Multi-representation ensembles and delayed SGD updates improve syntax-based NMT

UNIVERSITY OF CAMBRIDGE

Danielle Saunders[†] Felix Stahlberg[†] Adrià de Gispert[‡†] Bill Byrne[‡†]

[†]Department of Engineering, University of Cambridge, UK [‡]SDL Research, Cambridge, UK

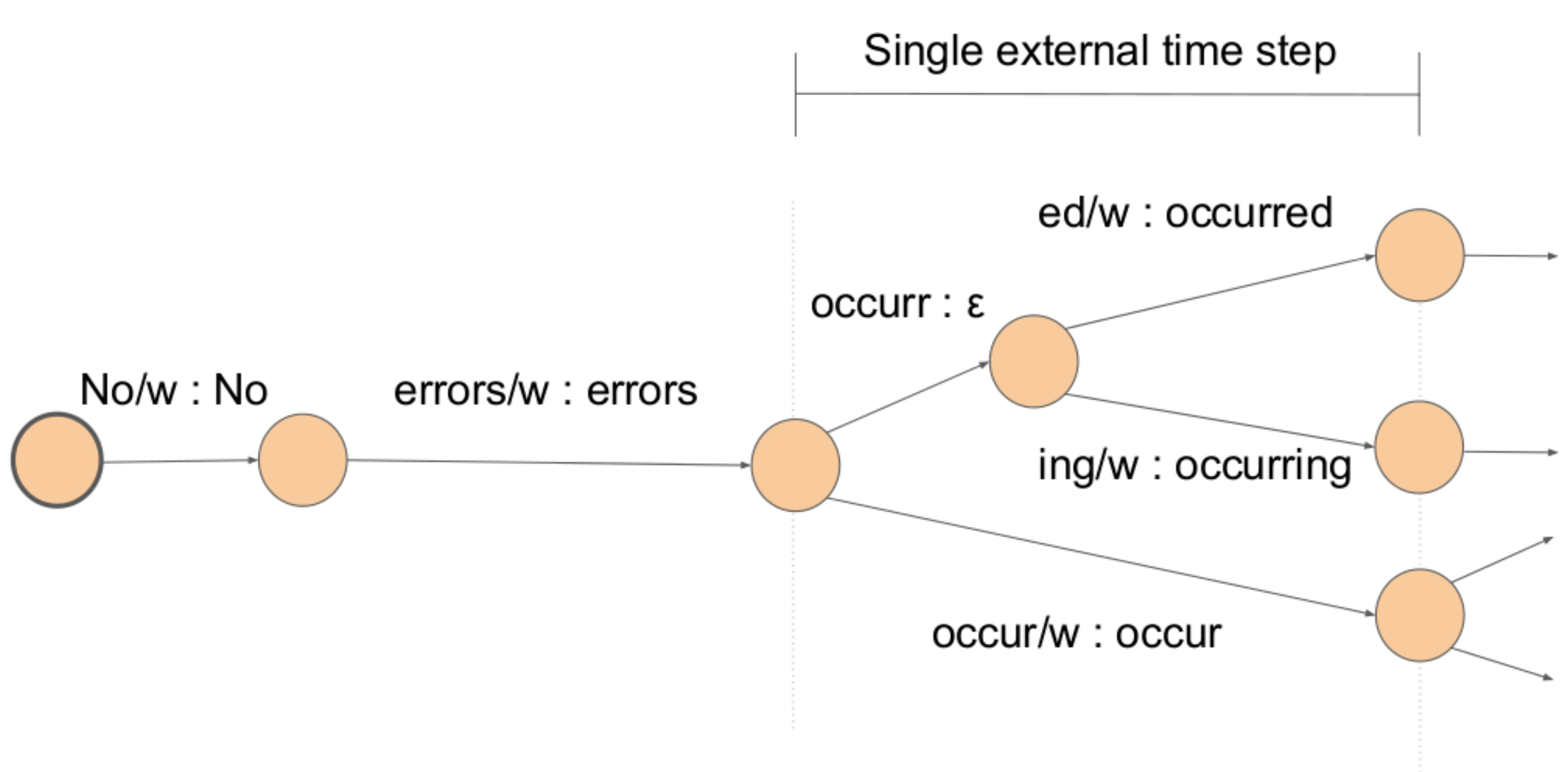
Multi-representation ensembling with FSTs

- Problem: ensemble models with different target representations which may not be synchronized, e.g.:

Words	No errors occurred
Subwords	No/w errors/w occur ed/w
POS/plain	DT No NNS errors VBD occurred
Derivation	S/R NP VP/R DT NNS/R No errors VBD/R occurred
Tree	(S (NP (DT No) (NNS errors)) (VP (VBD occurred)))

- Use FSTs for a synchronized search over two representations such that paths $p \in \mathcal{P}$ through the FST map between representations:

$$i(p) \rightarrow o(p)$$



- Accumulate scores at the path level via a 2-level beam search
- An ideal equal-weight ensembling of two models P_i and P_o yields:

$$p^* = \text{argmax}_{p \in \mathcal{P}} \; P_i(i(p))\, P_o(o(p))$$

with $o(p^*)$ as the external representation of the translation.

- Partial hypothesis in $o(p)$, $h = h_1 \dots h_j$, has current token score:

$$P(h_j|h_{<j}) = P_o(h_j|h_{<j}) \times \max_{(x,y)\in M(h)} P_i(i(y)|i(x))$$

Where set of partial paths yielding h is given by:

$$M(h) = \{(x,y)|xyz \in \mathcal{P}, o(x) = h_{<j}, o(xy) = h\}$$

- Implementation: `https://github.com/ucam-smt/sgnmt`

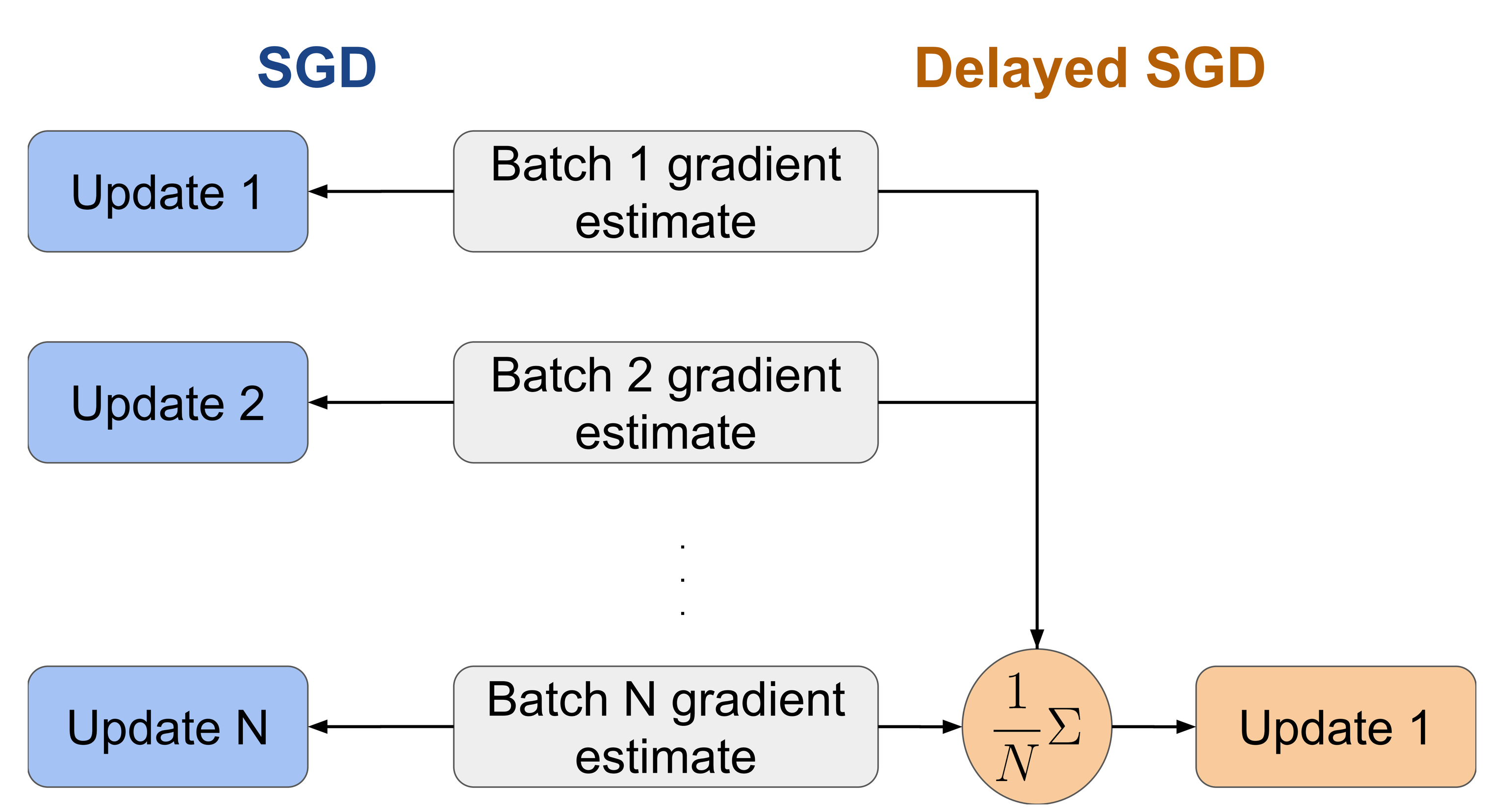
Delayed SGD updates

- Gradients for NMT training updates usually estimated every batch
- Long sequences (e.g. syntax representations) mean fewer sequences per batch: could cause noisier updates

Representation	Mean length
Plain subwords (BPE)	27.5
POS/plain	53.3
Derivation	73.8
Tree	120

Lengths for representations from first 1M training sentences of English ASPEC

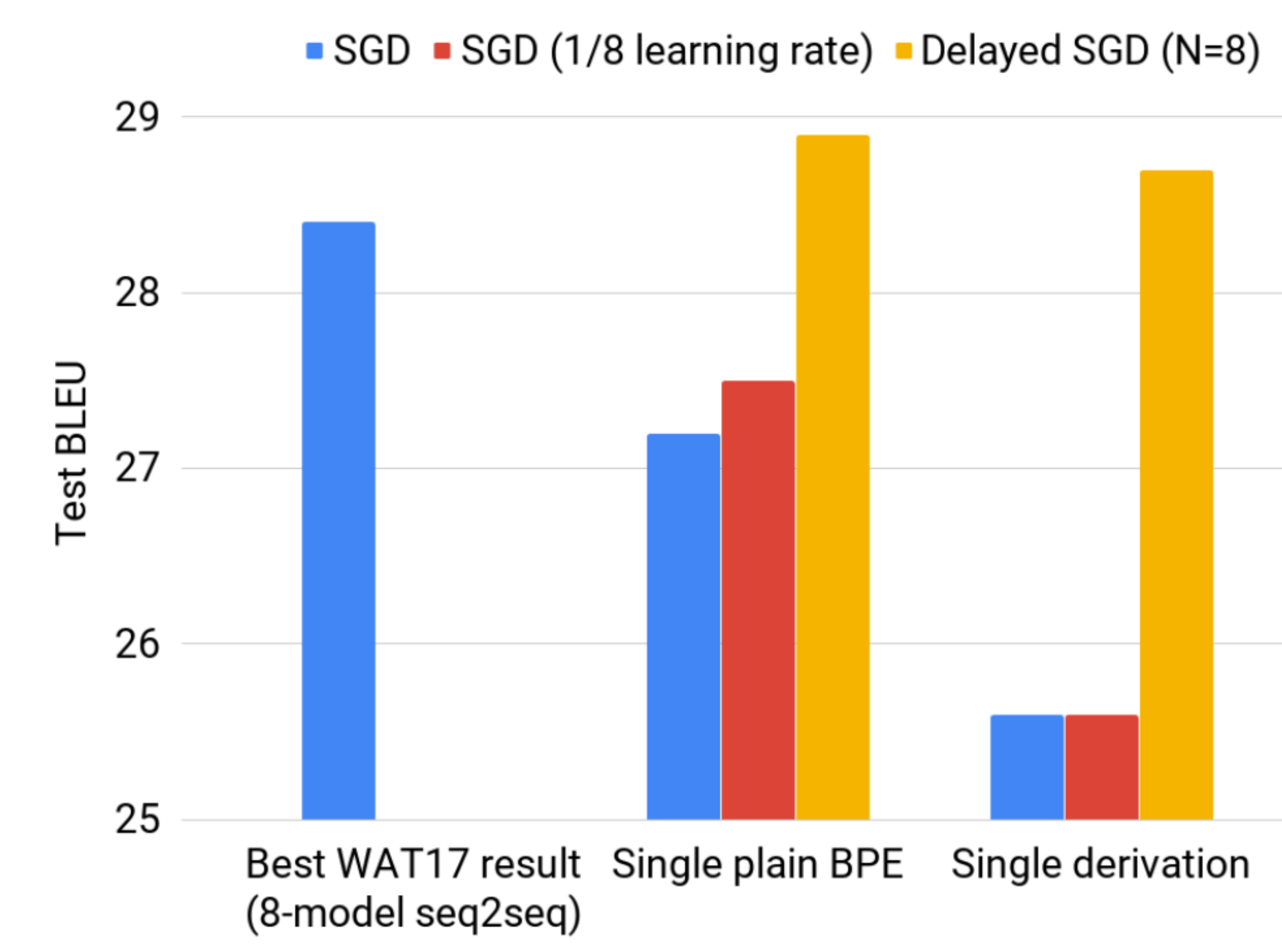
- Delayed SGD accumulates estimates over several batches per update on one GPU
- Decouples maximum batch size from available memory / GPUs
- Implementation: `multistep_optimizer` in `https://github.com/tensorflow/tensor2tensor`



Experiments

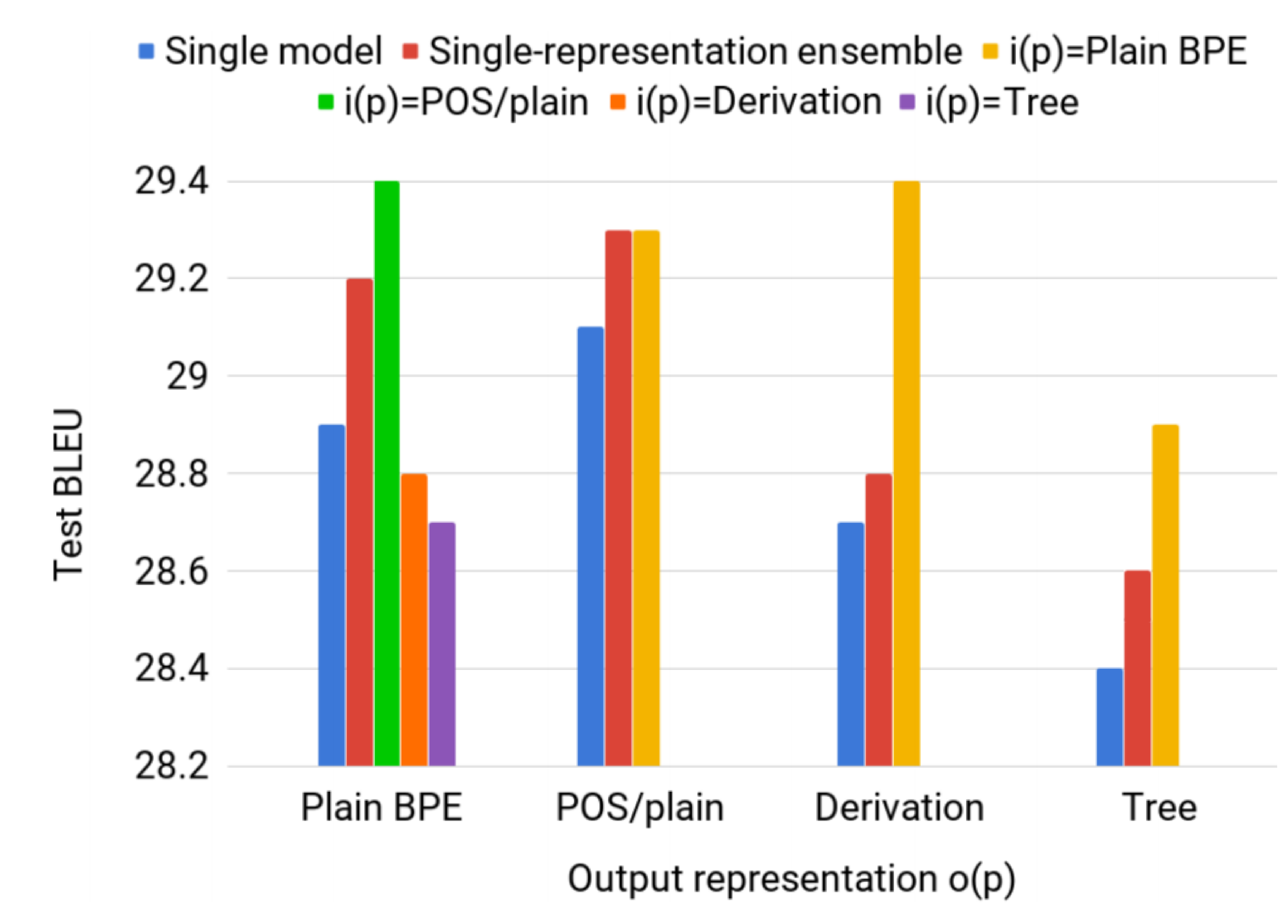
- All models trained with the first 1M sentences of ASPEC Ja-En
- Source and target sentences use BPE (30K vocab)
- All models use the Tensor2Tensor Transformer architecture
- All ensembles contain two models

Delayed SGD improves long representations



- Syntax performance severely lags plain BPE without delayed SGD
- Reduced learning rate alone does not provide the same gains

Gains from multi-representation ensembles



- Denser syntax representations have better single model performance
- Choice of internal / external representation affects result

Acknowledgements

This work was supported by EPSRC grant EP/L027623/1.
Contact: {ds636,fs439,ad465,wjb31}@cam.ac.uk