Cross-lingual Knowledge Transfer in Multi-lingual Language Models

Soumen Kumar Mondal
23m2157@iitb.ac.in

Guide: Prof. Preethi Jyothi

Indian Institute of Technology Bombay

May 8, 2024

Introduction

Problem Statement

- **Objective:**
  - To explore and demonstrate the effectiveness of fine tuning methods in enhancing the adaptability and performance of multilingual language models across various languages and tasks.
  - To develop methodologies that can improve the efficacy of factual knowledge transfer in multilingual setting.
- **Importance:** Multilingual models are essential for information exchange. Enhancing their efficiency and transferability without extensive retraining is important for practical usage.

Presentation Outline

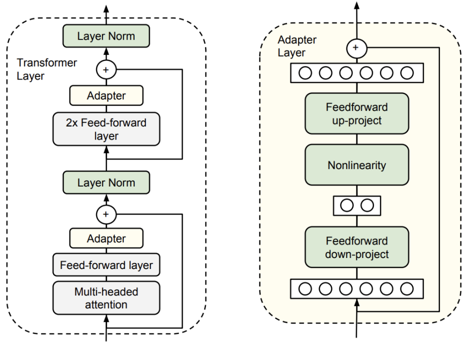
- **Language Model Fine Tuning:** Adapters, MAD-X, Composable SFT
- **Fact Representation:** Task Vectors, Cross Lingual Fact Representation
- **Geometry of Language Models:** Affine Language Subspaces

Language Model Fine Tuning: Adapters

What is Adapter?

- Adapters are small neural networks inserted between the layers of a pre-trained model.
- Each adapter only learns task-specific or language-specific features, leaving the original model weights untouched.
- This method is particularly advantageous for multilingual models because it enables the customization of a single foundational model for a variety of languages and tasks without the need for extensive retraining.

Adapter Based Fine Tuning[1]



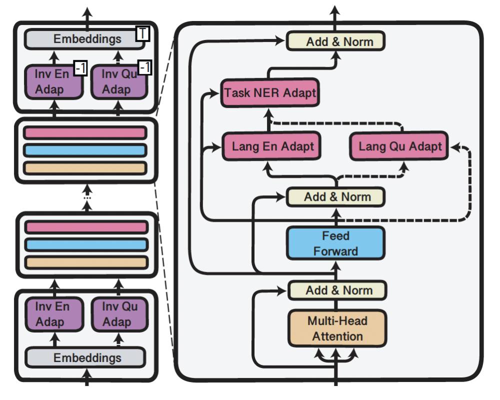
[1] Neil Houlsby et al. "Parameter-efficient transfer learning for NLP". In: *Proceedings of the 36th International Conference on Machine Learning*. 2019, pp. 2790–2799.

Language Model Fine Tuning: MAD-X

Overview of MAD-X

- Language adapters are designed to adapt the model to the specificities of a given language. These adapters are trained using MLM on unlabelled data from the target language.
- Task adapters are used to fine-tune the model for a specific task. They are applied after the language adapters
- The success of transfer learning heavily relies on the quality and comprehensiveness of the source language models.

MAD-X Architecture[2]



[2] Jonas Pfeiffer et al. "MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

Language Model Fine Tuning: Composable SFT

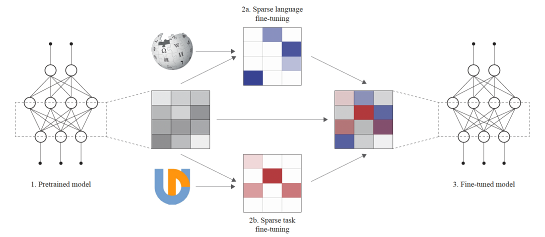
Overview of Composable SFT

- Composable SFT selects a subset of parameters that exhibit significant changes during initial training and fine-tune only these parameters in subsequent phases as per the LTH.
- Initially, the full model parameters $\theta^{(0)}$ are trained on target data, resulting in updated parameters $\theta^{(1)}$.
- Parameters are reset to their original values $\theta^{(0)}$ and only those marked by μ (top K based on absolute change) are updated in the subsequent training.

Equations of SFT

The sparse fine-tuning can be represented as $\phi = \theta^{(2)} - \theta^{(0)}$. Where $\theta^{(2)}$ are the parameters after sparse fine-tuning. The adaptation for a language and task can then be expressed as a function of the base model $F(\cdot; \theta + \phi_L + \phi_T)$.

Composable SFT Process[3]



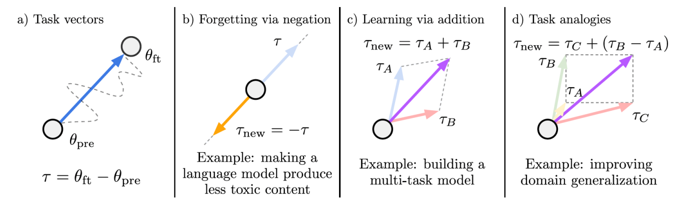
[3] Alan Ansell et al. "Composable Sparse Fine-Tuning for Cross-Lingual Transfer". In: (2023). arXiv: 2110.07560 [cs.CL].

Fact Representation: Model Editing

Overview of Task vectors

- A task vector τ_t is a vector that captures the changes made to a pre-trained model's weights when it is fine-tuned to perform a specific task ($\tau_t = \theta_{ft}^t - \theta_{pre}$)
- This vector τ_t can be used to adjust the model weights of another pre-trained model of the same architecture to improve its performance on task t, or combined with other task vectors or adjust the model's behavior such as unlearning a task.
- Task vectors involve element-wise operations on model weights, which assume a uniform structure across different model instances which could be a limitation.

Task Vectors in Model Editing[4]



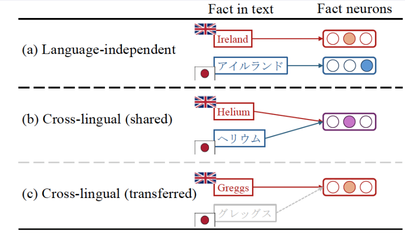
[4] Gabriel Ilharco et al. *Editing Models with Task Arithmetic*. 2023. arXiv: 2212.04089 [cs.LG].

Fact Representation: Cross Lingual

Overview of Fact Representation

- **Language-Independent:** Each language has a unique set of neurons responsible for the representation of facts, independent of other languages.
- **Cross-Lingual Shared:** ML-LMs use the same set of neurons to represent the same facts across multiple languages.
- **Cross-Lingual Transferred:** This representation type involves transferring factual knowledge from one language to others, typically from a high-resource language to low-resource languages.

Cross Lingual Fact Representation[5]



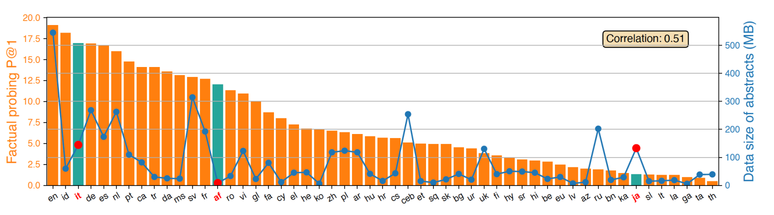
[5] Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. "Tracing the Roots of Facts in Multilingual Language Models: Independent, Shared, and Transferred Knowledge". In: (2024). arXiv: 2403.05189 [cs.CL].

Fact Representation: Training Dataset

Effect of Training Data Size

- It has been found that the highest correlation (0.51) with probing accuracy (P@1) is observed for data-size of abstracts.
- Italian and Japanese have P@1 score of 16.94% and 1.34% even though both of these languages are high resource with more than 100 MB of abstracts.
- Afrikaans language despite being low resource (<20 MB), shows high precision score (12.05%) for factual probing.

Factual Probing Results[6]



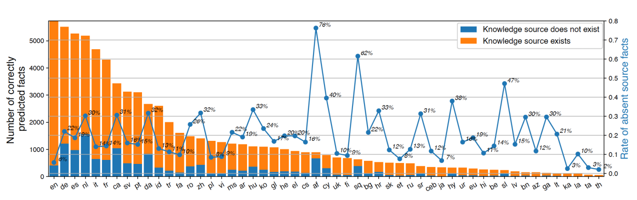
[6] Zhao, Yoshinaga, and Oba, "Tracing the Roots of Facts in Multilingual Language Models: Independent, Shared, and Transferred Knowledge".

Fact Representation: Tracing Roots

Formation of Cross Lingual Fact Representation

- To check if a given fact originates from the training data (Wikipedia), the roots is traced. If they co-occur, the fact is considered present, otherwise, it is considered as absent.
- Many of the facts that were absent in the knowledge source but correctly predicted were relatively easy to predict because of entity tokens and naming cues.
- Facts in low resource languages are correctly predicted despite not being verifiable in the training corpus indicating a possibility of cross-lingual transfer.

Predicted Facts Results[7]



[7] Zhao, Yoshinaga, and Oba, "Tracing the Roots of Facts in Multilingual Language Models: Independent, Shared, and Transferred Knowledge".

Geometry of Language Model Representation

Affine Language Subspaces[8]

- The space contains embeddings or vectors assigned to tokens based on their semantic or syntactic properties.
- Subspaces are low-dimensional vector space within the high-dimensional embedding space that capture certain linguistic properties.
- The language sensitive axes are axes (basic vectors) that are within the subspace that capture language-specific information (e.g. grammar).
- The language neutral axes are axes that are within the subspace encode information that is common across languages, such as word position or parts of speech.
- Languages tend to occupy similar linear subspaces in high-dimensional embedding spaces, after mean-centering.

[8] Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. "The Geometry of Multilingual Language Model Representations". In: (2022). arXiv: 2205.10964 [cs.CL].

Geometry of Language Model Representation

Identification of Affine Language Subspaces

For a particular *language A*, 512 input sentences are taken(each consists of 512 tokens) — therefore giving a total 262K contextual tokens.

$$\mathbf{c}^{(i)} \in \mathbb{R}^d \text{ where } i \in \{1, 2, \ldots, 262K\} \tag{1}$$

$$\boldsymbol{\mu}_A = \frac{1}{262K} \sum_{i=1}^{262K} \mathbf{c}^{(i)} \in \mathbb{R}^d \; ; \; S = \frac{1}{262K} \sum_{i=1}^{262K} (\mathbf{c}^{(i)} - \boldsymbol{\mu}_A)(\mathbf{c}^{(i)} - \boldsymbol{\mu}_A)^T \in \mathbb{R}^{d \times d} \tag{2}$$

After performing the eigenvalue decomposition on S, the top k eigenvectors of S are given by $V_A \in \mathbb{R}^{d \times k}$. The language subspace is identified by k eigenvector of S. k is selected such that $q/p = 0.9$.

$$E = (\lambda_1, \lambda_2, \ldots, \lambda_d) \text{ s.t. } \lambda_i \geq \lambda_{i+1} \quad \forall i \in \{1, 2, \ldots, d\} \; ; \; q = \sum_{i=1}^{k} E[i] \; ; \; p = \sum_{i=1}^{d} E[i] \tag{3}$$

The dimension of the context vector d was originally 768 and if the value of reduced dimension k is considered from each of the 12 layers of the transformer then the median value of k was found to be 335.

Geometry of Language Model Representation

Perplexity Ratio

The perplexity of the LLM is defined as:

$$pp(\mathbf{t}^{(i)}, \mathbf{t}^{(i-1)}, \ldots, \mathbf{t}^{(1)}) = \prod_{i=1}^{N} \left[\frac{1}{\mathbb{P}(\mathbf{t}^{(i)} \mid \mathbf{t}^{(i-1)}, \ldots, \mathbf{t}^{(1)})} \right]^{\frac{1}{N}} \tag{4}$$

$$\mathbf{u} = V_A^T(\mathbf{x} - \boldsymbol{\mu}_A) \text{ ; } \hat{\mathbf{x}} = V_A V_A^T(\mathbf{x} - \boldsymbol{\mu}_A) + \boldsymbol{\mu}_A \tag{5}$$

$$\text{Language-A:} \quad \mathbf{x}_A^{(i)[l]} \quad \forall i \in \{1, 2, \ldots, N\},\ \forall l \in \{1, 2, \ldots, 12\} \Longrightarrow pp_A^{[l]} \tag{6}$$

$$\text{Language-A:} \quad \hat{\mathbf{x}}_A^{(i)[l]} = V_A^{[l]} {V_A^{[l]}}^T (\mathbf{x}_A^{(i)[l]} - \boldsymbol{\mu}_A^{[l]}) + \boldsymbol{\mu}_A^{[l]} \quad \forall i,\ \forall l \Longrightarrow \hat{pp}_A^{[l]} \tag{7}$$

$$\text{Language-A:} \quad r_A^{[l]} = \frac{\hat{pp}_A^{[l]}}{pp_A^{[l]}} \quad \forall l \in \{0, 1, \ldots, 12\} \tag{8}$$

The average perplexity ratio over all the 88 languages for each of the layer is calculated as:

$$r^{[l]} = \frac{1}{88} \sum_{i=A}^{CJ} r_i^{[l]} \quad \forall l \in \{0, 1, \ldots, 12\} \tag{9}$$

Geometry of Language Model Representation

Key Findings

- Affine subspaces can be used for language modeling: To reconstruct the vectors of a particular *language A* from its corresponding *language subspace A*, the following equation can be used.

$$Proj_A(\mathbf{x}_A) = V_A V_A^T(\mathbf{x}_A - \boldsymbol{\mu}_A) + \boldsymbol{\mu}_A \qquad (10)$$

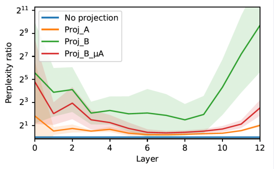
- Language subspaces differed from one another: To reconstruct the vectors of a particular *language A* from another *language subspace B*, the following equation can be used.

$$Proj_B(\mathbf{x}_A) = V_B V_B^T(\mathbf{x}_A - \boldsymbol{\mu}_B) + \boldsymbol{\mu}_B \qquad (11)$$

- Mean-shifted subspaces were similar to one another: To reconstruct the vectors of a particular *language A* after mean centering from another *language subspace B*, the following equation can be used.

$$Proj_{B,\boldsymbol{\mu}_A}(\mathbf{x}_A) = V_B V_B^T(\mathbf{x}_A - \boldsymbol{\mu}_A) + \boldsymbol{\mu}_A \qquad (12)$$

Results[9]



[9] Chang, Tu, and Bergen, "The Geometry of Multilingual Language Model Representations".

Conclusion

Summary

- This study highlights the significance of language model fine-tuning techniques such as adapter-based fine-tuning and composable sparse fine-tuning.
- Moreover, the analysis of fact representation in language models highlight on the importance of task-specific model editing, cross-lingual fact representation, and the underlying geometry of language model representation.

Future Directions

- We can explore different versions of the Lottery Ticket algorithm to improve efficiency. Additionally, experimenting with other pruning methods like DiffPruning[10]and ChildTuning[11]could help refine our approach.
- We can plan to improve how multilingual language models represent factual knowledge across languages. This includes developing better methods for cross-lingual fact representation learning and creating more accurate datasets for probing factual knowledge.

[11] Demi Guo, Alexander M. Rush, and Yoon Kim. "Parameter-Efficient Transfer Learning with Diff Pruning". In: (2021). arXiv: 2012.07463 [cs.CL].

[11] Runxin Xu et al. "Raise a Child in Large Language Model: Towards Effective and Generalizable Fine-tuning". In: (2021). arXiv: 2109.05687 [cs.CL].

References

- Houlsby, Neil et al. "Parameter-efficient transfer learning for NLP". In: *Proceedings of the 36th International Conference on Machine Learning*. 2019, pp. 2790–2799.
- Pfeiffer, Jonas et al. "MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.
- Ansell, Alan et al. "Composable Sparse Fine-Tuning for Cross-Lingual Transfer". In: (2023). arXiv: 2110.07560 [cs.CL].
- Ilharco, Gabriel et al. *Editing Models with Task Arithmetic*. 2023. arXiv: 2212.04089 [cs.LG].
- Zhao, Xin, Naoki Yoshinaga, and Daisuke Oba. "Tracing the Roots of Facts in Multilingual Language Models: Independent, Shared, and Transferred Knowledge". In: (2024). arXiv: 2403.05189 [cs.CL].
- Chang, Tyler A., Zhuowen Tu, and Benjamin K. Bergen. "The Geometry of Multilingual Language Model Representations". In: (2022). arXiv: 2205.10964 [cs.CL].
- Guo, Demi, Alexander M. Rush, and Yoon Kim. "Parameter-Efficient Transfer Learning with Diff Pruning". In: (2021). arXiv: 2012.07463 [cs.CL].
- Xu, Runxin et al. "Raise a Child in Large Language Model: Towards Effective and Generalizable Fine-tuning". In: (2021). arXiv: 2109.05687 [cs.CL].

Thank You!