Google Research

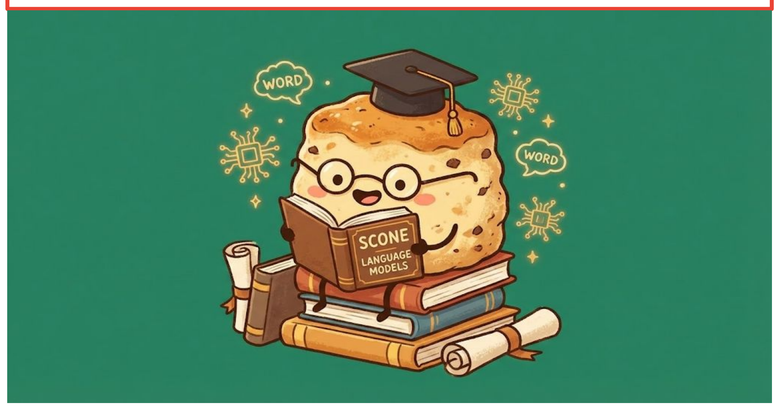
SCONE: Scaling Embedding Layers in Language Models

Da Yu, Edith Cohen, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Daogao Liu, Chiyuan Zhang

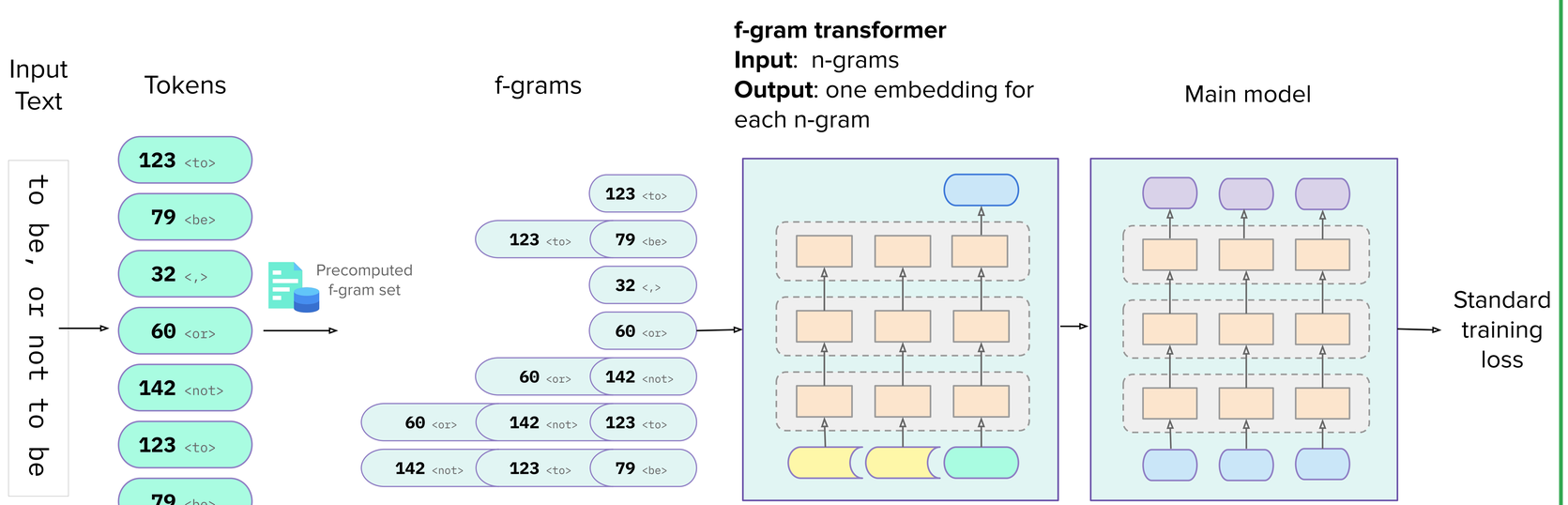
Summary

SCONE (**S**calable, **C**ontextualized, **O**ffloaded, **N**-gram **E**mbedding) enables two new scaling strategies, while keeping inference FLOPS and accelerator memory fixed:

1. Saving a larger number of embeddings with O(1) access complexity.
2. Using larger models to learn those embeddings.



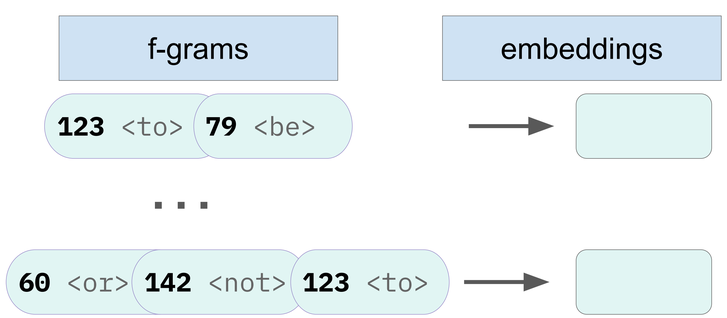
Learning f-gram Embeddings in Standard Training Process



Frequent n-gram Embedding Layer

We introduce an **f-gram embedding layer** as an addition to the standard token embedding layer.

This layer maps '**f-grams**', defined as frequent n-grams identified in the training corpus, to their corresponding embedding vectors.



New Scaling Axes

- **f-gram model scaling**: Scaling up the f-gram transformer improves embedding quality and performance, without increasing any inference cost.
- **Embedding Scaling:** Increasing the number of cached f-gram embeddings.
  While storage costs rise, the O(1) access complexity allows efficient offloading to main memory or disk.
- We observe meaningful scaling curves for both axes.

Model Performance

Zero-shot evaluation after pretraining

Model	MMLU
OLMo-1B	37.6
OLMo-1.9B	38.6 (+1.0)
SCONE-1B **10 million f-grams**, 1.8B f-gram model	39.3 (+1.7)
SCONE-1B **1 billon f-grams**, 1.8B f-gram model	39.9 (+2.3)

See paper for additional results and ablation studies.

Notably, we also demonstrate that:
SCONE can also be applied during post-training.