# Deep Probabilistic Ensembles:
## Approximate Variational Inference through KL Regularization

Kashyap Chitta[1], Jose M. Alvarez[2], Adam Lesnikowski[2]
[1] Carnegie Mellon University | [2] NVIDIA

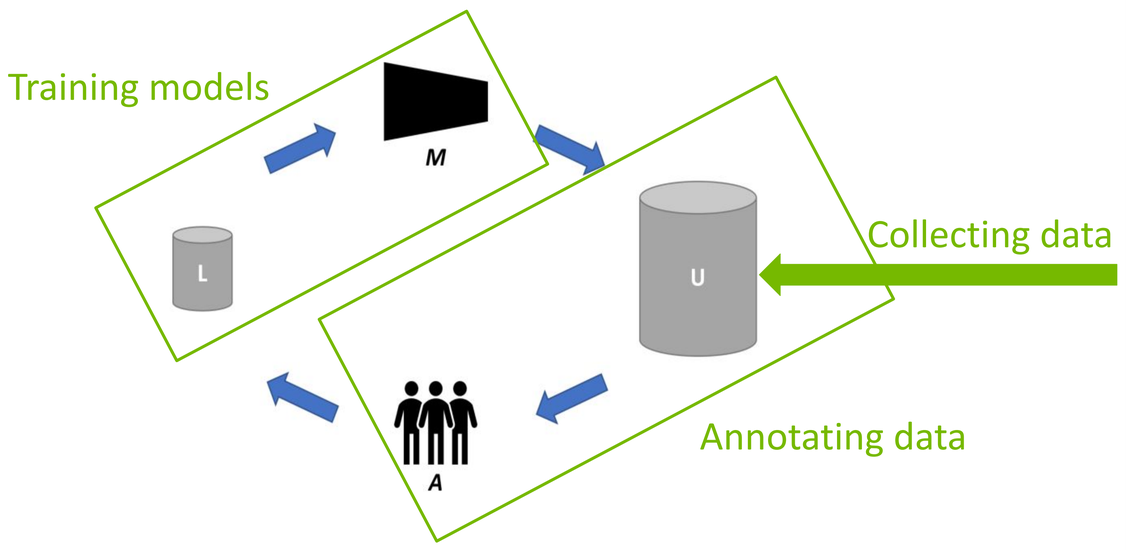
## BNNs

- Principled uncertainty estimates at all layers of network
- **Hard** to train (MCMC, variational inference)
- Requires drawing a **large number of samples** for reliable uncertainty estimates in practice

## Ensembles

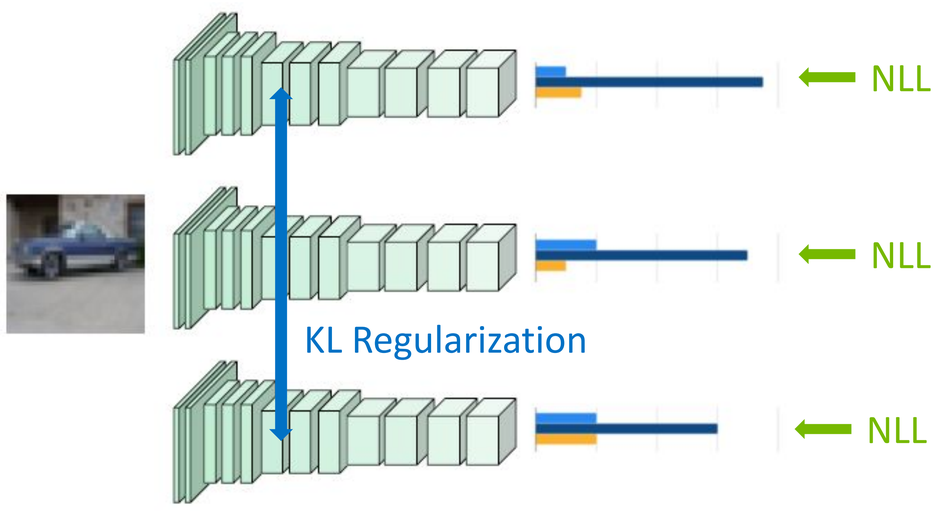
- Practical uncertainty estimates at layers where loss function is applied
- **Simple** to train (SGD + $L_2$/$L_1$ regularization)
- **Small number of ensembles** sufficient for reliable uncertainty estimates in practice

## Active Learning




- Cost of data annotation far exceeds cost of data collection, so can we improve models with smarter annotation?
- Yes! Only label those samples with **highest uncertainty** as per chosen acquisition function (eg: Entropy)

## Deep Probabilistic Ensembles




- Traditional variational inference: attempting to minimize NLL for *any generic network sampled from a BNN*, complicating optimization
- Proposed approach: **independent NLL loss** for each model in an ensemble, simplifying training
- KL Regularization applied to set of values taken by a parameter over all models

## KL Regularization

- Apply variational inference objective function to ensembles

$$-ELBO = \underbrace{KL(q(w)||p(w))}_{\text{KL Regularization}} \underbrace{- \mathbb{E}[\log p(x|w)]}_{\text{NLL}},$$

- Select Gaussian prior values for each weight (using initialization technique of He et al. [1])
- Penalty for each weight: analytical calculation

$$KL(q||p) = \frac{1}{2}\left(\log\frac{\sigma_q^2}{\sigma_p^2} + \frac{\sigma_p^2 + (\mu_q - \mu_p)^2}{\sigma_q^2} - 1\right)$$

- Penalty for layer: sum over all weights

$$\Omega^l = \sum_{i=1}^{n_i n_o w h}\left(\log\sigma_i^2 + \frac{2}{n_o w h \sigma_i^2} + \frac{\mu_i^2}{\sigma_i^2}\right)$$

## Experiments and Results

- Experimental setup: labels 'hidden away' initially and revealed to model through active learning
- Datasets: CIFAR-10, CIFAR-100

Table 1: Validation accuracies (in %) on an active learning task. DPEs give consistent improvements in results over both random sampling and standard ensembles on both datasets. Relative performance to the upper bound performance for that data sampling strategy is given in (brackets).

| Task | Data Sampling | Accuracy @8% | Accuracy @16% | Accuracy @32% |
|---|---|---|---|---|
| CIFAR-10 | Random | 80.60 (84.66) | 86.80 (91.18) | 91.08 (95.67) |
| | Ensemble | 82.41 (86.56) | 90.05 (94.59) | 94.13 (98.87) |
| | DPE (Ours) | **82.88 (87.06)** | **90.15 (94.70)** | **94.33 (99.09)** |
| CIFAR-100 | Random | 39.57 (50.18) | 54.92 (69.64) | 66.65 (84.51) |
| | Ensemble | 40.49 (51.34) | 56.89 (72.14) | 69.68 (88.36) |
| | DPE (Ours) | **40.87 (51.83)** | **56.94 (72.20)** | **70.12 (88.92)** |

Table 2: **CIFAR-10:** Comparison of our proposed approach to active learning baselines for learning with limited labels. DPEs show state-of-the-art performance. Relative performance refers to the performance for each method using a limited labeling budget compared to the performance using the whole dataset.

| Method | Accuracy @20% | Accuracy @100% | Relative Performance |
|---|---|---|---|
| Core-set [15] | 74% | 90% | 82.2% |
| Ensemble [1] | 85% | **95.5%** | 89.0% |
| Deterministic Network | 87.5% | 94.4% | 92.7% |
| DPE (Ours) | **92%** | 95.2% | **96.3%** |

## Current Work: Segmentation

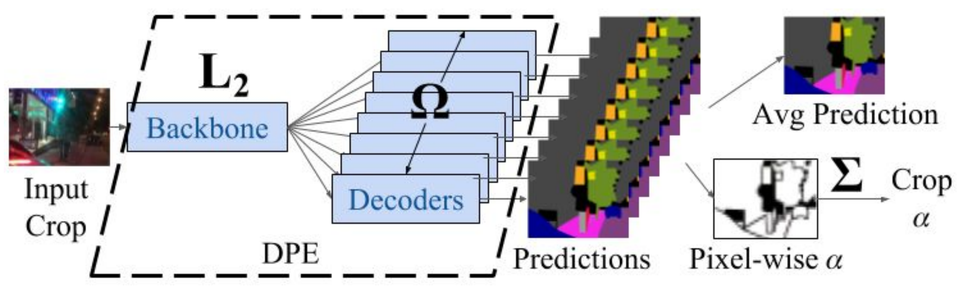



Table 7. **BDD100k:** mIoU (in %) comparing the proposed approach to standard ensembles for active segmentation. Initial 3.3k crops are randomly sampled. In our setup, DPEs improve upon the mIoU of ensembles by up to 1%.

| Data Sampling | 6.7k (8%) | 13.4k (16%) | 26.9k (32%) |
|---|---|---|---|
| Random | 45.37 | 46.92 | 48.41 |
| Ensemble | 45.60 | 47.62 | 49.14 |
| **Ours** (DPE) | **45.80** | **48.10** | **50.12** |

## References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.* In ICCV, 2015.

Carnegie Mellon University NVIDIA