Compression of NLP-Domain Deep Learning Models

Yan Lashchev, Jorge Murillo, Nathan Roll, Lawrence Su, and Yangyi Zhang (Mentor: Erika McPhillips)

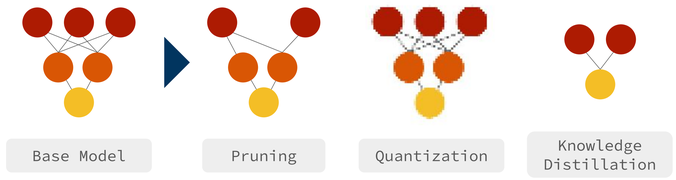
{ydl, jorgemurillo, nroll, lawrencesu, yangyi}@ucsb.edu

UC SANTA BARBARA | pwc

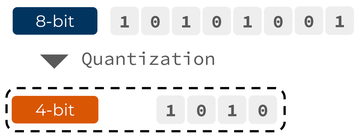
Abstract

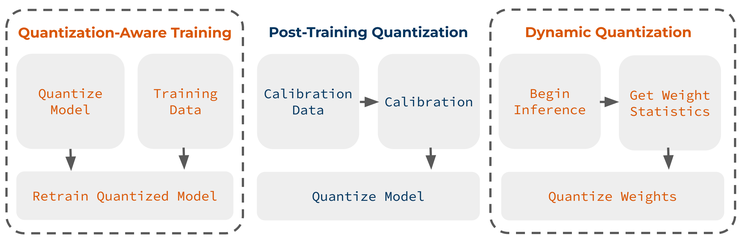
The field of model compression has enjoyed many advancements in recent years, yet few reliable methods have been developed specifically for the natural language processing (NLP) domain. In this presentation, we showcase a survey on model compression techniques and implement custom compression methods on an emotion classification task.

Compression



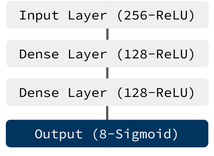
Compressed models run **faster**, **cheaper**, and are more **accessible** than their uncompressed counterparts, often at cost of some level of accuracy. There are three general ways to accomplish model compression: strategically remove unneeded components, reduce the bit representation of the model's parameters, or approximate the base model with a less complex version. These methods are known as pruning, quantization, and knowledge distillation respectively. Quantization can be further broken down into three main types:



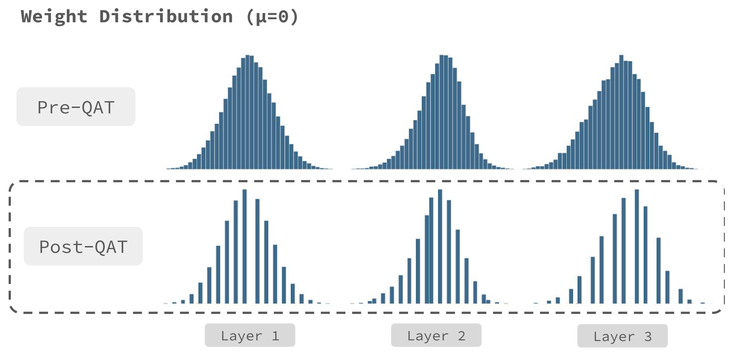


Model 1 Setting the Baseline

Utilizing the GoEmotions[1] dataset and GloVe pre-trained embeddings to create mean comment vectors, we trained a **dense sequential ANN** with 10% interlayer dropout and eight output neurons. The multilabel nature of the task lends itself to a final **sigmoid activation**.



Quantization-Aware Training (QAT)

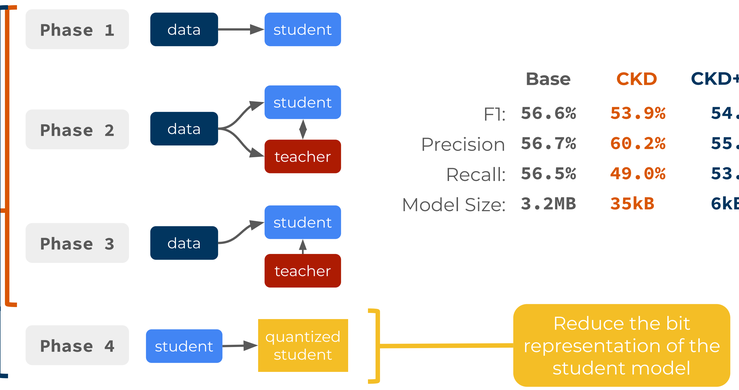


Cluster preserving quantization-aware training (CQAT) was performed on model 1, which resulted in a **75.4% size reduction** and minimal loss in precision. The model's recall actually increased, and with it the overall F1-score.

F1:	46.3% → 46.9%
Precision:	63.0% → 61.5%
Recall:	36.6% → 38.0%
Model Size:	589kB → 145kB

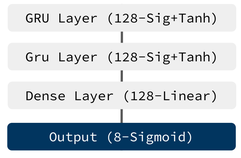
Implementation of CKD

Implementation of CKD with APOT quantization



Model 2 Adding Recurrent Layers

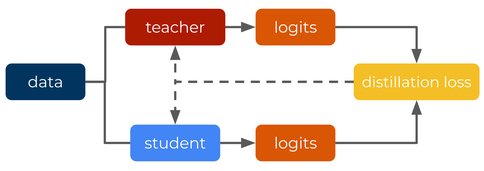
Mean comment vectors used in the base model were replaced with padded sequences of 27 word vectors. The first two dense layers were replaced with **GRU layers**. Interlayer dropout was preserved to help prevent overfitting.



Standard Knowledge Distillation (KD)

A less complex model was trained using the same inputs, but using the predictions as labels. This model consistently outperformed a similarly complex model trained on the raw data, yet (by design) slightly underperforms the non-distilled version.

Use a more complex teacher model to train a smaller student model.



F1:	54.0% → 52.6%
Precision:	63.6% → 68.3%
Recall:	46.9% → 42.8%
Model Size:	1.2MB → 242kB

Additive Powers of Two (APOT) + Collaborative KD

As you can see from the figure to the left, weight distributions are generally normal. Using powers of two we represent a **nonuniform quantization** of weights with higher precision around the middle of the bell curve by specifying smaller quantization **levels** around the median. By sacrificing minimal memory we can **calculate higher precision for the majority of the weights**. Furthermore, we can combine this with Knowledge Distillation for an optimized model using both methods. Our teacher model has two GRU layers with hidden size 256. Our student model is has one GRU layer with hidden size 8.

	Base	CKD	CKD+APOT
F1:	56.6%	53.9%	54.4%
Precision	56.7%	60.2%	55.4%
Recall:	56.5%	49.0%	53.3%
Model Size:	3.2MB	35kB	6kB

Looking Ahead

We want to optimally compress our GRU-RNN using APOT and KD. Yet a current limitation of our methods is that we are not quantizing the activations, nor optimizing weight clipping thresholds for dynamic Apot quantization. Our next steps are to **optimize the clipping thresholds** for the weights, **make the student quantization aware**, and **insert the quantization aware student into Phase 2 and Phase 3**. Our phases then align with those of QKD, and we will be able to more fully optimize compressing the model using APOT.

Acknowledgements

Rob Bernard, Ilana Golbin, Hannah Moran, Dr. Arit Kumar Bishwas & PwC.

References

[1] Dorottya Demszky et al. GoEmotions: A Dataset of Fine-Grained Emotions, 2020.
[2] John Frankle, Michael Carbin, The Lottery Ticket Hypothesis: Finding Sparse Trainable Neural Networks, 2019.
[3] Amir Gholami et al. A Survey of Quantization Methods for Efficient Neural Network Inference, 2021.
[4] Jangho Kim et al. QKD: Quantization-aware Knowledge Distillation, 2019.
[5] Yuhang Li et al. APOT: An Efficient Non-Uniform Discretization for Neural Networks, 2020.