Efficient Test Time Adapter Ensembling for Low-resource Language Varieties
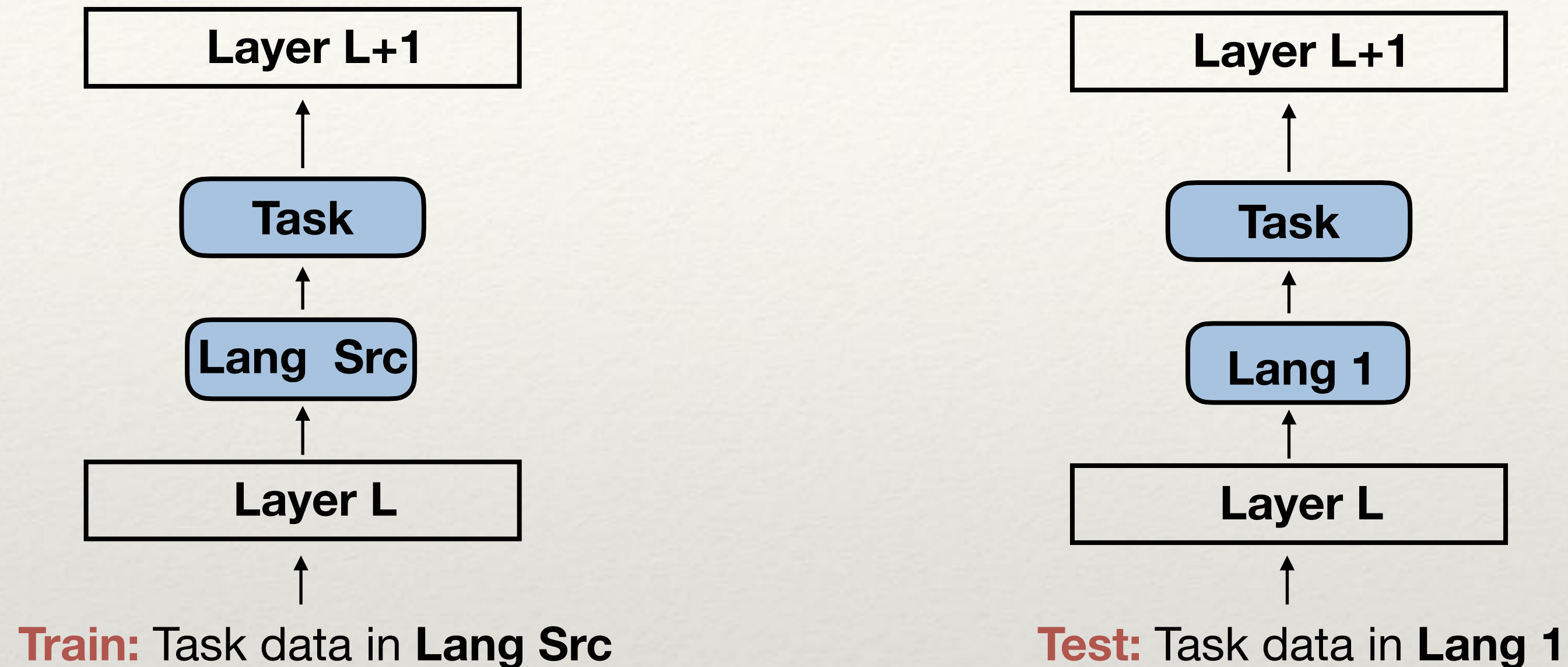
Xinyi Wang[1]

Yulia Tsvetkov[2]

Sebastian Ruder[3]

Graham Neubig[1]

1. Language Technologies Institute, CMU
2. Paul G. Allen School of Computer Science & Engineering, University of Washington
3. DeepMind

Adapters for Cross-lingual Transfer

Layer L+1

↑

Task

↑

Lang Src

↑

Layer L

↑

Train: Task data in **Lang Src**

Layer L+1

↑

Task

↑

Lang 1

↑

Layer L

↑

Test: Task data in **Lang 1**

Language/task adapters

- Let M be the pretrained model

- Then we can use $\langle T_j, \mathrm{Lang}_i, M \rangle$ to do zero-shot inference on any language Lang_i on task T_j

Supporting Language Varieties

7000+ languages in the world

AdapterHub

How are you? $\longrightarrow$ **English**
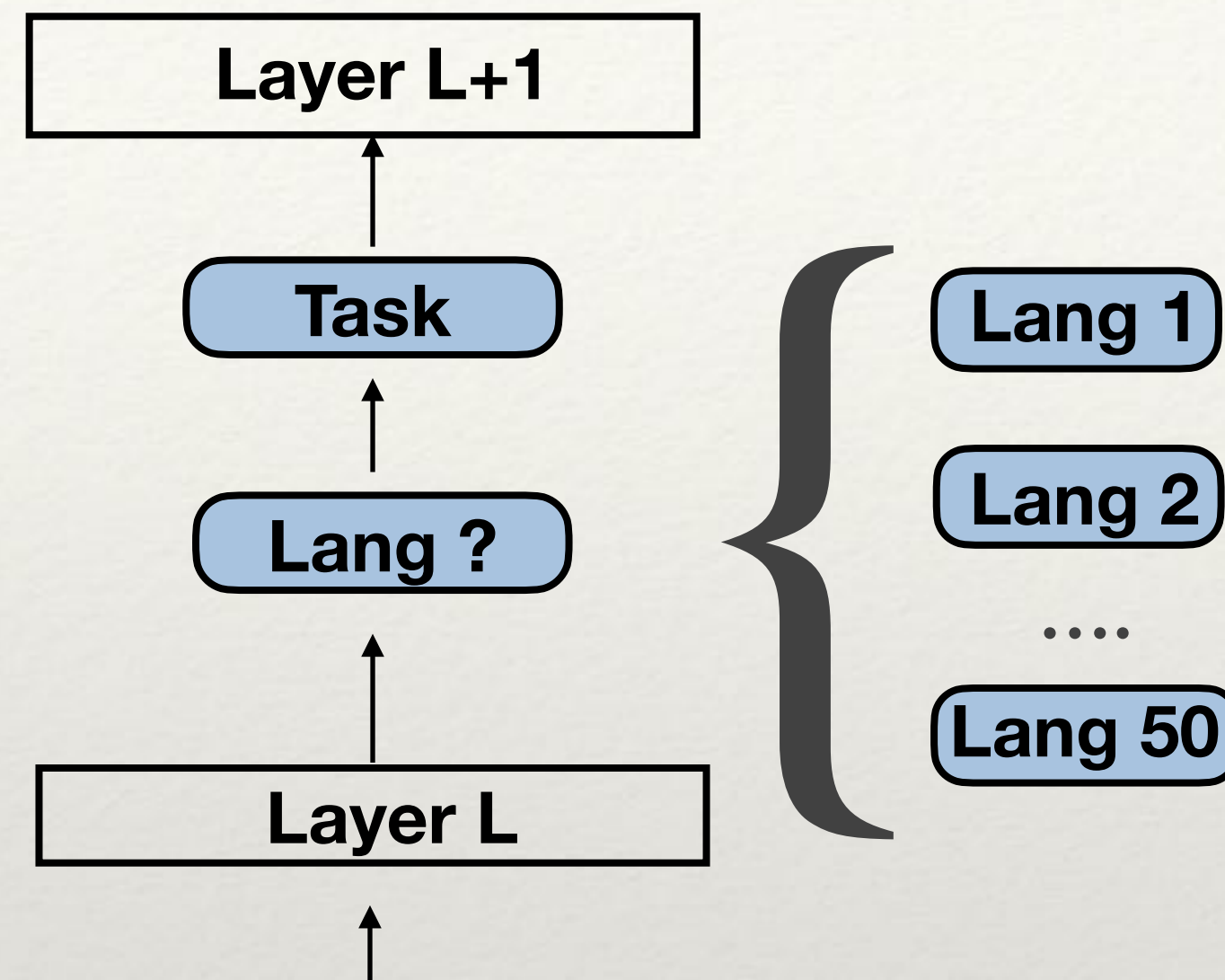
How you dey? $\longrightarrow$ **Nigerian Pidgin**

Currently only supports 50 languages

Language ID is hard for low-resource languages

https://adapterhub.ml/
Language ID in the wild: Unexpected challenges on the path to a thousand language web text corpus. (Caswell et al. 2020)
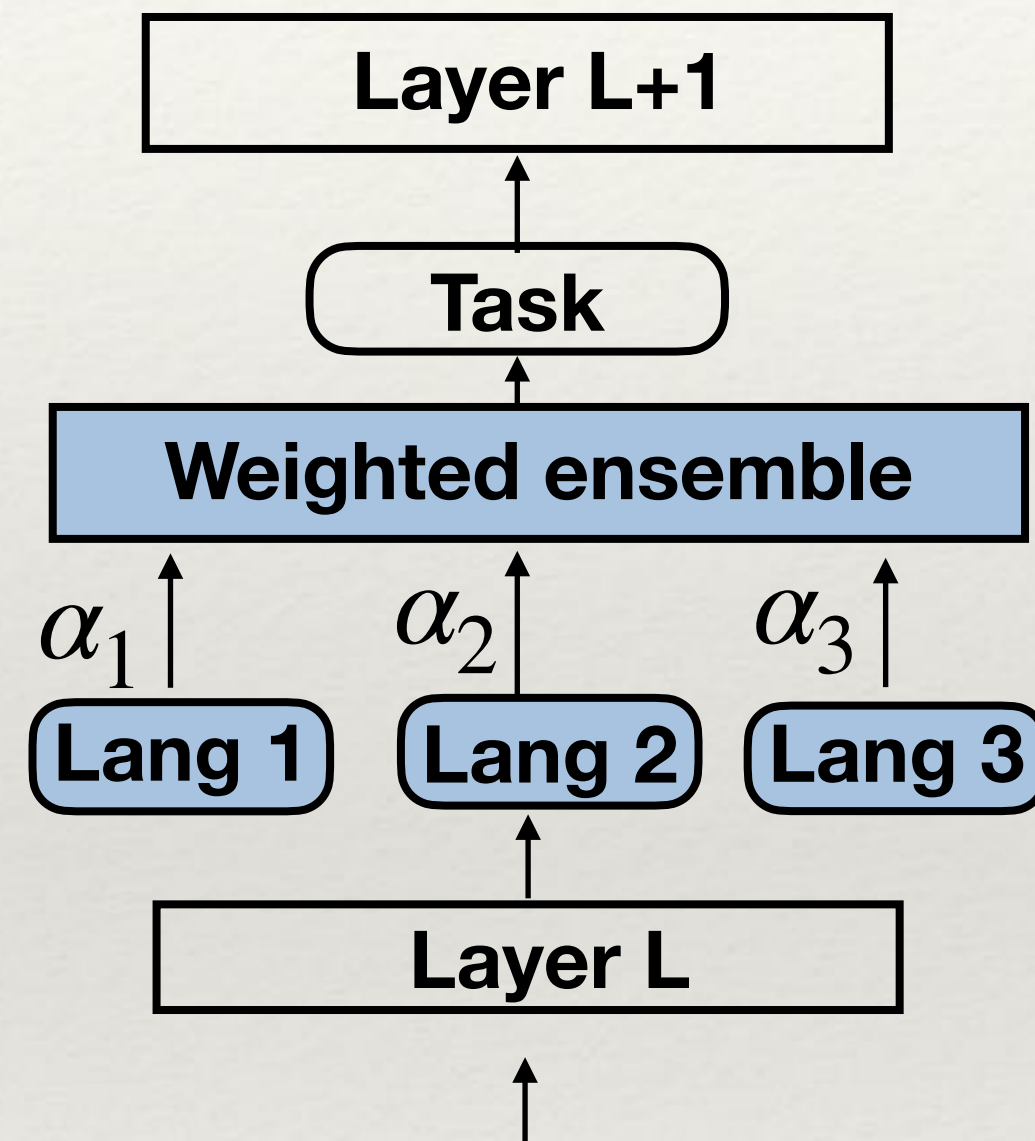
Adapters for Unseen Language Variety

Test: Task data in **unseen** language varieties Lang_i

❖ What if we encounter a **new language variety** that does not have a language adapter?

    ❖ We found using adapter from the related language is not competitive

❖ Can we adapt to unseen language varieties at **test time**?
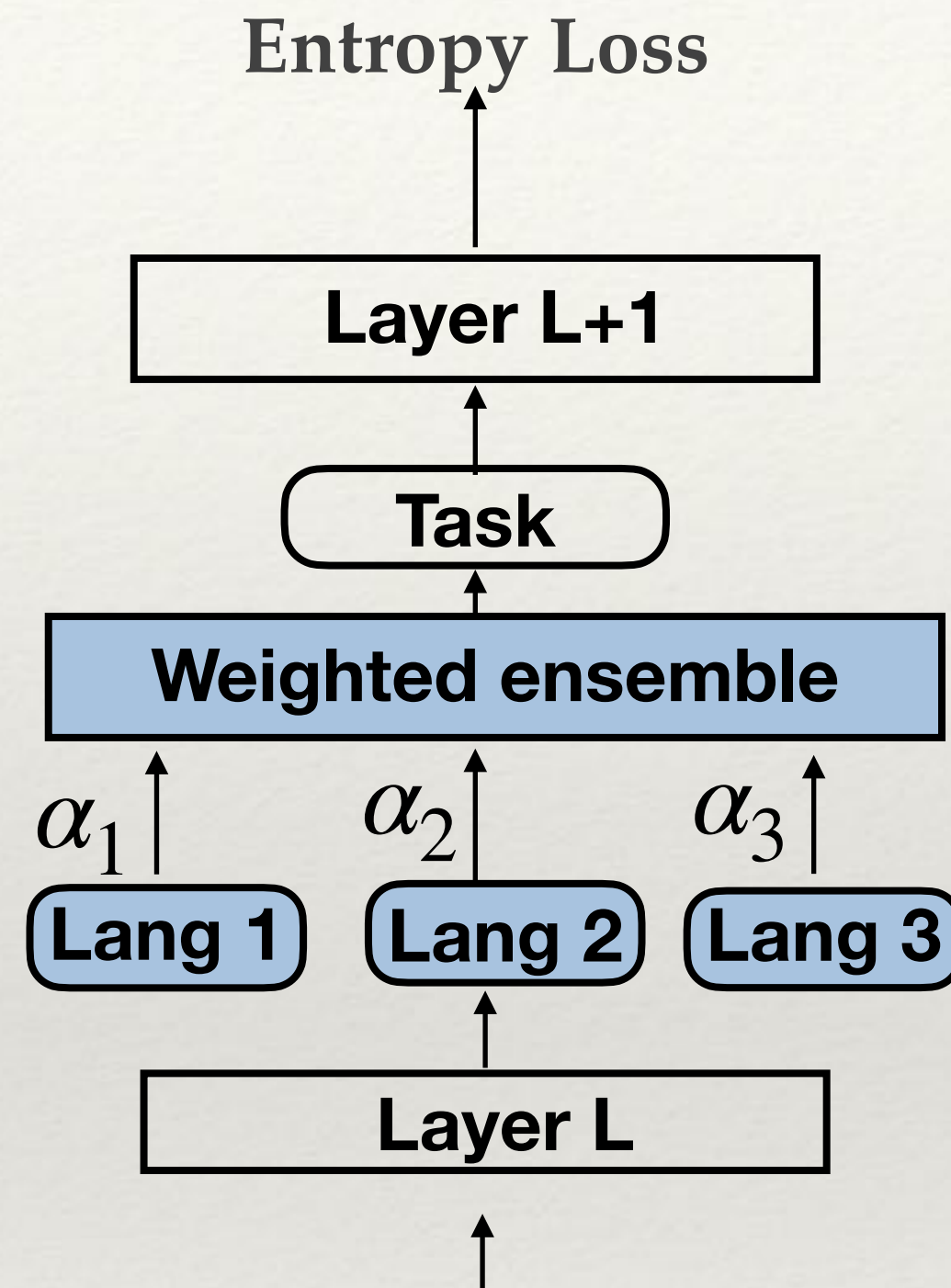
Ensemble of Language Adapters

❖ Equal weighting of related language adapters and source language adapters

Entropy Minimized Ensemble of Language Adapters

Entropy Loss

Layer L+1

Task

Weighted ensemble

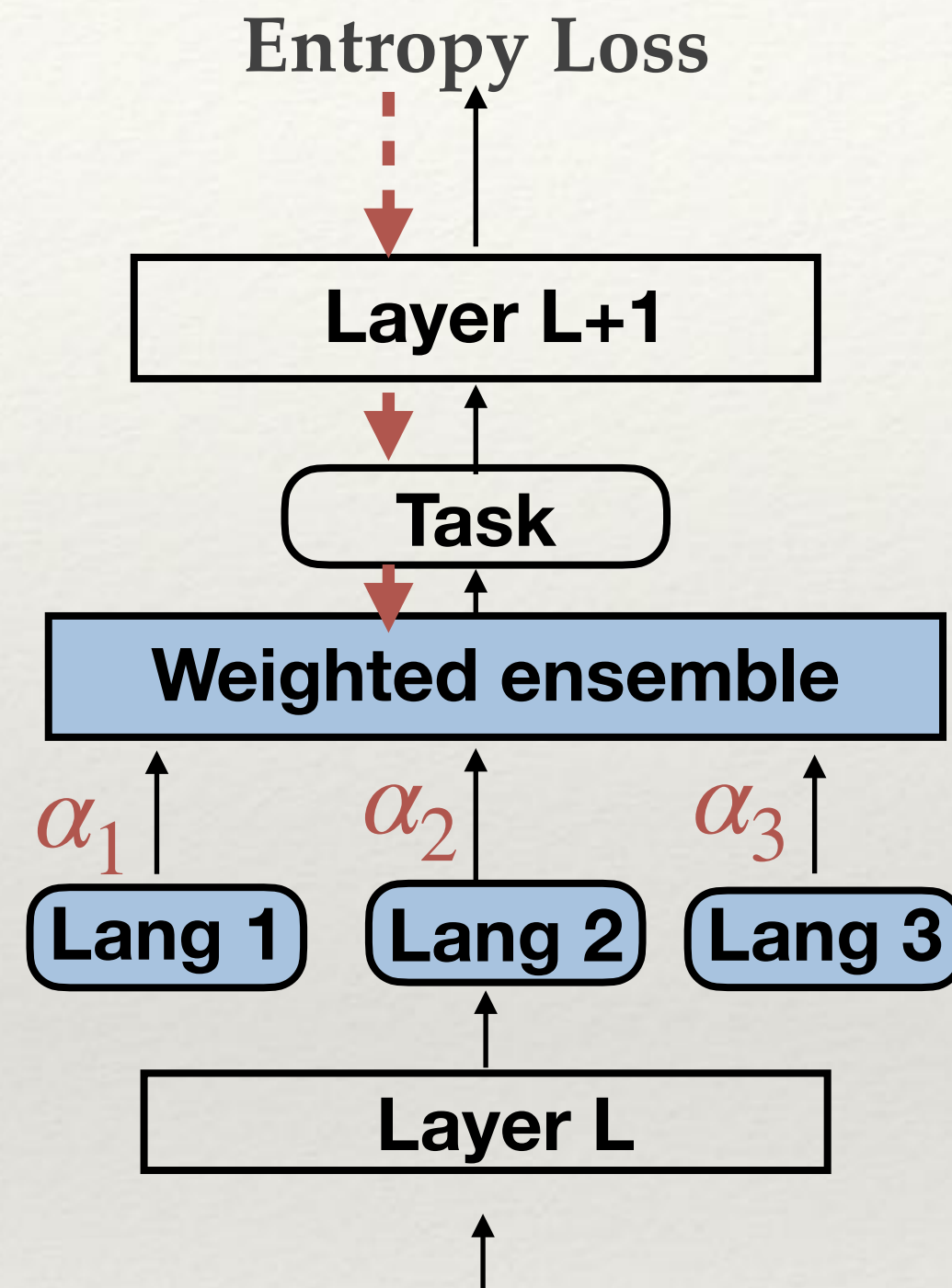α_1 α_2 α_3

Lang 1 Lang 2 Lang 3

Layer L

Test: Task data in **unseen** language variety Lang_i

❖ Optimize α_i to minimize test entropy $H(x; \alpha)$

Entropy Minimized Ensemble of Language Adapters

Entropy Loss

Layer L+1

Task

Weighted ensemble

α_1 α_2 α_3

Lang 1   Lang 2   Lang 3

Layer L

Test: Task data in **unseen** language variety Lang_i

❖ Optimize α_i to minimize test entropy $H(x; \alpha)$

Experiment & Results

❖ **Tasks**

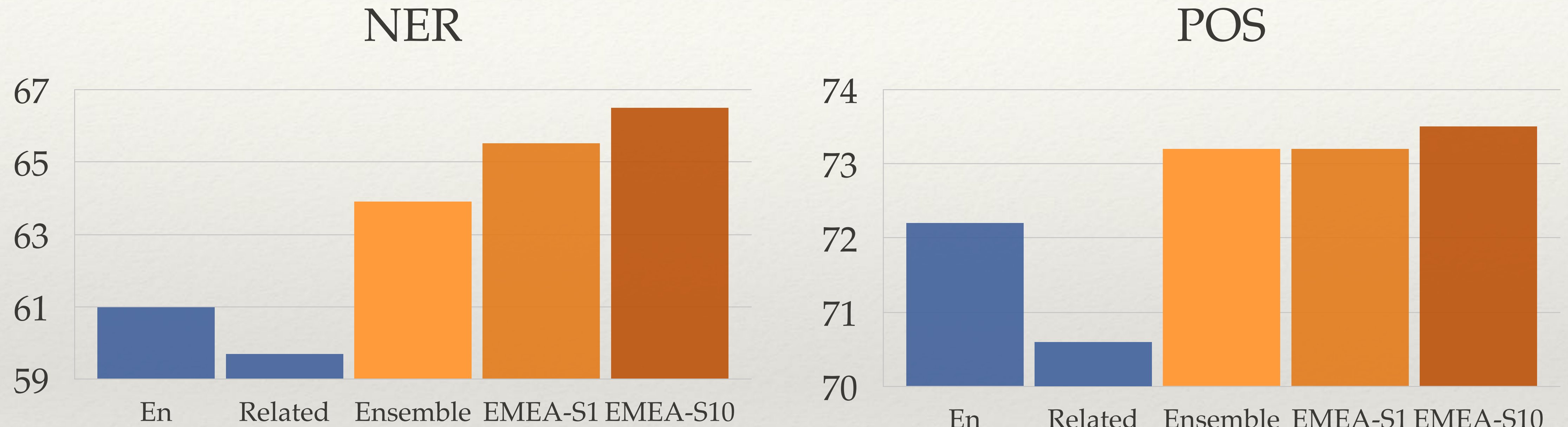 ❖ NER/POS tagging

❖ **Languages** (3 groups of related language)

Test Languages	Adapter Languages
Marathi, Bengali, Tamil	Hindi, Arabic, English
Faroese, Norwegian, Denmark	Icelandic, German, English
Belarusian, Ukrainian, Bulgarian	Russian, English

Experiment & Results

NER

POS

- Ensemble performs much better than using any single language adapter
- 1 step of EMEA already performs better than Ensemble on NER
- 10 steps of EMEA further improves the performance

Conclusion and Limitation

❖ NLP models should be adaptive to personalized language varieties

❖ We propose a **test time adaptation algorithm** for language varieties

❖ Future work: Curate high quality **dataset** for language varieties

❖ Future work: Reduce **inference cost** for test time adaptation algorithms

Code: https://github.com/cindyxinyiwang/emea
Contact: xinyiw1@cs.cmu.edu