Towards Improved Sentence Representations using Token Graphs

Krishna Sri Ipsit Mantri, Carola-Bibiane Schönlieb, Zorah Lähner, Moshe Eliasof
kmantri@uni-bonn.de, cbs31@cam.ac.uk, laehner@uni-bonn.de, eliasof@bgu.il

ICLR International Conference On Learning Representations · UNIVERSITÄT BONN · LAMARR INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE · UNIVERSITY OF CAMBRIDGE · Ben-Gurion University of the Negev

Scan for Code

Goal: Produce robust sentence representations from LLMs without requiring any fine-tuning
Key Idea: Reframe pooling as relational learning over latent token similarity graphs, followed by aggregation
Technique: Use a GNN to explictly learn multi-hop linguistic token relationships from frozen LLMs

Motivation

- Semantic textual understanding requires collapsing a multi-token sequence into a single global vector $\mathbf{z}$
- Standard pooling (like Mean) discards the rich relational context captured by LLM's self-attention
- This results in signal dilution i.e., crucial information is drowned out by distractor tokens
- **Goal:** Can we generate robust sentence embeddings from frozen LLMs that actually preserve multi-hop token relationships?

Relational Learning over Graphs

- Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with node features $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times d}$
- The initial node representations $\mathbf{H}^{(0)} = \mathbf{X}\mathbf{W}_{in}$ are fed to a K-layer GNN:

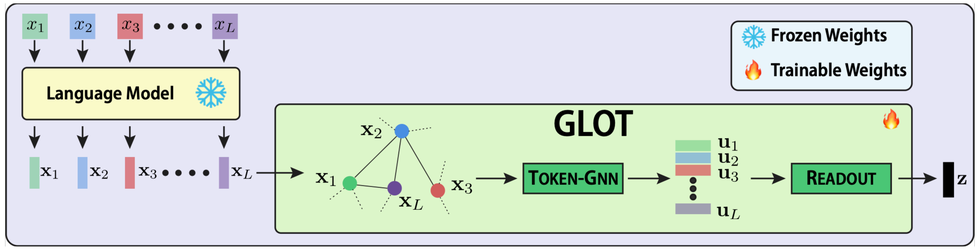
$$\mathbf{a}_i^{(\ell)} = \underset{j \in \mathcal{N}_i}{\text{AGGREGATE}} \left(\mathbf{h}_j^{(\ell)} \right),$$

$$\mathbf{h}_i^{(\ell+1)} = \sigma \left(\mathbf{W}^{(\ell)} \text{CONCAT}(\mathbf{h}_i^{(\ell)}, \mathbf{a}_i^{(\ell)}) \right)$$

  where $\mathbf{W}^{(\ell)} \in \mathbb{R}^{p \times 2p}$ is a learnable weight matrix, σ is a nonlinear activation function and $\ell = 1, \cdots, K$ denotes the current layer
- GNNs are theoretically more powerful than deepsets learning frameworks (Bronstein et al., 2021)

GLOT

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L] \in \mathbb{R}^{L \times d}$ be token hidden states from a frozen LLM



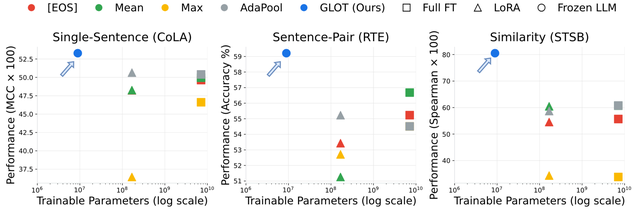
- Define token graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where nodes $|\mathcal{V}| = L$ correspond to tokens
- Create edge $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{E}$ if the cosine similarity exceeds a threshold τ
- Capture multi-hop linguistic dependencies and refine the tokens using a lightweight Token-GNN to obtain $\mathbf{H}^{(K)} = \mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_L] \in \mathbb{R}^{L \times p}$
- Aggregate structurally refined tokens using a learnable readout mechanism to get a single, robust sentence embedding $\mathbf{z}$

Computational Efficiency using Mistral-7B

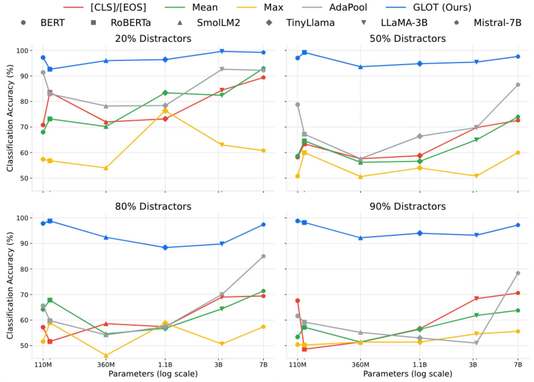
Full Fine-Tuning	LoRA ($r = 64$)	GLOT
7.11 B Parameters	167.8 M Parameters	8.9 M Parameters
32.59 GB Memory	33.50 GB Memory	0.42 GB Memory
1318 ms Runtime	1454 ms Runtime	13.4 ms Runtime

Results

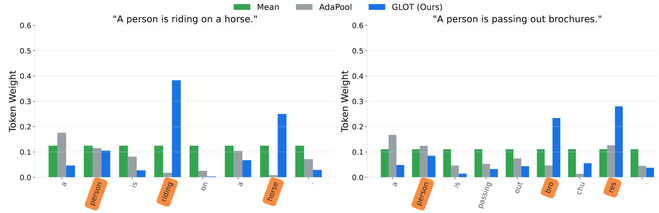
✓ Powerful sentence embeddings from frozen decoder-only LLMs like Mistral-7B



✓ GLOT sustains >97% accuracy even under extreme 90% signal dilution



Visualizing Token Contributions



Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021. URL https://arxiv.org/abs/2104.13478.