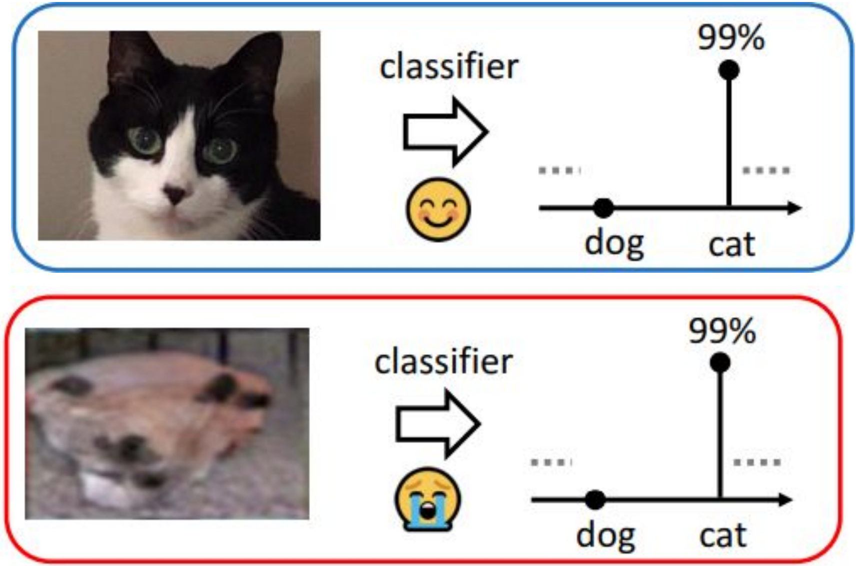
Master Thesis:
Generalization of MSP based out-of-distribution detection to intermediate convolutional layers

Praveen Annamalai Nathan
nathan@rhrk.uni-kl.de

Reviewers: Prof. Dr. Didier Stricker, DFKI, Kaiserslautern
Prof. Dr. Bernt Schiele, MPII, Saarbrücken

Supervisor: Max Maria Losch, MPII, Saarbrücken

Out-of-distribution detection (OOD)

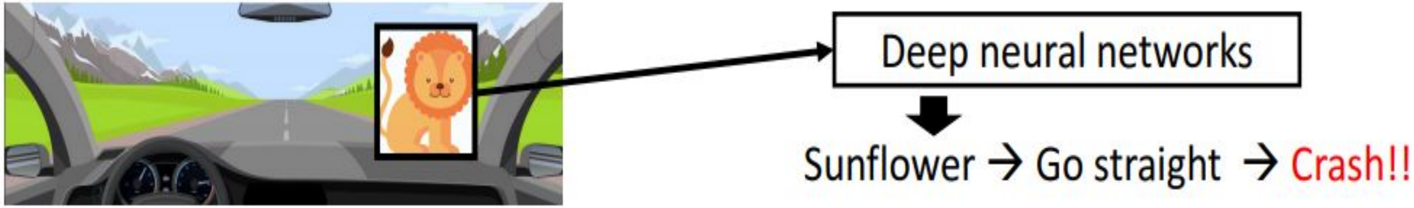


- DNN classifiers perform well. But..

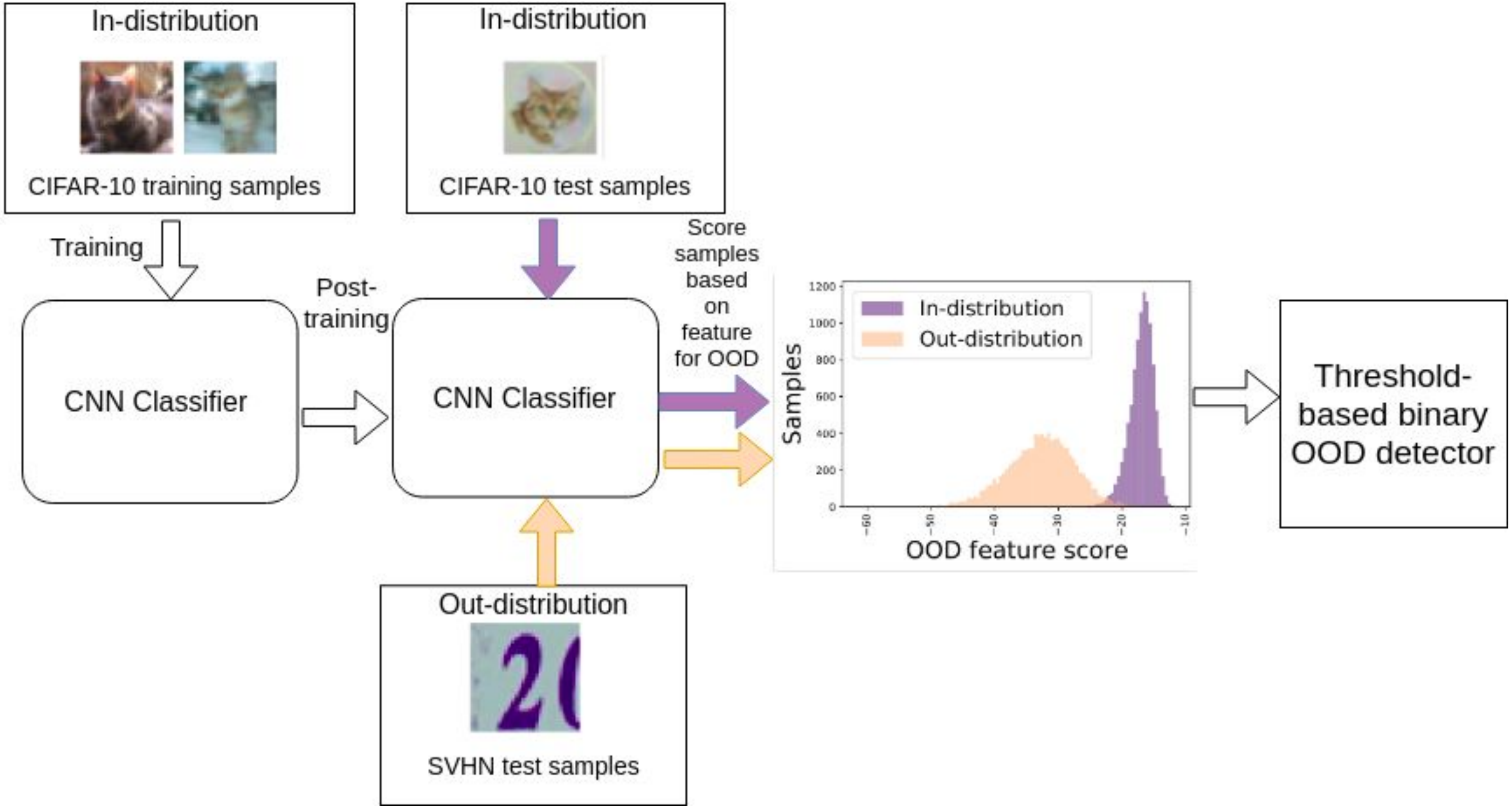
- Overconfident predictions: for samples different from training distribution.

- Fundamental requirement:

  out of distribution sample detection



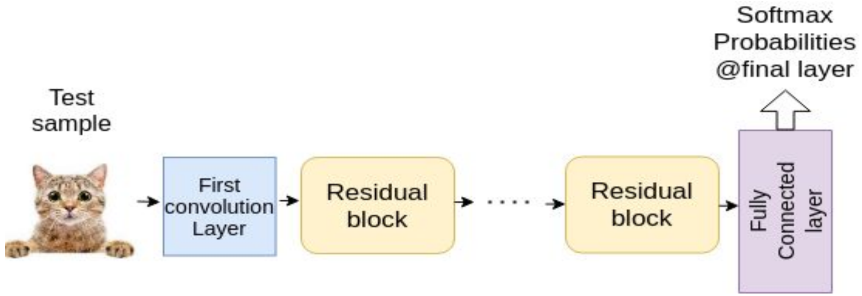
General OOD pipeline



Softmax as a feature for OOD

Softmax probabilities:

- final layer softmax studied in literature.
  E.g. Maximum softmax probability
- Many softmax-based extensions available.
- Competitive to SOTA methods.



What is missing?

- Softmax methods does not explore intermediate layers.

What we propose:

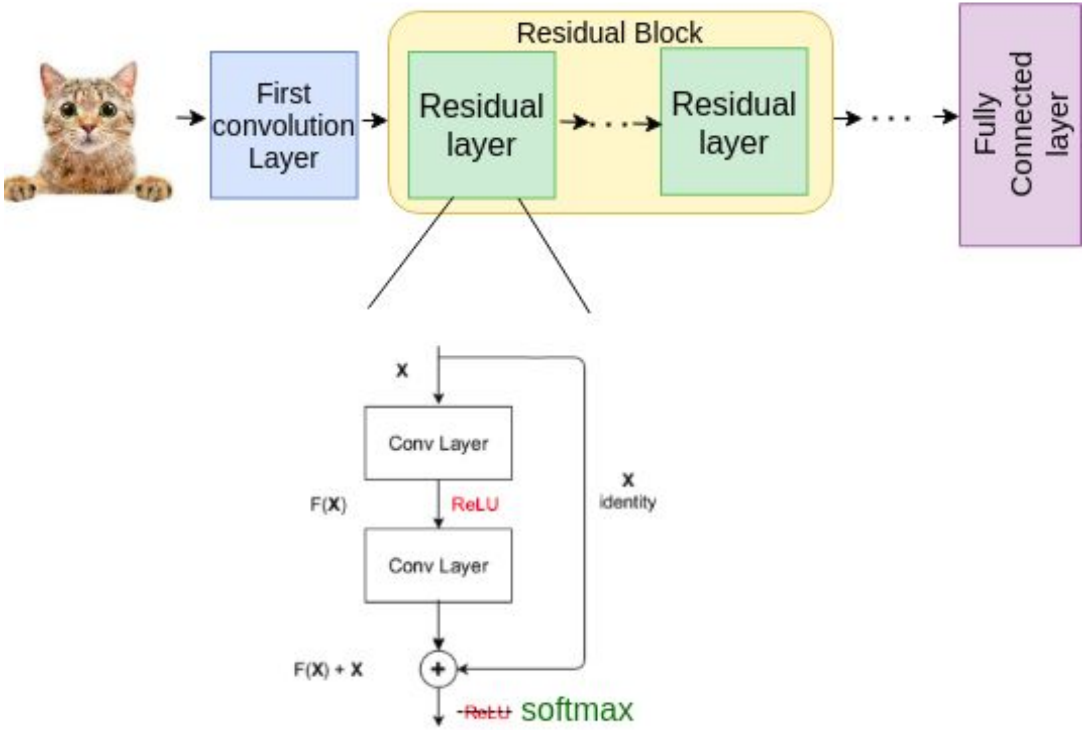
- Softmax as an intermediate layer activation.
- Generalization of softmax based OOD methods to intermediate layer.

Motivation-Softmax as intermediate layer activation

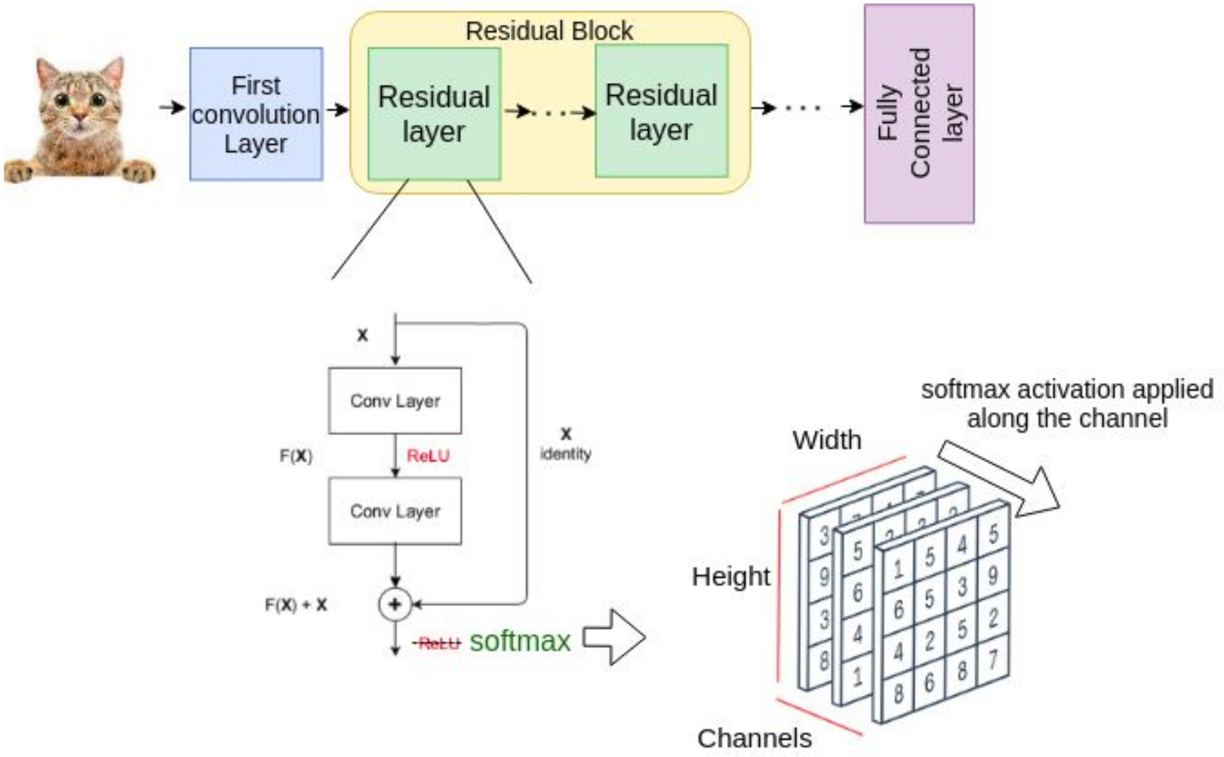
Where do we introduce

softmax activation?

Replace ReLU after the residual skip connection.



Motivation-Softmax as intermediate layer activation



How do we apply softmax activation?

Activation along the channel.

Motivation- softmax as intermediate activation

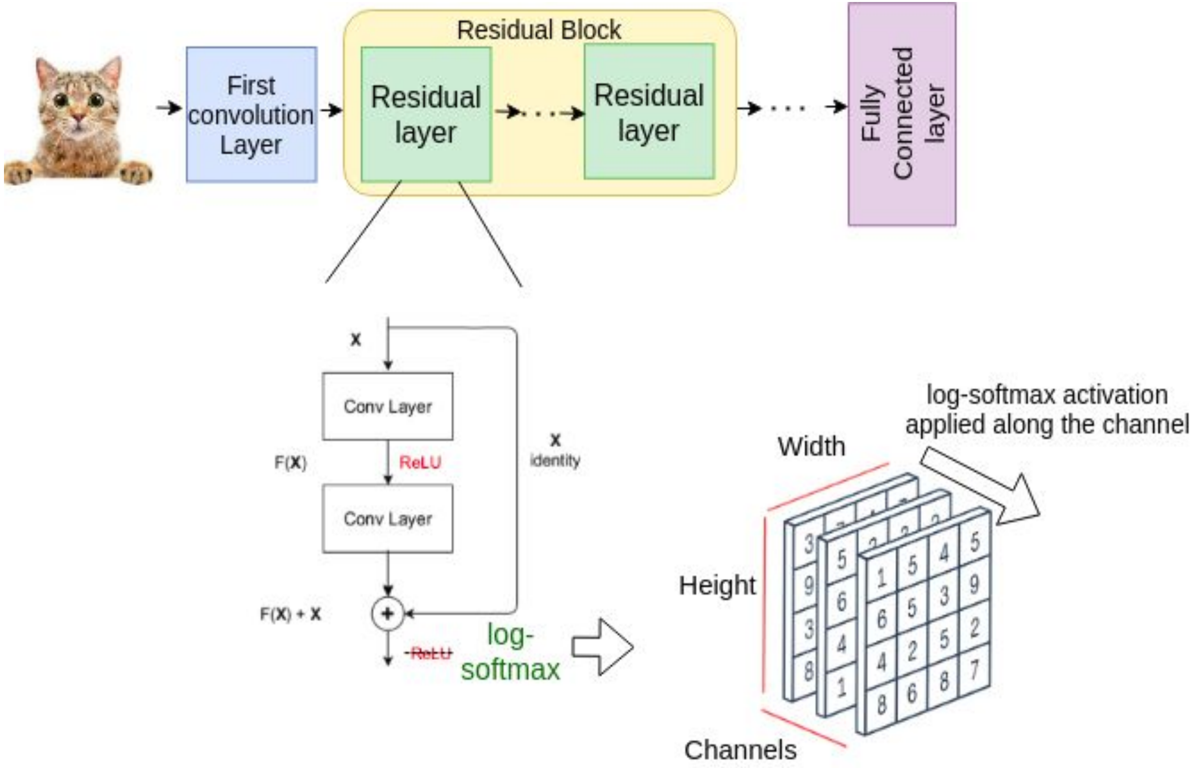
Apply log-softmax instead of softmax

Why?

Classification performance
(Model:ResNet-34 Dataset: CIFAR-10)

Softmax activation: 89.75%
Log-softmax activation: 93.5%
ReLU activation:94.31%

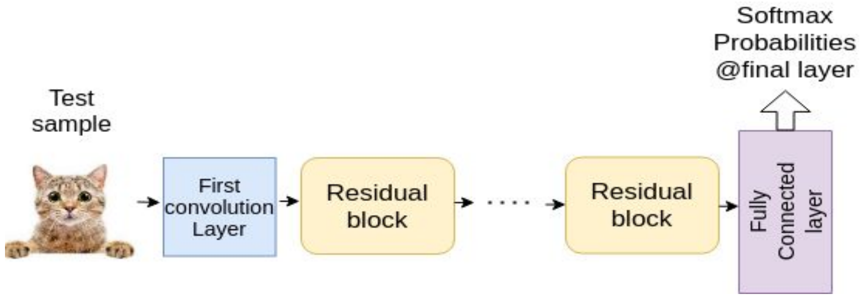
Log-softmax retains classification performance.

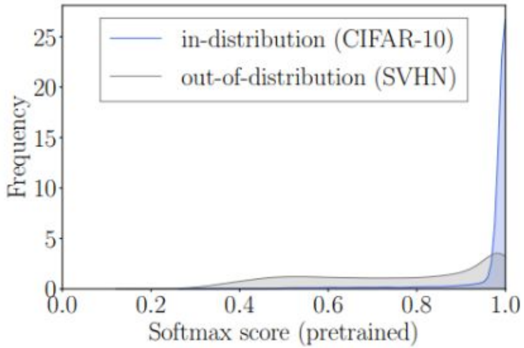


Related work-softmax based methods

Maximum Softmax Probability- **MSP** [1]

Scoring function: Maximum Softmax probability





ODIN[2]

Improves **MSP** method by:

- Applying softmax temperature scaling.
- Input preprocessing – adding small perturbations to input.

$$S_i(\boldsymbol{x}; T) = \frac{\exp\left(f_i(\boldsymbol{x})/T\right)}{\sum_{j=1}^{N} \exp\left(f_j(\boldsymbol{x})/T\right)},$$

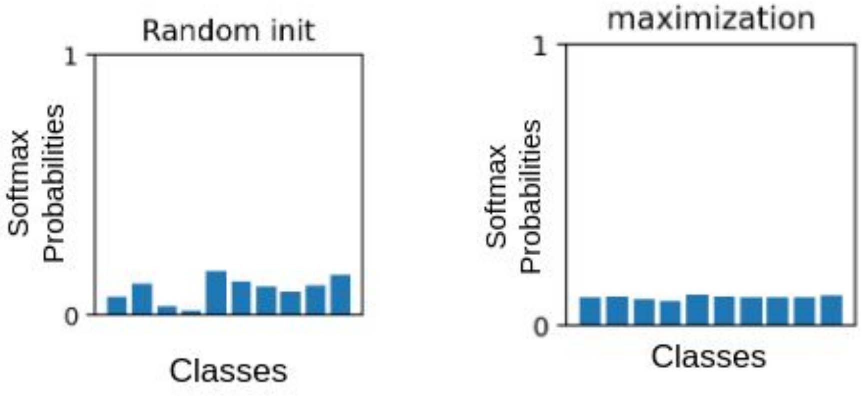
Temperature scaling

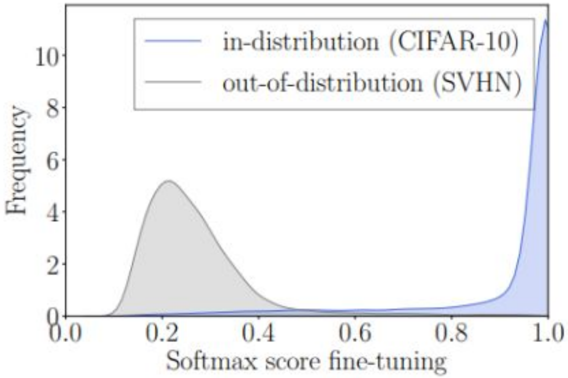
[1] Hendrycks, Dan, and Kevin Gimpel. "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks." ICLR. 2016.
[2] Liang, Shiyu, Yixuan Li, and R. Srikant. "Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks." ICLR. 2018.

Related work-softmax based methods

Outlier Exposure[3]

- Improves over **MSP** method.
- Fine-tuning with an auxiliary OOD dataset. E.g. 80 Million Tiny images dataset
- Forces auxiliary dataset predictions to be uniform (Entropy Maximization).





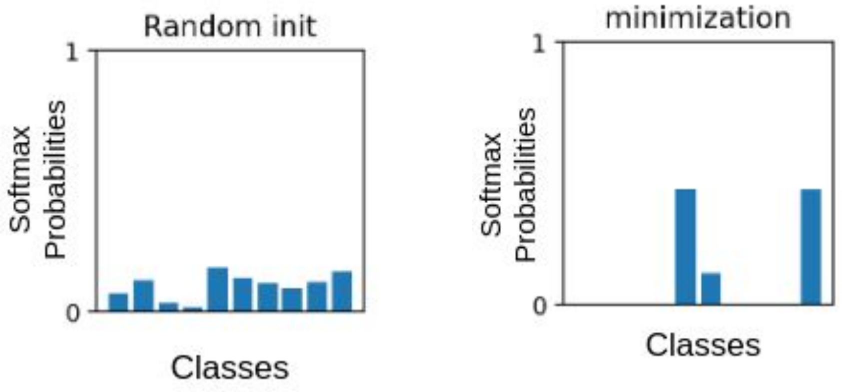
[3] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. “Deep Anomaly Detection with Outlier Exposure.”, ICLR 2019

Related work-softmax based methods

Additional possibility for log-softmax models

Entropy Minimization:

- cross-entropy@final layer beneficial for **MSP** method
- Mimic cross-entropy at final layer for intermediate layer log-softmax outputs.
- Reduce entropy per pixel for the log-softmax outputs.
- Entropy loss optimized with cross-entropy loss.



Log-softmax model benchmarking for OOD

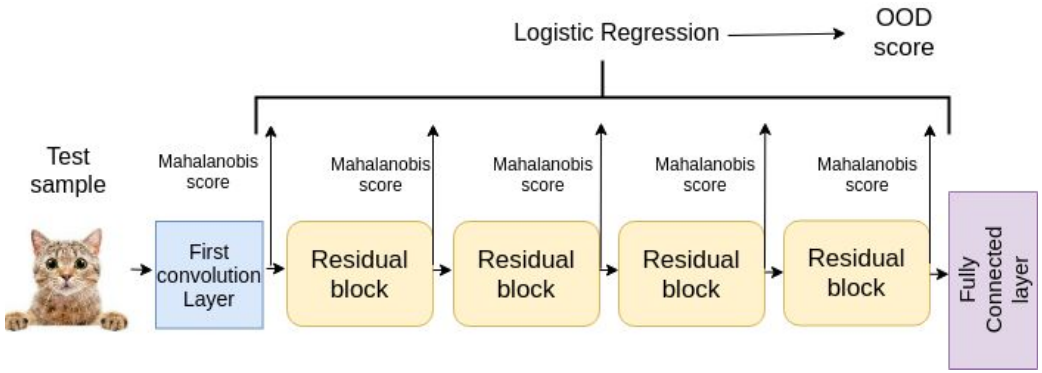
Softmax based methods :

- **MSP**@final layer serves as baseline.

Baseline missing for intermediate layer.

Mahalanobis-distance based [4]

- Fit pretrained features by a class conditional Gaussian distribution.
- **Scoring function**: Mahalanobis distance between a test sample and the closest Gaussian.
- **Mahalanobis-ensemble**:use feature from five intermediate blocks.
- Mahalanobis score per layer can be used as a baseline for intermediate layers.

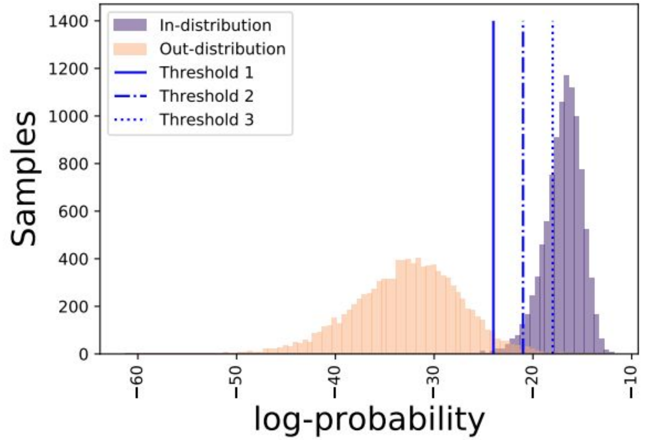


[4] Lee, Kimin et al. "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks." NeurIPS (2018).

OOD evaluation metrics

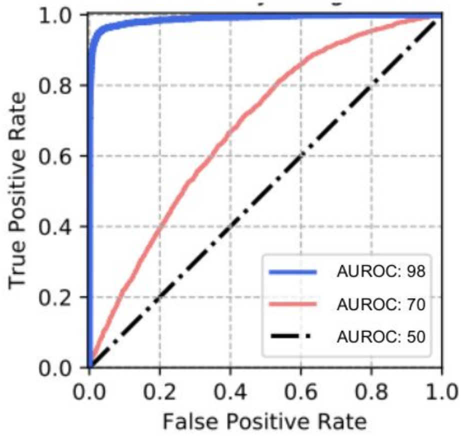
- OOD methods derive a confidence score. E.g. Softmax based/intermediate feature based.

- Threshold needs to be selected for measuring OOD detection accuracy considering trade off between False Positive and False Negatives allowed.



AUROC: Area under Receiver Operating Characteristic curve.

- Threshold independent metric
- Area under TPR vs FPR for all thresholds.
- Higher the better.



Performance comparison of methods from literature

- ODIN/Outlier Exposure improves over MSP.
- Fine tuning methods offer good results.
- Mahalanobis method is competitive.

Network Achitecture (In-Dist: CIFAR10)	OOD dataset	AUROC(↑)					
		MSP	ODIN	Mahalanobis	Energy Based	MSP+ Outlier Exposure (Fine tuning)	Energy Based (Fine tuning)
WideResNet	SVHN	91.89[1]	91.96[1]	97.62[1]*	90.96[1]	98.6[1]	**99.41**[1]
	LSUN-crop	95.65[1]	97.04[1]	94.15[1]*	98.35[1]	**99.49**[1]	99.32[1]
	LSUN-resize	91.37[1]	94.57[1]	93.23[1]*	94.24[1]	98.94[1]	**99.39**[1]
ResNet-34	SVHN	89.9[2]	96.7[2]	**99.1**[2]	-NA-	-NA-	-NA-
	LSUN-crop	-NA-	99.2[3]	**99.3**[3]	-NA-	-NA-	-NA-
	LSUN-resize	91.0[2]	94.1[2]	**99.7**[2]	-NA-	-NA-	-NA-

Note: Some values are -NA-(not available) since that configuration is not reported in the literature.

[1] Weitang Liu et al. "Energy-based Out-of-distribution Detection." NeurIPS '20
[2] Lee, Kimin et al. "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks." NeurIPS (2018)
[3] Yen-Chang Hsu et al. "Generalized ODIN:Detecting Out-of-distribution Image without Learning from Out-of-distribution Data". CVPR '20
*Mahalanobis score calculated using only features of second to last layer.

Experimental setup

Classifier Architecture: ResNet-34

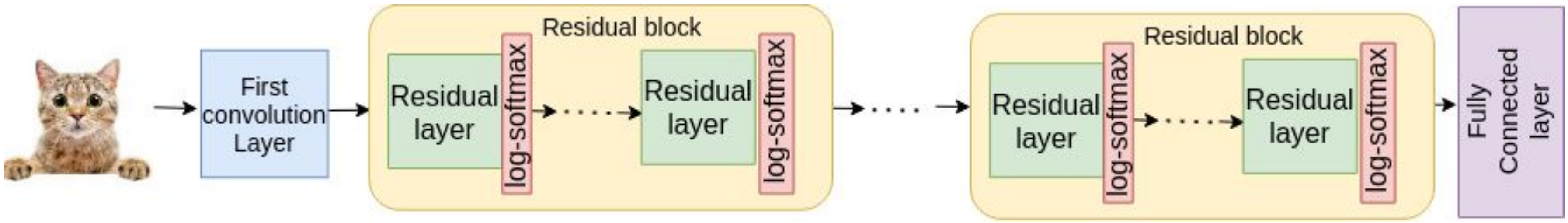
- Residual skip connection based architecture.

Train 2 types of model:

1) Baseline ReLU architecture
2) Log-softmax architecture

Generalization of log-softmax introduction

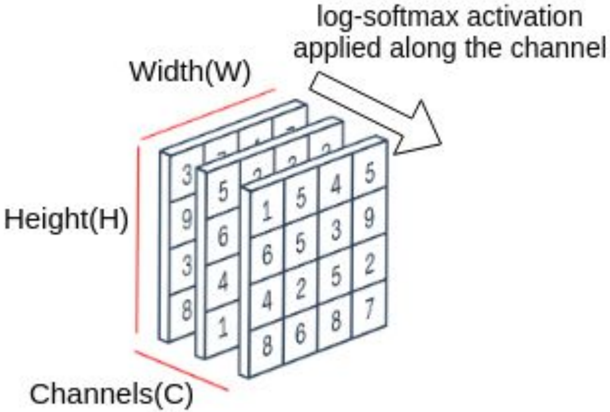
- Introduce after all residual layers.
- Log-softmax introduced after 16 residual layers of ResNet-34.



Experimental setup

For log-softmax models:

- Entropy minimization for log-softmax outputs from all blocks.
- Entropy loss added to cross entropy loss with a factor.



$$\mathcal{L} = CE_{loss} + \lambda \underbrace{\sum_{i=1}^{N} \sum_{j=1}^{H_i \times W_i} \sum_{k=1}^{C_i} -s(x_{ijk}) log(s(x_{ijk}))}_{\text{Entropy loss}}$$

N - number of layers with log-softmax activation
λ - entropy loss factor
s(.) - softmax
CE_{loss} - Final layer cross entropy loss

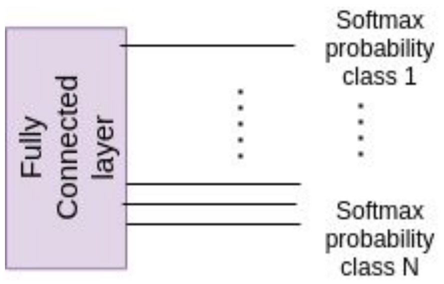
Experimental setup - Intermediate layer OOD evaluation

Challenges in direct application of softmax methods for intermediate log-softmax outputs.

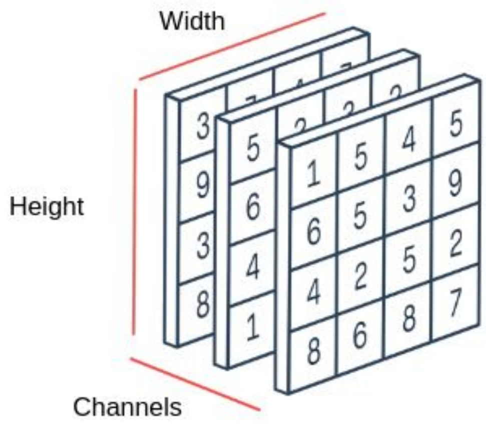
- Final layer softmax probabilities are 1-dimensional.
- Intermediate layer log-softmax outputs are 3-dimensional.

Need to define new evaluations for the intermediate block activations.

- Generalize MSP method at final layer by aggregating 3-dimensional outputs.



Intermediate logsoftmax output



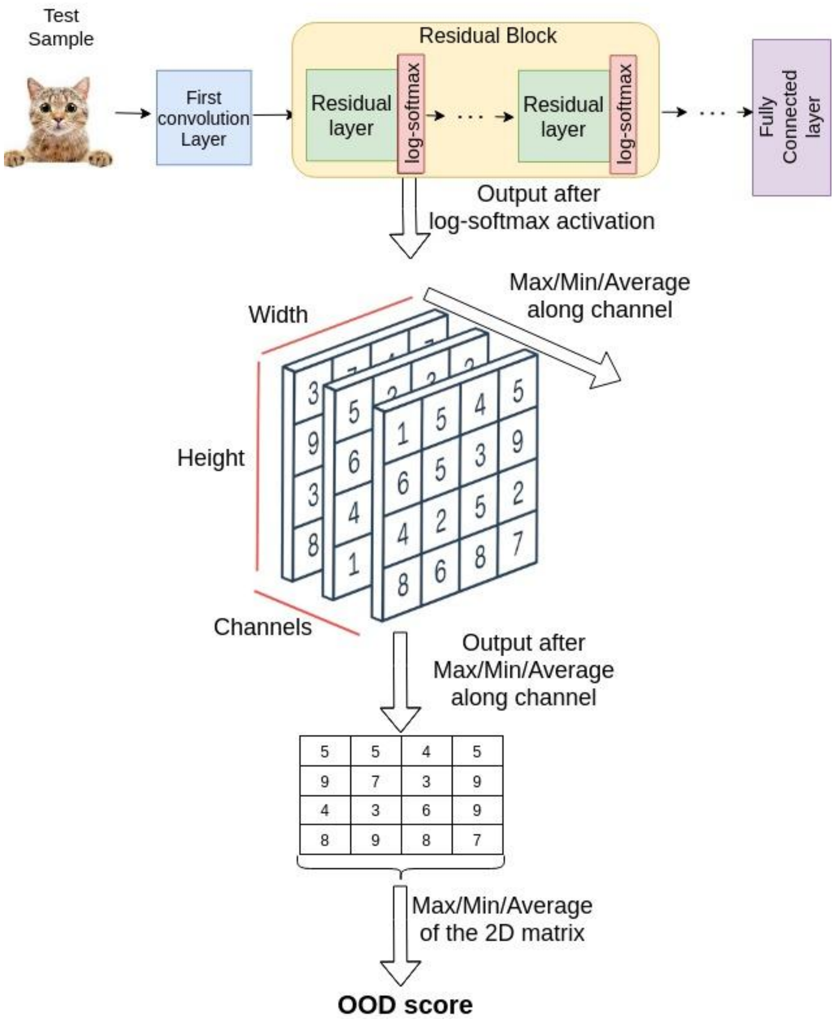
Experimental setup- Intermediate layer OOD evaluation

Generalization of MSP

Total evaluations – 9 types

- max/min/avg of MSP(Maximum Softmax Probability)
- max/min/avg of MiSP(Minimum Softmax Probability)
- max/min/avg of ASP(Average Softmax Probability)

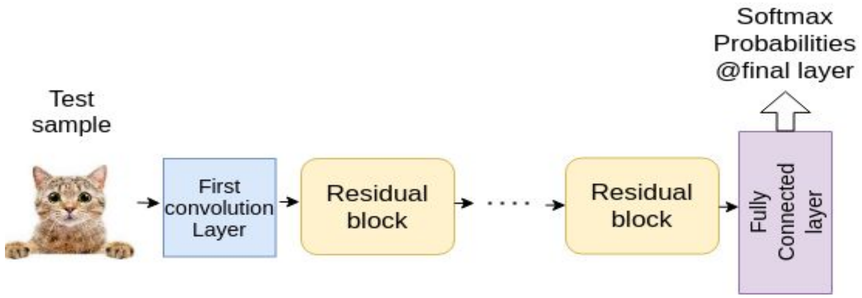
Cheap: no additional training for OOD detector.



Experimental setup-extension for final layer

Extension possible@final layer:

- Current methods only explore maximum softmax probability@final layer.
- We propose to explore Minimum Softmax Probability(**MiSP**) as an OOD score.
- Average Softmax Probability(**ASP**)@final layer not explored since it is constant.

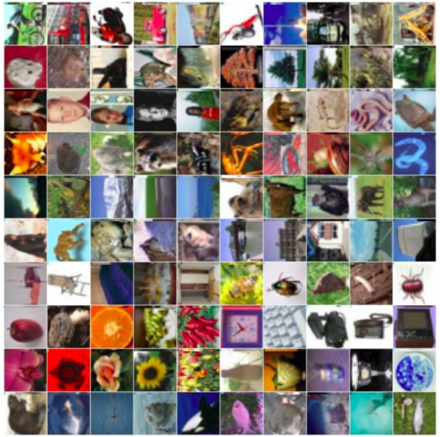


Experimental setup-Datasets

In-distribution dataset (Training dataset)

CIFAR-10

Near out distribution[5]



CIFAR-100

Far out distribution[5]

LSUN

SVHN

[5] Winkens, Jim et al. “Contrastive Training for Improved Out-of-Distribution Detection.” (2020).

Results

Log-softmax model benchmark:

- Classification accuracy
- MSP vs MiSP
- Mahalanobis-ensemble vs MiSP

Intermediate layer evaluation

- Mahalanobis per layer vs min of MiSP.
- min of MiSP trends based on dataset distance.
- Comparison of aggregation methods vs baseline methods.

Classification Accuracy

- Does softmax activation integrated models retain performance?
  - Yes

- Higher entropy loss factor leads to degradation in accuracy.

Network: ResNet34	In-distribution: CIFAR-10	
log-softmax	Entropy Loss Factor	Top-1 accuracy(%)
-NA-	-NA-	94.31
yes	0	93.5
yes	1e-3	93.37
yes	5e-3	93.09
yes	1e-2	92.64
yes	5e-2	92.48
yes	1e-1	91.92

OOD evaluation benchmark for log-softmax models

- MiSP better than MSP
- MSP for log-softmax models: comparable
- Entropy minimization: no significant effect.

Network Architecture: ResNet34 In-distribution: CIFAR-10				
Log-softmax	Entropy Loss factor	Out-distribution dataset	AUROC(↑)	
			MSP	MiSP
-NA-(Baseline)		CIFAR100	84.89	88.19
		LSUN	93.48	95.66
		SVHN	92.08	96.54
yes	0	CIFAR100	85.20	88.39
		LSUN	88.48	93.53
		SVHN	90.27	94.70
yes	1e-2	CIFAR100	83.26	86.23
		LSUN	89.89	95.32
		SVHN	92.43	97.41

OOD evaluation benchmark for log-softmax models

- Mahalanobis:

  comparable performance for baseline and log-softmax models.

  .

- MiSP close to Mahalanobis-ensemble performance.

Network Architecture: ResNet34 In-distribution: CIFAR-10				
Log-softmax	Entropy Loss factor	Out-distribution dataset	AUROC(↑)	
			Mahalanobis ensemble	MiSP
-NA-(Baseline)		CIFAR100	66.06	88.19
		LSUN	99.75	95.66
		SVHN	98.62	96.54
yes	0	CIFAR100	73.51	88.39
		LSUN	99.46	93.53
		SVHN	98.85	94.70
yes	1e-2	CIFAR100	74.07	86.23
		LSUN	99.45	95.32
		SVHN	98.19	97.41

Results

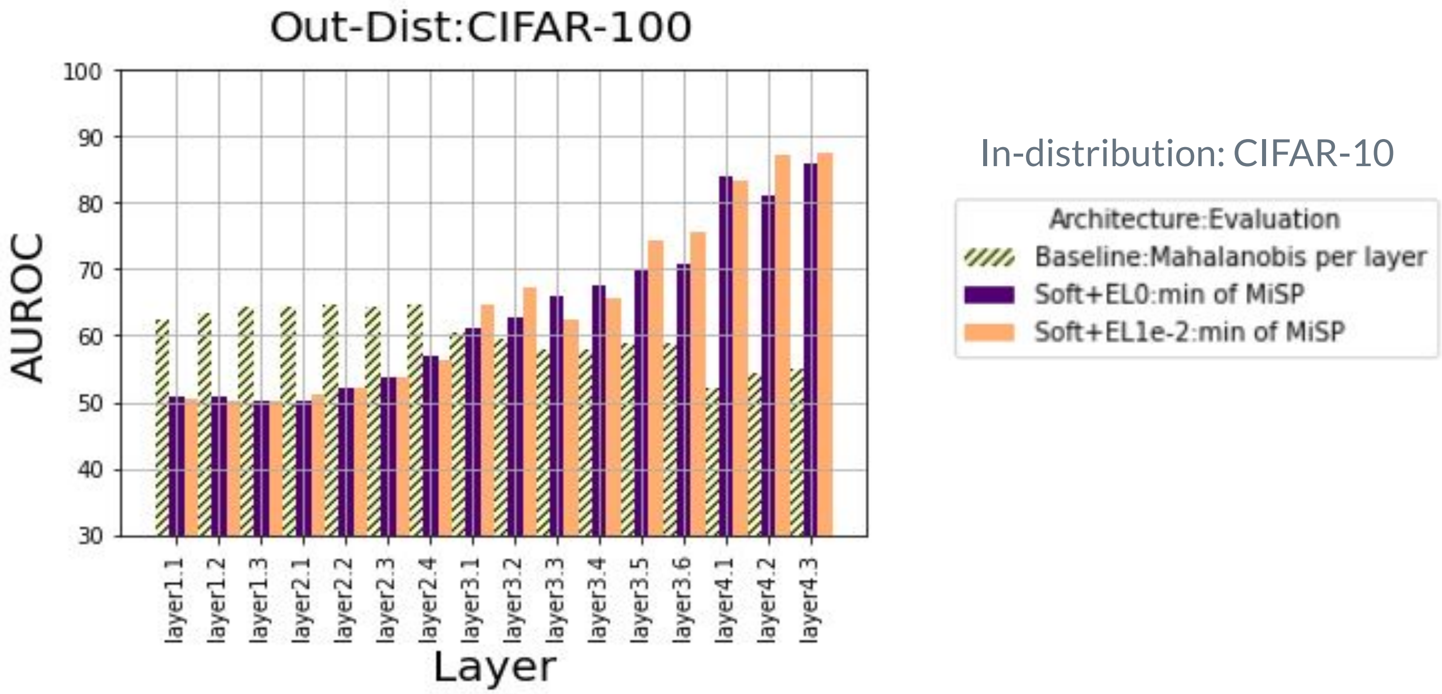
Log-softmax model benchmark:

- Classification accuracy
- MSP vs MiSP
- Mahalanobis-ensemble vs MiSP

Intermediate layer evaluation

- Mahalanobis per layer vs min of MiSP.
- min of MiSP trends based on dataset distance.
- Comparison of aggregation methods vs baseline methods.

Comparison of aggregation methods with Mahalanobis per layer

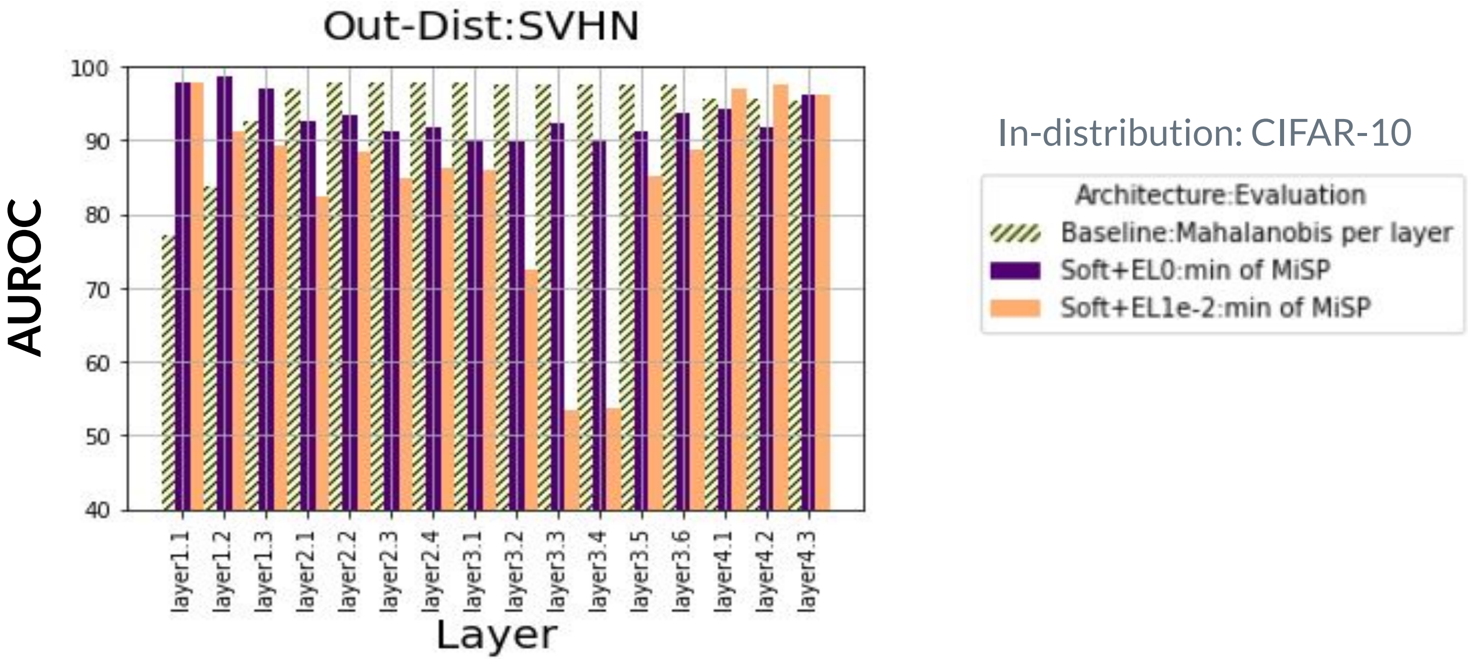


In-distribution: CIFAR-10

- Different layer enable different performances for OOD detection.
- min of MiSP observed to be the best in aggregation methods.

- Mahalanobis performs well for layers in block 1 & 2.
- min of MiSP better than Mahalanobis for final residual blocks.
- Entropy loss model: minor improvements in final block.

Comparison of aggregation methods with Mahalanobis per layer

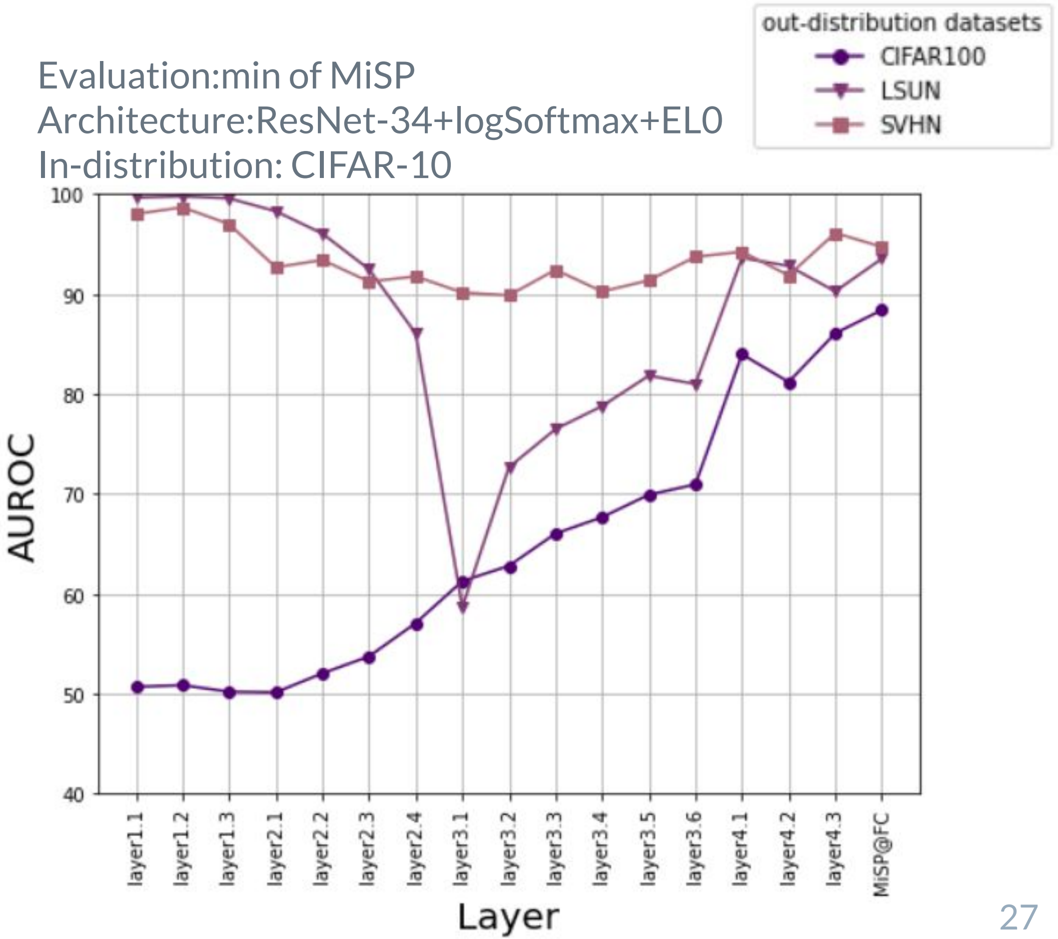


- Different layer enable different performances for OOD detection.
- Mahalanobis performs well for layers in block 2 & 3.
- min of MiSP better than Mahalanobis for initial and final residual blocks.

min of MiSP trends based on dataset distances

- Initial layers: beneficial for far datasets- SVHN & LSUN
- Final layers: beneficial for near dataset- CIFAR-100
- min of MiSP better for initial layers compared to MiSP@FC for LSUN and SVHN.
- MiSP@FC better for CIFAR100 than min of MiSP at intermediate layer.

Evaluation:min of MiSP
Architecture:ResNet-34+logSoftmax+EL0
In-distribution: CIFAR-10



Comparison of aggregation methods vs baseline methods

- The cheap evaluation method min of MiSP have competitive performance with Mahalanobis-ensemble.

Network Architecture: ResNet34 In-distribution: CIFAR-10				
Log-softmax	Entropy Loss factor	Out-distribution dataset	AUROC(↑)	
			MSP/ Mahalanobis-ensemble (best)	Intermediate layer aggregation (best)
-NA-(Baseline)		CIFAR100	84.89(MSP)	-NA-
		LSUN	99.75(Mahalanobis)	-NA-
		SVHN	98.62(Mahalanobis)	-NA-
yes	0	CIFAR100	85.20(MSP)	86.72(avg of ASP layer4.3)
		LSUN	99.46(Mahalanobis)	99.76(min of MiSP layer1.2)
		SVHN	98.85(Mahalanobis)	98.66(min of MiSP layer1.2)
yes	1e-2	CIFAR100	83.26(MSP)	87.44(min of MiSP layer4.3)
		LSUN	99.45(Mahalanobis)	99.66(min of MiSP layer1.2)
		SVHN	98.19(Mahalanobis)	97.89(min of MiSP layer1.1)

Discussion

- Classifiers with intermediate log-softmax activation retain performance.
- We observe MiSP as a better feature than MSP for OOD detection at final layer.
- This cheap generalization of MSP to intermediate log-softmax outputs, especially min of MiSP, has competitive performance with the Mahalanobis-ensemble method.
- We find early layers of the network are more beneficial for far out-distribution datasets and later layers for near out-distribution datasets.

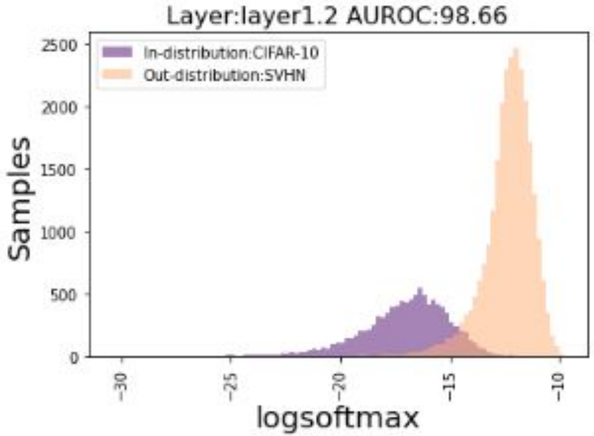
Discussion

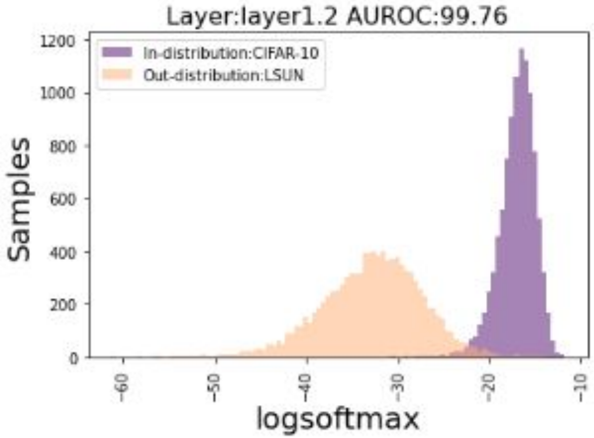
Evaluation:min of MiSP
Architecture:ResNet-34+logSoftmax+EL0

Challenges in adopting the MSP generalization for intermediate layers in an open world setting .

- Cannot reuse pretrained networks: Architecture should be integrated with log-softmax and retrained.

- min of MiSP spectrum: it is not consistent across datasets. OOD threshold cannot be predefined in an open world setting.





Conclusion & Future Work

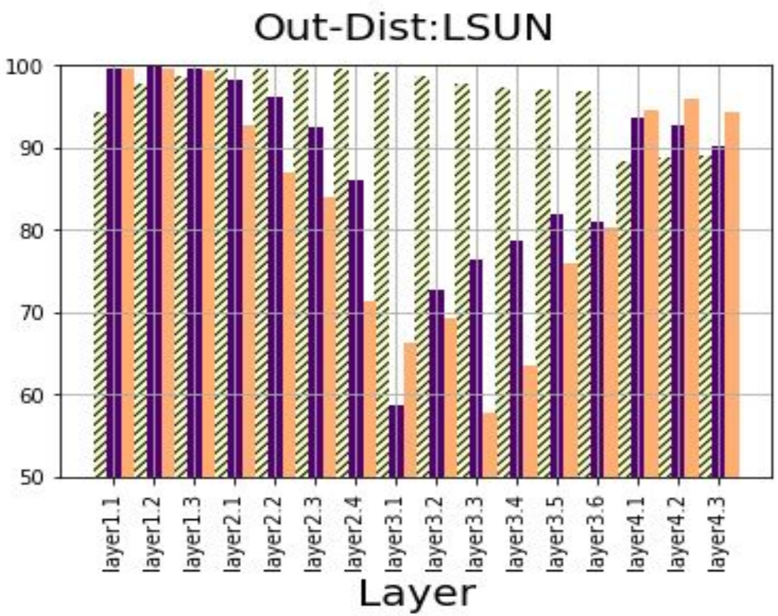
Developed a softmax based baseline for intermediate layer OOD evaluation.

- Extension of intermediate layer evaluation methods with ODIN/Outlier Exposure.
- Verification for other classifier architectures e.g. WideResNet and DenseNet.
- Further study on why MiSP works.

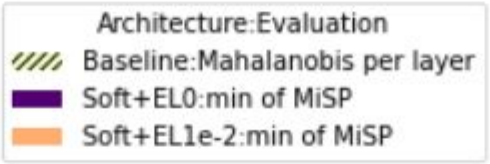
Thanks!

Any questions?

Comparison of aggregation methods with Mahalanobis per layer



In-distribution: CIFAR-10

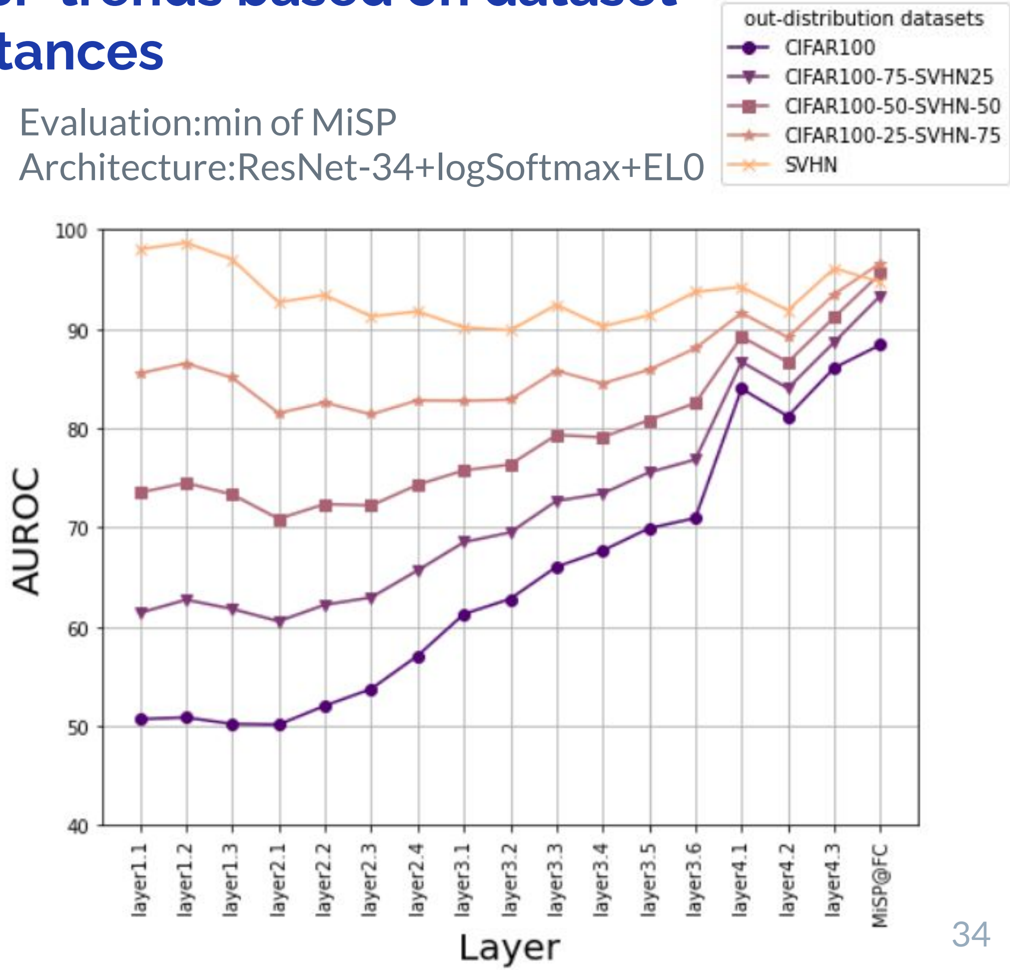


- Min of MiSP performance drops in block 2 & 3.
- Mahalanobis performs well for layers in block 2 & 3.
- General observation: min of MiSP better than Mahalanobis for initial and final residual blocks.
- Model with entropy loss have minor improvements in final block.

Experiments-min of MiSP trends based on dataset distances

- Introduce mixture datasets to study layer benefits.
- Mix samples from CIFAR-100 and SVHN datasets in some percentages.
- Initial layers gets more benefited for OOD detection as dataset moves from CIFAR-100 to SVHN.

Evaluation:min of MiSP
Architecture:ResNet-34+logSoftmax+EL0



Results generalization to CIFAR-100

Classification Performance

Network Architecture: ResNet34 In-distribution: CIFAR-100		
log-softmax	Entropy Loss Factor	Top-1 accuracy(%)
-NA-	-NA-	75.93
yes	0	74.25
yes	1e-2	73.39

Cifar100 OOD evalaution

Network Architecture: ResNet34 In-distribution: CIFAR-10						
Log-softmax	Entropy Loss factor	Out-distribution dataset	AUROC(↑)			
			MSP	Mahalanobis -ensemble	MiSP	Intermediate layer aggregation (best)
-NA-(Baseline)		CIFAR10	73.31	57.71	72.43	-NA-
		LSUN	75.15	99.21	78.3	-NA-
		SVHN	8.50	96.73	94.28	-NA-
yes	0	CIFAR10	70.56	61.29	70.49	62.5(avg of MiSP layer4.3)
		LSUN	71.4	99.05	73.69	99.84(min of MiSP layer1.1)
		SVHN	73.95	97.32	86.59	98.41(min of MiSP layer1.1)
yes	1e-2	CIFAR10	71.25	57.49	71.44	63.77(min of MiSP layer4.3)
		LSUN	69.27	98.88	69.36	99.65(min of MiSP layer1.2)
		SVHN	75.15	96.63	85.53	96.82(min of MiSP layer1.2)