minBERT using PALs with Gradient Episodic Memory

Stanford CS224N Default Project

Christopher Nguyen
Department of Computer Science
Stanford University
cnguye29@stanford.edu

Abstract

The project aims to extend the minBERT model to allow for multi-task learning with the main purpose of improving or matching the performance to separately fine-tuned models. To perform multi-task learning, the model is adapted with Projected Attention Layers (PALs) and the training is adapted using gradient episodic memory (GEM) for a more robust and diversely applicable model.

1 Key Information

- Mentor: Arvind Mahankali

2 Approach

2.1 Baselines

Two baselines are used to evaluate the extended model. The first is the GLUE benchmark test results and corresponding accuracy scores presented by Stickland and Murray (2019) which contains tasks and datasets used in this project. The second baseline is the implemented minBERT model fine-tuned on individual tasks and comparing its separate task accuracy results to PALs and GEM.

2.2 Main Approach

The main approach is to retrofit the implemented minBERT model for PALs and GEM following the methods of Stickland and Murray (2019) and Lopez-Paz and Ranzato (2017) respectively. With their original open-source code as reference, I incorporated the PALs architecture and GEM algorithm into my pre-exisiting code.

2.2.1 Projected Attention Layers (PALs)

The objective of PALs is to incorporate a task-specific low-dimensional multi-head attention layer in parallel to each BERT layer. The proposed layer is given by a task-specific function (TS) of the form

$$TS(\mathbf{h}) = V^D(SA(V^E(\mathbf{h}))) \tag{1}$$

where V^E is a $d_s \times d_m$ encoder matrix and V^D is a $d_m \times d_s$ decoder matrix. The function takes the hidden states h to the BERT layer as input and V^E projects h to a lower dimension with a linear layer. V_E is transformed by multi-head self attention then decoded back to the original hidden size by the linear decoder layer. This allows the parameters, V_E and V_D, to be shared across layers not tasks. The parameters are added in parallel with the second residual connection ($\mathbf{h_{att}}$) and the feedforward network (FFN) before layer normalization(LN):

$$\mathbf{h}^{\ell+1} = LN(\mathbf{h_{att}}^{\ell} + FFN(\mathbf{h_{att}}) + TS(\mathbf{h})) \tag{2}$$

where ℓ indexes the layer.

2.2.2 Gradient Episodic Memory for Continual Learning (GEM)

Gradient Episodic Memory proposed by Lopez-Paz and Ranzato (2017) is a model that alleviates forgetting while allowing knowledge transfer to previous tasks. GEM maintains an episodic memory $\mathcal{M}_t$ that stores a subset of observed examples from each task t using integer task descriptors to index the episodic memory. The model utilizes predictors f_θ parameterized by memories from each task to ensure the loss ℓ at previous tasks don't increase after each update. This is done by using losses as inequality constraints formulated as the following:

$$\langle g, g_k \rangle := \langle \frac{\partial \ell(f_\theta(x,t),y)}{\partial \theta}, \frac{\partial \ell(f_\theta, \mathcal{M}_k)}{\partial \theta} \rangle \geq 0, \forall k < t \tag{3}$$

where the proposed parameter update is project to minimize violations in the constraints. Recovery of the projected gradient update which is biased towards beneficial backward transfer is done using Quadratic Programming. The primal problem is formulated on the proposed parameter update and the dual problem is based in terms of observed tasks.

3 Experiments

3.1 Data

Datasets are provided for the CS224n default project. Sentiment analysis uses the Stanford Sentiment Treebank, paraphrase detection uses Quora Question Pairs, and semantic textual similarity uses SemEval STS Benchmark dataset.

3.2 Evaluation Method

An evaluation metric to be used is the F1/accuracy score and Pearson correlation score for the STS dataset; and as stated in the baseline, I will be evaluating my extension of PALs against the score and accuracy obtained by Stickland and Murray (2019) and the accuracy of each task run individually on the base minBERT model.

3.3 Experimental Details

The experiment was run on the Google Console Platform deep learning VM instance with a Nvidia T4 GPU. Learning rate for pretraining was 1e-3 and finetuning was 1e-5 for both original minBERT and extended minBERT. Full training and testing takes approximately 40-50 minutes with an epoch of 10, steps per epoch of 1200, and batch size of 16.

3.4 Results

Table 1: Dev Results Comparison

Method	SST (dev)	QQP (dev)	STS-B (dev)	Av. (dev)
Fine-tuned BERT	0.515	0.780	0.384	0.726
PALs w/o GEM	0.515	0.754	0.329	0.645

These results are expected as given the results of Stickland and Murray (2019) where they saw roughly equal performance on the SST and QQP datasets. Also, my results differ as my training uses different hyperparameters such as the learning rate

4 Future work

My plan is to implement the GEM algorithm into the training and evaluate the change. Additionally, adjust hyperparameters for better performance.

References

David Lopez-Paz and Marc' Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR.