CNN for sequences: speed benefit

- They work faster than RNN:
  - During **training** we can process all time steps in parallel
  - During **testing** encoder can do the same
  - During **testing** we get higher throughput thanks to convolution optimizations in GPUs

	BLEU	Time (s)
GNMT GPU (K80)	31.20	3,028
GNMT CPU 88 cores	31.20	1,322
GNMT TPU	31.21	384
ConvS2S GPU (K40) $b = 1$	33.45	327
ConvS2S GPU (M40) $b = 1$	33.45	221
ConvS2S GPU (GTX-1080ti) $b = 1$	33.45	142
ConvS2S CPU 48 cores $b = 1$	33.45	142

Translation generation speed during testing

https://arxiv.org/pdf/1705.03122.pdf