G-Loss: Graph-Guided Fine-Tuning of Language Models

Traditional loss functions (e.g., cross-entropy (CE), triplet and supervised contrastive loss (SCL) for fine-tuning pre-trained language models (e.g., `BERT`) follow a *two-stage* paradigm: unsupervised pre-training on large unlabeled corpora, then supervised fine-tuning with task-specific losses. While effective, these losses require substantial labeled data and compute, and mostly operate in the local neighborhood, ignoring global semantic structure. They optimize pairwise relations independently, so learning that A is close to B and distinct from C does not guarantee proper alignment of B and C, limiting generalization. We address this by shifting from local optimization to global structural alignment through a *graph-based fine-tuning strategy*, introducing *G-Loss*. *G-Loss* integrates the semi-supervised Label Propagation Algorithm (LPA)[2] into language model fine-tuning, diffusing labels from labeled to unlabeled nodes. This leverages both explicit supervision and implicit graph structure, consistent with the manifold assumption that proximity in feature space implies label similarity.

G-Loss introduces two key innovations: (1) dynamic mini-batch graphs built from evolving embeddings, enabling inductive learning with low memory cost, and (2) direct integration of LPA into the loss, eliminating separate pseudo-labeling steps while preserving embedding consistency. LPA's parameter-free nature ensures scalability across models (`BERT`, `RoBERTa`, `DistilBERT`) and benchmark datasets (`MR`, `R8`, `R52`, `20NG`, `Ohsumed`). Finally, our approach forms a self-reinforcing system, where improved embeddings reshape the graph, which in turn guides further refinement, yielding robust global alignment. For a multi-class classification problem, a document set $\mathcal{D} = \{d_1, \ldots, d_n\}$ spans $\mathcal{C}$ classes and is split into training, validation, and test sets. Our task is to fine-tune a pre-trained language model on training data and evaluate performance by classifying each test document into one of $\mathcal{C}$ classes. For each minibatch, the language model $\Phi(.)$ generates embeddings that are passed to both a linear classifier (logits) and a Gaussian similarity-based graph, where nodes represent documents and edge weights w_{ij} capture semantic similarity. A subset of labels is masked, and the Label Propagation Algorithm (LPA) infers them, yielding graph loss $\mathcal{L}_G$ as shown in below figure. This is combined with cross-entropy loss $\mathcal{L}_{CE}$ via a weighting factor λ, and hybrid loss backpropagates to update both the model and classifier. Iterative adaptation ensures that evolving embeddings maintain graph-consistent relationships, improving label propagation and classification accuracy.

Fine-tuning with *G-Loss* yields substantial improvements in accuracy over both standard cross-entropy-based fine-tuning of pre-trained models (BERT-base, RoBERTa-large) and graph-based baselines (TextGCN). With BERT-base, it consistently outperforms vanilla BERT across all datasets, achieving gains of up to $+1.84\%$ on MR and $+0.98\%$ on Ohsumed, underscoring the benefits of graph-driven structural supervision. When paired with RoBERTa-large, *G-Loss* surpasses the hybrid SOTA model BertGCN [1] on MR with $(1.12 + \%)$, R52 $(+0.5\%)$, and Ohsumed $(+2.96\%)$. Unlike BertGCN, which requires costly full data graph construction and Bert and GCN co-training, *G-Loss* attains comparable or superior accuracy with lightweight mini-batch dynamic graphs, ensuring scalability and inductive generalization.

Overall, these results show that *G-Loss* not only enhances language model fine-tuning but also effectively bridges graph-based and transformer-based paradigms in a more efficient manner.

Keywords: Fine-tuning, Label propagation, Semi-supervised learning, Text classification

References

[1] Yuxiao Lin et al. "BertGCN: Transductive Text Classification by Combining GCN and BERT." *CoRR*, abs/2105.05727, 2021.

[2] Xiaojin Zhu and Zoubin Ghahramani. "Learning from Labeled and Unlabeled Data with Label Propagation." 2002.