Logit-Based Ensemble Distribution Distillation for Robust Autoregressive Sequence Uncertainties (Supplementary Material)

Yassir Fathullah[1] **Guoxuan Xia**[2] **Mark J. F. Gales**[1]

[1]Engineering Department, University of Cambridge, UK
[2]Department of Electrical & Electronic Engineering , Imperial College London, UK

A EXPERIMENTAL CONFIGURATION

This section will provide detailed information about the datasets used for training, development, evaluation and detection. It will also give the exact training and various hyperparameters used for all models.

A.1 DATASETS

We utilise two training sets WMT16/20, each with a pair of development and evaluation datasets based on newstest13/14 and newstest19/20. Additionally, we utilise three out-of-domain datasets for evaluating detection performance of a wide range of transformer models, see Table 1. As stated previously, all data is cleaned and tokenized using Moses[1]. For WMT16, a shared dictionary is learned using BPE with 32,000 merge operations. On WMT20 we learn disjoint dictionaries using BPE with 40,000 merge operations. A consequence of the larger disjoint dictionary on WMT20 is the significantly lower number of unknown tokens in the OOD datasets.

Table 1: Dataset information together with average source and target sentence sizes post tokenization and processing. The OOD testsets Khresmoi, MTNT and KFTT have two quoted numbers for each field as they were processed using either the En-De WMT16 or En-Ru WMT20 BPE based dictionaries. Additionally, only source side information is provided for OOD sets as these are only used for unsupervised uncertainty estimation.

Dataset	Type	Number of Sentences	Tokens per Sentence		Fraction of Unknown Tokens in Source
			Source	Target	
En-De WMT16	policy, news, web	4.5M	29.5	30.6	0.01%
En-De newstest13	news	3.0K	26.0	28.0	0.00%
En-De newstest14		3.0K	27.6	29.1	0.00%
En-Ru WMT20	policy, news, web	58.4M	27.8	27.5	0.00%
En-Ru newstest19	news	2.0K	29.9	33.4	0.00%
En-Ru newstest20		2.0K	30.9	32.5	0.00%
Khresmoi	medical	1.0K	30.9/30.3	—	0.78%/0.00%
MTNT	noisy reddit	1.4K	21.1/21.3	—	0.45%/0.06%
KFTT	encyclopedia	1.2K	35.4/35.2	—	1.46%/0.01%

[1]`github.com/moses-smt/mosesdecoder`

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

A.2 EN-DE WMT16 TRAINING

We use the base transformer from [Vaswani et al., 2017] implemented in `fairseq` [Ott et al., 2019] and train it using 4 NVIDIA© A100 with an update frequency of 32. This is virtually equivalent to training on $4 \times 32 = 128$ GPUs. A per-gpu batch has a maximum of 3584 tokens. Models are optimized with Adam [Kingma and Ba, 2015] using β_1 = 0.9, β_2 = 0.98, and ϵ= 1e-8. We use a similar learning rate schedule to Vaswani et al. [2017], i.e., the learning rate increases linearly for 4000 warmup steps to a learning rate dependent on $d_{\texttt{model}}$ after which it is decayed proportionally to the inverse square root of the number of steps:

$$\eta = (\texttt{step} \cdot d_{\texttt{model}})^{-0.5} \min\left(1, \frac{\texttt{step}}{\texttt{warmup}}\right)^{1.5}$$

We use label smoothing with 0.1 weight for the uniform prior distribution over the vocabulary. The last 10 weight checkpoints were averaged. Training was stopped after 31 epochs corresponding to approximately a total of 18 GPU-hours. At inference, a beam of 4 with a length-penalty of 0.6 is used for all models. The Deep Ensemble consists of 5 of such models.

KD of Deep Ensemble: Knowledge distilled models are first initialised by one of the teacher members and then trained using the knowledge distillation loss $\mathcal{L}_{\texttt{KD}}$ provided in Section 2.2 with $\lambda = 0.50$. The student was trained with a warmup of 1026 steps (3 epochs), from $\eta = 4.0 \times 10^{-4}$ to $\eta = 7.0 \times 10^{-4}$ after which it decays for a total of 24 epochs. A temperature of $T = 0.8$ was used in the KL-divergence loss as this was found to be mildly beneficial. All other hyperparameters match the standard case above.

Snapshot Ensemble: The Snapshot Ensemble was generated by first starting from the last checkpoint of a standard trained transformer. At this point, a cyclic triangular learning rate schedule [Smith, 2017] was employed oscillating between the values of $\eta_{\texttt{min}} = 1.0 \times 10^{-4}$ and $\eta_{\texttt{max}} = 1.0 \times 10^{-3}$ with a period of 3 epochs. Note that the maximum learning rate in this cyclic phase is notably larger than the peak learning rate (7.0×10^{-4}) during standard training This setting was run for 15 epochs generating an ensemble with 5 members.

KD of Snapshot Ensemble: This system was trained using the same parameters as the Deep Ensemble distilled students but was however, trained for only 12 epochs since it converged faster.

EDD & L-EDD: All of the EDD and L-EDD systems were distribution distilled from the Snapshot Ensemble using the same setup as "KD of Snapshot Ensemble". We chose $\beta = 0.10$ by evaluating the translation performance of a range of values $\beta \in \{0.05, 0.10, 0.20, 0.50\}$ on the development newstest-13 set, see Section 3.1.

A.3 EN-RU WMT20 TRAINING

We use the big transformer from Vaswani et al. [2017] again implemented in `fairseq` and trained using 4 NVIDIA© A100 with an update frequency of 32. A per-gpu batch has a maximum of 5120 tokens. Dropout was set to a value of 0.10 and weight decay to 0.0001. In this case we train the model for 20 epochs, corresponding to 53960 update steps and approximately 230 GPU-hours. The last 5 checkpoints were averaged leading to improved performance. At inference, a beam of 5 with a length-penalty of 1.0 is used for all models.

Snapshot Ensemble: Based on the last checkpoint of a standard trained big transformer, a triangular cyclic learning rate is utilised, oscillating between $\eta = 5.0 \times 10^{-5}$ and $\eta = 5.0 \times 10^{-4}$ every 2 epochs for 10 epochs. This results in an ensemble with 5 members.

KD of Snapshot Ensemble: Similar to the previous section, the distillation student is initialised from its teacher but is trained using a learning rate warmup of 2698 steps (one epoch) from $\eta = 2.0 \times 10^{-4}$ to $\eta = 4.0 \times 10^{-4}$ after which it decays for a total of 12 epochs. The last 3 or 5 epochs are averaged, based on development newstest19 performance.

L-EDD: Following distillation, L-EDD (Laplace) models are trained using the same parameters. The best-found parameter $\beta = 0.10$ in the WMT'16 experiments is to be used here. No hyperparameter search is performed at this stage.

B ABLATION STUDY: ENSEMBLE SIZE

Following the experimental setup in Section 5.1, we perform out-of-distribution detection of both Deep and Snapshot Ensembles of increasing size, see Table 2. This shows that increasing the size of an ensemble, regardless of its nature, does not improve its out-of-distribution detection notably.

Table 2: OOD detection performance (%AUROC ↑) for base transformer with ID dataset newtest-14 and OOD dataset Khresmoi. **Bold** indicates best in a column, underline second best.

Ensemble Size	Deep Ensemble		Snapshot Ensemble	
	TU	KU	TU	KU
2	48.3	61.5	48.7	61.5
3	48.2	61.7	48.8	61.9
5	48.0	61.9	49.0	62.6
7	48.0	61.9	49.0	62.6
10	48.0	62.7	49.1	62.6

References

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Leslie N. Smith. Cyclical learning rates for training neural networks. In *Winter Conference on Applications of Computer Vision*, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems*, 2017.