Failure Prediction by Confidence Estimation of Uncertainty-Aware Dirichlet Networks

Theodoros Tsiligkaridis, PhD

MIT Lincoln Lab - Artificial Intelligence Technology Group

June 6-11, 2021

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

ICASSP2021

Research was sponsored by the United States Air Force Research Laboratory and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. 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 United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Motivation

- **Deep neural networks have achieved SOTA performance in image classification, object detection, speech recognition, etc.**
- **Safety is a concern when DNN systems are deployed in the real world**
- **When are neural networks likely to make errors? Important for high-risk applications such as healthcare, autonomous driving, and cybersecurity**

Source: https://techinsight.com.vn/en/when-health-care-moves-online-many-patients-are-left-behind/

Source: https://techinsight.com.vn/en/5-ways-artificial-intelligence-is-impacting-the-automotive-industry/

Prior Work on Failure Prediction

- **Hendrycks & Gimpel (2017): baseline based on *maximum class probability (MCP)***
  - **inherently flawed since it assigns high confidence scores even for failure cases**
- **Jiang et al (2018): *Trust Score***
  - **measures agreement between classifier and modified nearest-neighbor classifier on test example**
  - **not scalable since nearest-neighbor computations are expensive for large datasets**
  - **use of metric is limited as local distances in high dimensions are less meaningful**
- **Corbiere et (2019): *true class probability (TCP) learning***
  - **TCP attractive candidate metric for failure prediction**
  - **Estimate TCP using an unconstrained confidence network**
  - **Performance hinges on how well TCP scores can be learnt**

Prior work on uncertainty-aware learning

- It is known that conventional DNNs provide poorly calibrated scores and overconfident predictions
- Bayesian neural networks: learn weight distributions and estimate posterior predictive distributions by approximate integration
  - Variational inference (Blundell et al (2015), Kingma et al. (2015))
  - Laplace approximation (MacKay (1992), Ritter et al. (2018))
  - Expectation propagation (Hernandez-Lobato & Adams (2015), Sun et al. (2017))
  - Hamiltonian Monte Carlo (Chen et al. (2014))
- Dirichlet networks: explicit modeling of predictive distribution on probability simplex
  - With out-of-distribution training (Malinin & Gales (2019))
  - Within-distribution training only & regularization (Sensoy et al (2018), Tsiligkaridis (2020))

Uncertainty-Aware Learning

- **Explicit Dirichlet prior on class composition vectors** $\mathbf{p}_i \sim f(\cdot|\mathbf{x}_i, \boldsymbol{\theta})$

- **Predictive uncertainty of model**

$$P(y=j|\mathbf{x}^*, \mathcal{D}) = \int P(y=j|\mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$$
$$= \int \int P(y=j|\mathbf{p}) f(\mathbf{p}|\mathbf{x}^*, \boldsymbol{\theta}) d\mathbf{p} \cdot p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$$

- **Assume proper regularization control and large training set**

$$f(\mathbf{p}|\mathbf{x}^*, \mathcal{D}) = \int f(\mathbf{p}|\mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \approx f(\mathbf{p}|\mathbf{x}^*, \bar{\boldsymbol{\theta}})$$

- **Dirichlet model governed by concentration parameters** $\boldsymbol{\alpha} = (\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_K})$

$$f(\mathbf{p}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1}, \quad \mathbf{p} \in \mathcal{S} \qquad \alpha_0 = \sum_k \alpha_k$$

Training Uncertainty-Aware Dirichlet Networks

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{F}_i + \lambda \mathcal{R}_i$$

Minimize upper bound to Bayes risk of prediction error in L_∞ space

$$\mathcal{F}_i = \left(\mathbb{E}_{\mathbf{p}_i \sim f(\cdot;\boldsymbol{\alpha}_i)}\left[\sum_k (y_{ik} - p_{ik})^p\right]\right)^{1/p}$$
$$\geq \mathbb{E}_{\mathbf{p}_i \sim f(\cdot;\boldsymbol{\alpha}_i)}[\|\mathbf{y}_i - \mathbf{p}_i\|_p] \geq \mathbb{E}_{\mathbf{p}_i \sim f(\cdot;\boldsymbol{\alpha}_i)}[\|\mathbf{y}_i - \mathbf{p}_i\|_\infty]$$

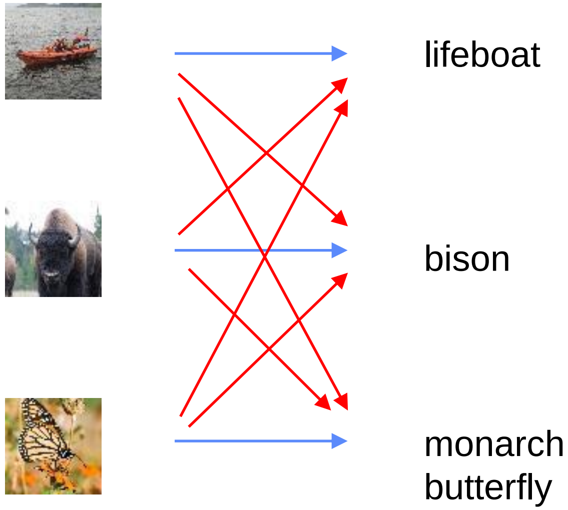
$$\mathcal{F}_i \stackrel{\text{def}}{=} \frac{1}{\mu(\alpha_{i0})^{1/p}}\left(\mu\left(\sum_{k\neq c_i} \alpha_{ik}\right) + \sum_{k\neq c_i} \mu(\alpha_{ik})\right)^{1/p}$$

where $\mu(\alpha) \stackrel{\text{def}}{=} \Gamma(\alpha+p)/\Gamma(\alpha).$

$$\mathcal{R}_i \stackrel{\text{def}}{=} \frac{1}{2}\sum_{k\neq c_i} (\alpha_{ik} - 1)^2 [J(\tilde{\boldsymbol{\alpha}}_i)]_{kk}$$

$$\tilde{\boldsymbol{\alpha}}_i = (1 - \mathbf{y}_i) \odot \boldsymbol{\alpha}_i + \mathbf{y}_i$$
$$J(\tilde{\boldsymbol{\alpha}}) = -\mathbb{E}_{\mathbf{p}\sim f(\cdot;\tilde{\boldsymbol{\alpha}})}[\nabla^2 \log f(\mathbf{p};\tilde{\boldsymbol{\alpha}})]$$
$$= \operatorname{diag}(\{\psi^{(1)}(\tilde{\alpha}_{ik})\}_k) - \psi^{(1)}(\tilde{\alpha}_{i0})\mathbf{1}_{K\times K}$$

Minimize information associated with incorrect outcomes



Model Confidence (TCP) Distributions

True class probability (TCP)

$$c^*(\mathbf{x}_i, c_i) = \mathbb{E}_{\mathbf{p}\sim f(\cdot|\mathbf{x}_i;\bar{\boldsymbol{\theta}})}[P(y = c_i|\mathbf{p})] = \frac{\alpha_{i,c_i}}{\alpha_{i,0}}$$

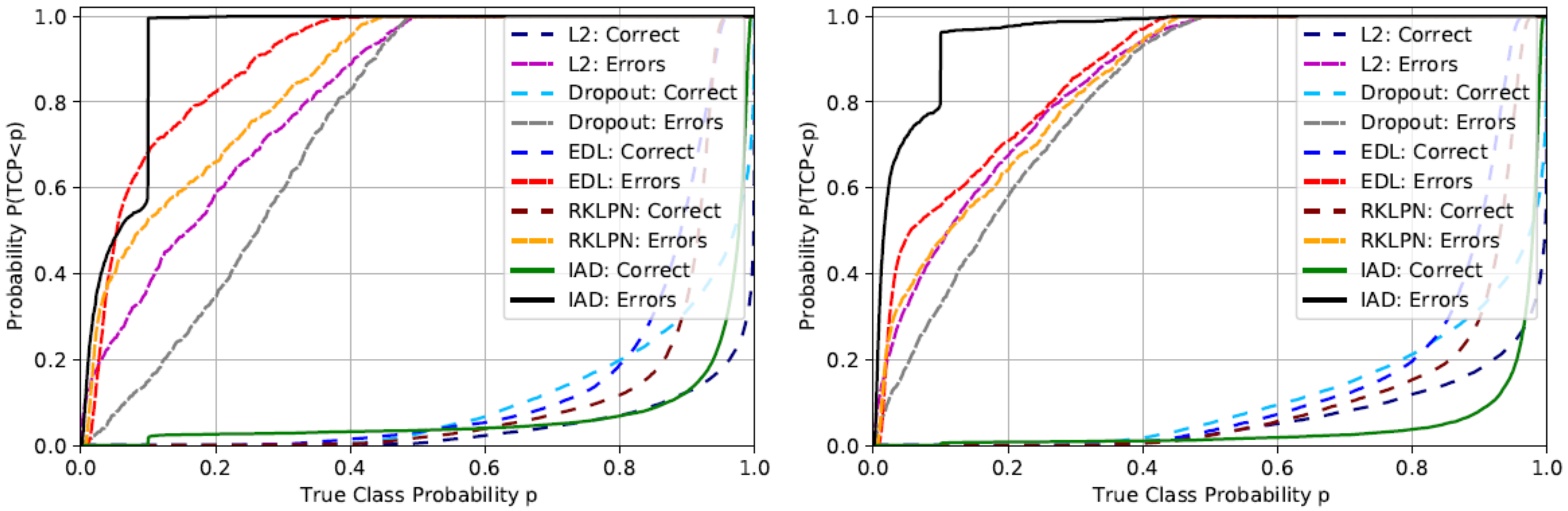


Fig. 1: Empirical cumulative density function (CDF) of TCP scores for various deep learning methods on Fashion-MNIST (left) and CIFAR-10 (right) test sets.

Learning Model Confidence

$$\mathcal{L}_c(\boldsymbol{\theta}_c) = \frac{1}{N}\sum_{i=1}^{N}\Big\{\mathbb{1}_{\{\hat{y}_i = y_i\}}\Big(\mathrm{MSE}(\mathbf{x}_i;\boldsymbol{\theta}_c) + \lambda_c \boxed{\phi(\mathbf{x}_i, -M, t_c;\boldsymbol{\theta}_c)}\Big) + \boxed{\zeta}\mathbb{1}_{\{\hat{y}_i \neq y_i\}}\left(\mathrm{MSE}(\mathbf{x}_i;\boldsymbol{\theta}_c) + \lambda_c \boxed{\phi(\mathbf{x}_i, M, t_e;\boldsymbol{\theta}_c))}\right\}$$

Constraints

Failure case weighting

Mean-square error function

$$\mathrm{MSE}(\mathbf{x}_i;\boldsymbol{\theta}_c) = \left(\hat{c}(\mathbf{x}_i,\boldsymbol{\theta}_c) - c^*(\mathbf{x}_i, c_i)\right)^2$$

Constraint function

$$\phi(\mathbf{x}_i, M, t;\boldsymbol{\theta}_c) = \sigma(M(\hat{c}(\mathbf{x}_i,\boldsymbol{\theta}_c) - t))$$

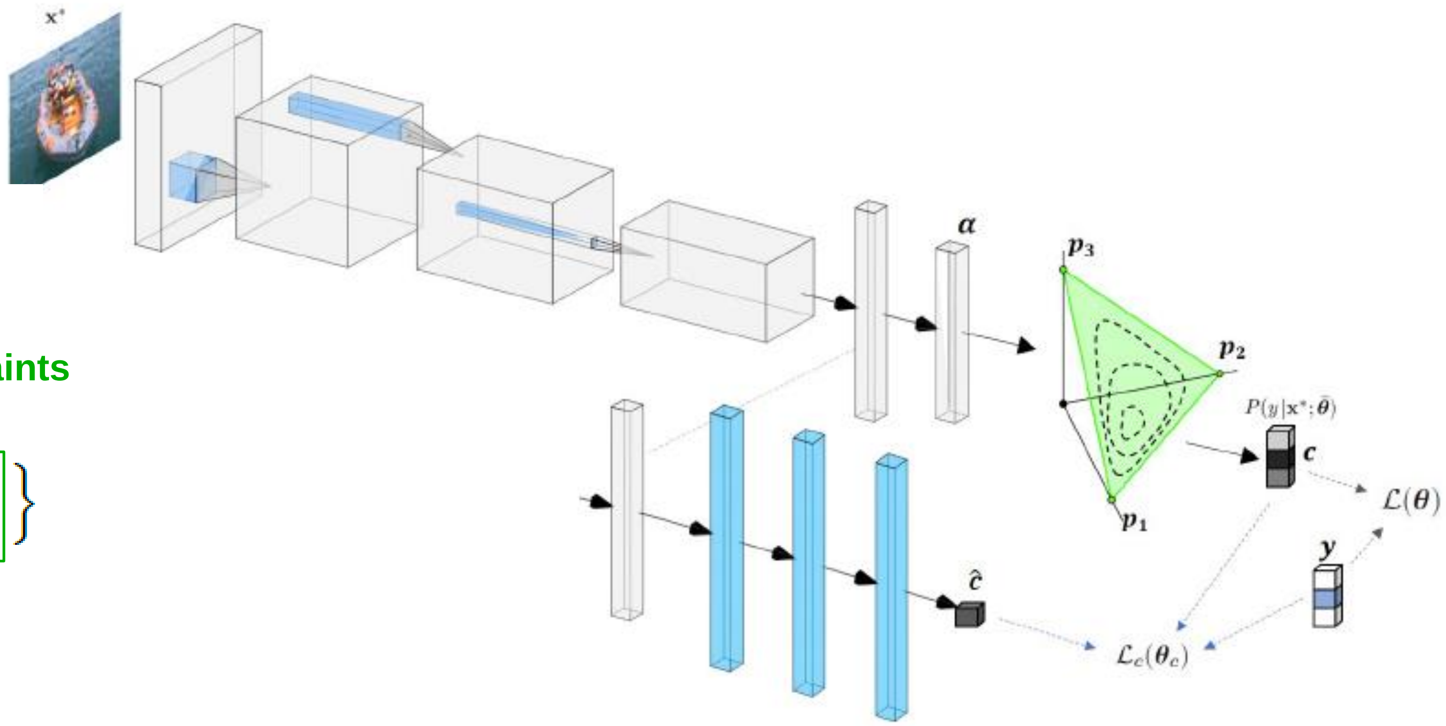


Fig. 2: The constrained confidence network (blue) transfers knowledge from the primary classification network (gray) to learn confidence scores $\hat{c}(\mathbf{x}, \boldsymbol{\theta}_c)$.

Results

ICASSP 2021

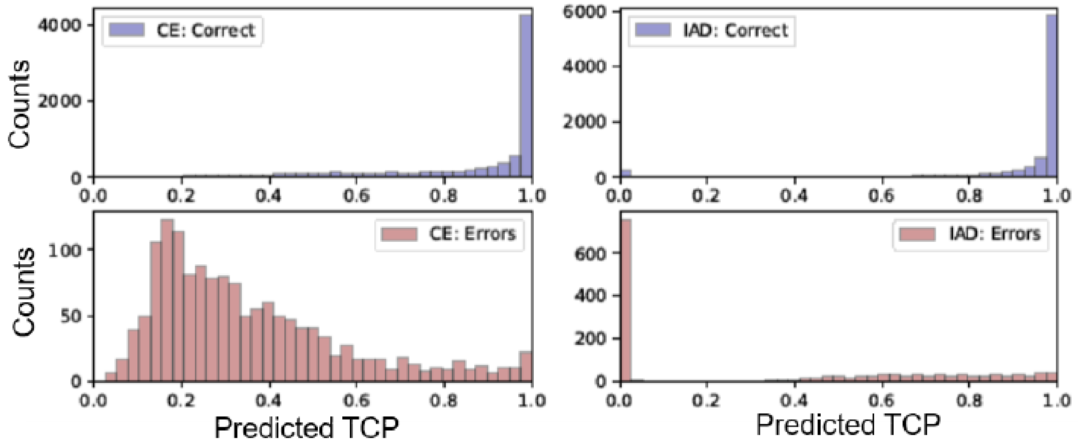


Fig. 3: Histogram of predicted TCP scores for CE confidence network [13] (left) and our proposed IAD constrained confidence network (right) on CIFAR-10 test set.

Table 1: Comparison of failure prediction methods on various image classification datasets and network architectures.

Method	AUROC	AUPRC-Success	AUPRC-Error	FPR at 85% TPR
Fashion-MNIST: LeNet 20{5}-50{5}-500				
CE-MCP	91.85	99.23	47.09	16.69
CE-TCP	91.94	99.23	46.36	15.68
IAD-MCP	92.01	99.05	61.59	16.65
IAD-TCP*	**93.42**	**99.26**	**63.80**	**13.87**
CIFAR-10: VGG 64{3}-128{3}-256{3}-256				
CE-MCP	89.22	97.97	51.99	24.61
CE-TCP	92.92	98.59	71.55	13.93
IAD-MCP	90.54	98.29	61.45	20.69
IAD-TCP*	**93.79**	**98.91**	**74.31**	**10.81**
CIFAR-100-coarse: VGG 128{3}-256{3}-512{3}-512{3}-1024-1024				
CE-MCP	86.78	95.72	62.62	33.33
CE-TCP	**88.71**	**96.27**	69.36	27.43
IAD-MCP	86.48	95.12	67.48	33.33
IAD-TCP*	88.13	95.63	**73.15**	**26.38**
Tiny-ImageNet: VGG 128{3}-256{3}-512{3}-512{3}-1024-1024				
CE-MCP	84.90	85.68	83.47	35.65
CE-TCP	**87.06**	**87.85**	85.85	30.30
IAD-MCP	82.97	81.55	83.85	40.29
IAD-TCP*	86.96	85.62	**87.30**	**29.72**

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Conclusions

- New method presented for estimated confidence of Deep Dirichlet networks
  - Transfer knowledge from classification network to confidence estimation network to estimate TCP score

- Confidence network trained for failure prediction task by matching TCP distribution in a flexible manner and taking TCP constraints into account for correct predictions and failures

- Empirical results show improvements for image classification tasks over SOTA