**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

---

**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.

**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.

---

**4-bit quantisation**

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