Efficient Machine Translation with Model Pruning and Quantization
Edinburgh's Submission to WMT21 Efficiency Shared Task

Maximiliana Behnke[†], Nikolay Bogoychev[†], Alham Fikri Aji[†], Kenneth Heafield[†], Graeme Nail[†], Qianqian Zhu[†], Svetlana Tchistiakova[†], Jelmer van der Linde[†], Pinzhen Chen[†], Sidharth Kashyap[‡], Roman Grundkiewicz[†§]

[†]University of Edinburgh, [‡]Intel Corporation, [§]Microsoft

Teachers

Edinburgh's constrained English-German system from the WMT21 news task:

- Parallel data cleaning
- Back-translation
- Fine-tuning
- Ensembling

Students

Speed- and size-optimized models trained on teacher's data:

- Knowledge distillation
- SSRU decoder
- Shortlisting
- Structural pruning
- Quantisation (8bit, 4bit)
- Fine-tuning

	Depth		Dimensions			
Model	Enc	Dec	Emb.	FFN	Att.	Heads
teacher x 3	6	6/6/8	1024	4096	1024	16
12-1.large	12	1	1024	3072	256	8
12-1.base	12	1	512	2048	256	8
12-1.tiny	12	1	256	1536	256	8
12-1.micro	12	1	256	1024	256	8
8-4.tied.tiny	8	4	256	1536	256	8
6-2.tied.tiny	6	2	256	1536	256	8
6-2.base	6	2	512	2048	512	8
6-2.tiny	6	2	256	1536	256	8

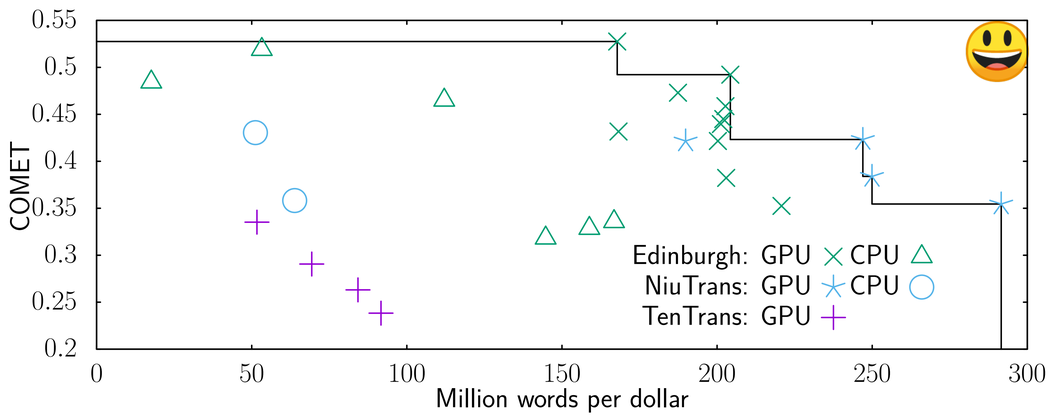

Figure: Pareto trade-off between quality and million words translated per dollar.

Structural pruning

We structurally pruned our transformer student models using group lasso. We focused on pruning encoder only. Our most aggressive pruning removed parameters from both feedforward and attention layers.

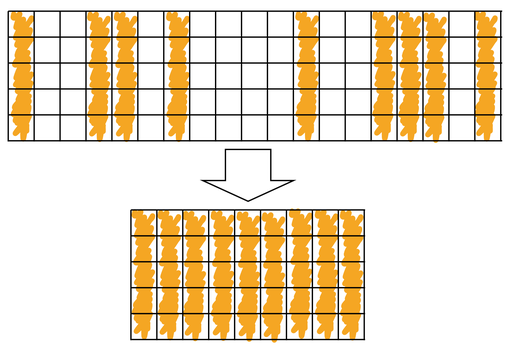

Figure: Structural pruning of nodes in FFN layers.

8-bit quantisation

We applied 8-bit quantisation to our pruned models to speed-up inference. We fine-tuned the models to recover some quality lost due to lower precision.

	BLEU	COMET	Sparsity		
	WMT21	WMT21	Att.	FFN	Speed (s)
12-1.tiny	27.6	41.9	0%	0%	19.2
+ prune	27.0	38.8	3%	75%	14.5
+ 8bit	26.2	33.6	3%	75%	9.3
+ ft	26.7	33.0	3%	75%	9.3
12-1.micro	27.6	40.2	0%	0%	17.1
+ prune	26.4	35.1	60%	59%	12.0
+ 8bit	25.5	29.1	60%	59%	7.5
+ ft	25.9	30.5	60%	59%	7.5

Table: A performance of pruned and quantised models with additional fine-tuning.

4-bit quantisation

We quantised our models into 4-bits to compress the model sizes. The inference is still in fp32.

	BLEU WMT21	COMET WMT21	Size
12-1.base	28.3	45.1	195MB
+ 4bit	27.7	43.2	25MB
12-1.tiny	28.0	42.5	85MB
+ 4bit	27.6	38.3	11MB
8-4.tied.tiny	27.5	43.6	69MB
+ 4bit	26.4	38.2	9MB

Table: 4-bit model performance, fine-tuned and run on fp32.

GPU specific work

For our GPU submission we used experimental patch set from nvidia that improves inference performance. On top of that we tried to take advantage of tensor cores using CUTLASS 8-bit integer GEMM, but the overall results were not better than plain fp16 decoding.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (grant EP/L01503X/1), EPSRC Centre for Doctoral Training in Pervasive Parallelism at the University of Edinburgh, School of Informatics.