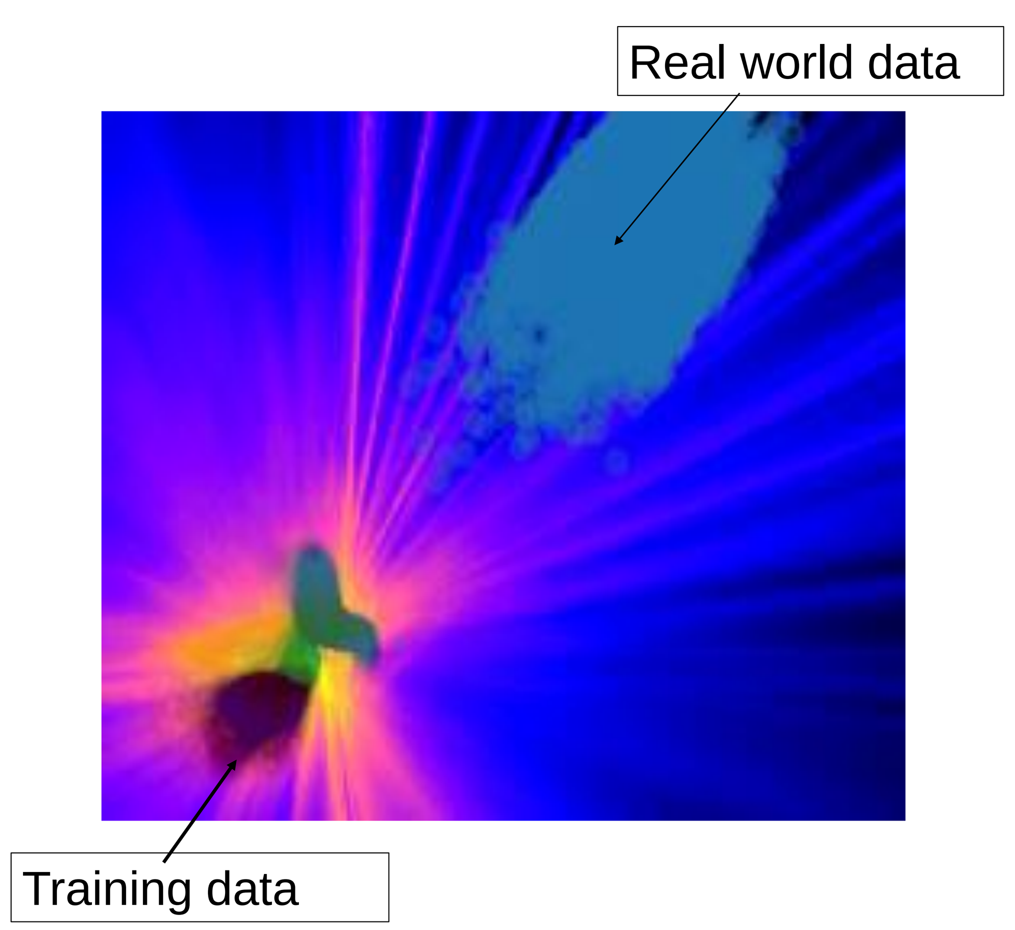
Accurate Layerwise Interpretable Competence Estimation (ALICE)

Vickram Rajendran, William LeVine

The Johns Hopkins University Applied Physics Laboratory

Overview

- Models are trained in isolation and then deployed on some real-world domain.
  - How can we quantify how well a model is performing when we don't have access to the ground truth labels?
  - Test set metrics tell us how well the model generalizes *to the test set*, not on any individual *point in the real world.*
  - A machine learning model "doing well" depends on the particular use-case...
  - The model's scores are often overconfident and hard to interpret.
- We need a way to know when the model is performing competently without having access to ground truth.
  - This should encompass ALL use-cases and work on ANY trained model.



- We want an accurate uncertainty estimate that generalizes to all classifiers and is interpretable.

Competence

Defining Competence

- Some problems have different risk and acceptance thresholds.
- Some problems have different performance metrics (Cross Entropy/IOU/PR).
- We model this by defining competence to be the probability that the value of some "error function" $\mathcal{E}$ is less than some "acceptance threshold" δ.
- A model is ***competent*** on a point if the competence is greater than some "risk threshold" ϵ.
- The error function, delta, and epsilon are all modular.

$$p(\mathcal{E}(f(x), \hat{f}(x)) < \delta | x, \hat{f}) > \epsilon$$

Evaluating Competence Estimators

- Competence is known when ground truth is available – this makes competence estimation a binary classification problem.
- We evaluate competence estimators with two metrics:
  - **Mean Average Precision:** This is the standard binary classification average precision metric for competence estimation at a particular delta, averaged over 100 deltas.

Model	Accuracy	Softmax	TrustScore	Ablated ALICE	ALICE
MLP (U)	.121 ± .048	.0486 ± .015	.505 ± .27	.0538 ± .031	**.999 ± .0015**
RF (U)	.563 ± .078	.824 ± .16	.504 ± .33	.290 ± .322	**.999 ± .0011**
SVM (O)	.258 ± .023	.200 ± .16	.215 ± .12	.252 ± .16	**.981 ± .028**
VGG16 (U)	.0878 ± .0076	.899 ± .014	.292 ± .049	.0369 ± .0041	**.913 ± .012**
VGG16 (W)	.498 ± .012	.975 ± .013	.604 ± .104	.0863 ± .0071	**.978 ± .0082**
VGG16 (O)	.282 ± .15	.659 ± .024	.665 ± .0080	.257 ± .018	**.738 ± .019**

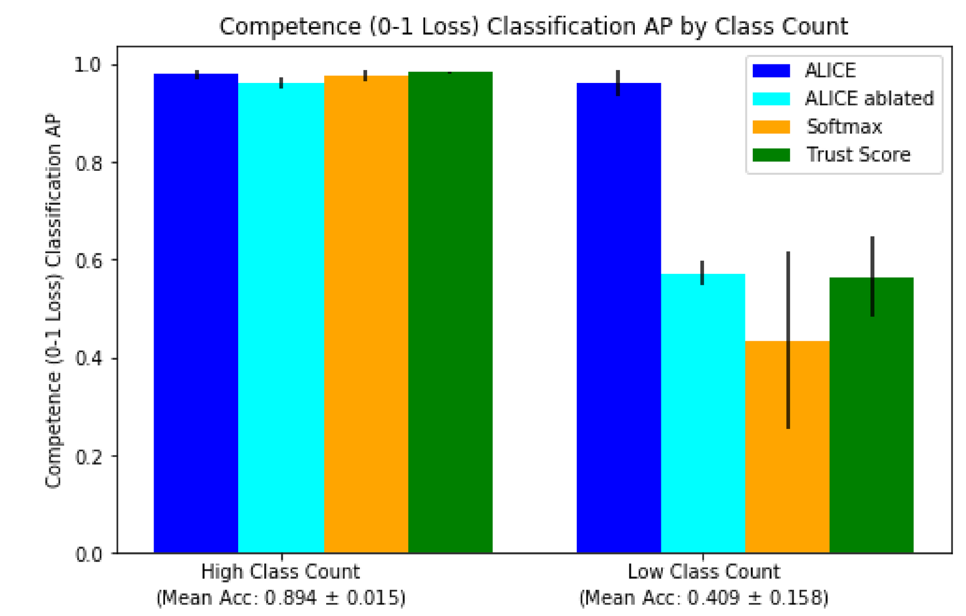
Model	Accuracy	Softmax	TrustScore	ALICE
SVM (RBF)	.147 ± .032	.394 ± .066	.361 ± .046	**.985 ± .011**
SVM (Poly)	.988 ± .007	.999 ± .0018	.990 ± .0045	**.998 ± 0.0012**
SVM (Linear)	.971 ± .011	1.00 ± .00065	.994 ± .0037	**.999 ± .0013**
RF	.928 ± .013	.996 ± .0016	.956 ± .012	**.999 ± .00034**
MLP (5 Iterations)	.158 ± .056	.384 ± .11	.746 ± .049	**.992 ± .015**
MLP (200 Iterations)	.925 ± .017	.986 ± .0069	.985 ± .012	**.997 ± .0027**
LR	.946 ± .017	.995 ± .0025	.989 ± .0051	**.998 ± .0015**

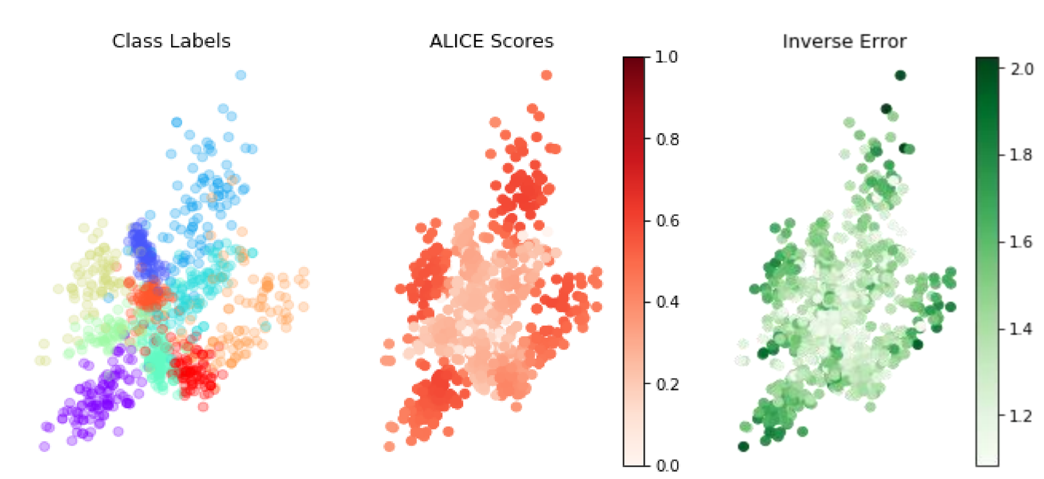
  - **Calibration**: We bin the ALICE scores into ten bins and compute the average competence of each of the bins.

Results

Accurate

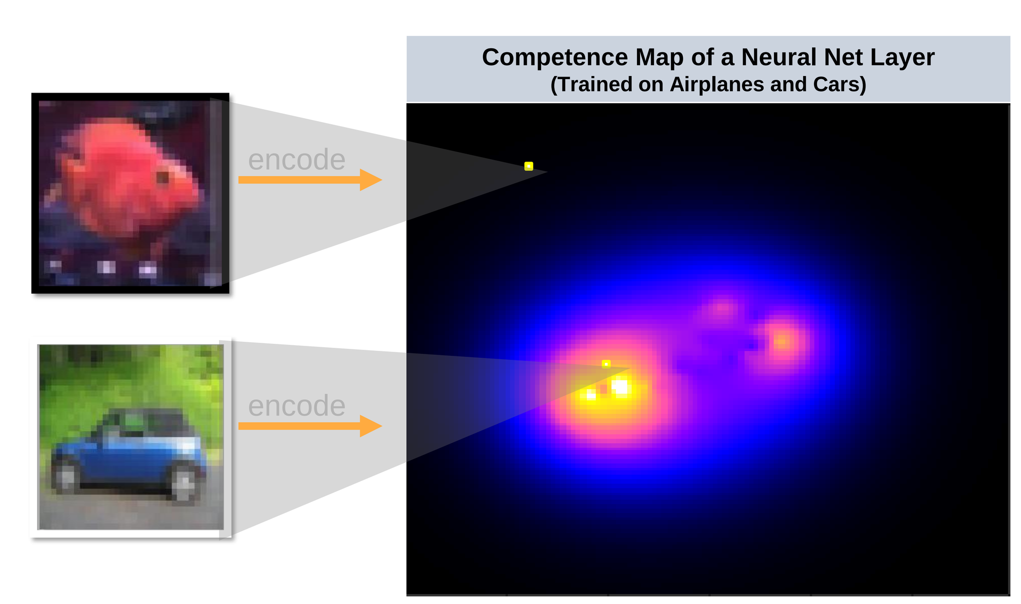
- ALICE accurately predicts which points are competent, regardless of type of uncertainty.





Layerwise

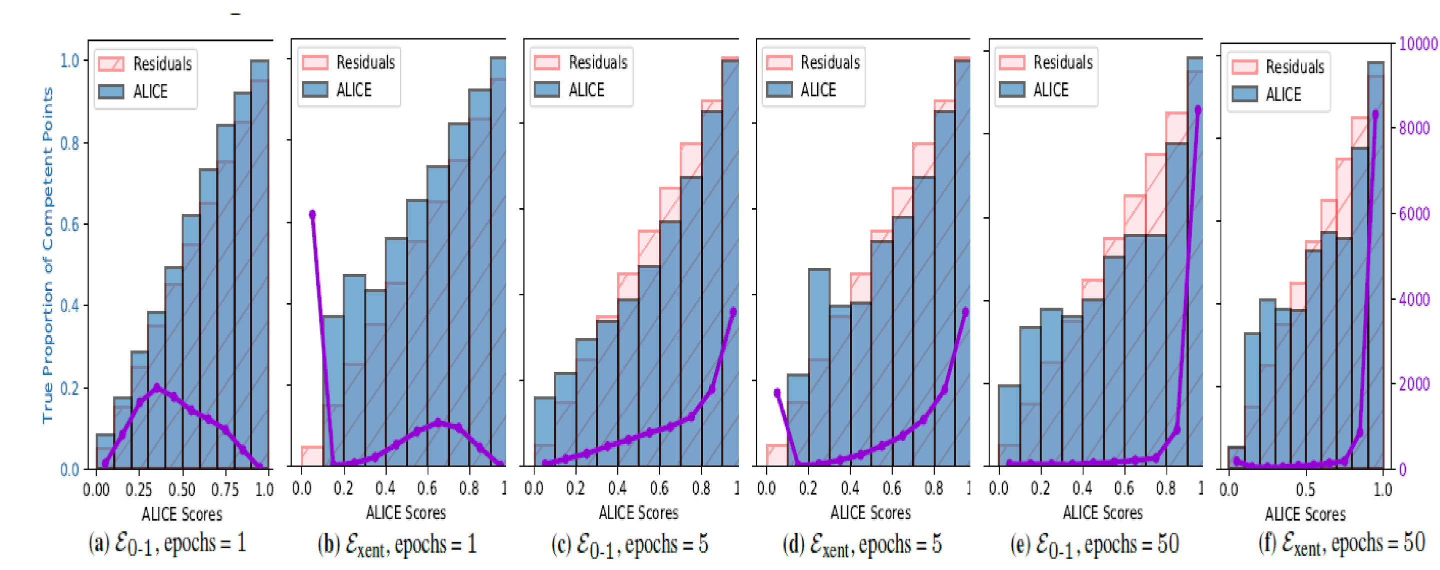
- ALICE can be performed at any layer of deep models.



Results

Interpretable

- ALICE scores are automatically calibrated at all stages of training.



Competence Estimation

- We approximate competence by estimating the distributional, model, and data uncertainty of the model on a particular point.
- **Distributional:** How similar is this new data point to data the model has seen before?
- **Model:** How well matched is the model to data like this new data point?
- **Data:** How intrinsically difficult is it to classify this new data point, based entirely on the data?

$$p(\mathcal{E}(f(x), \hat{f}(x)) < \delta | x, \hat{f})$$
$$\geq p(D|x)\Big[\sum_{c_j \in C} \mathcal{I}(\mathcal{E} < \delta) p(c_j | x, D)\Big]$$

- We approximate each of these terms in order to compute the ALICE score.

Can we determine what a classification model really knows?