PowerBERT: Improving BERT with a Power Set Ensemble of Fine-Tuned Single and Multitask Models

Eric Lee, Jeanette Han, Kevin Song
Department of Computer Science, Stanford University

Problem

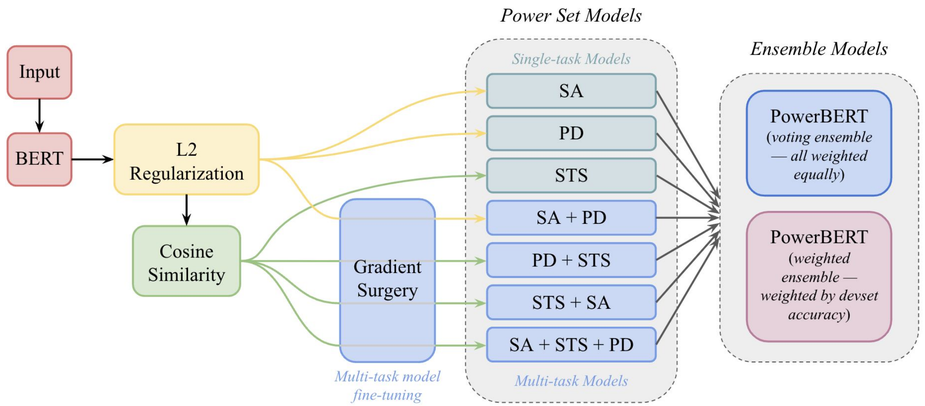
- Many recent proposals offer novel solutions to address limitations of large LMs.
- However, there is minimal research into the combination of such optimization techniques to further enhance model performance within an extensive ensemble to train intermediate combinations of tasks.
- We wanted to see if we could combine the beneficial effects of specialization in the single-task models with insights about complex interactions learned in the multitask models for more robust predictions.

Background

- RoBERTa (Liu et al. (2019)) provides a robust transformer-based model that introduces the idea of ensembling which yields better results due to reduction in variance/noise and better generalization.
- Recall that the **power set** is the set of all subsets. The power set of the set of tasks, excluding the empty set $\emptyset$: {*STS*}, {*PD*}, {*SA*}, {*STS* + *PD*}, {*PD* + *SA*}, {*SA* + *STS*}, {*STS* + *PD* + *SA*}.
- We test a novel approach to ensembling by creating an individual model for each member of the power set of tasks and ensembling their results with PowerBERT.
- We tested 2 ensembling schema:
  - **Voting**: logits for each model relevant to the given task are evenly averaged (arithmetic mean)
  - **Weighted**: logits from models with higher accuracy are assigned greater weights.

Methods

- To maximize PowerBERT's performance, we implemented extensions:
  - **L2 Regularization** - summing the squares of all n feature weights
  - **Cosine-Similarity Fine Tuning** - cosine similarity used as metric for semantic similarity comparing embeddings: $\frac{u \cdot v}{||u||||v||}$
  - **Gradient Surgery** - eliminates gradient conflicts between tasks during training: $g_i = g_i - \frac{g_i \cdot g_j}{||g_j||^2} * g_j$



Experiments

- All models were trained using an Adam optimizer with **weight decay 1e-4, L2 regularization,** and a **learning rate 1e-5**. For models trained for the STS task, **cosine similarity fine-tuning** was implemented. For all multitask training models, **gradient surgery** was applied after each alternating loop.
- After creating the 7 individual models, each for a different subset of tasks, they were loaded into the ensemble model. For each task in the **voting ensemble** the logits from each relevant individual model **each had weight ¼.** For each task in the **weighted ensemble**, the logits from the relevant models each had **weight proportional to that model's accuracy on the development set.** The final test results are shown below:

Model	Sentiment Analysis	Paraphrase	STS	Overall
Single-Model Ensemble	0.524	0.858	0.796	0.760
Voting PowerBERT Ensemble	0.527	**0.872**	**0.809**	0.768
Weighted PowerBERT Ensemble	**0.534**	**0.872**	0.806	**0.770**

Analysis

- **87%** of PowerBERT's errors deviated by only one rating value away from the correct labels for **SST** development set.
- The model accurately identifies general sentiment (positive/negative) but struggled to identify strong sentiment. This was most common for very negative reviews with advanced semantics (idioms, sarcasm).
- Some dataset labels may be ambiguously determined — that is, in some cases, our model's classification mismatched the dataset's, but both were subjectively plausible.
- For **STS**, PowerBERT generally predicted scores ±1 accurate rating but struggled with lower scores.
  - Our model tends to assign higher similarity scores when sentences have the same word tokens, even if their meaning and use is different. Our model might rely too heavily on token similarities rather than evaluating semantic embeddings.

Conclusion

- PowerBERT was able to achieve **significant improvements** through our exploration of advanced fine-tuning techniques (Cosine Similarity Fine-Tuning, Gradient Surgery, L2 Regularization) and a novel power set ensemble structure.
- Our final model generated competitive results, yielding an overall **test set result of 0.77** and an **STS score** matching the original Reimers and Gurevych (2019) paper.
- Our ensemble model successfully combined the strengths and mitigated the weakness of individual models in the **power set** to create a robust model demonstrating proficiency on multiple tasks.