Transformer to CNN: Label-scarce distillation for efficient text classification

RED DRAGON AI

Yew Ken Chia
ken@RedDragon.ai

Sam Witteveen
sam@RedDragon.ai

Martin Andrews
martin@RedDragon.ai

Summary

Task :

▲ Efficient text classification

Industry Constraints :

▲ Inference speed/latency
▲ Memory footprint
▲ Lack of labelled data

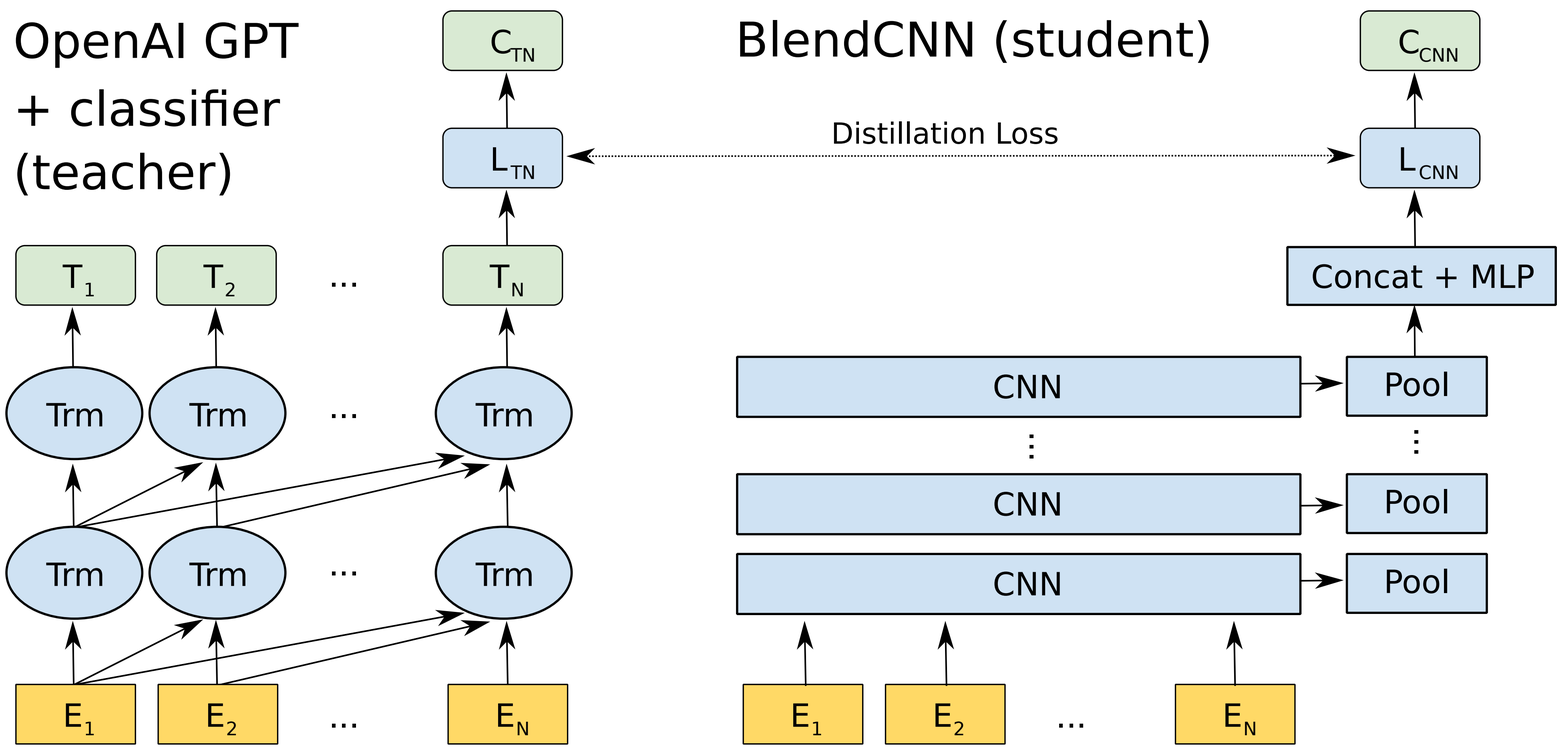
Ideas :

▲ Train SOTA Transformer-based model
▲ Benefit from including unlabelled data
▲ Develop high-efficiency CNN student

Results :

▲ Same accuracy as Transformer model
▲ 39x fewer parameters
▲ 300x faster inference speed

Teacher / Student Model Architecture



Results

	AG News		DBpedia		Yahoo Answers	
TRAINED ON 100 LABELLED EXAMPLES PER CLASS						
TFIDF + SVM	81.9		94.1		54.5	
fastText	75.2		91.0		44.9	
8-Layer BlendCNN	87.6		94.6		58.3	
OpenAI Transformer	88.7		97.5		70.4	
TRAINED BY DISTILLATION[1] OF OPENAI TRANSFORMER						
2-Layer BiLSTM	91.2		97.0		70.5	
KimCNN	90.9		97.6		70.4	
3-Layer BlendCNN	91.2	/ 88.4[2]	98.2	/ 95.5	71.0	/ 63.4
8-Layer BlendCNN	91.2	/ 89.9	98.5	/ 96.0	70.8	/ 63.4

[1] Distillation training used 100 labelled examples per class, plus 10 times as many unlabelled examples as pseudo-labelled by the OpenAI LM
[2] Small figures are results where distillation was conducted without unlabelled data

Parameter counts and inference timing

	Total parameters[2]	Sentences per second[3]
2-Layer BiLSTM[1]	2,406,114	173.01
KimCNN	2,124,824	3154.57
OpenAI Transformer	116,534,790	11.76
8-layer BlendCNN	3,617,426	2392.34
3-layer BlendCNN	2,975,236	3676.47

Discussion

Motivation :

▲ Low number of labelled examples
▲ Newer techniques include fine-tuning
▲ Large LM inference cost is prohibitive

Initial Training of Teacher :

▲ Leverage pre-trained LM
▲ Fine-tune Transformer classifier with labelled data

Student Features :

▲ Pure CNN architecture
▲ Hierarchical outputs from CNN
▲ Pseudo distillation (incl. Unlabelled) > Labelled-only distillation
> Labelled-only training from scratch

Future directions :

▲ Explore distillation on other NLP tasks
▲ Test benefits of hierarchical outputs
▲ Update to use BERT as teacher

Source code available:

▲ http://RedDragon.ai/research

Key References

"Distilling the knowledge in a neural network." - Hinton et al. (2015)
"Deep contextualized word representations" - Peters et al. (2018)
"Improving language understanding with unsupervised learning" - Radford et al. (2018)

Contact

martin@RedDragon.ai
+65 8585 1750
http://RedDragon.ai