazon Dataset

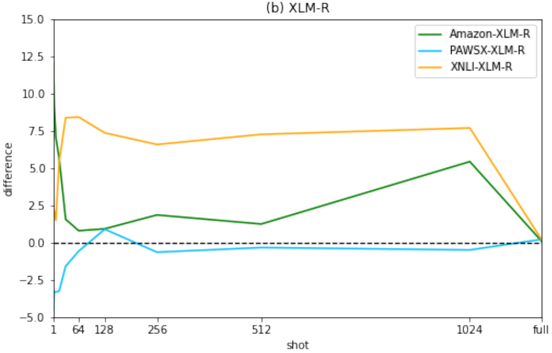
Main Findings

- Zero-shot cross-lingual results on **full** source language fine-tuning: Slight, but consistent improvement.

	Amazon	PAWS-X	XNLI	Avg.
MAJ	20	55.81	33.33	36.17
Direct-mBERT	20.21	45.05	35.05	33.44
Vanilla-mBERT	42.97	80.24	65.05	62.75
ProFiT-mBERT	**43.98**	**82.16**	**65.79**	**63.98**
Direct-XLM-R	21.98	51.10	35.68	36.25
Vanilla-XLM-R	54.56	82.51	73.61	70.22
ProFiT-XLM-R	**54.66**	**82.73**	**73.82**	**70.40**

Scaling Effect of Few-Shot Samples

- Zero-shot cross-lingual results on **few-shot** source language fine-tuning:
  Large improvements for Amazon/XNLI.



Multilingual Adaptation: Summary

- In zero-shot cross-lingual transfer:
  prompt-based fine-tuning $>$ vanilla fine-tuning
- Performance improvement is larger in **few-shot learning** scenarios.

Outline

Motivation

- Prompt design for **sentence classification** tasks is **not complex**, given that these tasks typically assign a single label to each sentence, requiring only **one prompt per task**.

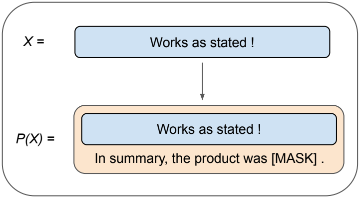
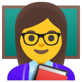


Figure: Prompting example for sequence classification.

Generalize prompt-based fine-tuning from **sentence-level** to **token-level**

ToPro: Token-Level prompt decomposition



Please give the pos tags of the sentence: "**Works as stated!**".

The pos tags of the sentence: "**Works as stated!**" are: ???

"Works", "as", "stated", "!"

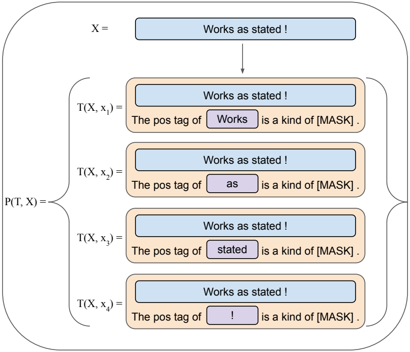
The pos tag of "**Works**" is "**VERB**".
The pos tag of "**as**" is "**CCONJ**".
The pos tag of "**stated**" is "**VERB**".
The pos tag of "**!**" is "**PUNCT**".

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

ToPro: Method

Token-Level Prompt Decomposition

1. Given an input sentence $X = x_1, x_2, \cdots, x_n$.
2. Decompose the sentence X into n tokens.
3. Apply the token level prompt function $T(X, x_i)$ n times such that each token x_i has a prompt.



The prompt pattern used in this example:

$T(X, x_i) =$ “ X The POS tag of x_i is a kind of [MASK] .”

Experiments

ToPro fine-tuning for zero-shot cross-lingual transfer

- **Tasks**
  - PAN-X for named entity recognition **(NER)** in 41 languages (Pan et al., 2017)
  - UDPOS for **POS tagging** in 38 languages (Nivre et al., 2020)
- **Models**
  - **Encoder-only Models:**
    - Multilingual BERT model (M) (Devlin et al., 2019)
    - XLM-R model (X) (Conneau et al., 2020)
  - **Encoder-decoder Model:**
    - Multilingual T5 model (T) (Xue et al., 2021)

Experiments

Baselines

- Vanilla Fine-Tuning (Vanilla):
  predicts the token labels through the **hidden states of each token** in the output layer without using a prompt pattern.
- Prompt Tuning (PT):
  only trains the parameters of **continuous prefix prompts** (Tu et al., 2022).

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Results

- ToPro Fine-Tuning **outperforms** Vanilla Fine-Tuning and Prompt-Tuning substantially across both tasks.
- ToPro with **mT5** model achieves **SOTA** performance.

Model	Method	PAN-X	UDPOS
mBERT	Vanilla Fine-Tuning	62.73	70.89
	Prompt-Tuning	56.76	69.91
	ToPro Fine-Tuning	**81.91**	**76.16**
XLM-R	Vanilla Fine-Tuning	61.30	72.42
	Prompt-Tuning	53.05	71.86
	ToPro Fine-Tuning	**80.03**	**76.16**
mT5	Vanilla Fine-Tuning	64.19	71.38
	Prompt-Tuning	-*	-*
	ToPro Fine-Tuning	**92.82**	**86.11**

Generalization to Complex Task Tasks: Summary

- ToPro extends prompt-based fine-tuning to sequence labeling tasks.
- ToPro outperforms two baselines on NER/POS on a zero-shot cross-lingual evaluation.

Outline

Background

- Prompt-based fine-tuning methods introduced so far are **effective**, especially in **low-data settings**.

Background

- Prompt-based fine-tuning methods introduced so far are **effective**, especially in **low-data settings**.
- These methods traditionally employ *Full-Parameter Fine-Tuning* (FPFT), which involves adjusting the **entirety** of a model's parameters.

Background

- Prompt-based fine-tuning methods introduced so far are **effective**, especially in **low-data settings**.
- These methods traditionally employ *Full-Parameter Fine-Tuning* (FPFT), which involves adjusting the **entirety** of a model's parameters.
- However,
  - **Large Language Models (LLMs)** have billions of parameters.
  - Updating all these parameters poses a practical **challenge**.

Background

- Prompt-based fine-tuning methods introduced so far are **effective**, especially in **low-data settings**.
- These methods traditionally employ *Full-Parameter Fine-Tuning* (FPFT), which involves adjusting the **entirety** of a model's parameters.
- However,
  - **Large Language Models (LLMs)** have billions of parameters.
  - Updating all these parameters poses a practical **challenge**.

$\Rightarrow$ *Parameter-Efficient Fine-Tuning* (PEFT):
optimizes a relatively **small subset** of an LLM's parameters

Question

How to develop an effective
prompt-based
parameter-efficient fine-tuning (PEFT) method?

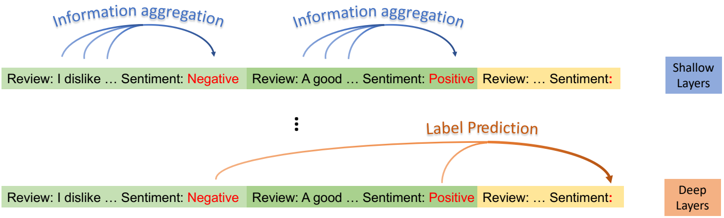
Question

How to develop an effective
prompt-based
parameter-efficient fine-tuning (PEFT) method?

$\Rightarrow$ **GNN For NLP**: Navigating Information Flow

Motivation

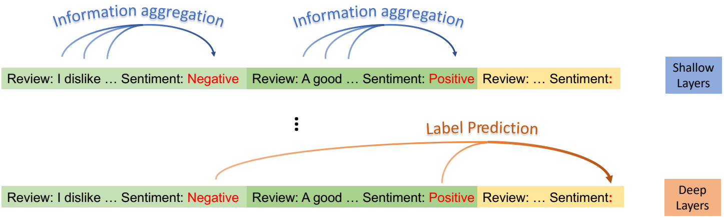
Label words are anchors: Understanding the mechanism of In-Context Learning (ICL) from an **information flow** perspective (Wang et al., 2023).



Wang et al. (2023)

Motivation

Label words are anchors: Understanding the mechanism of In-Context Learning (ICL) from an **information flow** perspective (Wang et al., 2023).



Wang et al. (2023)

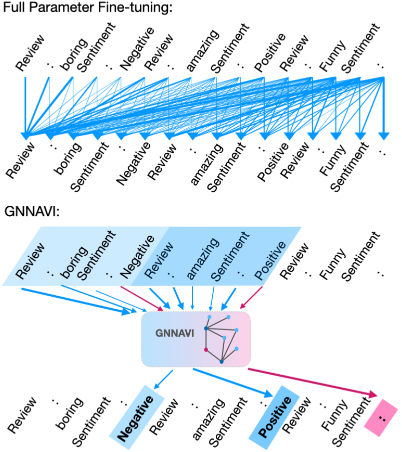
Two roles of label words as anchors:

- Information aggregation:
  aggregating information from preceding words.
- Information distribution:
  propagating information to last token for label prediction.

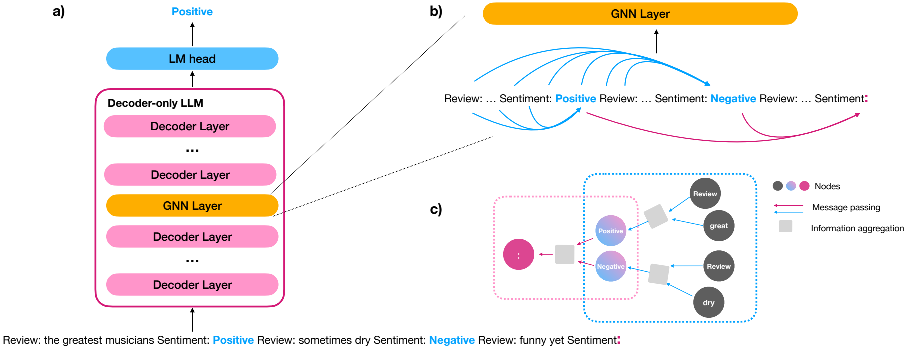
Idea: GNNavi

GNNavi: Navigating the information flow in prompt-based fine-tuning

- Inspired by the **information flow** perspective of ICL, we proposed a novel prompt-based PEFT method **GNNavi**.
- **GNNavi** is able to:
  - **navigate** the information flow
  - **save** the training resources
  - **outperform** FPFT and other PEFT methods (LoRA, Adapter, Prefix-tuning) in few-shot settings.

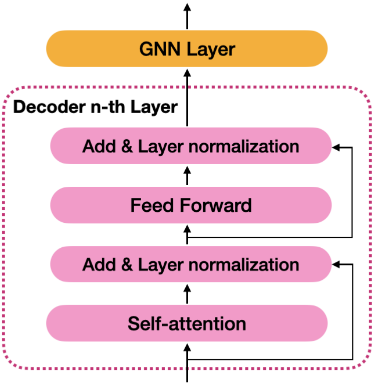


Pipeline: GNNavi



a) A GNN layer is inserted into LLM, taking a sentiment analysis task as example. (**Note:** Only **parameters in the GNN layer** are updated in fine-tuning.)
b) The input text is transformed into a **graph**, with tokens as nodes and information flow paths as edges.
c) Visualization of the working mechanism of the **GNN**.

GNN Structure in GNNavi



We adopted **two types** of GNN architecture in the GNN Layer.

- **GNNavi-GCN**:
  Graph Convolutional Network (GCN) (Kipf and Welling, 2017)

$$h_v^{(l)} = \sigma\left(W \sum_{v' \in N(v)} \frac{h_{v'}^{(l)}}{|N(v)|}\right)$$

- **GNNavi-SAGE**:
  GraphSAGE (Hamilton et al., 2017) generates node embeddings for previously unseen data using node feature information.

$$h_v^{(l)} = \sigma\left(W\left(h_v^{(l)} \oplus \mathrm{AGG}(\{h_{v'}^{(l)}, \forall v' \in N(v)\})\right)\right)$$

$h_v^{(l)}$ denotes the updated node representation of v, $h_{v'}^{(l)}$ denotes the token representation of its neighbouring nodes from l-th decoder layer, σ is the activation function, W is the trainable parameter of GNN, $N(v)$ includes all the neighbouring nodes of v.

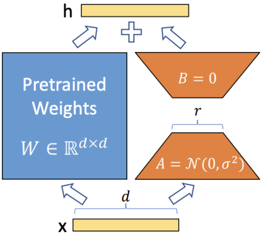
Experimental Setup

GNNavi for sentence classification with few-shot fine-tuning

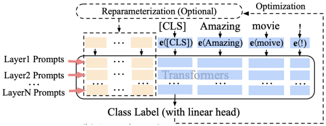
- **Tasks**
  - **SST-2**: Stanford Sentiment Treebank Binary for sentiment analysis (Socher et al., 2013)
  - **EmoC**: EmoContext for 4-label emotion classification (Chatterjee et al., 2019)
  - **TREC**: Text REtrieval Conference Question Classification for question type classification containing 6 types (Li and Roth, 2002)
  - **Amazon**: Binary classification for Amazon reviews (McAuley and Leskovec, 2013)
  - **AGNews**: AG's news topic classification dataset with 4 labels (Zhang et al., 2015)
- **Models**
  - **GPT2-XL (1.6B)** (Radford et al., 2019)
  - **LLaMA2 (7B)** (Touvron et al., 2023)

Baselines

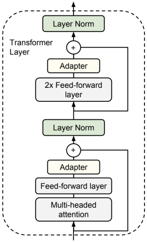
- **ICL:** In-context learning with one- or few-shot demonstrations per class.
- **FPFT:** Full-Parameter Fine-Tuning.
- **PEFT** (Paramter-Efficient Fine-Tuning):



(a) LoRA (Hu et al., 2022)



(b) Prefix Tuning (Li and Liang, 2021)



(c) Adapter (Houlsby et al., 2019)

a) **LoRA:** Low-Rank Adaptation, reducing training parameters by injecting trainable rank decomposition matrices into each layer (Hu et al., 2022).
b) **Prefix Tuning:** incorporating virtual tokens into the LLM and updating only the parameters of the virtual tokens (Li and Liang, 2021).
c) **Adapter:** inserting an adapter module to each layer (Houlsby et al., 2019).

Results: GNNavi

Overall Performance

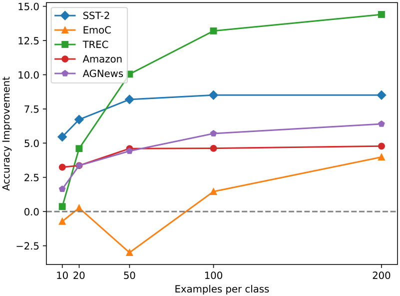
- GNNavi **outperforms** all the baselines on **average**.
- The performance **improves** as training examples increase.

Method	#Param	SST-2	EmoC	TREC	Amazon	AGNews	Average	#Param	SST-2	EmoC	TREC	Amazon	AGNews	Average
				GPT2-XL							**Llama2**			
							$k = 0$							
ICL	-	55.44	6.48	54.68	53.32	72.12	48.41	-	67.55	9.60	70.36	94.98	84.14	65.33
							$k = 5$							
ICL	-	63.17	6.30	57.68	53.67	50.43	46.25	-	86.93	20.18	45.72	92.30	80.16	65.06
LoRA	2.5M	91.98	50.60	75.20	88.80	**85.20**	78.36	4.2M	**95.42**	64.20	**88.40**	91.80	86.60	85.28
Prefix	6.1M	59.13	73.46	32.92	60.00	75.40	60.18	39.3M	50.96	58.56	21.36	49.36	25.78	41.20
Adapter	15.4M	79.82	76.00	**79.60**	**91.45**	81.25	81.62	198M	50.92	**84.05**	18.80	49.45	24.80	45.60
FPFT	1.6B	62.13	61.30	65.28	73.00	80.82	68.51	6.7B	94.63	61.92	81.72	**95.86**	**87.58**	84.34
GNNavi-CGN	2.6M	**84.31**	75.48	76.72	90.90	83.16	**82.11**	16.8M	94.56	78.30	83.2	94.00	86.25	86.63
GNNavi-SAGE	5.1M	81.95	**78.70**	77.92	88.66	82.88	82.02	33.6M	92.91	80.12	80.80	95.66	86.06	**87.11**
							$k = 200$							
LoRA	2.5M	**90.83**	80.80	90.80	82.00	86.20	86.13	4.2M	91.29	**86.80**	93.60	95.80	90.40	91.32
Prefix	6.1M	50.92	80.18	69.80	59.80	79.08	67.96	39.3M	48.35	81.72	45.68	52.28	27.54	51.11
Adapter	15.4M	88.65	80.70	**96.60**	92.30	**89.80**	89.61	198M	50.92	85.05	88.20	49.45	81.50	67.57
FPFT	1.6B	68.97	73.70	80.16	74.82	85.34	76.60	6.7B	**95.64**	79.90	**96.76**	96.12	**91.44**	91.97
GNNavi-GCN	2.6M	90.67	78.82	91.88	92.94	89.20	88.70	16.8M	95.36	82.85	95.50	**96.45**	91.05	**92.24**
GNNavi-SAGE	5.1M	90.46	**82.68**	92.32	**93.44**	89.28	**89.64**	33.6M	95.30	81.94	94.76	95.96	90.68	91.73

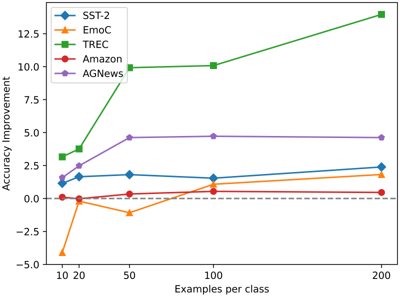
Results: GNNavi

Influence of Training Sample Size

- The improvement is particularly pronounced in **low-data settings**.



GPT2-XL



LLaMA2

Improvement gained by adding training examples for GNNavi-SAGE, compared to performance of **5**-shot fine-tuning.

Results: GNNavi

Efficiency analysis

Method	GPT2-XL	LLaMA2
LoRA	2.5M	4.2M
Predix	6.1M	39.3M
Adapter	15.4M	198M
FPFT	1.6B	6.7B
GNNavi-GCN	2.6M	16.8M
GNNavi-SAGE	5.1M	33.6M

Size of training parameters.

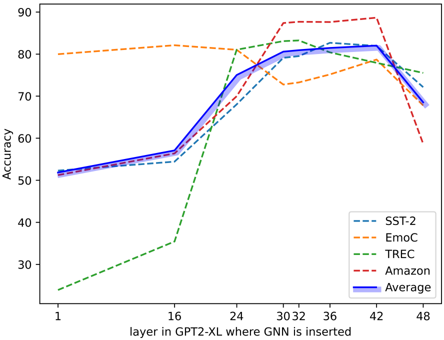
	SST-2	EmoC	TREC	Amazon	Agnews
GPT2-XL	4.7×	6.3×	4.1×	3.9×	3.4×
Llama2	4.3×	2.4×	1.6×	1.4×	1.2×

Training acceleration of GNNavi-GCN compared to **FPFT**.

- GNNavi **reduces** the number of training parameters.
- GNNavi **speeds up** the training process by a factor of up to 6 compared to FPFT.

Ablation Study: GNNavi

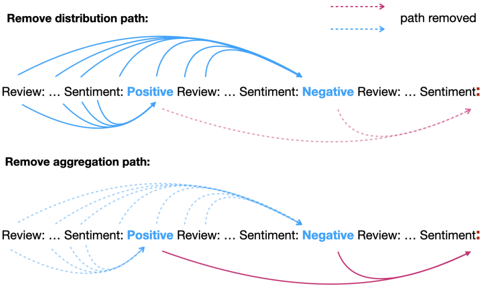
Position of GNN Layer



- The **insertion position** of the GNN layer greatly affects the model's performance.
- Adding the GNN layer within the **first 10 layers** results in lower performance, except for EmoC.
- Performance peaks at around the **44th layer**, then declines.

Ablation Study: GNNavi

Removal of Information Flow



	SST-2	EmoC	TREC	Amazon	Agnews	Average
	81.95	78.70	77.92	88.66	82.88	82.02
-aggregation	-0.07	-1.10	-0.68	+0.56	-0.08	-0.27
-distribution	+3.07	-12.88	-2.44	+1.64	-1.44	-2.41

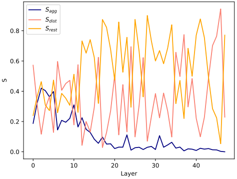
- Both aggregation and distribution paths impact performance.
- Except for SST-2 and Amazon binary tasks, removing the **distribution** path leads to a **larger** performance drop.
- These findings suggest the **distribution path** is more crucial for information flow, particularly in multi-label tasks.

Further Discussion on Information Flow

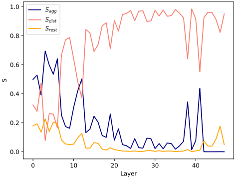
Saliency score:

$$I_l = \sum_h \left| A_{h,l}^\top \frac{\partial L(x)}{\partial A_{h,l}} \right|$$

$$S = \frac{\sum_{(i,j)\in C} I_l(i,j)}{|C|}$$



FPFT



GNNavi

Comparison of information flow between FPFT and GNNavi for SST-2. Both models are trained with 5 training examples per class.

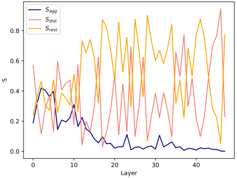
- In FPFT, token interactions with all previous words can cause **information flow confusion** without guided navigation.
- GNNavi follows a GNN-guided information flow, producing stable curves that represent **consistent information aggregation**.

Further Discussion on Information Flow

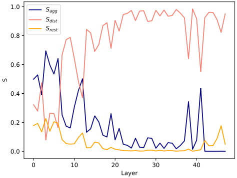
Saliency score:

$$I_l = \sum_h \left| A_{h,l}^\top \frac{\partial L(x)}{\partial A_{h,l}} \right|$$

$$S = \frac{\sum_{(i,j)\in C} I_l(i,j)}{|C|}$$



FPFT



GNNavi

Comparison of information flow between FPFT and GNNavi for SST-2.
Both models are trained with 5 training examples per class.

- In FPFT, token interactions with all previous words can cause **information flow confusion** without guided navigation.
- GNNavi follows a GNN-guided information flow, producing stable curves that represent **consistent information aggregation**.

$\Rightarrow$ GNNavi's role as a **navigator**, directing information flow in specific directions.

Integration with GNNs - Summary

- Inspired by the "Labels are anchors" theory of in-context learning, we propose **GNNavi**, a novel prompt-based parameter-efficient fine-tuning method that incorporates a GNN layer.
- In language understanding tasks, GNNavi demonstrates superior **efficacy** and **efficiency** over FPFT and other PEFT methods.

Thanks for your attention!

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Bolei Ma, Ercong Nie, Helmut Schmid, and Hinrich Schuetze. 2023. Is prompt-based finetuning always better than vanilla finetuning? insights from cross-lingual language understanding. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 1–16, Ingolstadt, Germany. Association for Computational Lingustics.

Bolei Ma, Ercong Nie, Shuzhou Yuan, Helmut Schmid, Färber Michael, Frauke Kreuter, and Hinrich Schütze. 2024. Topro: Token-level prompt decomposition for cross-lingual sequence labeling tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, Malta, Malta. Association for Computational Linguistics.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lifu Tu, Caiming Xiong, and Yingbo Zhou. 2022. Prompt-tuning can be much better than fine-tuning on cross-lingual understanding with multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5478–5485, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Shuzhou Yuan, Ercong Nie, Bolei Ma, Michael Färber, Helmut Schmid, and Hinrich Schütze. 2024. Gnnavi: Navigating the information flow in large language models by graph neural network. *preprint*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.