Investigation of the behavior of single-model uncertainty estimation approaches on large-scale tasks

Author: Mihail Solotchii
Advisors: Andrey Malinin, Artem Babenko

Master's thesis
Higher School of Economics, Faculty of Computer Science

June 8, 2021

Uncertainty estimation

- We want our model to make predictions along with some uncertainty measure

Uncertainty estimation

- We want our model to make predictions along with some uncertainty measure

- We want uncertainty measures to correlate with quality of predictions

Uncertainty estimation

- We want our model to make predictions along with some uncertainty measure
- We want uncertainty measures to correlate with quality of predictions
- All theory behind ML models assumes the training and test distributions are the same

Uncertainty estimation

- We want our model to make predictions along with some uncertainty measure
- We want uncertainty measures to correlate with quality of predictions
- All theory behind ML models assumes the training and test distributions are the same
- We validate uncertainty estimation on Out-of-Distribution detection task (AUROC for quality measurement)

Known approaches

- Ensembles
  - Measure models' disagreement
  - Expensive

Known approaches

- Ensembles
  - Measure models' disagreement
  - Expensive
- Prior Networks
  - Emulate ensembles with a Dirichlet over predictions
  - Need both ID and OoD data for training

Known approaches

- Ensembles
  - Measure models' disagreement
  - Expensive
- Prior Networks
  - Emulate ensembles with a Dirichlet over predictions
  - Need both ID and OoD data for training
- Evidential Networks
  - Interpret networks' outputs as parameters of a Dirichlet
  - Don't use OoD data on training
  - Are tested using a simple architecture (LeNet) on simple datasets (e.g. MNIST)

Goals and questions

- Do Evidential models scale to large-scale tasks?
- We need to understand this method better
  (ideally, we would like to know why it works)

Evidential Networks

- Parameters of a Dirichlet are inferred as follows

$$\boldsymbol{\alpha} = \mathrm{ReLU}(f_{\boldsymbol{\theta}}(\boldsymbol{x})) + 1$$

- Uncertainty is measured by

$$u(\boldsymbol{x}) = \frac{K}{\sum_{j=1}^{K} \alpha_j(\boldsymbol{x})}$$

Last-layer activation

Table: Validation accuracy (%)

Train dataset	Exponent	ReLU	Softplus
CIFAR10	**96.17**	59.31	95.79
CIFAR100	**81.00**	32.17	67.30

Evidential Networks

- Parameters of a Dirichlet are now inferred as follows

$$\boldsymbol{\alpha} = \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x})) + 1$$

- Uncertainty is measured by

$$u(\boldsymbol{x}) = \frac{K}{\sum_{j=1}^{K} \exp\left(f_{\boldsymbol{\theta}}(x)\right)}$$

Evidential Networks

- Parameters of a Dirichlet are now inferred as follows

$$\boldsymbol{\alpha} = \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x})) + 1$$

- Uncertainty is measured by

$$u(\boldsymbol{x}) = \frac{K}{\sum_{j=1}^{K} \exp\Big(f_{\boldsymbol{\theta}}(x)\Big)}$$

- The loss function is

$$L(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{L}(\boldsymbol{x}, y) + \lambda\, R_{\mathrm{KL}}(\boldsymbol{x}, y)$$

Regularization and loss function

Table: OOD detection performance (AUROC %)

Dataset		Reg	EoE			MI		
ID	OOD		CE	Gibbs	L_2	CE	Gibbs	L_2
C100	SVHN	On	43.5	41.8	50.0	39.0	47.6	50.0
		Off	**79.9**	77.2	78.1	**80.4**	77.8	78.8
	C10	On	51.6	59.2	50.0	50.6	59.0	50.0
		Off	**81.3**	**81.4**	78.2	**80.8**	**81.1**	77.0
	LSUN	On	54.3	53.4	50.0	56.6	57.2	50.0
		Off	**75.5**	**75.8**	71.3	**74.0**	**74.8**	69.1
	TiM	On	51.2	58.4	50.0	51.2	58.2	50.0
		Off	**81.7**	**81.6**	78.6	**81.1**	**81.1**	76.6

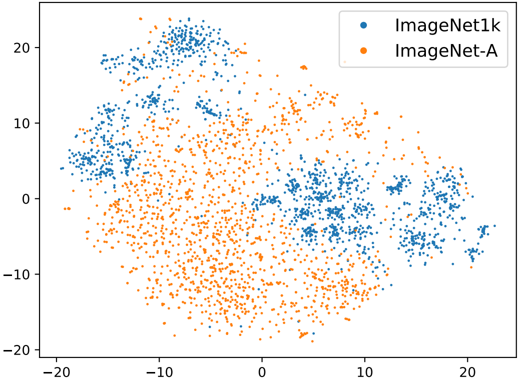
Large-scale experiments

Table: OOD detection performance (AUROC %) on ImageNet

ID set	OOD set	Evidential	MI	
		Single	Single	Ens
ImNet-1k	ImNet-O	58.1	57.4	**60.9**
	ImNet-A	85.8	85.7	**87.0**
	ImNet-R	**86.2**	86.1	84.8
	ImNet-C1	**68.1**	67.9	67.3
	ImNet-C2	**75.4**	**75.3**	**75.4**
	ImNet-C3	**81.2**	**81.1**	**81.0**
	ImNet-C4	**87.3**	**87.3**	86.0
	ImNet-C5	**91.6**	**91.6**	88.4

Embedding space

Figure: Representations of the network's penultimate layer reduced to 2 dimensions with t-SNE



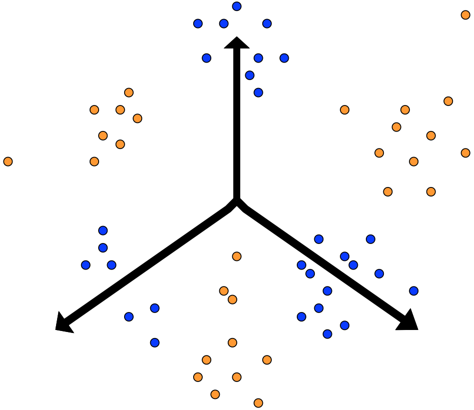
Maximum cosine correlation with model performance

Table: Confidence measures and model performance on ImageNet-C

Corruption level	C0	C1	C2	C3	C4	C5
Evidential	17.4	14.4	13.1	12.0	10.8	10.0
Maximum cosine	0.45	0.38	0.35	0.32	0.29	0.27
Model accuracy (%)	75.9	59.7	48.7	38.4	27.1	17.8

Maximum cosine

Figure: ID embeddings and OoD embeddings around the prototypes (arrows)



Large-scale experiments

Table: OOD detection performance (AUROC %) on ImageNet

ID set	OOD set	MaxCos	Evidential	MI	
		Single	Single	Single	Ens
ImNet-1k	ImNet-O	**68.0**	58.1	57.4	60.9
	ImNet-A	**88.1**	85.8	85.7	87.0
	ImNet-R	**87.1**	**86.2**	86.1	84.8
	ImNet-C1	66.7	**68.1**	67.9	67.3
	ImNet-C2	74.6	**75.4**	**75.3**	**75.4**
	ImNet-C3	80.5	**81.2**	**81.1**	**81.0**
	ImNet-C4	86.4	**87.3**	**87.3**	86.0
	ImNet-C5	90.6	**91.6**	**91.6**	88.4

Conclusions

- Do Evidential methods scale?
    - Evidential methods as they are do not scale
    - A simple modification to Evidential methods does scale

Conclusions

- Do Evidential methods scale?
  - Evidential methods as they are do not scale
  - A simple modification to Evidential methods does scale
- Properties of the method:
  - ID embeddings are closer to the prototypes by cosine distance
  - Maximum cosine between an embedding and prototypes is a good OoD detector