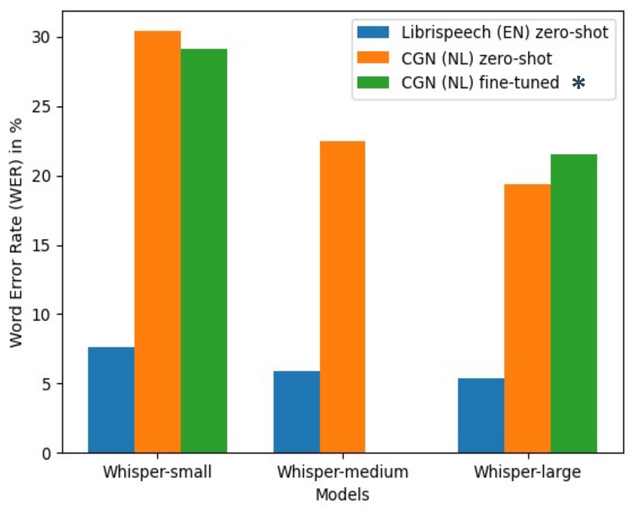
Primal-OWSM: Speech Foundation Model with Parameter-efficient Primal Attention for Low-resource Dutch Speech Recognition

Pu Wang, Hugo Van hamme
KU Leuven, Dept. of Electrical Engineering, PSI, Leuven, Belgium

Motivation

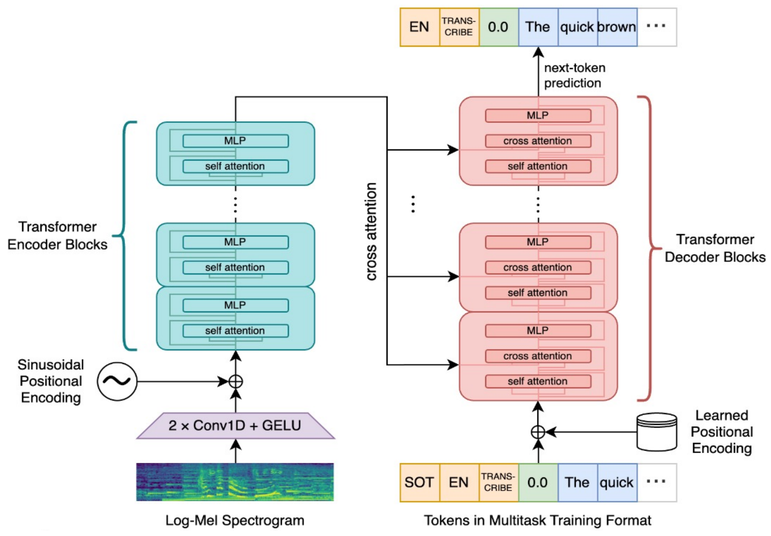
- Speech foundation models, such as OpenAI Whisper, achieve human-level accuracy in English automatic speech recognition (ASR);
- These models are **not robust** with **non-English audio or low-resouce languages**;
- Fine-tuning can help enhance performance;
- Whisper (medium/large) models tend to **overfit** quickly on **small datasets** during **fine-tuning.**

	# of parameters†		
	Small	Medium	Large
Whisper	244 M	769 M	1550 M
Primal-OWSM	272 M	712 M	889 M



* Fine-tuned with the Dutch subset of the Multilingual Librispeech (MLS) dataset.

ASR Speech Foundation Model



Whisper Source: https://github.com/openai/whisper

† The amount of training data increases as the model size scales from small to large.

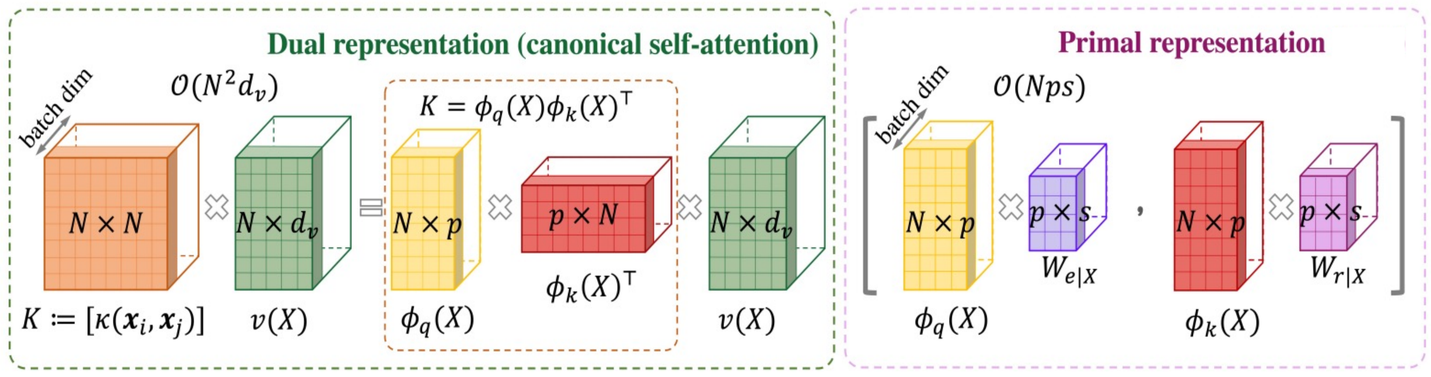
Aim

- To **build** the **speech foundation model**, focusing on **efficient fine-tuning** for **low-resource Dutch ASR**;
- Whisper is not open-source and its training resources are not fully accessible;
- To promote transparency, the efficient fine-tuning is based on the **Open-source Whisper style Speech Model (OWSM)**.

Primal-OWSM

- The ***Softmax*** **self-attention** in the **last layers** of the encoder (approximately $^1/_5$ to $^1/_3$ of the layers) is replaced with **Primal Attention** to reduce computations;
- Configured with **three model scales**, similar to Whisper;
- Initialized with OWSM V1, V2 and V3.
- Projection directions $s = 48$, $p = d_k$ =64.

Primal Attention



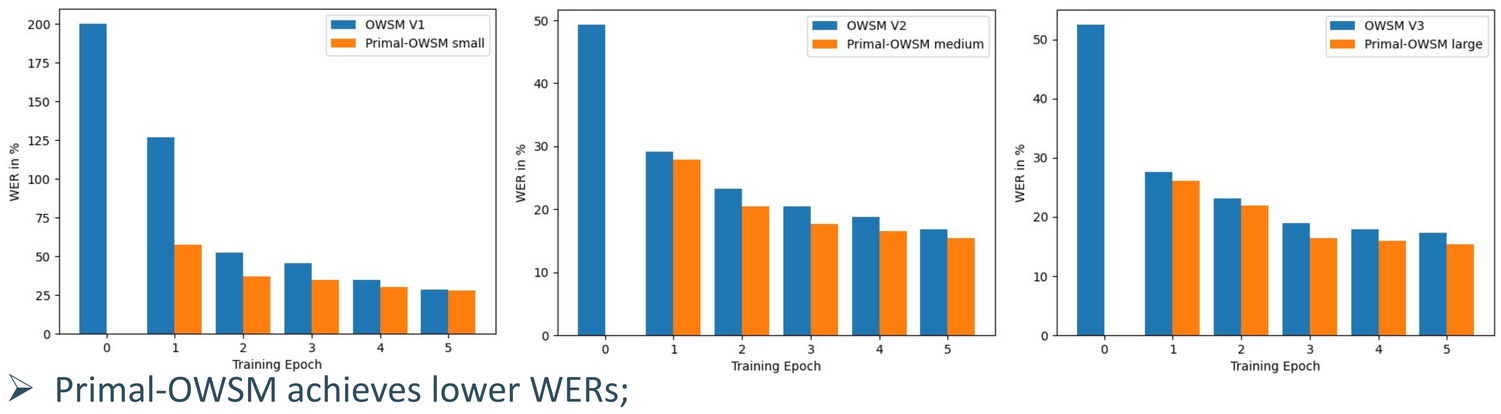
Source: Primal-Attention: Self-attention through Asymmetric Kernel SVD in Primal Representation

- **Low-rank or sparsity** properties of the attention matrices are observed in the deeper encoder layers;
- Implement **singular value decomposition (SVD)** on the *Softmax* attention;
- Self-attention can be interpreted as a **kernel method**:

$$Attn_i^n = \sum_j S_{ij}^n V^n x_j \text{ with } S_{ij}^n = \text{Softmax}\left(\frac{x_j^T K^{nT} Q^n x_i}{\sqrt{d_k^n}}\right) \quad K \coloneqq [\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)]$$

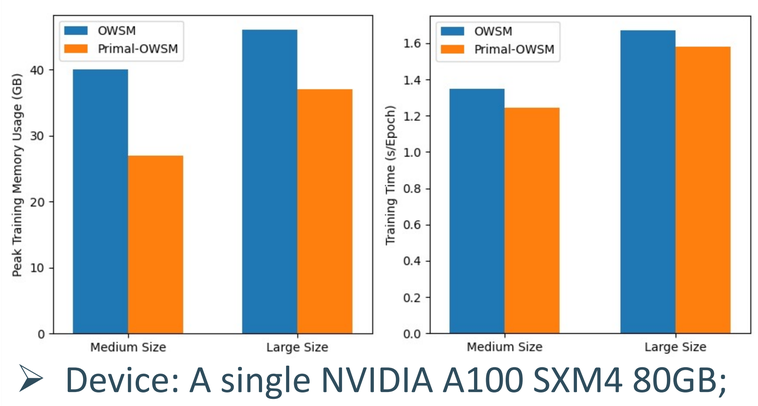
- Dual representation computation is expensive and less informative;
- Primal representation: Use explicit feature maps to approximate the attention mechanism, jointly maximizing the variances of the two projections;
- Replace the attention output $Attn_i^n$ with the cosine similarity of feature maps $\phi_q(X)$ and $\phi_k(X)$.

ASR Performance (CGN 236 hours)



- Primal-OWSM achieves lower WERs;
- To achieve a similar WER, Primal-OWSM saves more than 40% of training time (fewer training epochs).

Efficiency



- Device: A single NVIDIA A100 SXM4 80GB;
- Up to 40% training memory savings;
- 10% training time savings per epoch.

Acknowledgments

The research was supported by the Flemish Government under “Onderzoeksprogramma AI Vlaanderen” and FWO-SBO grant S004923N: NELF.