Logit-Based Ensemble Distribution Distillation for Robust Autoregressive Sequence Uncertainties

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

Abstract

Efficiently and reliably estimating uncertainty is an important objective in deep learning. It is especially pertinent to autoregressive sequence tasks, where training and inference costs are typically very high. However, existing research has predominantly focused on tasks with static data such as image classification. In this work, we investigate Ensemble Distribution Distillation (EDD) applied to large-scale natural language sequence-to-sequence data. EDD aims to compress the superior uncertainty performance of an expensive (teacher) ensemble into a cheaper (student) single model. Importantly, the ability to separate knowledge (epistemic) and data (aleatoric) uncertainty is retained. Existing probability-space approaches to EDD, however, are difficult to scale to large vocabularies. We show, for modern transformer architectures on large-scale translation tasks, that modelling the ensemble *logits*, instead of softmax probabilities, leads to significantly better students. Moreover, the students surprisingly even *outperform Deep Ensembles* by up to $\sim$10% AUROC on out-of-distribution detection, whilst matching them at in-distribution translation.

1 INTRODUCTION

The ability to produce reliable estimates of uncertainty is important to many tasks in deep learning. When it is costly to make mistakes, a model should know when to discard a prediction or defer it to a human expert. Although there is a significant body of research in this area [Ovadia et al., 2019, Hullermeier and Waegeman, 2021, Yang et al., 2021], it tends to focus on *static* data, where outputs are of a fixed dimension, with approaches most commonly evaluated on image classification [Geifman and El-Yaniv, 2019, Liu et al., 2020, Yang et al., 2022, Moon et al., 2020, Xia and Bouganis, 2022b]. In contrast, in this work, we aim to investigate uncertainty estimation for sequence prediction tasks, such as machine translation, which is a relatively under-explored domain [Malinin and Gales, 2021].

Large attention-based autoregressive neural networks have recently emerged as the most competitive approach to many structured sequence-prediction tasks, especially in translation [Bahdanau et al., 2015, Vaswani et al., 2017, Ott et al., 2018], and are increasingly being used in practice. However, as the computational and memory costs of these modern approaches are typically very large, it is particularly important that approaches for improving the quality of uncertainties should be *efficient*. We will focus on one such efficient approach to better uncertainties, Ensemble Distribution Distillation (EDD) [Malinin et al., 2020].

Ensembling multiple neural networks trained using different random seeds is a well-established approach for boosting uncertainty performance [Lakshminarayanan et al., 2017]. Deep Ensembles have been shown to be effective over a wide range of data, tasks, and evaluation metrics [Ovadia et al., 2019, Malinin and Gales, 2021, Kim et al., 2021, Gustafsson et al., 2020]. Moreover, they are naturally able to decompose total uncertainty into knowledge (epistemic) and data (aleatoric) uncertainty [Hullermeier and Waegeman, 2021], which can be useful for different tasks such as active learning [Gal et al., 2017, Radmard et al., 2021], reinforcement learning [Depeweg et al., 2018], and out-of-distribution detection [Malinin and Gales, 2021]. However, Deep Ensembles suffer from costs that scale linearly with the number of members. EDD aims to tackle this by using Knowledge Distillation (KD) [Hinton et al., 2014] to compress the (teacher) ensemble into a more efficient (student) single model. Crucially, EDD not only has the student learn the predictions of the ensemble but also the *distribution* over individual ensemble member outputs. By explicitly modelling the diversity over the ensemble, the student is able to express knowledge and data uncertainty independently just like the teacher ensemble [Malinin et al., 2020].

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

However, EDD is not without its challenges. Prior work has shown that EDD suffers from optimisation issues, meaning it can be difficult to scale to confident ensembles with large label spaces. Thus, EDD requires a number of practical modifications in order to be applied to large-scale tasks such as machine translation [Fathullah et al., 2021, Ryabinin et al., 2021]. Despite these challenges, the concept behind EDD remains a promising approach for training single autoregressive models with smaller footprints and the ability to estimate high-quality, robust uncertainties.

Summary of contributions: In this paper, we focus on an underexplored area of uncertainty estimation: robust and efficient autoregressive sequence uncertainties. Specifically, we address the drawbacks of sequence EDD by using *logit-based* ensemble distribution distillation (L-EDD). Instead of training a student to distribution distil the information from an ensemble in the probability/softmax space, we teach it to perform the same task in the pre-softmax logit space. Experiments on the En-De WMT'16 and En-Ru WMT'20 machine translation tasks show that L-EDD, in particular when using a Laplace distribution, produces strong estimates of sequence uncertainty. L-EDD is able to outperform EDD and surprisingly even Deep Ensembles on out-of-distribution (OOD) detection and match them for translation quality. Furthermore, by using Snapshot Ensembles [Huang et al., 2017], we are able to greatly reduce the overall training costs of EDD compared to using a Deep Ensemble teacher.

2 BACKGROUND

In this section, we review ensemble-based uncertainty estimation. We follow with a discussion of how the limitations of ensembles can be addressed using recently developed distillation techniques for autoregressive sequence tasks such as machine translation.

2.1 UNCERTAINTY ESTIMATION

We adopt a Bayesian perspective on ensembles as this offers a flexible framework within which uncertainties have an information-theoretic justification. The posterior over model parameters $\mathtt{p}(\boldsymbol{\theta}|\mathcal{D})$ is derived given some observed (training) data, $\mathcal{D}$. Unfortunately, the posterior is often intractable and cannot be derived for large non-linear networks. Alternatively an approximation $\mathtt{q}(\boldsymbol{\theta}) \approx \mathtt{p}(\boldsymbol{\theta}|\mathcal{D})$ can be used. Samples from this approximate distribution can then be drawn to generate an ensemble of models.

Take an ensemble $\{\mathtt{P}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}^{(m)})\}_{m=1}^{M}$ sampled from an approximate posterior $\mathtt{q}(\boldsymbol{\theta})$ where each model maps a *variable-length* input $\boldsymbol{x} \in \mathcal{X}$ into a *variable-length* output $\boldsymbol{y} \in \mathcal{Y}$ of discrete units. The predictive distribution is obtained by:

$$\mathtt{P}(\boldsymbol{y}|\boldsymbol{x},\mathcal{D}) = \mathbb{E}_{\mathtt{q}(\boldsymbol{\theta})}\big[\mathtt{P}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\big]. \tag{1}$$

From this predictive distribution, a measure of total uncertainty can be estimated using the entropy:

$$\mathcal{H}\big[\mathtt{P}(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})\big] = \mathbb{E}_{\mathtt{P}(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})}\big[-\ln\mathtt{P}(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})\big]. \tag{2}$$

Furthermore, a measure of disagreement between models, also referred to as *knowledge* or *epistemic* uncertainty, can be estimated by using mutual information between $\boldsymbol{y}$ and $\boldsymbol{\theta}$:

$$\mathcal{I}\big[\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{x},\mathcal{D}\big] = \mathbb{E}_{\mathtt{q}(\boldsymbol{\theta})}\Big[\mathtt{KL}\big(\mathtt{P}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) \,\big\|\, \mathtt{P}(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})\big)\Big]. \tag{3}$$

This estimate can also be decomposed into a measure of total and data (aleatoric) uncertainty, as mentioned in Malinin and Gales [2021]. There are also many other potential measures of knowledge uncertainty such as expected pair-wise KL-divergence or reverse mutual information [Malinin and Gales, 2021], however, for the sake of simplicity we restrict our focus to the already mentioned eq. (2) and (3) since these represent uncertainties of differing natures.

Limitations: The discussion has so far assumed one can enumerate all possible variable-length outputs $\boldsymbol{y} \in \mathcal{Y}$ which is not tractable in autoregressive sequence tasks. Instead, one can approximate the uncertainties by monte-carlo methods [Notin et al., 2021] and utilising the autoregressive structure of predictions [Malinin and Gales, 2021]:

$$\mathtt{P}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \prod_{l=1}^{L}\mathtt{P}(y_l|\boldsymbol{y}_{<l},\boldsymbol{x},\boldsymbol{\theta}). \tag{4}$$

We refer to Malinin and Gales [2021] for an in-depth discussion and analysis of approximations for predictive entropy and mutual information for autoregressive prediction.

2.2 KNOWLEDGE DISTILLATION

Ensembles $\{\mathtt{P}(\boldsymbol{y}|\boldsymbol{x},\ \boldsymbol{\theta}^{(m)})\}_{m=1}^{M}$ sampled from some posterior can be computationally demanding. One approach to efficiently exploit the information of the ensemble is to use Knowledge Distillation (KD) to yield a single student model [Hinton et al., 2014, Kim and Rush, 2016].

Given a reference data pair $(\boldsymbol{x},\boldsymbol{y}) \sim \tilde{\mathtt{p}}(\boldsymbol{x},\boldsymbol{y})$, a standard model might be trained using negative log-likelihood (NLL):

$$\mathcal{L}_{\mathtt{NLL}}(\boldsymbol{\theta}) = -\frac{1}{L}\sum_{l=1}^{L}\ln\mathtt{P}(y_l|\boldsymbol{y}_{<l},\boldsymbol{x},\boldsymbol{\theta}). \tag{5}$$

This is referred to as teacher-forcing since during training the model makes predictions at step l conditioned on the true output $\boldsymbol{y}_{<l}$ (rather than its own previous predictions). Similarly, a student model with parameters $\boldsymbol{\phi}$ can be trained to emulate a teacher ensemble by additionally using *average* ensemble categorical/softmax outputs $\boldsymbol{\pi}_l, \pi_{l,k} = \mathtt{P}(y_l = k|\boldsymbol{y}_{<l},\boldsymbol{x},\mathcal{D})$ as soft labels:

$$\mathcal{L}_{\mathtt{KL}}(\boldsymbol{\phi}) = \frac{1}{L}\sum_{l=1}^{L}\mathtt{KL}\big(\boldsymbol{\pi}_l \,\big\|\, \mathtt{P}(y_l|\boldsymbol{y}_{<l},\boldsymbol{x},\boldsymbol{\phi})\big). \tag{6}$$

However in practice, one optimises a convex combination of the likelihood and KL-divergence losses $\mathcal{L}_{\texttt{KD}}(\boldsymbol{\phi}) = \lambda\mathcal{L}_{\texttt{NLL}}(\boldsymbol{\phi}) + (1-\lambda)\mathcal{L}_{\texttt{KL}}(\boldsymbol{\phi}), \lambda \in [0,1]$ for added supervision and stability. The probability mass functions in the KL-divergence can also be temperature scaled to improve optimisation [Hinton et al., 2014]. Note that this criterion is only considered for the teacher-forcing case, more sophisticated distillation approaches exist, by sampling $(\boldsymbol{x},\boldsymbol{y})$ from alternative distributions, but is outside the scope for this work, see Kim and Rush [2016], Wong et al. [2016], Lee et al. [2018], Malinin et al. [2017] for details.

2.3 ENSEMBLE DISTRIBUTION DISTILLATION

Whilst KD has been successful in many sequence tasks, the resulting student is not able to estimate knowledge uncertainty, since it only models the average ensemble output. To avoid this issue, Malinin et al. [2020], Ryabinin et al. [2021] consider the task of distilling the *distribution* of sequence ensemble predictions onto a single student. This allows the student to retain both predictive performance and information about ensemble diversity.

To explain the mechanics behind Ensemble Distribution Distillation (EDD), consider modelling a distribution over autoregressive ensemble predictions, in which all M ensemble members share the same back-history $\boldsymbol{y}_{<l}$:

$$\{\boldsymbol{\pi}_l^{(m)}\}_{m=1}^M, \pi_{l,k}^{(m)} = \texttt{P}(y_l = \omega_k | \boldsymbol{y}_{<l}, \boldsymbol{x}, \boldsymbol{\theta}^{(m)}). \tag{7}$$

Now let an autoregressive student predict the parameters $\boldsymbol{\alpha}_l$ of a Dirichlet distribution $\texttt{Dir}(\boldsymbol{\pi}_l|\boldsymbol{\alpha}_l) = \texttt{p}(\boldsymbol{\pi}_l|\boldsymbol{y}_{<l}, \boldsymbol{x}, \boldsymbol{\phi})$. Since the Dirichlet models a distribution over categorical distributions it is an ideal candidate for this task. The distribution distillation loss of such a model is then simply the result of (negative) log-likelihood:

$$\begin{aligned}\mathcal{L}_{\texttt{NLL}}^{\texttt{DD}}(\boldsymbol{\phi}) &= -\frac{1}{MLK}\sum_{m,l}\ln\texttt{Dir}(\boldsymbol{\pi}_l^{(m)}|\boldsymbol{\alpha}_l) \\ &\equiv \frac{1}{LK}\sum_l\Big(\ln B(\boldsymbol{\alpha}_l) - \sum_k \alpha_{l,k}\ln\tilde{\pi}_{l,k}\Big),\end{aligned} \tag{8}$$

where K is the number of classes, $B(\boldsymbol{\alpha})$ is the beta function and $\tilde{\pi}_{l,k}$ is the geometric average of the individual ensemble softmax probabilities in Equation (7).

Whilst this approach was shown to be promising on a small-scale image classification task in Malinin et al. [2020], following work [Fathullah et al., 2021, Ryabinin et al., 2021] found that direct application of Equation (8) encounters optimisation issues when scaled to larger label spaces. This arises from the way classwise loss gradients are related to teacher class probabilities. It turns out that, unlike standard distillation, the loss in Equation (8) induces small gradients for (important) high-probability classes and large gradients for (unimportant) low-probability classes. This negatively affects convergence as the number of low-probability classes increases. Ryabinin et al. [2021] proposed an approach for scaling Dirichlet EDD, where the student aims to minimise a normalized reverse KL-divergence to a *proxy* Dirichlet, which will be used as a baseline in this work.

3 SEQUENCE LOGIT-BASED EDD

In sequence tasks with a large number of classes, which commonly occurs in speech recognition and machine translation, the output categorical distributions are often very sparse and concentrated. Therefore, it often becomes highly challenging to apply EDD to tasks of this nature. On the other hand, KD has been shown to work well for larger tasks [Kim and Rush, 2016, Gaido et al., 2020, Tan et al., 2019, Jiao et al., 2019], but since it only models the average teacher predictions, it cannot estimate data and knowledge uncertainties that are important for many downstream tasks such as out-of-distribution detection.

In this section, we describe a *Logit-based* Ensemble Distribution Distillation (L-EDD) approach for autoregressive models which addresses the drawbacks of both KD and EDD in a single consistent framework and is scalable to sequence problems with a large number of classes. Consider a set of logits produced by an ensemble:

$$\{\boldsymbol{z}_l^{(m)}\}_{m=1}^M, \boldsymbol{\pi}_l^{(m)} = \texttt{Softmax}(\boldsymbol{z}_l^{(m)}). \tag{9}$$

Traditional distillation approaches thereafter use the logits to produce categorical probability distributions by applying the softmax function. However, instead of operating in the probability space, we propose training a student, with model parameters $\boldsymbol{\phi}$, to directly model the logit space by predicting the mean $\boldsymbol{\mu}_l$ and scale $\boldsymbol{\sigma}_l$ parameters of a diagonal Laplace distribution:

$$\begin{aligned}\texttt{p}(\boldsymbol{z}|\boldsymbol{y}_{<l}, \boldsymbol{x}, \boldsymbol{\phi}) &= \texttt{Lap}(\boldsymbol{z}|\boldsymbol{\mu}_l, \boldsymbol{\sigma}_l) \\ &= \prod_k \frac{1}{2\sigma_{l,k}}\exp\left\{-\frac{|z_k - \mu_{l,k}|}{\sigma_{l,k}}\right\}.\end{aligned} \tag{10}$$

Because we opt for a diagonal distribution, sampling is parallelisable, highly efficient, and straightforward and allows for the estimation of uncertainties in exactly the same manner as in standard ensembles. Additionally, significantly fewer parameters are required compared to using a fully-specified covariance matrix. Another benefit of the chosen distribution is the long tails which make the Laplace robust to outliers, unlike the Gaussian distribution. This robustness also makes it a natural choice for handling the early stages of training when the student model is randomly initialised and its output distribution substantially differs from the ensemble logits.

Furthermore, given the set of logits produced by an ensemble, the student model $\texttt{p}(\boldsymbol{z}|\boldsymbol{y}_{<l}, \boldsymbol{x}, \boldsymbol{\phi})$ can be trained

by straightforward application of log-likelihood training:

$$\begin{aligned}\mathcal{L}_{\texttt{NLL}}^{\texttt{L}-\texttt{EDD}}(\boldsymbol{\phi}) &= -\frac{1}{MLK}\sum_{m,l}\ln\texttt{Lap}(\boldsymbol{z}_l^{(m)}|\boldsymbol{\mu}_l,\boldsymbol{\sigma}_l) \\ &\equiv \frac{1}{MLK}\sum_{m,l,k}\frac{\left|z_{l,k}^{(m)}-\mu_{l,k}\right|}{\sigma_{l,k}}+\ln\sigma_{l,k}.\end{aligned} \tag{11}$$

We also perform experiments with a student (diagonal) Gaussian distribution on the logits, variations of which have been explored in static image classification [Fathullah and Gales, 2022, Lindqvist et al., 2020] but remained unexplored for autoregressive sequence tasks:

$$\begin{aligned}\texttt{p}(\boldsymbol{z}|\boldsymbol{y}_{<l},\boldsymbol{x},\boldsymbol{\phi}) &= \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_l,\boldsymbol{\sigma}_l^2) \\ &= \prod_k \frac{1}{(2\pi\sigma_{l,k}^2)^{\frac{1}{2}}}\exp\left\{-\frac{(z_k-\mu_{l,k})^2}{2\sigma_{l,k}^2}\right\}.\end{aligned} \tag{12}$$

Similar to all of the mentioned approaches, this system is also trained using the log-likelihood objective:

$$\mathcal{L}_{\texttt{NLL}}^{\texttt{L}-\texttt{EDD}}(\boldsymbol{\phi}) = -\frac{1}{MLK}\sum_{m,l}\ln\mathcal{N}(\boldsymbol{z}_l^{(m)}|\boldsymbol{\mu}_l,\boldsymbol{\sigma}_l^2). \tag{13}$$

The Gaussian distribution, which induces an L2-norm loss function is much more sensitive to outliers in the ensemble outputs. This student could potentially be more challenging to train, but should still be more stable than Dirichlet EDD.

3.1 PRACTICAL CONSIDERATIONS

Since the softmax activation function is shift invariant,

$$\texttt{Softmax}(\boldsymbol{z}-\mathbf{1}b) = \texttt{Softmax}(\boldsymbol{z})\ \forall b\in\mathbb{R},$$

one has to consider this property when performing distribution distillation. Ensemble members are unconstrained along $\mathbf{1}$, and so can potentially vary wildly in the logit space, even if they give consistent softmax predictions. Therefore, logits are normalised by $\tilde{\boldsymbol{z}} = \boldsymbol{z}-\mathbf{1}\texttt{LogSumExp}(\boldsymbol{z})$. This particular normalisation scheme is not special and any choice of the normalisation constant such as $\texttt{Max}(\boldsymbol{z})$ or $\texttt{Mean}(\boldsymbol{z})$ would be valid. Next, to ensure that the student can be trained reliably, we interpolate the knowledge and distribution distillation losses $\mathcal{L}_{\texttt{KD}}(\boldsymbol{\phi})+\beta\mathcal{L}_{\texttt{NLL}}^{\texttt{L}-\texttt{EDD}}(\boldsymbol{\phi})$ (see Eq. 11 and Sec. 2.2).

Furthermore, distributions in logit space often lead to analytically intractable expectations in probability space. The standard approach to circumvent this issue is by sampling from the distribution using monte-carlo approximations. However, in this paper, we opt for an approximative deterministic approach when computing the predictive distribution (e.g. when decoding):

$$\begin{aligned}\texttt{P}(y_l|\boldsymbol{y}_{<l},\boldsymbol{x},\boldsymbol{\phi}) &= \mathbb{E}_{\texttt{p}(\boldsymbol{z}_l|\boldsymbol{y}_{<l},\boldsymbol{x},\boldsymbol{\phi})}\left[\texttt{Softmax}(\boldsymbol{z}_l)\right] \\ &\approx \texttt{Softmax}(\boldsymbol{\mu}_l),\end{aligned}$$

in which we approximate the expectation by just using the mean of the logit distribution. When performing downstream tasks that require uncertainties we revert to a stochastic sampling scheme to generate multiple predictions from the distribution.

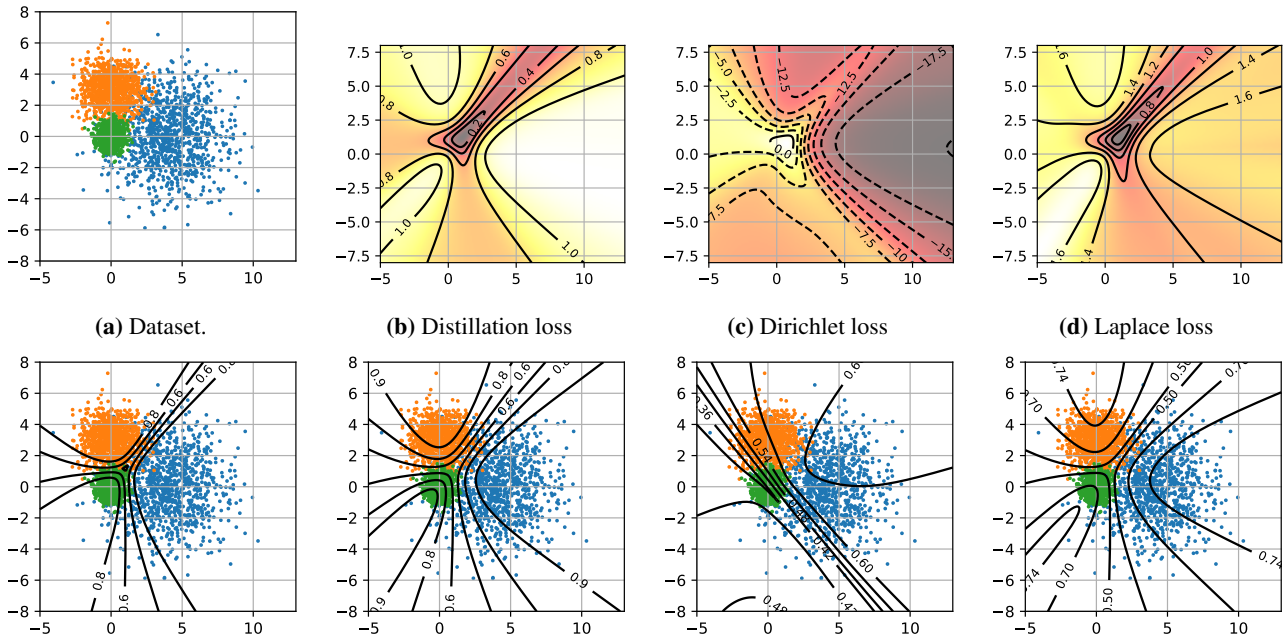
4 EXPERIMENTS ON ARTIFICIAL DATA

This section investigates the proposed Laplace logit-based ensemble distribution distillation (L-EDD) technique on a static artificial dataset, see Figure 1a. The dataset was generated by sampling 3000 data points from three isotropic Gaussian distributions equally. The location and standard deviation of the Gaussians were chosen such that there would be regions with significant overlap and regions where models can be highly confident.

In these exploratory experiments, an ensemble of 10 small neural networks is first trained by randomly initialising each member. Thereafter, the ensemble is distilled using KD, EDD and Laplace L-EDD. We perform a qualitative comparison of these methods by displaying both the loss surface (see Figures 1c-1d) of each approach and the resulting confidence (maximum softmax probability) contours ((see Figures 1e-1h)). The loss surface shows how the student distillation loss varies over the input space, thus providing useful information about which regions of the data are successfully optimised. The confidence contours are useful to understand if the system can separate between each of the three classes. EDD training had to be terminated early as it diverged due to large gradients originating from the high-confidence regions (as discussed in Ryabinin et al. [2021]).

Figure 1e shows the ensemble confidence contours which clearly trace out class boundaries and partially separate the three classes. The confidence also increases as one moves further away from regions of overlap since there is less uncertainty. This is the behaviour we expect from a properly trained system and further, expect distilled students to behave similarly. Next, we knowledge distil the ensemble onto a single student model, see Figures 1b and 1f. The loss surface shows that the student can optimise the distillation objective over regions with high overlap well and generate confidence score contours that are consistent with the teacher ensemble.

However, Ensemble Distribution Distillation completely fails on this very simple task. Observing the loss surface in Figure 1c, one can infer that the Dirichlet student is unable to optimise regions where there is significant data overlap, instead displaying extremely small losses in regions for which the teacher ensemble is already confident. This links back to a result in Ryabinin et al. [2021] in which they find that highly confident teachers can induce extremely large gradients. This also translates into inaccurate confidence contours which are unable to separate between classes,



(a) Dataset. **(b)** Distillation loss **(c)** Dirichlet loss **(d)** Laplace loss

(e) Ensemble conf. **(f)** Distillation conf. **(g)** Dirichlet conf. **(h)** Laplace conf.

Figure 1: An artificial three-class classification problem with 1000 examples per class. The top row shows the loss surface contours for various distillation approaches; darker colours imply lower losses. The bottom row shows the corresponding confidence contours (the confidence scores are reported on the contour). This shows that Dirichlet-based EDD is unable to learn properly whilst our proposed Laplace L-EDD can imitate an ensemble.

especially in regions of overlap, see Figure 1g.

The Laplace L-EDD approach circumvents these issues by operating in the logit space. The resulting loss surface is much more consistent with the knowledge-distilled student, optimising regions of high overlap, see Figure 1d. Similarly, the confidence contours trace out boundaries consistent with the ensemble but with lower overall confidence. Since the log-likelihood (similar to KL-divergence) is a mode-covering objective [Minka et al., 2005], the Laplace student ends up predicting distributions that overestimate the range of ensemble logits. Couple this with the long tails of the distributions and the Laplace will often overestimate the variance in logits and produce lower overall confidence.

5 MACHINE TRANSLATION

This section reports on the performance of base transformers [Vaswani et al., 2017] trained on the En-De WMT'16 dataset, consisting of 4.5 million sentence pairs covering topics such as news & policy. We use newstest-13 for validation and newstest-14 for predictive evaluation. For the main task of investigating out-of-distribution (OOD) detection, we compare the in-distribution (ID) newstest-14 with one of the publically available Khresmoi-Summary (Khresmoi) [Dušek et al., 2017], MTNT [Michel and Neubig, 2018] and Kyoto Free Translation Task (KFTT) [Neubig, 2011] datasets. These datasets relate to medical articles, Reddit-based noisy conversational text and specialised Wikipedia articles, respectively. Furthermore, we apply insights from training these systems to big transformers [Vaswani et al., 2017] trained on the larger En-Ru WMT'20 consisting of 58 million pairs after processing. In this case, we use newstest-19 for validation and newstest-20 for evaluation. The OOD detection task uses newstest-20 as ID and the same OOD datasets as above for the base transformer.

Data is tokenized using Moses, following Ott et al. [2018]. For WMT'16, a shared dictionary is trained using Byte Pair Encoding (BPE) with 32,000 merge operations [Sennrich et al., 2016]. For WMT'20 we learn disjoint dictionaries using BPE with 40,000 merge operations. The predictive performance (translation quality) will be evaluated using corpus-level (Sacre)BLEU [Post, 2018], with no post-processing of outputs before being scored. For the main task of detection, we use the ubiquitous threshold-independent AUROC metric [Manning and Schütze, 1999], with baseline random detection corresponding to a score of 50%.

All standard transformers are trained using an inverse square root with a linear warmup stage. A stronger Deep Ensemble baseline is formed by taking $M = 5$ such models. To avoid the high training cost of Deep Ensembles, we also train Snapshot Ensembles [Xie et al., 2013, Huang et al., 2017] with a cyclic learning rate [Smith, 2017] to showcase that

distribution distillation can be achievable with smaller training budgets. Since building Deep Ensembles is expensive, and the performance difference to Snapshot Ensembles was shown to be small, we opted to perform most distillation experiments on Snapshot Ensembles. We compare the proposed L-EDD approaches with KD and EDD and repeat each experiment 5 times, each with a different Snapshot Ensemble. Finally, similar to a range of prior work on uncertainty estimation tasks [Malinin et al., 2020, Malinin and Gales, 2021, Corbière et al., 2019], we do not aim to achieve state-of-the-art predictive performance but opt for a simpler setup with a focus on achieving better uncertainty estimation. All setup details are provided in Appendix A. Hyperparameters were determined on ID validation sets.

Table 1: Model parameter size, relative training time and translation performance on newstest-14 ±2std (BLEU) for base transformer. We include two different KD baselines, one for each teacher ensemble.

Model	**Size**	**Train Time**	**BLEU ↑**
Standard	60.9M	1.0	25.85 ± 0.17
Deep Ensemble	304.5M	5.0	26.72
KD (Categorical)	60.9M	5.9	26.70 ± 0.26
Snapshot Ensemble	304.5M	1.5	26.54 ± 0.16
KD (Categorical)	60.9M	1.9	27.02 ± 0.19
EDD (Dirichlet)	60.9M	2.0	26.96 ± 0.06
L-EDD (Gaussian)	77.8M	1.9	26.90 ± 0.28
L-EDD (Laplace)	77.8M	1.9	27.08 ± 0.20

5.1 BASE TRANSFORMER RESULTS

Table 1 shows both the efficiency and performance of a wide range of systems on newstest-14. As expected the performance of the Deep Ensemble trumps both the Snapshot Ensemble and a standard trained system. Surprisingly, Snapshot Ensemble distilled students achieve better performance, a pattern also observed in self-distilled systems and is explored in more detail in Allen-Zhu and Li [2021].

Next, we compare the threshold-independent out-of-distribution detection performance of baseline systems with L-EDD models. From Table 2, we observe that Snapshot Ensembles are able to compete with the Deep equivalent whilst being more than 3 times cheaper to train. Furthermore, the knowledge-distilled students are able to match the detection performance of their Deep and Snapshot ensemble teachers using total uncertainty (TU). This is a natural result since they were specifically designed to capture the predictive distribution of their teacher ensemble. However, since KD students are unable to estimate knowledge uncertainty (KU), they fail to reach ensemble-level detection performance in all but the MTNT dataset. Similarly, the modified Dirichlet baseline as described by Ryabinin et al. [2021] is able to achieve similar detection performance using TU but with the added ability to estimate KU. And whilst the Dirichlet KU are often better than its TU estimates, they often fall short when compared to ensembles.

On the other hand, the Laplace & Gaussian L-EDD models are (surprisingly) able to outperform both ensembles in all three detection splits, producing either similar or significantly better TU and KU estimates. This may partially be due to the fact that diagonal Laplace and Gaussian distributions have more parameters and are more flexible than the Dirichlet, and also because they do not suffer from the same optimisation issues. Nonetheless, neither reason explains why L-EDD models can outperform ensembles in detection. We explore this pattern in Section 6.

Additionally, many models are worse than a random detector, especially for the KFTT and partially for the Khresmoi dataset. A partial explanation could be that these datasets contain longer sequences. When decoding, the transformer models produce more and more confident predictions further along in the output sequence, causing lower uncertainty scores. Section 6 investigates this effect and isolates a possible reason behind Laplace's success in

Table 2: OOD detection performance (%AUROC ↑ ± 2 std) for base transformer with ID dataset newtest-14 and OOD datasets Khresmoi, MTNT and KFTT. **Bold** indicates best in a column, underline second best. Laplace L-EDD with knowledge uncertainty (KU), shows superior performance for all OOD datasets even compared to the Deep Ensemble.

Model	**Khresmoi**		**MTNT**		**KFTT**	
	TU	**KU**	**TU**	**KU**	**TU**	**KU**
Standard	47.5 ± 0.8	✗	63.5 ± 1.3	✗	30.6 ± 1.2	✗
Deep Ensemble	48.0	61.9	64.5	63.7	30.1	44.0
KD (Categorical)	47.9 ± 1.1	✗	64.5 ± 1.3	✗	29.8 ± 0.7	✗
Snapshot Ensemble	49.0 ± 0.6	62.6 ± 1.1	63.8 ± 1.2	63.1 ± 0.7	31.7 ± 0.9	47.4 ± 2.5
KD (Categorical)	48.0 ± 1.4	✗	64.6 ± 0.9	✗	31.3 ± 0.5	✗
EDD (Dirichlet)	49.6 ± 1.3	57.1 ± 1.4	65.1 ± 1.7	65.6 ± 2.0	31.0 ± 0.9	36.2 ± 1.4
L-EDD (Gaussian)	59.5 ± 1.1	71.7 ± 1.9	**66.3** ± 1.6	64.0 ± 2.1	35.8 ± 1.2	44.0 ± 0.2
L-EDD (Laplace)	**65.1** ± 1.8	**73.1** ± 1.7	65.1 ± 1.5	**66.8** ± 1.8	**37.8** ± 0.2	**48.8** ± 1.4

outperforming ensembles.

As an aside, we observe that estimates of knowledge uncertainty are clearly important for OOD detection for autoregressive sequence tasks (and this is corroborated in prior work Malinin and Gales [2021], Ryabinin et al. [2021]). This is in contrast to recent empirical results on image classification data, which show the opposite, that measures of knowledge uncertainty are not useful for indicating distributional shifts [Xia and Bouganis, 2022a, Abe et al., 2022].

Table 3: Model parameter size, relative training time and translation performance on newstest-20 ±2std (BLEU) for the big transformer.

Model	Size	Train Time	BLEU ↑
Standard	271M	1.0	26.28 ± 0.34
Deep Ensemble	1.35B	5.0	26.81
Snapshot Ensemble	1.35B	1.5	26.42 ± 0.23
KD (Categorical)	271M	2.1	26.73 ± 0.16
EDD (Dirichlet)	271M	2.2	26.66 ± 0.19
L-EDD (Laplace)	320M	2.2	26.71 ± 0.18

5.2 BIG TRANSFORMER RESULTS

In this section, we take the best-performing systems from the previous section and apply them to the 'big transformer' on the larger En-Ru WMT'20 dataset. Table 3 shows the efficiency and predictive performance on newstest-20. Again, we observe that the Deep Ensemble outperforms its Snapshot equivalent. Furthermore, the KD and L-EDD students, distilled from the Snapshot Ensemble were able to outperform their teacher. However, unlike the smaller-scale experiments, these students were only able to reach Deep Ensemble performance within a standard deviation, but were able to do so with a single forward pass.

From Table 4 we observe a similar pattern in which Deep and Snapshot ensembles perform equivalently whilst L-EDD (Laplace) is able to significantly outperform both ensembles in all but the MTNT dataset. Interestingly, unlike in Table 2 where no model was able to beat a random detector on the KFTT detection, the larger En-Ru WMT'20 based models are able to differentiate between newstest-20 and KFTT; switching the ID dataset to newstest-14 does not affect the results notably.

6 ANALYSIS: ENSEMBLE VS LAPLACE

6.1 AUGMENTED ENSEMBLE UNCERTAINTIES

Both Sections 5.1 and 5.2 found that L-EDD models overall significantly outperformed their teacher Snapshot Ensemble and a Deep Ensemble. Therefore, we propose an alternative experiment to understand the source of L-EDD's superior performance. We fit an auxiliary Laplace distribution to a Deep Ensemble during inference and use the samples from this proxy to perform the detection task.

Consider a Deep Ensemble which produces a set of normalised logits $\{\tilde{\boldsymbol{z}}_l^{(m)}\}_{m=1}^M$ as in Equation (9). In traditional uncertainty estimation, these logits would be transformed into categorical distributions. However, in this experiment, we estimate an auxiliary Laplace distribution using maximum likelihood (which is the loss-minimising distribution for a Laplace L-EDD student):

$$\tilde{\boldsymbol{\mu}}_l,\tilde{\boldsymbol{\sigma}}_l = \underset{\boldsymbol{\mu},\boldsymbol{\sigma}}{\text{argmax}} \sum_m \ln \texttt{Lap}(\tilde{\boldsymbol{z}}_l^{(m)}|\boldsymbol{\mu},\boldsymbol{\sigma}). \tag{14}$$

By sampling new points from this auxiliary distribution, we can estimate total and knowledge uncertainty:

$$\tilde{\boldsymbol{\pi}} = \texttt{Softmax}(\boldsymbol{z}), \boldsymbol{z} \sim \texttt{Lap}(\tilde{\boldsymbol{\mu}}_l,\tilde{\boldsymbol{\sigma}}_l). \tag{15}$$

The aim of this modified approach to ensemble-based uncertainty estimation is to investigate whether or not approximating the logits with a Laplace distribution is the reason behind L-EDD performing better.

Table 5 shows the detection performance of the Laplace-modified Deep Ensemble, following the detection setup in Section 5.1. Clearly, the augmented ensemble bridges the OOD detection performance gap between standard Deep Ensemble and L-EDD. This suggests that measures

Table 4: OOD detection performance (%AUROC ↑ ± 2 std) for big transformer with ID dataset newtest-20 and OOD datasets Khresmoi, MTNT and KFTT. **Bold** indicates best in a column, underline second best. Similar to Table 4, L-EDD (Laplace) with KU shows superior performance over all OOD datasets.

Model	Khresmoi		MTNT		KFTT	
	TU	KU	TU	KU	TU	KU
Deep Ensemble	39.3	53.2	70.8	69.0	51.0	60.3
Snapshot Ensemble	40.8 ± 0.5	55.0 ± 0.8	70.1 ± 0.5	69.3 ± 0.9	51.1 ± 0.6	60.9 ± 1.4
KD (Categorical)	40.4 ± 0.8	✗	70.9 ± 1.0	✗	50.9 ± 0.6	✗
EDD (Dirichlet)	41.0 ± 0.8	52.9 ± 1.3	71.3 ± 0.7	69.4 ± 0.7	51.0 ± 0.8	60.0 ± 1.3
L-EDD (Laplace)	**51.0** ± 0.9	**63.4** ± 1.2	**72.6** ± 0.8	**70.2** ± 0.6	**63.2** ± 1.0	**70.2** ± 1.1

Table 5: OOD detection performance (%AUROC ↑ ± 2 std) following the same setup as in Table 2. The Laplace augmented ensemble demonstrates much better performance in most cases compared to its standard counterpart.

Model	Khresmoi		MTNT		KFTT	
	TU	KU	TU	KU	TU	KU
Deep Ensemble	48.0	61.9	64.5	63.7	30.1	44.0
Deep Ensemble (Laplace)	62.5	72.1	63.8	56.1	34.8	**55.4**
L-EDD (Laplace)	**65.1** ± 1.8	**73.1** ± 1.7	**65.1** ± 1.5	**66.8** ± 1.8	**37.8** ± 0.2	48.8 ± 1.4

of uncertainty based on directly fit, logit-space models of an ensemble are better at indicating distributional shift *than directly using the ensemble logits*, for autoregressive sequence prediction. We remark that fully understanding why this is the case would be an interesting direction of future research, as it may enable further advancement in autoregressive out-of-distribution detection.

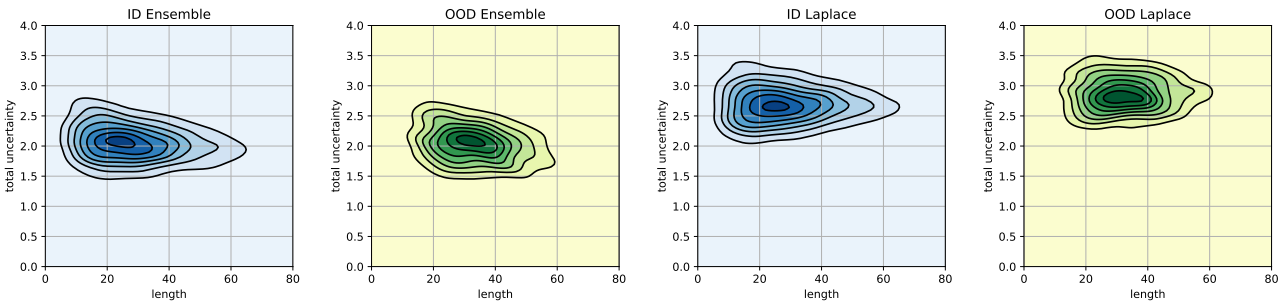
6.2 MODEL CONFIDENCE FOR LONGER SEQUENCES

Another reason for Laplace L-EDD's superior performance could be found in analysing behaviour with increasing sequence lengths. Figure 2 shows how Deep Ensemble and L-EDD total uncertainties scale with output sequence length for both ID (newstest-14) and OOD (Khresmoi) datasets. Under each figure is also the associated Pearson correlation.

Three observations can be made from this data. The first is that the L-EDD system consistently outputs much higher total uncertainties. The second is that the Deep Ensemble displays a negative correlation between total uncertainty and sequence length. This implies that the ensemble is more confident in translating longer sequences, but this is also why it fails in detecting OOD datasets which contain longer sequences. The third and possibly most significant observation is that the Laplace system shows almost no correlation between total uncertainty and sequence length for both ID and OOD datasets, effectively eliminating the length bias, and allowing it to better differentiate between ID and OOD datasets even when the OOD inputs differ in length from what the detection system was trained on.

7 CONCLUSION

In this work, we investigate the efficient estimation of uncertainties for large-scale autoregressive sequence prediction. To this end, we examine Ensemble Distribution Distillation (EDD) in the *logit*-space, in order to bypass optimisation issues found in softmax-space EDD. We perform experiments using modern transformer models trained to perform large-scale machine translation. They show that a student model trained to parameterise a *Laplace* distribution over logits is able to significantly outperform Deep Ensembles for OOD detection at a fraction of the inference cost, whilst matching the ensemble for translation quality. Moreover, we show that the use of Snapshot Ensembling can greatly reduce the training costs of EDD, without sacrificing translation performance. We hope that our work can encourage further investigation into the comparatively less well-explored domain of uncertainty estimation for structured sequence prediction, on tasks such as machine translation, image captioning, and automatic speech recognition.



(a) PCC = -20% **(b)** PCC = -29% **(c)** PCC = -3% **(d)** PCC = 1%

Figure 2: Length vs uncertainty density plots for ID (newstest-14) and OOD (Khresmoi) datasets. The left half corresponds to the Deep Ensemble and the right half to Laplace L-EDD. Each figure caption also shows the Pearson Correlation Coefficient (PCC) between sequence length and total uncertainty. Total uncertainty and length are negatively correlated for the ensemble, i.e. it is more confident on longer sequences, but they are uncorrelated for L-EDD.

Acknowledgements

Guoxuan Xia is funded jointly by Arm ltd. and EPSRC.

References

Taiga Abe, E. Kelly Buchanan, Geoff Pleiss, Richard Zemel, and John Patrick Cunningham. Deep ensembles work, but are they necessary? In *Advances in Neural Information Processing Systems*, 2022.

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *arXiv*, 2021.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems 32*. 2019.

Stefan Depeweg, Jose Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Learning Representations*, 2018.

Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Jindřich Libovický, Pavel Pecina, Aleš Tamchyna, and Zdeňka Urešová. Khresmoi summary translation test data 2.0. `http://hdl.handle.net/11234/1-2122`, 2017. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Yassir Fathullah and Mark J. F. Gales. Self-distribution distillation: efficient uncertainty estimation. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, 01–05 Aug 2022.

Yassir Fathullah, Mark J. F. Gales, and Andrey Malinin. Ensemble distillation approaches for grammatical error correction. In *International Conference on Acoustics, Speech, and Signal Processing*, 2021.

Marco Gaido, Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. End-to-end speech-translation with knowledge distillation: FBK@IWSLT2020. In *International Conference on Spoken Language Translation*, 2020.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 06–11 Aug 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 09–15 Jun 2019.

Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schön. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Conference on Neural Information Processing Systems Deep Learning Workshop*, 2014.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get m for free. In *International Conference on Learning Representations*, 2017.

Eyke Hullermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. In *Machine Learning*, 2021.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *International Conference on Learning Representations*, 2019.

Jihyo Kim, Jiin Koo, and Sangheum Hwang. A unified benchmark for the unknown detection capability of deep neural networks. *ArXiv*, abs/2112.00337, 2021.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Conference on Empirical Methods in Natural Language Processing*, 2016.

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

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Conference on Neural Information Processing Systems*, 2017.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018.

Jakob Lindqvist, Amanda Olmin, Fredrik Lindsten, and Lennart Svensson. A general framework for ensemble distribution distillation. In *International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Andrey Malinin and Mark J. F. Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021.

Andrey Malinin, Anton Ragni, Kate Knill, and Mark J. F. Gales. Incorporating uncertainty into deep learning for spoken language assessment. In *Association for Computational Linguistics*, 2017.

Andrey Malinin, Bruno Mlodozeniec, and Mark J. F. Gales. Ensemble distribution distillation. In *International Conference on Learning Representations*, 2020.

Chris Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

Paul Michel and Graham Neubig. Mtnt: A testbed for machine translation of noisy text. In *Conference on Empirical Methods in Natural Language Processing*, 2018.

Tom Minka et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *International Conference on Machine Learning*, 2020.

Graham Neubig. The Kyoto free translation task. `http://www.phontron.com/kftt`, 2011.

Pascal Notin, José Miguel Hernández-Lobato, and Yarin Gal. Improving black-box optimization in VAE latent space using decoder uncertainty. In *Advances in Neural Information Processing Systems*, 2021.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018.

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.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Matt Post. A call for clarity in reporting BLEU scores. In *Association for Computational Linguistics*, 2018.

Puria Radmard, Yassir Fathullah, and Aldo Lipani. Subsequence based deep active learning for named entity recognition. In *Association for Computational Linguistics*, 2021.

Max Ryabinin, Andrey Malinin, and Mark J. F. Gales. Scaling ensemble distribution distillation to many classes with proxy targets. In *Conference on Neural Information Processing Systems*, 2021.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Association for Computational Linguistics*, 2016.

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

Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*, 2019.

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.

Jeremy H. M. Wong, Mark J. F. Gales, and Yu Wang. Sequence-level knowledge distillation. In *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2016.

Guoxuan Xia and Christos-Savvas Bouganis. On the usefulness of deep ensemble diversity for out-of-distribution detection. *ArXiv*, abs/2207.07517, 2022a.

Guoxuan Xia and Christos-Savvas Bouganis. Augmenting softmax information for selective classification with out-of-distribution data. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, December 2022b.

Jingjing Xie, Bing Xu, and Zhang Chuang. Horizontal and vertical ensemble with deep representation for classification. In *International Conference on Machine Learning*, 2013.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Thirty-sixth Conference on*

Neural Information Processing Systems Datasets and Benchmarks Track, 2022.

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 6. 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 6: 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%

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 $4\times32=128$ 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 $\mathcal{L}_{\text{KD}}$ provided in Section 2.2 with $\lambda=0.50$. The student was trained with a warmup of 1026 steps (3 epochs), from $\eta=4.0\times10^{-4}$ to $\eta=7.0\times10^{-4}$ 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.

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

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.