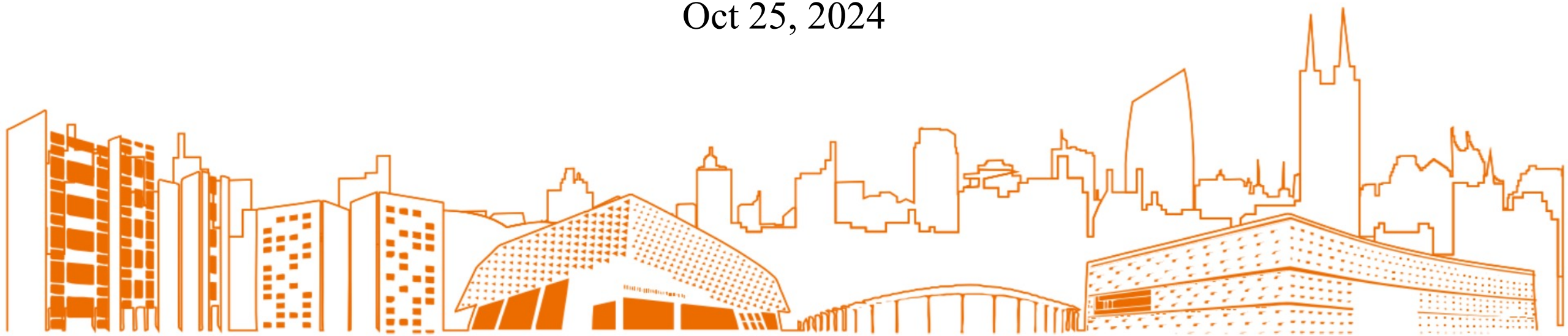
On the Interplay Between Fine-tuning and Sentence-level Probing for Linguistic Knowledge in Pre-trained Transformers

EMNLP 2020

Lao Leyi

South University of Science and Technology

Oct 25, 2024



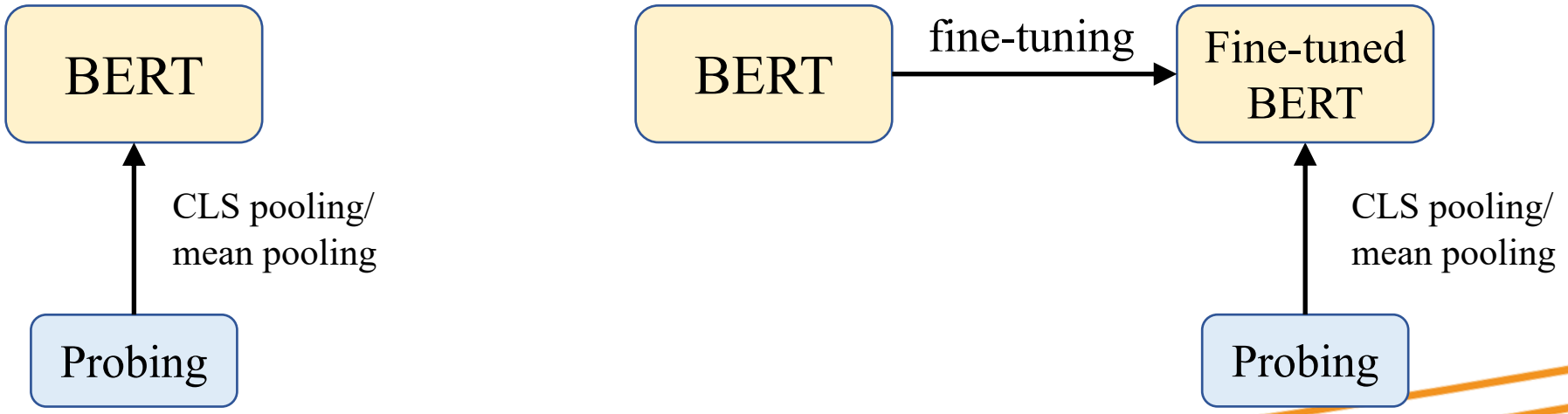
Motivation

- BERT, RoBERTa, and ALBERT perform well in various NLP tasks.
- It is still unclear how the **representations** of a pre-trained model change when fine-tuning on a downstream task.
- Further, little is known about whether and to what extent this process **adds or removes linguistic knowledge** from a pre-trained model.

Methods

- **Probing**
- **Investigating the following questions:**
  - How and where does fine-tuning affect the **representations** of a pre-trained model?
  - To which extent (if at all) can changes in **probing** accuracy be attributed to a change in linguistic knowledge encoded by the model.

Baseline



Methods

Pre-trained Models

BERT

RoBERTa

ALBERT

Fine-tuning tasks (Sentence-level)

- GLUE benchmark
  - CoLA (The Corpus of Linguistic Acceptability, focusing on **syntactic**)
  - SST-2 (Stanford Sentiment Treebank, focusing on **semantic and/or discourse**)
  - RTE (Recognizing Textual Entailment, focusing on **discourse**)
- SQuAD (Stanford Questions Answering Dataset, focusing on **discourse**)

Methods

Probing tasks (Sentence-level)

- bigram-shift (**Syntactic** Task)
  - Testing a model's sensitivity to word order.
- semantic-odd-man-out (**Semantic** Task)
  - Testing a model's sensitivity to semantic incongruity
- coordination-inversion (**Discourse** Task)
  - Testing for a model's broader discourse understanding.

Pooling

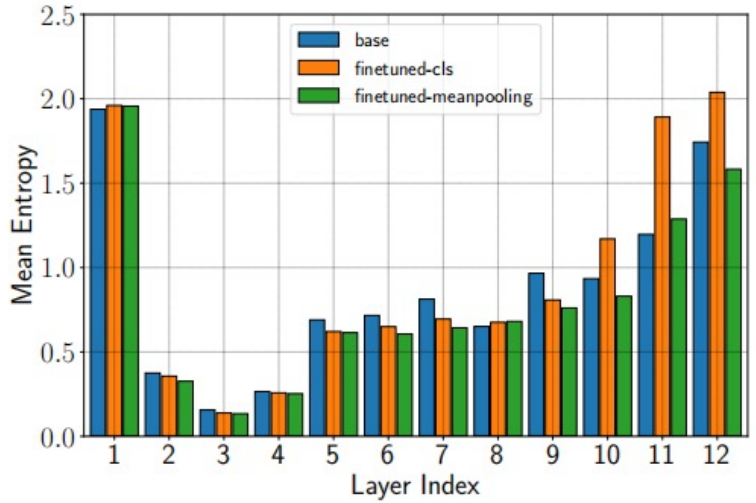
- CLS-Pooling
- mean-pooling

Experiment

Probing Task	BERT-base-cased							
	CLS-pooling				mean-pooling			
	CoLA		SST-2		CoLA		SST-2	
	0 – 6	7 – 12	0 – 6	7 – 12	0 – 6	7 – 12	0 – 6	7 – 12
bigram-shift	0.07	4.73	−1.02	−4.63	0.23	1.45	−0.37	−3.24
coordinate-inversion	−0.10	1.90	−0.25	−1.15	0.14	0.29	−0.48	−0.85
odd-man-out	−0.20	0.26	−0.02	−1.28	−0.34	−0.29	−0.30	−1.09

Probing Task	RoBERTa-base							
	CLS-pooling				mean-pooling			
	CoLA		SST-2		CoLA		SST-2	
	0 – 6	7 – 12	0 – 6	7 – 12	0 – 6	7 – 12	0 – 6	7 – 12
bigram-shift	0.58	5.35	−2.41	−7.22	0.69	1.74	−0.23	−4.87
coordinate-inversion	−0.72	1.84	−1.28	−0.63	−0.22	0.02	−0.18	−3.83
odd-man-out	−0.66	1.05	−1.09	−2.40	−0.08	−0.55	−0.46	−3.61

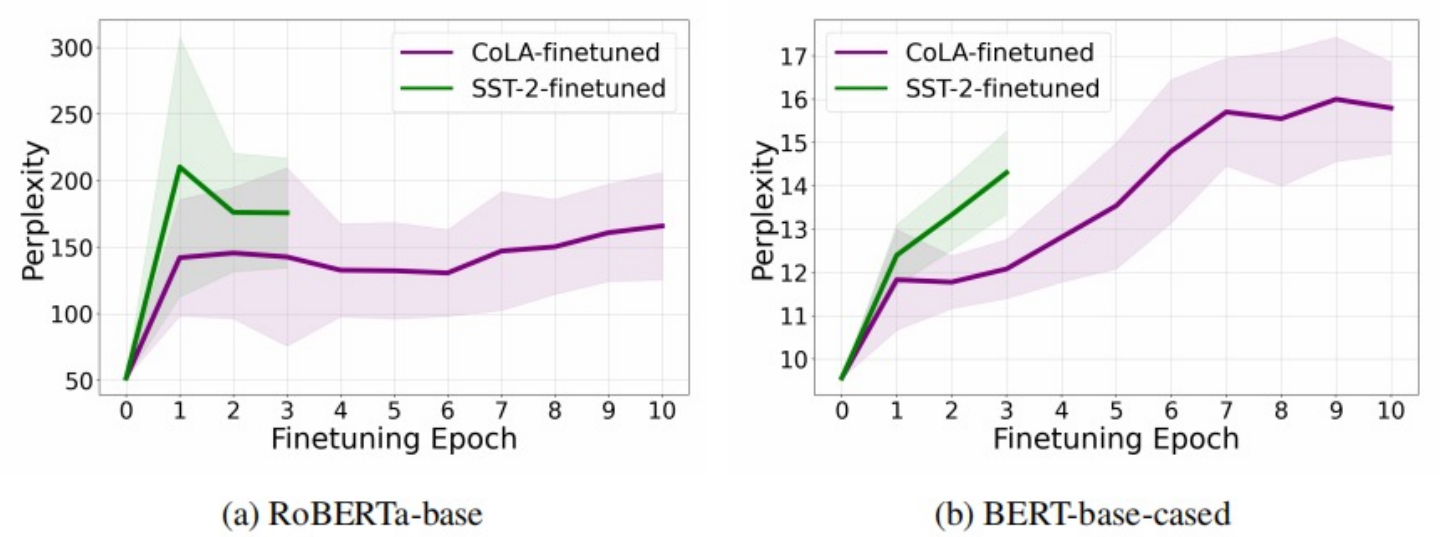
Probing Task	ALBERT-base-v1							
	CLS-pooling				mean-pooling			
	CoLA		SST-2		CoLA		SST-2	
	0 – 6	7 – 12	0 – 6	7 – 12	0 – 6	7 – 12	0 – 6	7 – 12
bigram-shift	1.55	3.39	−1.94	−5.15	0.26	0.66	−0.70	−2.73
coordinate-inversion	−0.69	−1.53	−1.07	-2.87	−0.07	−1.19	−0.35	−1.53
odd-man-out	−0.42	−1.39	−0.90	−2.75	−0.27	−1.40	−0.60	−2.82



(a) Entropy

- The CLS token learns to take more sentence-level information into account.
- The improvement in probing accuracy can not simply be attributed to the encoding of linguistic knowledge.

Experiment



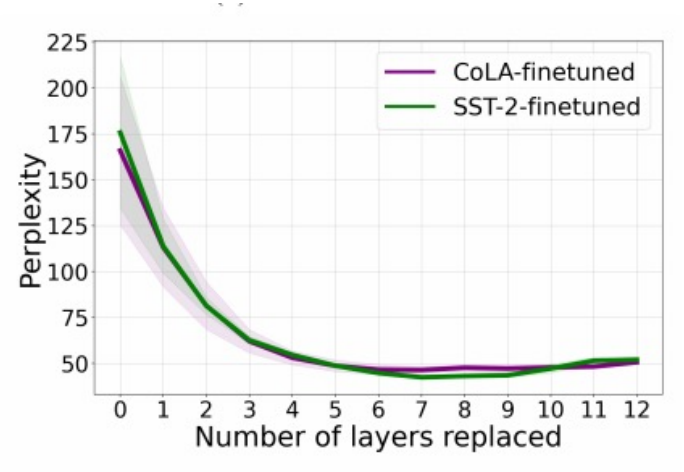
(a) RoBERTa-base

(b) BERT-base-cased

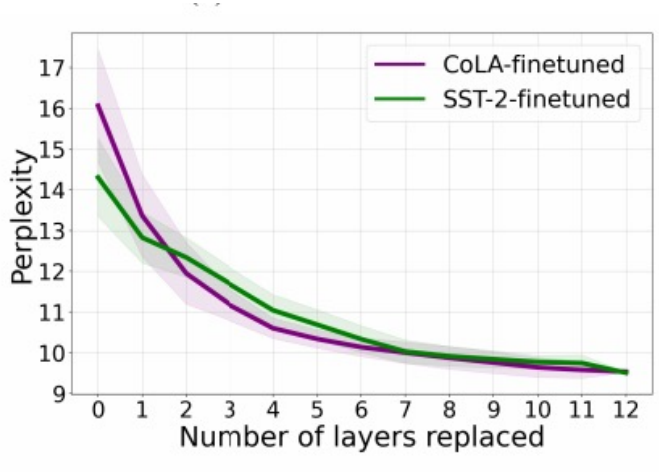
The **pretrained masked language model heads** are evaluated on the Wikitext-2 test set and compare it to the masked-language modeling **perplexity** of fine-tuned models.

Finetuning on SST-2 has indeed more dramatic effects on the representations of both models compared to fine-tuning on CoLA.

Experiment



(c) RoBERTa-base



(d) BERT-base-cased

Test which layers **contribute the most** to the change in perplexity and replace the layers of the fine-tuned encoder with pre-trained layers, starting with the last layer.

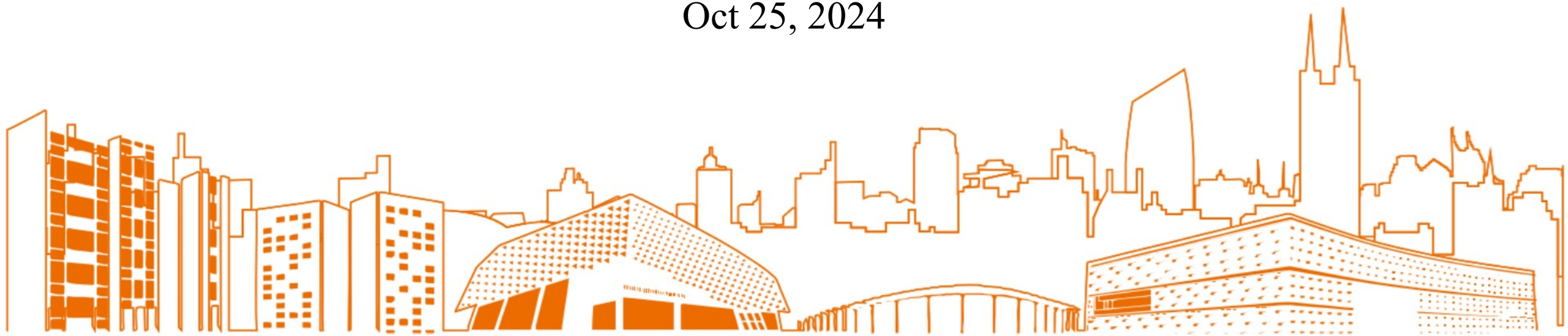
- Finding that the changes that lead to an increase in perplexity happen in the **last layers.**
- Fine-tuning indeed does affect the representations of a pre-trained model and in particular those of the **last hidden layers**.

Investigating the Representation Learning of Interjections in Fined-tuned Language Models

Lao Leyi

South University of Science and Technology

Oct 25, 2024



Motivation

- **Interjections** (e.g., uh, mmhmm) are important signals in conversation.
- Can Language model such as GPT-2 and Llama3 properly learn the representation of interjections through fine-tuning;
- Does the representation of the interjection show similarity across the t**hree chosen language** from the selected dialogue datasets
- Which fine-tuning task can best foster representation learning for interjections?

Methods

In order to verify how interjections are properly represented in the language model, we will focus on using **fine-tuning task** instead.

Pre-trained Models

- BERT
- GPT-2
- Llama3

Language

- Japanese
- English
- German

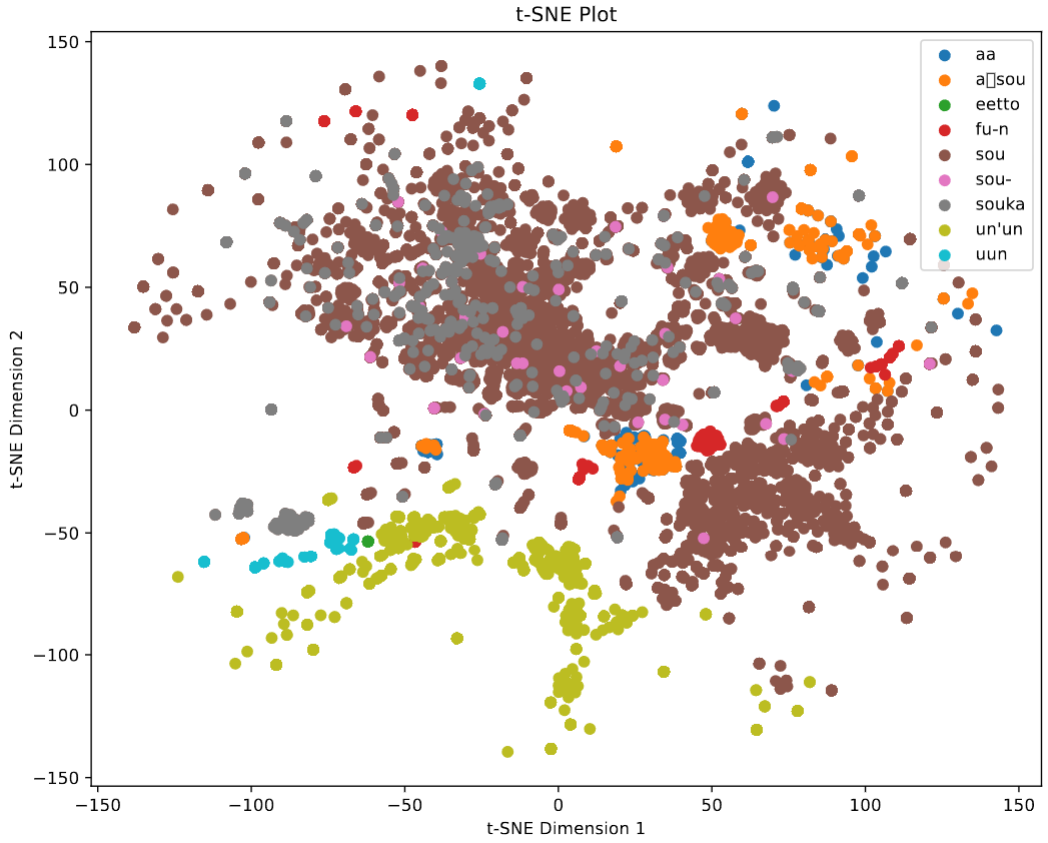
Methods

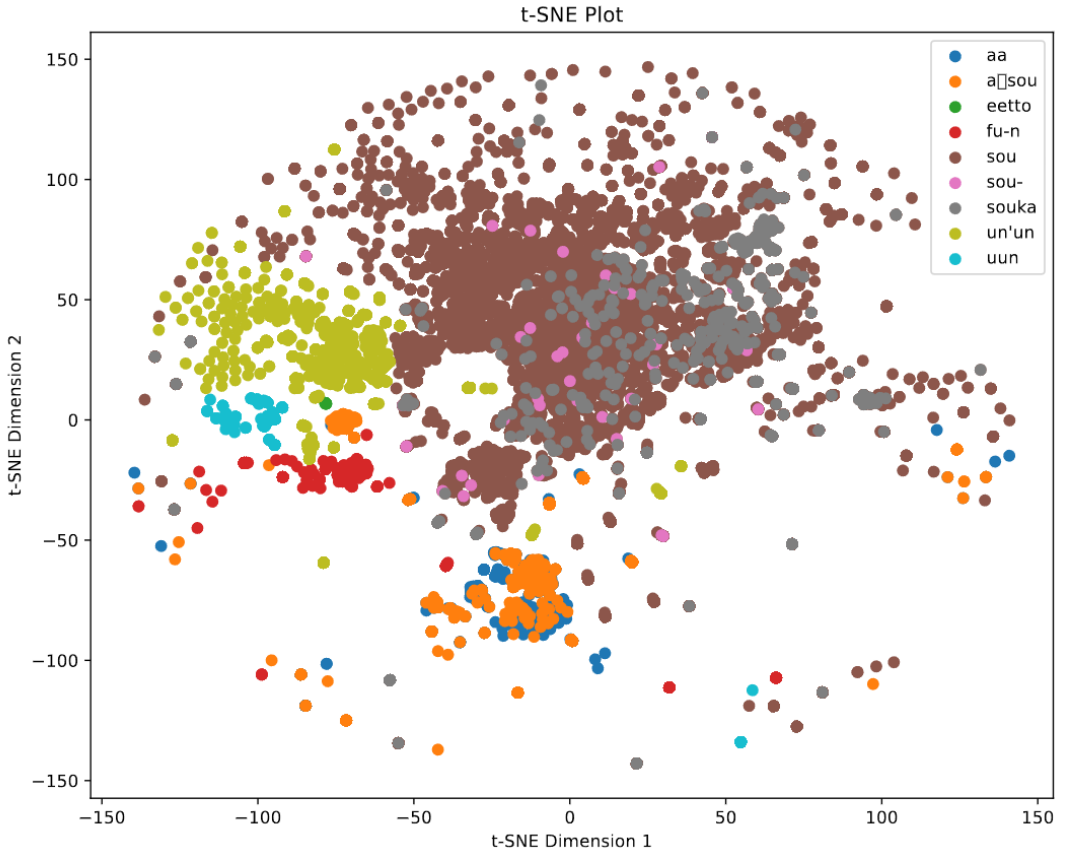
Fine-tuning tasks

- next-token prediction (baseline)
- dialogue-act prediction
- turn-taking prediction

Experiment

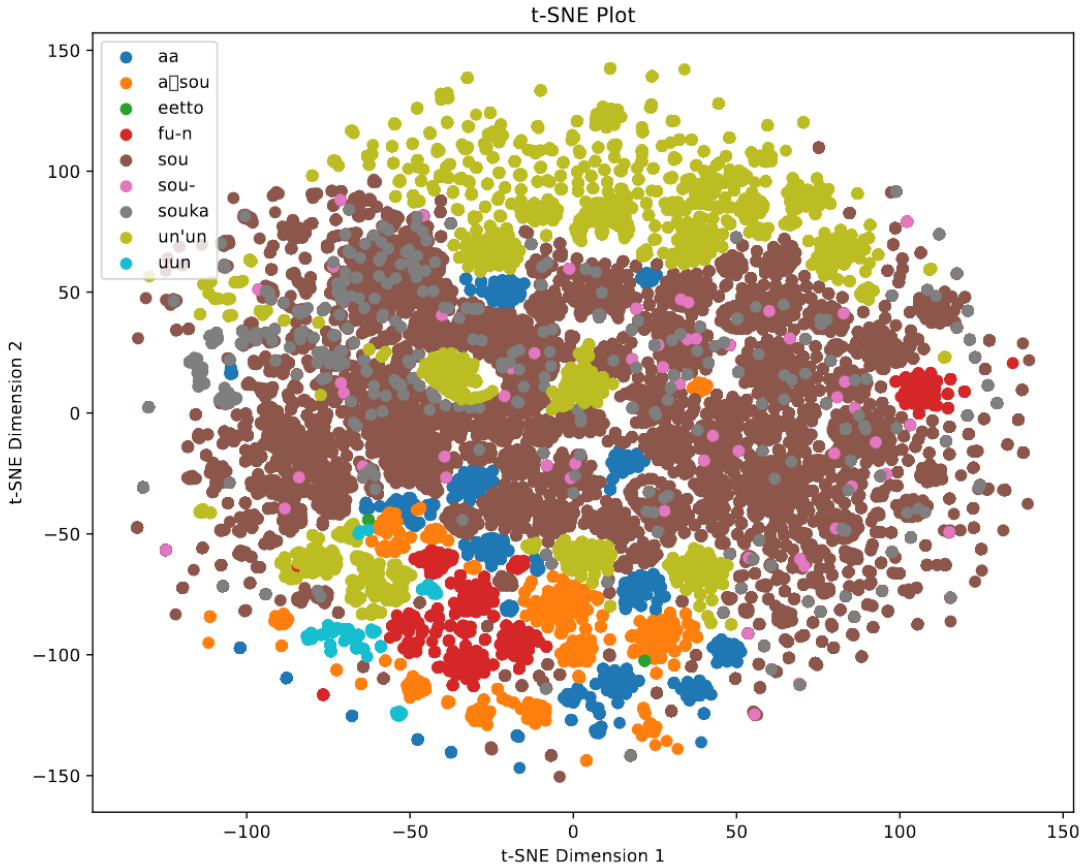
GPT-2_next-token prediction

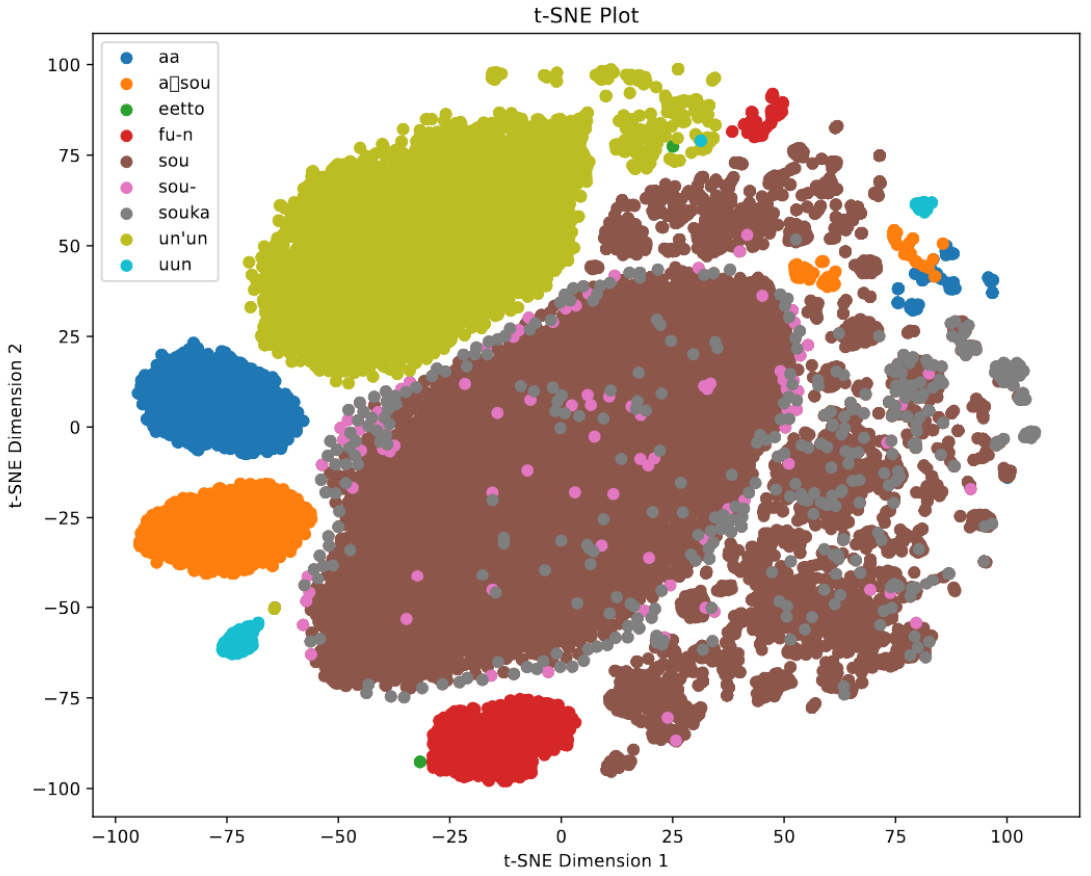




Experiment

Llama3_next-token prediction





SUSTech Southern University of Science and Technology

Thank You