
Generalized Bayesian Posterior Expectation Distillation for Deep Neural Networks

Meet P. Vadera¹, Brian Jalain², and Benjamin M. Marlin¹

¹University of Massachusetts Amherst, ²US Army Research Laboratory
{mvadera, marlin}@cs.umass.edu, brian.a.jalaian.civ@mail.mil

Abstract

In this paper, we present a general framework for distilling expectations with respect to the Bayesian posterior distribution of a deep neural network classifier, extending prior work on the Bayesian Dark Knowledge framework. The proposed framework takes as input “teacher” and “student” model architectures and a general posterior expectation of interest. The distillation method performs an online compression of the selected posterior expectation using iteratively generated Monte Carlo samples. We focus on the posterior predictive distribution and expected entropy as distillation targets. We investigate several aspects of this framework including the impact of uncertainty and the choice of student model architecture. We study methods for student model architecture search from a speed-storage-accuracy perspective and evaluate down-stream tasks leveraging entropy distillation including uncertainty ranking and out-of-distribution detection .

1 INTRODUCTION

Deep learning models have shown promising results in the areas including computer vision, natural language processing, speech recognition, and more (Graves et al., 2013; Huang et al., 2016; Devlin et al., 2018). However, existing point estimation-based training methods for these models may result in predictive uncertainties that are not well calibrated, including the occurrence of confident errors.

While Bayesian inference can often provide more robust posterior predictive distributions compared to point

¹Our PyTorch implementation can be found at: <https://github.com/meetvadera/GPED>

estimation-based training, the integrals required to perform Bayesian inference in neural network models are well-known to be intractable. Monte Carlo methods provide one solution to represent neural network parameter posteriors as ensembles of networks, but this requires large amounts of both storage and compute time (Neal, 1996; Welling and Teh, 2011).

To help overcome these problems, Balan et al. (2015) introduced a model training method referred to as *Bayesian Dark Knowledge* (BDK). BDK attempts to compress (or distill) the Bayesian posterior predictive distribution induced by the full parameter posterior of a “teacher” network (represented via a set of Monte Carlo samples) into a significantly more compact “student” network. The major advantage of BDK is that the computational complexity of prediction at test time is drastically reduced compared to directly computing predictions via Monte Carlo averages over the set of teacher network samples (the teacher ensemble). As a result, such posterior distillation methods have the potential to be much better suited to learning models for deployment in resource constrained settings.

However, the posterior predictive distribution is not the only statistic of the posterior distribution that is of interest. Indeed, recent work including Wang et al. (2018) and Malinin et al. (2020) has investigated leveraging multiple statistics of ensembles (both general ensembles and Monte Carlo representations of Bayesian posteriors) for performing tasks that leverage uncertainty quantification and uncertainty decomposition including out-of-distribution detection and uncertainty-based ranking.

In this paper, we propose a Bayesian posterior distillation framework for the classification setting that generalizes the BDK approach by directly distilling general posterior expectations. We further generalize the BDK approach by proposing methods for efficiently searching the space of speed-storage-accuracy trade-offs for the student model, enabling more fine grained control over model size, test time, speed and predictive performance. The primary

empirical contributions of this work are (1) evaluating the distillation of the posterior predictive distribution and the posterior expected entropy across a range of models, data sets, and levels of uncertainty; (2) evaluating the impact of the student model architecture and architecture search methods on distillation performance; and (3) evaluating the utility of generalized expectation distillation through the study of down-stream tasks including out-of-distribution detection and uncertainty ranking that leverage entropy distillation. We show that distillation performance can be very sensitive to student model capacity and that the proposed architecture search methods effectively expose the space of speed-storage-accuracy trade-offs. We further show that our direct generalized posterior distillation framework outperforms an adaptation of the approach of Malinin et al. (2020) both on terms of distillation performance and in terms of several downstream tasks that leverage uncertainty quantification.

In the next section, we present background material and related work. In Section 3, we present the proposed framework. In Section 4, we present experiments and results. Additional details regarding data sets and experiments can be found in Appendix A, with supplemental results included in Appendix B.

2 BACKGROUND AND RELATED WORK

In this section, we present background and related work.

Bayesian Neural Networks: Let $p(y|\mathbf{x}, \theta)$ represent the probability distribution induced by a deep neural network classifier over classes $y \in \mathcal{Y} = \{1, \dots, C\}$ given feature vectors $\mathbf{x} \in \mathbb{R}^D$. The most common way to fit a model of this type given a data set $\mathcal{D} = \{(\mathbf{x}_i, y_i) | 1 \leq i \leq N\}$ is to use maximum conditional likelihood estimation, or equivalently, cross entropy loss minimization (or their penalized or regularized variants). However, when the volume of labeled data is low, there can be multiple advantages to considering a full Bayesian treatment of the model. Instead of attempting to find the single (locally) optimal parameter set θ_* according to a given criterion, Bayesian inference uses Bayes rule to define the posterior distribution $p(\theta|\mathcal{D}, \theta^0)$ over the unknown parameters θ given a prior distribution $P(\theta|\theta^0)$ with prior parameters θ^0 as seen in Equation 1.

$$p(\theta|\mathcal{D}, \theta^0) = \frac{p(\mathcal{D}|\theta)p(\theta|\theta^0)}{\int p(\mathcal{D}|\theta)p(\theta|\theta^0)d\theta} \quad (1)$$

$$\begin{aligned} p(y|\mathbf{x}, \mathcal{D}, \theta^0) &= \int p(y|\mathbf{x}, \theta)p(\theta|\mathcal{D}, \theta^0)d\theta \\ &= \mathbb{E}_{p(\theta|\mathcal{D}, \theta^0)}[p(y|\mathbf{x}, \theta)] \end{aligned} \quad (2)$$

Posterior Expectations and Uncertainty Quantification: For prediction problems in machine learning, the quantity of interest is typically not the parameter posterior itself, but the posterior predictive distribution $p(y|\mathbf{x}, \mathcal{D}, \theta^0)$ obtained from it as seen in Equation 2.

However, the posterior predictive distribution is not the only statistic of the posterior distribution that is of interest. The decomposition of posterior uncertainty has also received recent attention in the literature. For example, Depeweg et al. (2017) and Malinin et al. (2020) describe the decomposition of the entropy of the posterior predictive distribution (the *total uncertainty*) into *expected data uncertainty* and *knowledge uncertainty*. These three forms of uncertainty are related by the equation shown below:

$$\underbrace{\mathcal{I}[y, \theta|\mathbf{x}, \mathcal{D}]}_{\text{Knowledge Uncertainty}} = \underbrace{\mathcal{H}[\mathbb{E}_{p(\theta|\mathcal{D})}[p(y|\mathbf{x}, \theta)]]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{p(\theta|\mathcal{D})}[\mathcal{H}[p(y|\mathbf{x}, \theta)]]}_{\text{Expected Data Uncertainty}} \quad (3)$$

Total uncertainty, as the name suggests, measures the total uncertainty in a prediction. Expected data uncertainty measures the uncertainty arising from class overlap. Knowledge uncertainty corresponds to the conditional mutual information between labels and model parameters and measures the disagreement between different models in the posterior. However, it can be efficiently computed as the difference between total uncertainty and expected data uncertainty, both of which are (functions) of posterior expectations. In recent work, Wang et al. (2018) and Malinin et al. (2020) have leveraged this decomposition to explore a range of down-stream tasks that rely on uncertainty quantification and decomposition.

Approximate Inference Methods for Bayesian Neural Networks: The primary problem with applying Bayesian inference to neural network models is that the distributions $p(\theta|\mathcal{D}, \theta^0)$ and $p(y|\mathbf{x}, \mathcal{D}, \theta^0)$ are not available in closed form, so approximations are required. We briefly review Bayesian inference approximations including variational inference (VI) (Jordan et al., 1999) and Markov Chain Monte Carlo (MCMC) methods (Neal, 1996; Welling and Teh, 2011).

In VI, an auxiliary distribution $q_\phi(\theta)$ is defined to approximate the true parameter posterior $p(\theta|\mathcal{D}, \theta^0)$. The variational parameters ϕ are selected to minimize the Kullback-Leibler (KL) divergence between $q_\phi(\theta)$ and $p(\theta|\mathcal{D}, \theta^0)$. Hinton and Van Camp (1993) did early work applying VI to neural networks. Graves (2011) later presented a method based on stochastic VI with improved scalability. In the closely related family of expectation propagation (EP) methods (Minka, 2001), Soudry et al. (2014) present an online EP algorithm for neural networks with

the flexibility of representing both continuous and discrete weights. [Hernández-Lobato and Adams \(2015\)](#) present the probabilistic backpropagation (PBP) algorithm for approximate Bayesian learning of neural network models, which is an example of an assumed density filtering (ADF) algorithm that, like VI and EP, generally relies on simplified posterior densities.

The main drawback of VB, EP, and ADF is that they typically result in biased posterior estimates for complex posterior distributions. MCMC methods provide an alternative family of sampling-based posterior approximations that are unbiased, but are often computationally more expensive to use. MCMC methods allow for drawing a correlated sequence of samples $\theta_t \sim p(\theta|\mathcal{D}, \theta^0)$ from the parameter posterior. These samples can then be used to approximate the posterior predictive distribution as a Monte Carlo average as shown in Equation 4

$$p(y|\mathbf{x}, \mathcal{D}, \theta^0) \approx \frac{1}{T} \sum_{t=1}^T p(y|\mathbf{x}, \theta_t) \quad (4)$$

$$\theta_t \sim p(\theta|\mathcal{D}, \theta^0) \quad (5)$$

[Neal \(1996\)](#) addressed the problem of Bayesian inference in neural networks using Hamiltonian Monte Carlo (HMC) to provide a set of posterior samples. This method uses the full dataset when computing the gradient needed by HMC, which is problematic for larger data sets. While this scalability problem has largely been solved for mid-sized models by more recent methods such as stochastic gradient Langevin dynamics (SGLD) ([Welling and Teh \(2011\)](#)), the problem of needing to compute over a large set of samples when making predictions at test or deployment time remains.

Distribution Distillation: As noted above, MCMC-based approximations are expensive in terms of both computation and storage. Bayesian Dark Knowledge ([Balan et al. \(2015\)](#)) is precisely aimed at reducing the test-time computational complexity of Monte Carlo-based approximations for neural networks. In particular, the method uses SGLD to approximate the posterior distribution using a set of posterior parameter samples. These samples can be thought of as an ensemble of neural network models with identical architectures, but different parameters.

This posterior ensemble is used as the “teacher” in a distillation process that trains a single “student” model to match the teacher ensemble’s posterior predictive distribution ([Hinton et al. \(2015\)](#)). The major advantage of this approach is that it can drastically reduce the test time computational complexity of posterior predictive inference relative to using a Monte Carlo average computed using many samples. A shortcoming of this approach is that it only distills the posterior predictive distribution, and thus, loses access to other posterior statistics.

Ensemble distribution distillation (EnD²) is a closely related approach that aims to distill the collective outputs of the models in an ensemble into a neural network that predicts the parameters of a Dirichlet distribution ([Malinin et al. \(2020\)](#)). The goal is to preserve information about distribution of outputs of the ensemble in such a way that multiple statistics of the ensemble’s outputs can be efficiently approximated. Our goal in this paper is broadly similar, although we focus specifically on distilling much larger Monte Carlo posterior ensembles and we avoid the parametric distribution assumptions of ([Malinin et al. \(2020\)](#)) by directly distilling posterior expectations of interest.

Finally, we note that with the advent of Generative Adversarial Networks ([Goodfellow et al. \(2014\)](#)), there has also been work on generative models for approximating posterior sampling. [Wang et al. \(2018\)](#) and [Henning et al. \(2018\)](#) both propose methods for learning to generate samples that mimic those produced by SGLD. However, while these approaches may provide a speed-up relative to running SGLD itself, the resulting samples must still be used in a Monte Carlo average to compute a posterior predictive distribution in the case of Bayesian neural networks. This is again a potentially costly operation and is exactly the computation that distillation-based methods seek to accelerate.

Model Compression and Pruning: As noted above, the problem that Bayesian Dark Knowledge attempts to solve is reducing the test-time computational complexity of using a Monte-Carlo posterior to make predictions. In this work, we are particularly concerned with the issue of enabling test-time speed-storage-accuracy trade-offs. The relevant background material includes methods for network compression and pruning.

Previous work has shown that over-parameterized deep learning models tend to show much better learnability. Further, it has also been shown that such over-parameterized models rarely use their full capacity and can often be pruned back substantially without significant loss of generality. [Hassibi et al. \(1993\)](#) use the second-order derivatives of the objective function to guide pruning network connections. More recently, [Han et al. \(2015\)](#) introduced a weight magnitude-based technique for pruning connections in deep neural networks using simple thresholding. [Guo et al. \(2016\)](#); [Jin et al. \(2016\)](#); [Han et al. \(2016\)](#) introduce thresholding methods which also support restoration of connections.

A related line of work includes pruning neurons, channels or filters instead of individual weights. Pruning these components explicitly reduces the number of computations by making the networks smaller. Group LASSO-based methods have the advantage of turning the pruning problem

into a continuous optimization problem with a sparsity-inducing regularizer. Zhang and Ou (2018); Alvarez and Salzmann (2016); Wen et al. (2016); He et al. (2017) are some examples that use Group LASSO regularization at their core. Similarly, Louizos et al. (2017) use hierarchical priors to prune neurons instead of weights. An advantage of these methods over connection-based sparsity methods is that they directly produce smaller networks.

3 PROPOSED FRAMEWORK

In this section, we describe our proposed framework.

3.1 Generalized Posterior Expectations

As described in the previous section, different statistics derived from the posterior distribution $p(\theta|\mathcal{D}, \theta^0)$ may be useful in different data analysis tasks. We consider the general case of inferences that take the form of posterior expectations as shown in Equation 6 where $g(y, \mathbf{x}, \theta)$ is an arbitrary function of y , \mathbf{x} and θ .

$$\mathbb{E}_{p(\theta|\mathcal{D}, \theta^0)}[g(y, \mathbf{x}, \theta)] = \int p(\theta|\mathcal{D}, \theta^0)g(y, \mathbf{x}, \theta)d\theta \quad (6)$$

Important examples of functions $g(y, \mathbf{x}, \theta)$ include $g(y, \mathbf{x}, \theta) = p(y|\mathbf{x}, \theta)$, which results in the posterior predictive distribution $p(y|\mathbf{x}, \mathcal{D}, \theta^0)$ as used in Bayesian Dark Knowledge. The choice $g(y, \mathbf{x}, \theta) = \sum_{y'=1}^C p(y'|\mathbf{x}, \theta) \log p(y'|\mathbf{x}, \theta)$ yields the expected data uncertainty introduced in the previous section. The choice $g(y, \mathbf{x}, \theta) = p(y|\mathbf{x}, \theta)(1 - p(y|\mathbf{x}, \theta))$ results in the posterior marginal variance of class y given \mathbf{x} . We use the posterior predictive distribution and expected data uncertainty as examples throughout this work.

3.2 Generalized Posterior Expectation Distillation

Our goal is to learn to approximate posterior expectations $\mathbb{E}_{p(\theta|\mathcal{D}, \theta^0)}[g(y, \mathbf{x}, \theta)]$ under a given teacher model architecture using a given student model architecture. The method that we propose takes as input the teacher model $p(y|\mathbf{x}, \theta)$, the prior $p(\theta|\theta^0)$, a labeled data set \mathcal{D} , an unlabeled data set \mathcal{D}' , the function $g(y, \mathbf{x}, \theta)$, a student model $f(y, \mathbf{x}|\phi)$, an expectation estimator, and a loss function $\ell(\cdot, \cdot)$ that measures the error of the approximation given by the student model $f(y, \mathbf{x}|\phi)$.² Similar to Balan et al. (2015), we propose an online distillation method based on the use of the SGLD sampler. We describe all of the components of the framework in the sections below, and provide a complete description of the resulting method in Algorithm 1 (presented in the appendix).

²Note that $f(y, \mathbf{x}|\phi)$ denotes the student’s output probability for class y given input \mathbf{x} and parameters ϕ .

SGLD Sampler: The prior distribution over the parameters $p(\theta|\theta^0)$ is chosen to be a spherical Gaussian distribution with mean $\mu = 0$ and precision τ (we thus have $\theta^0 = [\mu, \tau]$). We define \mathcal{S} to be a minibatch of size M drawn from \mathcal{D} . θ_t denotes the parameter set sampled for the teacher model at sampling iteration t , while η_t denotes the step size for the teacher model at iteration t . The Langevin noise is denoted by $z_t \sim \mathcal{N}(0, \eta_t I)$. The sampling update for SGLD is given by: $\theta_{t+1} \leftarrow \theta_t + \Delta\theta_t$ where $\Delta\theta_t$ is defined as:

$$\Delta\theta_t = \frac{\eta_t}{2} \left(\nabla_{\theta} \log p(\theta|\theta^0) + \frac{N}{M} \sum_{i \in \mathcal{S}} \nabla_{\theta} \log p(y_i|x_i, \theta_t) \right) + z_t \quad (7)$$

Distillation Procedure: For the distillation learning procedure, we make use of a secondary unlabeled data set $\mathcal{D}' = \{\mathbf{x}_i | 1 \leq i \leq N'\}$. This data set could use feature vectors from the primary data set \mathcal{D} , or a larger data set. We note that due to autocorrelation in the sampled teacher model parameters θ_t , we may not want to run a distillation update for every Monte Carlo sample drawn. We thus use two different iteration indices: t for SGLD iterations and s for distillation iterations.

On every distillation step s , we sample a minibatch \mathcal{S}' from \mathcal{D}' of size M' . For every data case i in \mathcal{S}' , we update an estimate \hat{g}_{yis} of the posterior expectation using the most recent parameter sample θ_t , obtaining an updated estimate $\hat{g}_{yis+1} \approx \mathbb{E}_{p(\theta|\mathcal{D}, \theta^0)}[g(y, \mathbf{x}_i, \theta)]$ (we discuss update schemes in the next section). Next, we use the minibatch of examples \mathcal{S}' to update the student model. To do so, we take a step $\phi_{s+1} \leftarrow \phi_s + \alpha_s \Delta\phi_s$ in the gradient direction of the regularized empirical risk of the student model as shown below where α_s is the student model learning rate at step s , $R(\phi)$ is the regularizer, and λ is the regularization hyper-parameter. We next discuss the estimation of the expectation targets \hat{g}_{yis} .

$$\Delta\phi_s = \frac{N'}{M'} \sum_{i \in \mathcal{S}'} \sum_{y \in \mathcal{Y}} \nabla_{\phi} \ell(\hat{g}_{yis+1}, f(y, \mathbf{x}_i|\phi_s)) + \lambda \nabla_{\phi} R(\phi_s) \quad (8)$$

Expectation Estimation: Given an explicit collection of posterior samples $\theta_1, \dots, \theta_s$, the standard Monte Carlo estimate of $\mathbb{E}_{p(\theta|\mathcal{D}, \theta^0)}[g(y, \mathbf{x}, \theta)]$ is simply $\hat{g}_{yis} = (1/S) \sum_{j=1}^s g(y, \mathbf{x}_i, \theta_j)$. However, this estimator requires retaining the sequence of samples $\theta_1, \dots, \theta_s$, which may not be feasible in terms of storage cost. Instead, we consider the application of an online update function. We define $m_{i,s}$ to be the count of the number of times data case i has been sampled up to and including distillation iteration s . An online update function

$U(\hat{g}_{yis}, \theta_t, m_{is})$ takes as input the current estimate of the expectation, the current sample of the model parameters, and the number of times data case i has been sampled, and produces an updated estimate of the expectation \hat{g}_{yis+1} . Below, we define two different versions of the function. $U_s(\hat{g}_{yis}, \theta_t, m_{is})$, updates \hat{g}_{yis} using the current sample only, while $U_o(\hat{g}_{yis}, \theta_t, m_{is})$ performs an online update equivalent to a full Monte Carlo average.

$$U_s(\hat{g}_{yis}, \theta_t, m_{is}) = g(y, \mathbf{x}_i, \theta_t) \quad (9)$$

$$U_o(\hat{g}_{yis}, \theta_t, m_{is}) = \frac{1}{m_{is+1}} (m_{is} \cdot \hat{g}_{yis} + g(y, \mathbf{x}_i, \theta_t)) \quad (10)$$

We note that both update functions provide unbiased estimates of $\mathbb{E}_{p(\theta|\mathcal{D}, \theta^0)}[g(y, \mathbf{x}, \theta)]$ after a suitable burn-in time B . The online update $U_o(\cdot)$ will generally result in lower variance in the estimated values of \hat{g}_{yis} , but it comes at the cost of needing to explicitly maintain the expectation estimates \hat{g}_{yis} across learning iterations, increasing the storage cost of the algorithm. It is worthwhile noting that the extra storage and computation cost required by U_o grows linearly in the size of the training set for the student. By contrast, the fully stochastic update is memoryless in terms of past expectation estimates, so the estimated expectations \hat{g}_{yis} do not need to be retained across iterations resulting in a substantial space savings.

General Algorithm and Special Cases: We show a complete description of the proposed method in Algorithm 1 in the appendix. The algorithm takes as input the teacher model $p(y|\mathbf{x}, \theta)$, the parameters of the prior $p(\theta|\theta^0)$, a labeled data set \mathcal{D} , an unlabeled data set \mathcal{D}' , the function $g(y, \mathbf{x}, \theta)$, the student model $f(y, \mathbf{x}|\phi)$, an online expectation estimator $U(\hat{g}_{yis}, \theta_t, m_{is})$, a loss function $\ell(\cdot, \cdot)$ that measures the error of the approximation given by $f(y, \mathbf{x}|\phi)$, a regularization function $R(\cdot)$ and regularization hyper-parameter λ , minibatch sizes M and M' , the thinning interval parameter H , the SGLD burn-in time parameter B and step size schedules for the step sizes η_t and α_s .

We note that the original Bayesian Dark Knowledge method is recoverable as a special case of this framework via the choices $g(y, \mathbf{x}, \theta) = p(y|\mathbf{x}, \theta)$, $\ell(p, q) = -p \log(q)$, $U = U_s$ and $p(y|\mathbf{x}, \theta) = f(y, \mathbf{x}, \phi)$ (i.e., the architecture of the student is selected to match that of the teacher). The original approach also uses a distillation data set \mathcal{D}' obtained from \mathcal{D} by adding randomly generated noise to instances from \mathcal{D} on each distillation iteration, taking advantage of the fact that the choice $U = U_s$ means that no aspect of the algorithm scales with $|\mathcal{D}'|$.

Our general framework allows for other trade-offs, including reducing the variance in the estimates of \hat{g}_{yis} at the cost of additional storage in proportion to $|\mathcal{D}'|$. We also

note that the loss function $\ell(p, q) = -p \log(q)$ and the choice $g(y, \mathbf{x}, \theta) = p(y|\mathbf{x}, \theta)$ are somewhat of a special case when used together as even when the full stochastic expectation update U_s is used, the resulting distillation parameter gradient is unbiased. To distill posterior expected entropy (e.g., expected data uncertainty), we set $g(y, \mathbf{x}, \theta) = \sum_{y \in \mathcal{Y}} p(y|\mathbf{x}, \theta) \log p(y|\mathbf{x}, \theta)$, $U = U_o$ and $\ell(h, h') = |h - h'|$.

3.3 Model Compression and Pruning

One of the primary motivations for the original Bayesian Dark Knowledge approach is that it provides an approximate inference framework that results in significant computational and storage savings at test time. However, a drawback of the original approach is that the architecture of the student is chosen to match that of the teacher. As we will show in Section 4, this will sometimes result in a student network that has too little capacity. On the other hand, if we plan to deploy the student model in a low resource compute environment, the teacher architecture may not meet the specified computational constraints. In either case, we need a general approach for selecting an architecture for the student model.

To begin to explore this problem, we consider two basic approaches to choosing student model architectures that enable trading off test time inference speed and storage for accuracy (or more generally, lower distillation loss). A helpful aspect of the distillation process relative to a de novo architecture search problem is that the architecture of the teacher model is available as a starting point. As a first approach, we consider wrapping the proposed GPED algorithm with an explicit search over a set of student models that are “close” to the teacher. Specifically, we consider a search space obtained by starting from the teacher model and applying a width multiplier to the width of every fully connected layer and a kernel multiplier to the number of kernels in every convolutional layer. While this search requires exponential time in the number of layers, it provides a baseline for evaluating other methods.

As an alternative approach with better computational complexity, we leverage the regularization function $R(\phi)$ included in the GPED framework to prune a large initial network using group ℓ_1/ℓ_2 regularization (Zhang and Ou, 2018; Wen et al., 2016). To apply this approach, we first must partition the parameters in the parameter vector ϕ across K groups \mathcal{G}_k . The form of the regularizer is $R(\phi) = \sum_{k=1}^K (\sum_{j \in \mathcal{G}_k} \phi_j^2)^{1/2}$. As is well-established in the literature, this regularizer causes all parameters in a group to go to zero simultaneously when they are not needed in a model. To use it for model pruning for a unit in a fully connected layer, we collect all of that unit’s inputs into a group. Similarly, we collect all of the

incoming weights for a particular channel in a convolution layer together into a group. If all incoming weights associated with a unit or a channel have magnitude below a small threshold ϵ , we can explicitly remove them from the model, obtaining a more compact architecture. We also fine-tune our models after pruning.

Finally, we note that any number of weight compressing, pruning, and architecture search methods could be combined with the GPED framework. Our goal is not to exhaustively compare such methods, but rather to demonstrate that GPED is sensitive to the choice of student model to highlight the need for additional research on the problem of selecting student model architectures.

4 EXPERIMENTS AND RESULTS

In this section, we present experiments and results evaluating the proposed approach using multiple data sets, posterior expectations, teacher model architectures, student model architectures, basic architecture search methods, and multiple down-stream tasks. We begin by providing an overview of the experimental protocols used.

4.1 Experimental Protocols

Data Sets: We use the MNIST (LeCun, 1998) and CIFAR10 (Krizhevsky et al., 2009) data sets as base data sets in our experiments. In the case of MNIST, posterior predictive uncertainty is very low, so we introduce two different modifications to explore the impact of uncertainty on distillation performance. The first modification is simply to subsample the data. The second modification is to introduce occlusions into the data set using randomly positioned square masks of different sizes, resulting in masking rates from 0% to 86.2%. For CIFAR10, we only use sub-sampling. Full details for both data sets and the manipulations applied can be found in Appendix A.1.

Models: We evaluate a total of three teacher models in this work: a three-layer fully connected network (FCNN) for MNIST matching the architecture used by Balan et al. (2015), a four-layer convolutional network for MNIST, and a five-layer convolutional network for CIFAR10. Full details of the teacher model architectures are given in Appendix A.2. For exhaustive search for student model architectures, we use the teacher model architectures as base models and search over a space of layer width multipliers K_1 and K_2 that can be used to expand sets of layers in the teacher models. A full description of the search space of student models can be found in Appendix A.2.

Distillation Procedures: We consider distilling both the posterior predictive distribution and the posterior entropy, as described in the previous section. For the posterior

Table 1: Results of posterior distillation when the student architecture is fixed to match the teacher architecture and base data sets are used with no sub-sampling or occlusion.

Model & Dataset	Teacher NLL	Student NLL	MAE (Entropy)
FCNN - MNIST	0.052	0.082	0.016
CNN - MNIST	0.022	0.053	0.016
CNN - CIFAR10	0.671	0.932	0.245

predictive distribution, we use the stochastic expectation estimator U_s while for entropy we experiment with both estimators. We allow $B = 1000$ burn-in iterations for MNIST and $B = 10000$ for CIFAR10, and total of $T = 10^6$ training iterations. The prior hyper-parameters, learning rate schedules and other parameters vary by data set or distillation target and are fully described in Appendix A.2.

4.2 Experiments

Experiment 1: Distilling Posterior Expectations For this experiment, we use the MNIST and CIFAR10 datasets without any subsampling or masking. For each dataset and model, we consider separately distilling the posterior predictive distribution and the posterior entropy. We fix the architecture of the student to match that of the teacher. To evaluate the performance while distilling the posterior predictive distribution, we use the negative log-likelihood (NLL) of the model on the test set. For evaluating the performance of distilling posterior entropy, we use the mean absolute difference between the teacher ensemble’s entropy estimate and the student model output on the test set. The results are given in Table 1. First, we note that the FCNN NLL results on MNIST closely replicate the results in Balan et al. (2015), as expected. We also note that the error in the entropy is low for both the FCNN and CNN architectures on MNIST. However, the student model fails to match the NLL of the teacher on CIFAR10 and the entropy MAE is also relatively high. In Experiment 2, we will investigate the effect of increasing uncertainty, while in Experiment 3 we will investigate the impact of student architectures.

Experiment 2: Robustness to Uncertainty We build on Experiment 1 by exploring methods for increasing posterior uncertainty on MNIST (sub-sampling and masking) and CIFAR10 (sub-sampling). We consider the cross product of four sub-sampling rates and six masking rates for MNIST and three sub-sampling rates for CIFAR10. We consider the posterior predictive distribution and posterior entropy distillation targets. For the posterior predictive distribution we report the negative

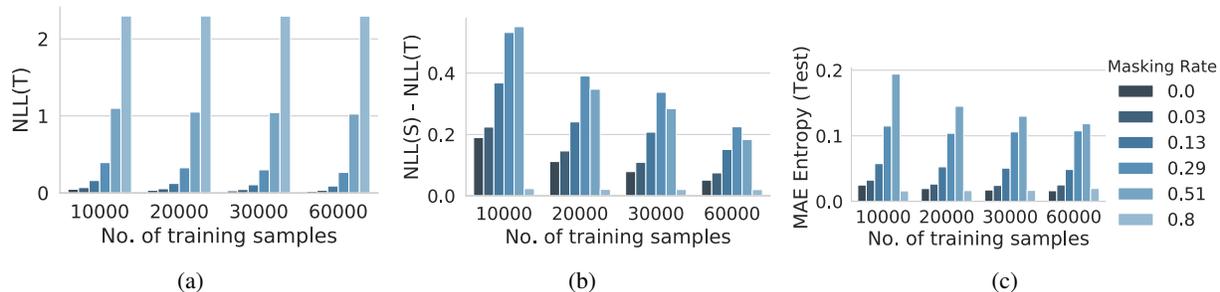


Figure 1: Distillation performance using CNNs on MNIST while varying data set size and masking rate. (a) Test negative log likelihood of the teacher posterior predictive distribution. (b) Difference in test negative log likelihood between student and teacher posterior predictive distribution estimates. (c) Difference between teacher and student posterior entropy estimates on test data set.

log likelihood (NLL) of the teacher, and the NLL gap between the teacher and student. For entropy, we report the mean absolute error between the teacher ensemble and the student. All metrics are evaluated on held-out test data. We also restrict the experiment to the case where the student architecture matches the teacher architecture, mirroring the Bayesian Dark Knowledge approach. In Figure 1 we show the results for the convolutional models on MNIST. The FCNN results are similar to the CNN results on MNIST and are shown in Figure 4 along with the CNN results on CIFAR10 in Figure 5 in Appendix B. In Appendix B we also provide a performance comparison between the U_o and U_s estimators while distilling posterior expectations.

As expected, the NLL of the teacher decreases as the data set size increases. We observe that changing the number of training samples has a similar effect on NLL gap for both CIFAR10 and MNIST. More specifically, for any fixed masking rate of MNIST (and zero masking rate for CIFAR10), we can see that the NLL difference between the student and teacher decreases with increasing training data. However, for MNIST we can see that the teacher NLL increases much more rapidly as a function of the masking rate. Moreover, the gap between the teacher and student peaks for moderate values of the masking rate. This fact is explained through the observation that when the masking rate is low, posterior uncertainty is low, and distillation is relatively easy. On the other hand, when the masking rate is high, the teacher essentially outputs the uniform distribution for every example, which is very easy for the student to represent. As a result, the moderate values of the masking rate result in the hardest distillation problem and thus the largest performance gap. For varying masking rates, we see exactly the same trend for the gap in posterior entropy predictions on MNIST. However, the gap for entropy prediction increases as a function of data set size for CIFAR10. Finally, as we would expect, the performance of distillation using the

U_o estimator is almost always better than that of the U_s estimator (see Appendix B).

The key finding of this experiment is that the quality of the approximations provided by the student model can significantly vary as a function of properties of the underlying data set. In the next experiment, we address the problem of searching for improved student model architectures.

Experiment 3: Student Model Architectures In this experiment, we compare exhaustive search to the group ℓ_1/ℓ_2 (group lasso) regularizer combined with pruning. For the pruning approach, we start with the largest student model considered under exhaustive search, and prune back from there using different regularization parameters λ , leading to different student model architectures. We present results in terms of performance versus computation time (estimated in FLOPs), as well as performance vs storage cost (estimated in number of parameters). As performance measures for the posterior predictive distribution, we consider accuracy and negative log likelihood. For entropy, we use mean absolute error. In all cases, results are reported on test data. We consider both fully connected and convolutional models.

Figure 2 shows results for the negative log likelihood (NLL) of the convolutional model on MNIST with masking rate 29% and 60,000 training samples. We select this setting as illustrative of a difficult case for posterior predictive distribution distillation. We plot NLL vs FLOPs and NLL vs storage for all points encountered in each search. The solid blue line indicates the Pareto frontier.

First, we note that the baseline student model (with architecture matching the teacher) from Experiment 2 on MNIST achieves an NLL of 0.469 at approximately 0.48×10^6 FLOPs and 0.03×10^6 parameters on this configuration of the data set. We can see that both methods for selecting student architectures provide a highly significant improvement over the baseline student architectures.

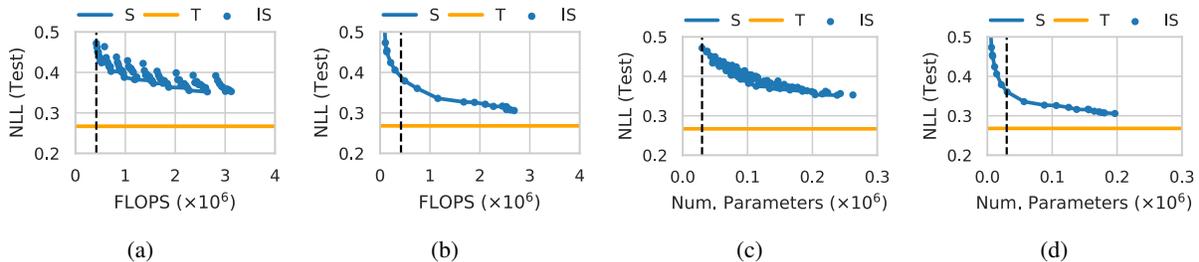


Figure 2: NLL-Storage-Computation tradeoff while using CNNs on MNIST with masking rate 29%. Test negative log likelihood of posterior predictive distribution vs FLOPS found using (a) exhaustive search and (b) group ℓ_1/ℓ_2 with pruning. Test negative log likelihood of posterior predictive distribution vs storage found using (c) exhaustive search and (d) group ℓ_1/ℓ_2 with pruning. The optimal student model for this configuration is obtained with group ℓ_1/ℓ_2 pruning. It has approximately $6.6\times$ the number of parameters and $6.4\times$ the FLOPS of the base student model. Notation: “S” - pareto frontier of the student models, “T” - Teacher, “IS” - Individual Student. The black dashed line denotes the FLOPS/number of parameters of the base student model having the same architecture as a teacher model.

Table 2: In-distribution Test set metrics comparison using U_s and largest student model obtained using width multiplier.

Model/ Dataset	NLL (Ensemble)	NLL (GPED)	NLL (EnD ²)	MAE Entropy (GPED)	MAE Entropy (EnD ²)
FCNN/ MNIST	0.362	0.408	0.415	0.069	0.105
CNN/ MNIST	0.269	0.296	0.321	0.086	0.106
CNN/ CIFAR10	0.799	0.859	0.907	0.146	0.328

On MNIST, the NLL is reduced to 0.30. Further, we can also see that the group ℓ_1/ℓ_2 approach is able to obtain much better NLL at the same computation and storage cost relative to the exhaustive search method. Lastly, the group ℓ_1/ℓ_2 method is able to obtain models on MNIST at less than 50% the computational cost needed by the baseline model with only a small loss in performance. Results for other models and distillation targets show similar trends and are presented in Appendix B. Additional experimental details are given in Appendix A.2.

In summary, the key finding of this experiment is that the capacity of the student model significantly impacts distillation performance, and student model architecture optimization methods are needed to achieve a desired speed-storage-accuracy trade-off.

Experiment 4: Uncertainty Quantification for Downstream Tasks As noted earlier, uncertainty quantification and decomposition is an important application of Bayesian posterior predictive inference. In this set of experiments, we evaluate our method on two downstream applications: out-of-distribution detection and uncertainty-

based ranking. We compare the GPED framework to the full Monte Carlo ensemble as well as to an adaptation of Ensemble Distribution Distillation (EnD²) (Malinin et al. 2020). In particular, Malinin et al. (2020) materialize a complete ensemble, which is not feasible in our case due to the large number of samples in the Bayesian ensemble ($\sim 10^5$ samples). We instead use Algorithm 1 with the Dirichlet log likelihood distillation loss used by Malinin et al. (2020) (see Appendix A.3 for EnD² implementation details). Additionally, we modify our student models to distill both the predictive distribution and expected data uncertainty in a single model.

Before assessing the performance of these methods on downstream tasks, we first compare their performance in terms of negative log likelihood and MAE on the posterior predictive distribution and expected data uncertainty distillation tasks. We use the same dataset augmentation as in the previous experiment. We compare the GPED and EnD² methods using U_o and U_s as well as for small and large model sizes. Note that for distilling entropy under our method in this section, we always use the U_o estimator. Wherever the U_s estimator is mentioned for our method in this section of experiments, it is only applied to distilling predictive means. In Table 2 we compare different distillation methods for different model-dataset combinations. These results correspond to the U_s estimator and the largest student model. As an illustration, we present joint and marginal expected data uncertainty distribution plots in Figure 15 that correspond to the results in Table 2. These figures show how GPED and EnD² compare against the Bayesian ensemble on a data case-by-data case basis. Additional results are presented in Tables 8-10 and Figure 14. The key result of these experiments is that the GPED framework consistently performs better than EnD² across all metrics on the test datasets.

Table 3: AUROC for OOD Detection using U_s and largest student model obtained using width multiplier.

Model & Train Data/ OOD Data	Uncertainty	Ensemble	GPED (ours)	EnD ²
FCNN-MNIST/ KMNIST	Total	0.929	0.867	0.816
	Knowledge	0.976	0.928	0.899
FCNN-MNIST/ notMNIST	Total	0.944	0.670	0.652
	Knowledge	0.990	0.762	0.681
CNN-MNIST/ KMNIST	Total	0.894	0.882	0.881
	Knowledge	0.956	0.932	0.952
CNN-MNIST/ notMNIST	Total	0.888	0.882	0.860
	Knowledge	0.946	0.934	0.939
CNN-CIFAR10/ TIM	Total	0.729	0.762	0.721
	Knowledge	0.796	0.808	0.792
CNN-CIFAR10/ LSUN	Total	0.790	0.779	0.747
	Knowledge	0.752	0.767	0.713

Out-of-distribution detection: OOD detection has garnered a lot of interest in the deep learning community as it is a practical challenge during deployment of deep models. In this experiment, we use measures of total uncertainty and knowledge uncertainty for detecting OOD inputs. OOD detection is a binary classification problem where we utilize a measure of uncertainty to classify an input as in-distribution or out-of-distribution based on a threshold. For our experiments, we use four OOD datasets: KMNIST (Clanuwat et al., 2018), notMNIST (Bulatov, 2011), TinyImageNet (TIM) (CS231N, 2017), and SVHN (Netzer et al., 2011). Additional experimental details are given in the Appendix A.4. We run our experiments for different combinations of models, in-distribution datasets, out-of-distribution datasets, model architectures, and estimators used for distilling the predictive distribution under the proposed framework as well as for the EnD² framework. We report example OOD detection results using the U_s estimator and the largest student model in Table 3. Our overall results show that GPED outperforms EnD² in 75% of cases across all experimental settings considered (additional results are given in Tables 11, 13 in Appendix B).

Uncertainty-Based Ranking: Another important application of Bayesian neural networks is ranking instances based on uncertainty. Such rankings are used in active learning and other human-in-the-loop decision systems to prioritize uncertain instances for labeling or analysis by human decision makers. This task is sensitive to the correct rank order of in-distribution instances by uncertainty level, whereas the OOD task is only sensitive to the existence of a threshold that separates in and out of distribution instances. To assess how well our distillation framework preserves the relative ranking between the inputs when compared to the full Bayesian ensemble, we compute the Normalized Discounted Cumulative Gain

Table 4: nDCG@20 out of 100 randomly selected test inputs using U_s estimator and largest student model. Results reported as mean \pm std. dev. over 500 trials.

Model & Data	Uncertainty	GPED (ours)	EnD ²
FCNN-MNIST	Total	0.954 ± 0.02	0.946 ± 0.021
	Knowledge	0.924 ± 0.03	0.941 ± 0.028
CNN-MNIST	Total	0.929 ± 0.034	0.916 ± 0.032
	Knowledge	0.888 ± 0.032	0.876 ± 0.045
CIFAR10	Total	0.935 ± 0.022	0.919 ± 0.027
	Knowledge	0.885 ± 0.033	0.889 ± 0.034

(nDCG) score (Järvelin and Kekäläinen, 2002) for total uncertainty and knowledge uncertainty. A higher nDCG score implies that the correct ranking of inputs is better preserved under the distillation framework. For our experiments, we assess nDCG@20. In Table 4, we report the nDCG scores using the U_s estimator and largest student model as example results. Overall, GPED outperforms EnD² in 91% of settings considered (additional ranking results are given in Tables 14, 16 in Appendix B).

5 CONCLUSIONS & FUTURE DIRECTIONS

We have presented a framework for distilling expectations with respect to the Bayesian posterior distribution of a deep neural network that significantly generalizes the Bayesian Dark Knowledge approach. Our results show that posterior distillation performance can be highly sensitive to the architecture of the student model, but that architecture search methods can identify student model architectures with improved speed-storage-accuracy trade-offs. We have also demonstrated that the proposed approach performs well on downstream tasks that leverage entropy distillation for uncertainty decomposition. There are many directions for future work including considering the distillation of a broader class of posterior statistics, developing more advanced architecture search methods, and applying the framework to larger models.

Acknowledgments

This work was partially supported by the US Army Research Laboratory under cooperative agreement W911NF-17-2-0196. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the US government.

References

- J. M. Alvarez and M. Salzmann. Learning the number of neurons in deep networks. In *NeurIPS*, 2016.
- A. K. Balan, V. Rathod, K. P. Murphy, and M. Welling. Bayesian dark knowledge. In *NeurIPS*, 2015.
- Y. Bulatov. notMNIST dataset. 2011.
- T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha. Deep learning for classical japanese literature. *arXiv:1812.01718*, 2018.
- S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of uncertainty for active learning and reliable reinforcement learning in stochastic systems. *ArXiv*, abs/1710.07283, 2017.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2018.
- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.
- A. Graves. Practical variational inference for neural networks. In *NeurIPS*, 2011.
- A. Graves, A. R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*. IEEE, 2013.
- Y. Guo, A. Yao, and Y. Chen. Dynamic network surgery for efficient dnns. In *NeurIPS*, 2016.
- S. Han, J. Pool, J. Tran, and W. J. Dally. Learning both weights and connections for efficient neural networks. In *NeurIPS*, 2015.
- S. Han, J. Pool, S. Narang, H. Mao, S. Tang, E. Elsen, B. Catanzaro, J. Tran, and W. J. Dally. Dsd: Regularizing deep neural networks with dense-sparse-dense training flow. *ArXiv*, abs/1607.04381, 2016.
- B. Hassibi, D. G. Stork, and G. J. Wolff. Optimal brain surgeon and general network pruning. *IEEE International Conference on Neural Networks*, 1993.
- Y. He, X. Zhang, and J. Sun. Channel pruning for accelerating very deep neural networks. *ICCV*, 2017.
- C. Henning, J. von Oswald, J. Sacramento, S. C. Surace, J.-P. Pfister, and B. F. Grewe. Approximating the predictive distribution via adversarially-trained hypernetworks. In *NeurIPS Bayesian Deep Learning Workshop*, 2018.
- J. M. Hernández-Lobato and R. Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *ICML*, 2015.
- G. Hinton and D. Van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *COLT*, 1993.
- G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.
- G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. *CVPR*, 2016.
- K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20: 422–446, 2002.
- X. Jin, X.-T. Yuan, J. Feng, and S. Yan. Training skinny deep neural networks with iterative hard thresholding methods. *ArXiv*, abs/1607.05423, 2016.
- M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.
- A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009.
- Y. LeCun. The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- C. Louizos, K. Ullrich, and M. Welling. Bayesian compression for deep learning. *ArXiv*, abs/1705.08665, 2017.
- A. Malinin, B. Mlodozieniec, and M. Gales. Ensemble distribution distillation. In *ICLR*, 2020.
- T. P. Minka. Expectation propagation for approximate Bayesian inference. In *UAI*, 2001.
- R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, 1996.
- Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. 2011.
- D. Soudry, I. Hubara, and R. Meir. Expectation backpropagation: Parameter-free training of multilayer neural networks with continuous or discrete weights. In *NeurIPS*, 2014.
- K.-C. Wang, P. Vicol, J. Lucas, L. Gu, R. Grosse, and R. Zemel. Adversarial distillation of Bayesian neural network posteriors. *arXiv:1806.10317*, 2018.
- M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.
- W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li. Learning structured sparsity in deep neural networks. In *NeurIPS*, 2016.
- Y. Zhang and Z. Ou. Learning sparse structured ensembles with stochastic gradient MCMC sampling and network pruning. *IEEE MLSP*, 2018.