Weekly Study Report

Wang Ma

2025-03-18

Electrical, Computer, and Systems Engineering Department
Rensselaer Polytechnic Institute

Outline

1. Credal Wrapper: Which Point to Predict?

1. Credal Wrapper: Which Point to Predict?

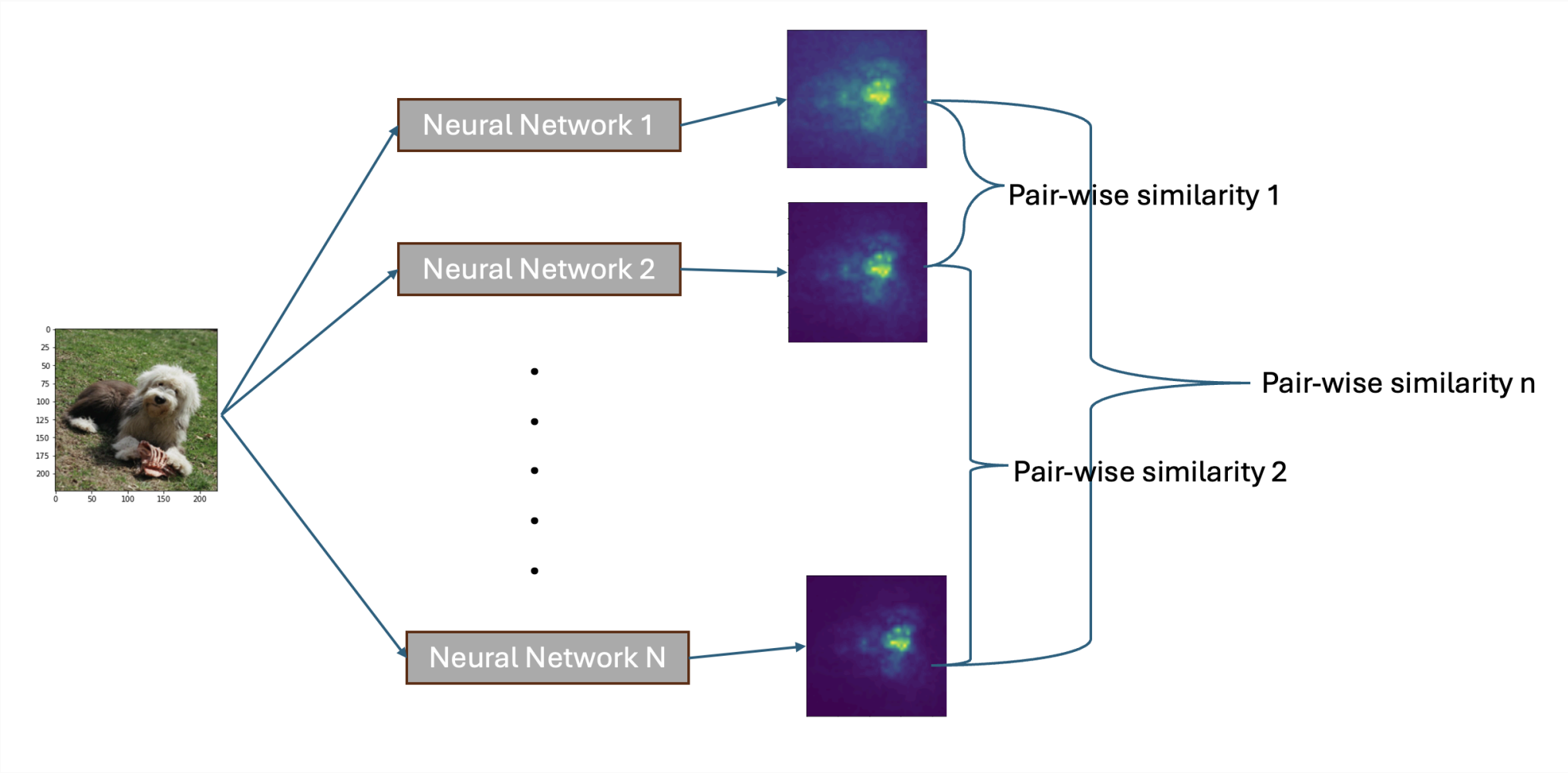
	Cifar 10 (Resnet 18)		Cifar 100 (Resnet 50)	
	Ensemble Size of 5	Ensemble Size of 30	Ensemble Size of 5	Ensemble Size of 30
Deep Ensemble (Simply Average)	**95.95%**	**96.10%**	**85.35%**	**86.05%**
Credal Wrapper (Mid Point)	95.91%	95.95%	**84.93%**	85.31%
Credal Wrapper (Lowest Entropy Point)	95.91%	95.95%	84.92%	85.31%
Credal Wrapper (Highest Entropy Point)	93.36%	88.35%	79.23%	70.07%
Credal Wrapper (Lower Bound)	95.89%	95.92%	84.84%	**85.44%**
Credal Wrapper (Upper Bound)	**95.98%**	**95.96%**	84.76%	85.22%

1. Credal Wrapper: Which Point to Predict?

1. If the test data is the same distribution of the training data, then different choices of the points has little influence on the accuracy (except for the “highest Entropy Point”).
2. When the ensemble size is large, the credal set is also larger (because we will have lower lower bound and higher upper bound), the entropy-related points will have large difference.
3. I think we can try the different points prediction on domain-shift/low-quality/failure data which have little difference from the training data, and we see which point’s prediction is the most robust.

2. Ensemble of Saliency Maps

2. Ensemble of Saliency Maps



2.1 Computing the Similarity

The choices of similarity scores:

1. Cosine Similarity: $\cos(I_1, I_2) = \dfrac{I_1 \cdot I_2}{\|I_1\|\|I_2\|}$
2. SSIM (Structural Similarity Index)
3. Intersection over Union(IoU): $\text{IoU} = \dfrac{|I_1(20\%) \cap I_2(20\%)|}{|I_1(20\%) \cup I_2(20\%)|}$, $I_{i(20\%)}$ means the most important 20% areas in image i.
4. Wasserstein Distance
5. Variance of Saliency Maps
6. Calculate the Similarity interval

2.2 OOD results

	Model Size of 5	Model Size of 30
Similarity Choice	AUROC	AUROC
Cosine Similarity	**0.9641**	**0.9789**
SSIM	0.9495	0.9728
IoU	0.9259	0.9719
Wasserstein	0.9192	0.9564
Variance	**0.9602**	**0.9766**

After we have the pairwise similarity scores (N models have $\frac{N(N-1)}{2}$ scores), we can

1. Average them
2. Calculate their varicances

Expectation:

- ID data: high average similarity, low variances, short interval
- OOD data: Low average similarity, high variances, large interval

2.2 OOD results

	Cifar 10 (Resnet 18) vs SVHN		
Models	Test Accuracy	AUROC	
Deep Ensemble	**95.95%**	0.9683	
Credal Ensemble	**96.43%**	**0.9813**	
Single Credal Nets	95.78%	0.8646	
Credal Wrapper (Mid Point)	95.91%	**0.9761**	
Credal Wrapper of Size 30	95.95%	0.9794	
Deep Ensemble of Size 30	96.10%	0.9794	
Ensemble of Saliency Maps of Size 5 (Average Pair-wise Similarity)	\	0.9641	
Ensemble of Saliency Maps of Size 30 (Average Pair-wise Similarity)	\	**0.9789**	
Ensemble of Saliency Maps (Similarity Variance)	\	0.971	
Ensemble of Saliency Maps (Similarity Interval Length)	\	0.9683	

2.3 Saliency Maps are robust on Corrupted Data

We treat Corrupted CIFAR-10 as OOD data, and see whether the saliency-map based methods can distinguish them.

- Uncertainty-based methods are sensitive to corruptions, and they can distinguish the corrupted data(most of corruption types have AUROC > 0.8)
- Saliency-Map-based methods are not sensitive to corruptions, that is, they cannot distinguish corrupted data from the clean data (AUROC $\approx$ 0.5)

2.3 Saliency Maps are robust on Corrupted Data

CIFAR-10-C OOD Detection Results (AUC-ROC)

Corruption	cosine	ssim	iou	emd	variance	entropy
brightness	0.5028	0.5024	0.5044	0.5047	0.5030	0.5055
impulse_noise	0.6789	0.6743	0.6431	0.6531	0.6808	0.5000
gaussian_blur	0.5041	0.4998	0.5057	0.5014	0.5035	0.5024
motion_blur	0.5975	0.5846	0.5785	0.5723	0.5939	0.4954
glass_blur	0.7116	0.7066	0.6490	0.7200	0.7138	0.5405
snow	0.5658	0.5576	0.5518	0.5559	0.5654	0.4947
fog	0.4974	0.4921	0.5008	0.4984	0.4962	0.4960
contrast	0.4977	0.4870	0.5058	0.4986	0.4953	0.5078
saturate	0.5532	0.5351	0.5285	0.5513	0.5488	0.5081
zoom_blur	0.6383	0.6221	0.6133	0.6051	0.6328	0.5171
elastic_transform	0.5994	0.5813	0.5569	0.5842	0.5913	0.5209
shot_noise	0.6033	0.5964	0.5734	0.6076	0.6058	0.4973
gaussian_noise	0.6508	0.6375	0.6095	0.6486	0.6522	0.4835
spatter	0.5443	0.5406	0.5324	0.5384	0.5443	0.5059
pixelate	0.5299	0.5201	0.5208	0.5271	0.5285	0.5012
frost	0.5593	0.5512	0.5466	0.5584	0.5582	0.5116
speckle_noise	0.6030	0.5957	0.5706	0.6017	0.6038	0.4948
jpeg_compression	0.6224	0.6069	0.5958	0.6107	0.6204	0.5045
defocus_blur	0.5043	0.4999	0.5051	0.5015	0.5036	0.5015
Average	0.5770	0.5680	0.5575	0.5705	0.5759	0.5047

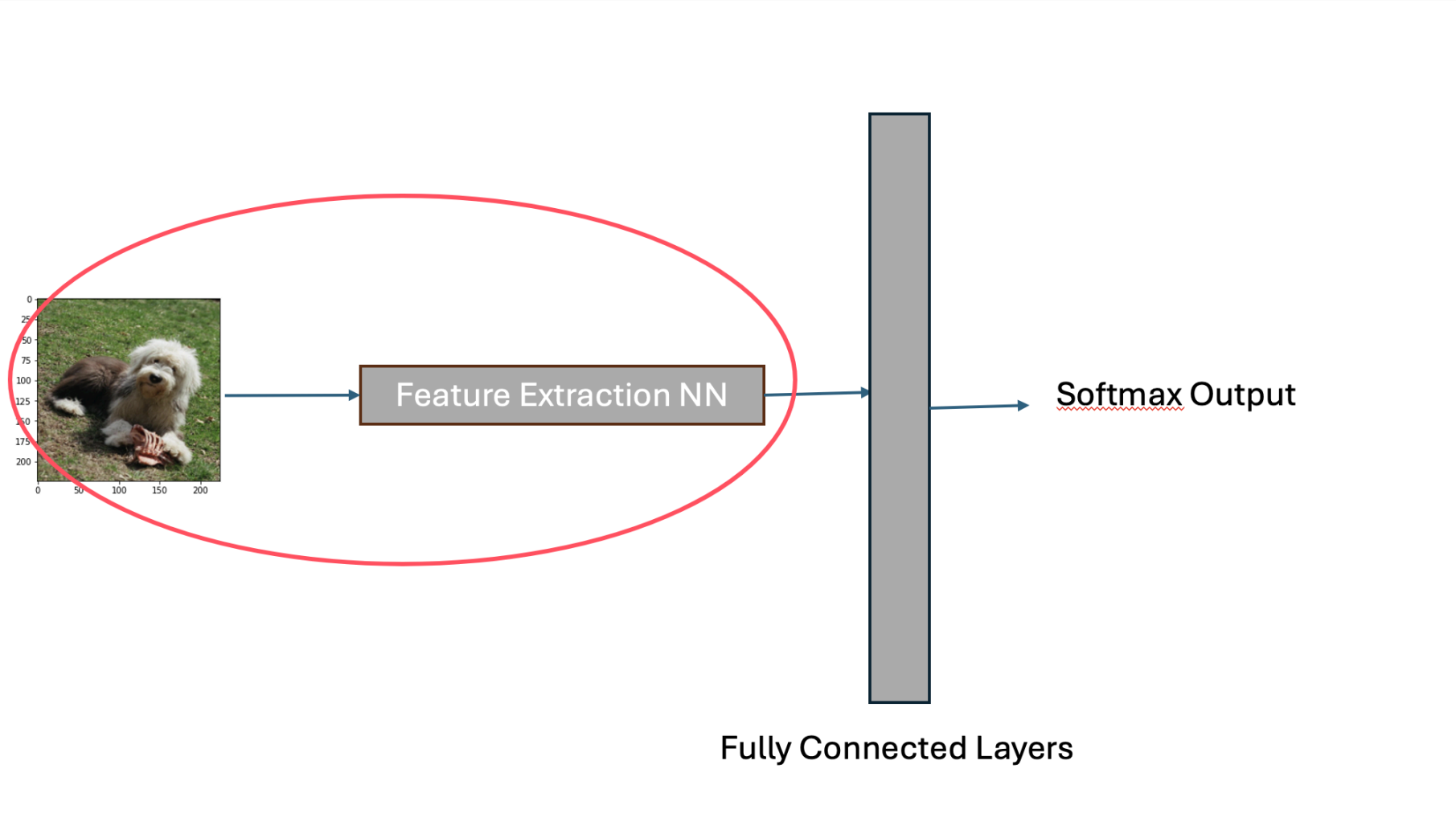
Best performing metric across all corruptions: cosine (Avg. AUC = 0.5770)

2.3 Saliency Maps are robust on Corrupted Data

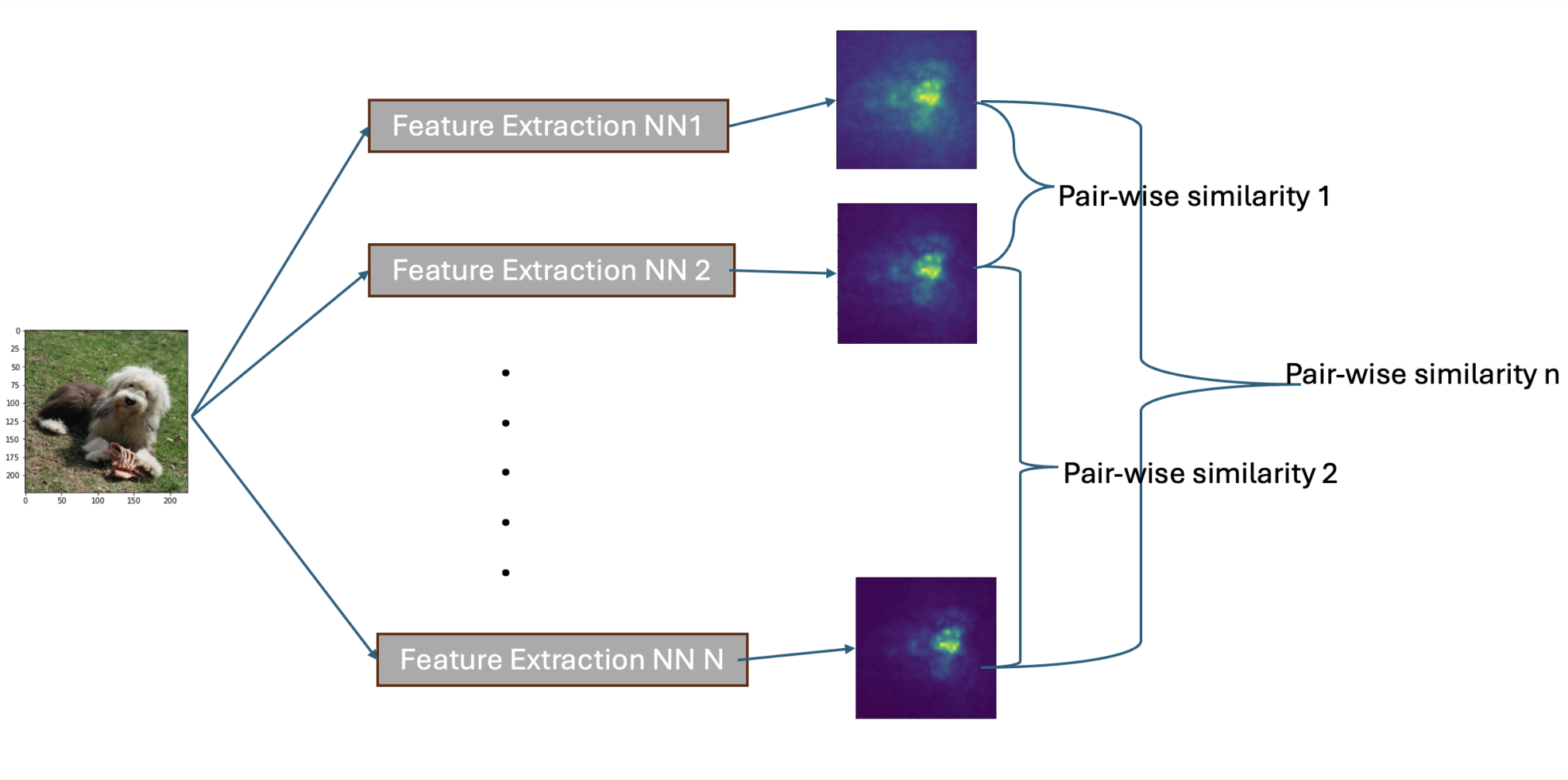
Thought: Can we use saliency maps to calibrate Epistemic Uncertainty so that EU can be distinct from AU?

3. Ensemble of Feature Map/Vector

3. Ensemble of Feature Map/Vector



3. Ensemble of Feature Map/Vector



3. Ensemble of Feature Map/Vector

Similarly we have different methods to computer similarity scores, but even the best results is not comparable. Because the Feature Extraction NN are trained with its own Fully-connected layers, different Feature Extraction NN can extract similar but not exactly the same features.

	Cifar 10 (Resnet 18) vs SVHN	
Models	Test Accuracy	AUROC
Deep Ensemble	**95.95%**	0.9683
Credal Ensemble	**96.43%**	**0.9813**
Single Credal Nets	95.78%	0.8646
Credal Wrapper (Mid Point)	95.91%	**0.9761**
Credal Wrapper of Size 30	95.95%	0.9794
Deep Ensemble of Size 30	96.10%	0.9794
Ensemble of Feature Map	\	0.8867
Ensemble of Feature Vector	\	0.8228

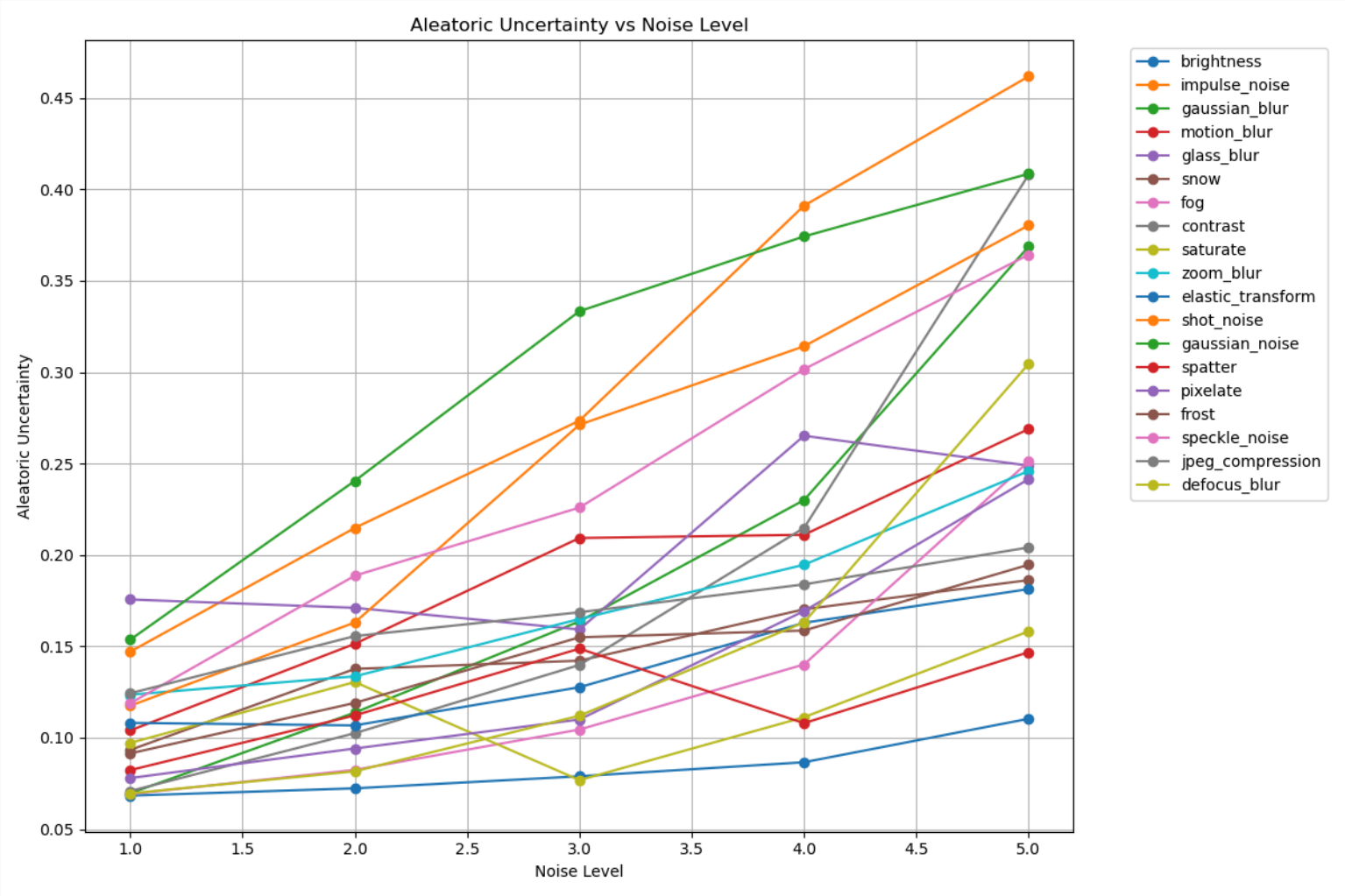
4. Rethinking Aleatoric and Epistemic Uncertainty

4.1 The Results on Corrupted CIFAR-10

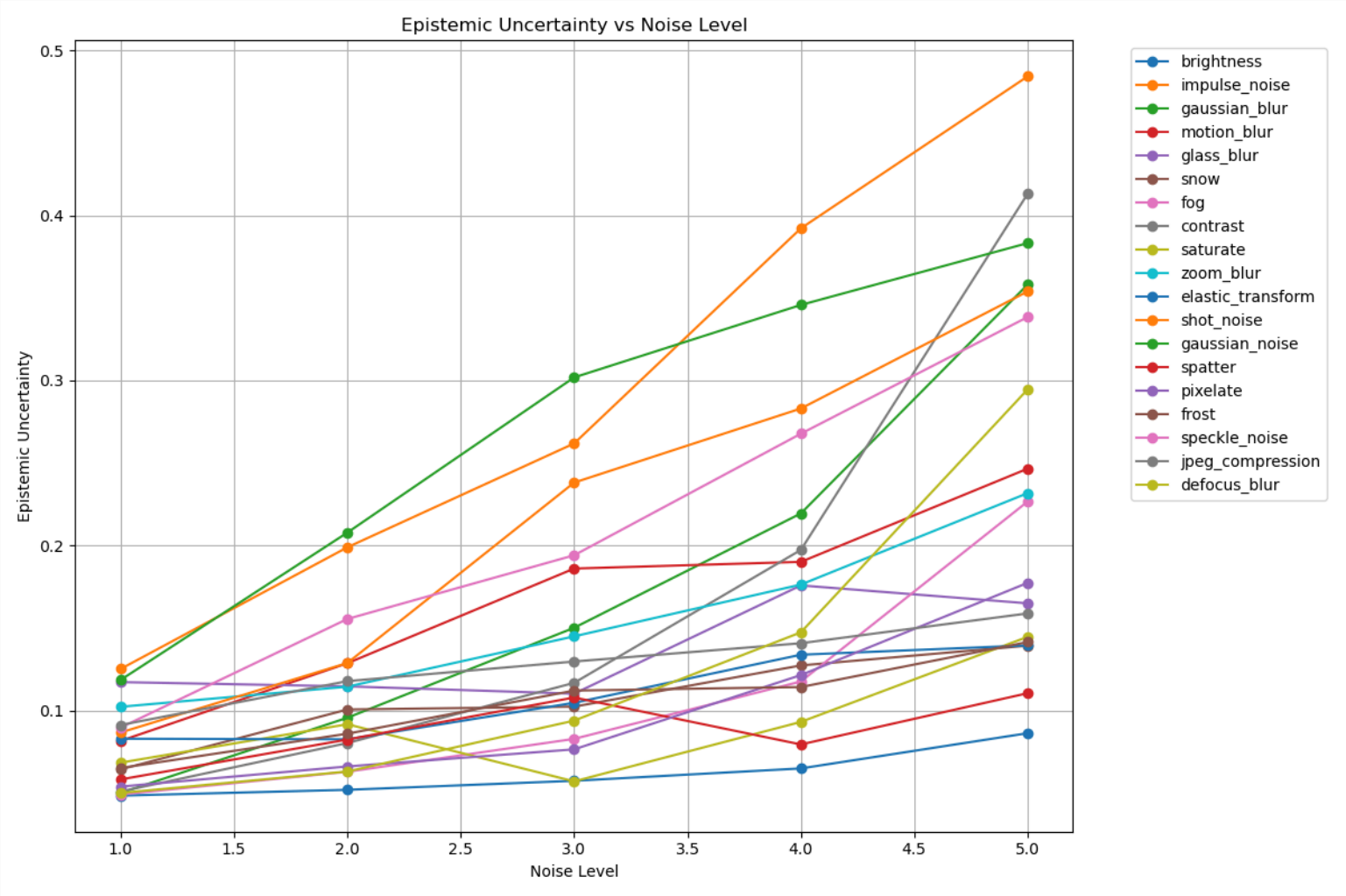
Previously, we trained model on Clean/high-quality data and test on the corrupted data, and we found that

- Aleatoric uncertainty and Epistemic uncertainty increase simultaneously, and they have a strong linear relationship, this phenomenon is also found by some published paper.
- Increasing AU and EU are reasonable, because AU measures the randomness in the data and EU measures whether the test data align withe the training data distribution, but the linear relationship seems problematic.

4.1 The Results on Corrupted CIFAR-10



4.1 The Results on Corrupted CIFAR-10



4.2 What I am Thinking?

- If I train a model on both clean data and low-quality(corrupted data), what will happen to the changes in AU and EU, compared with the model only trained on clean data.
- My expectation is
  - The EU should decrease, because the training data contains corrupted data, the model will have the knowledge of the corrupted version of data.
  - The AU whether keeps or decreases, but I expect that the relative ranking should keep. Like if *Impulse Noise* data has the highest AU in the original model (only trained on clean data), then it should also have highest AU in the new model (trained on both clean and corrupted data).
  - This way, even though the AU and EU are dependent, but at least AU can be robust.

4.3 Experiment

- We already have a model trained on clean data and the testing results the corrupted data.
- Train model on clean CIFAR-10 and Corrupted CIFAR-10 with Corruption Types: Gaussian Noise, Impulse Noise, Gaussian Blur and Contrast.
- Then we test the new model on Corrupted CIFAR-10, to see what's the ranking changes of the selected corruption types and what about others.

4.3.1 Preliminary Results

	Epistemic Uncertainty	Aleatoric Uncertainty	Total Uncertainty
Rank Correlation	0.6234	0.833	0.7662

- AU has high rank correlation, which means AU is robust and the rank mostly keeps.
- EU has lower rank correlation, this means after training on corrupted images, the model is can perform well on some corruption types, so the EU of them will decrease and change the ranking.

4.3.2 Unexpected Results

The model trained on different corruption types obatin a good generalization ability, the performance increases also on other corruption types.

4.3 Experiment

Corruption Type	Noise Level	Original Rank	Current Rank	Rank Change
contrast	4	25	63	-38
speckle_noise	2	31	61	-30
gaussian_blur	4	22	51	-29
shot_noise	3	14	40	-26
speckle_noise	3	23	49	-26
shot_noise	2	43	68	-25
gaussian_blur	3	41	64	-23
gaussian_blur	5	7	30	-23
contrast	3	57	80	-23
defocus_blur	5	11	34	-23
impulse_noise	2	24	46	-22
zoom_blur	4	29	50	-21
gaussian_noise	2	21	42	-21
impulse_noise	1	53	72	-19
impulse_noise	3	13	32	-19
gaussian_noise	1	50	69	-19
speckle_noise	4	12	31	-19
shot_noise	4	10	28	-18