Using Dynamic Neural Networks to Model the Speed-Accuracy Trade-Off in People Appendix

A Variation in performance and reaction time across human observers

In this section, we look at variation in accuracy and reaction time across all human observers in our 3 experiments that measured behavior under grayscale, noisy and blurry image perturbation conditions. Each block of trials in our experiments required participants to respond at a fixed reaction time chosen of 200, 400, 600, 800 or 1000 ms. Naturally, despite receiving training, observers showed some variance in their actual reaction time. Figure 1 plots mean and standard deviation of both accuracy and reaction time for each block of trials, separately for different image perturbations.

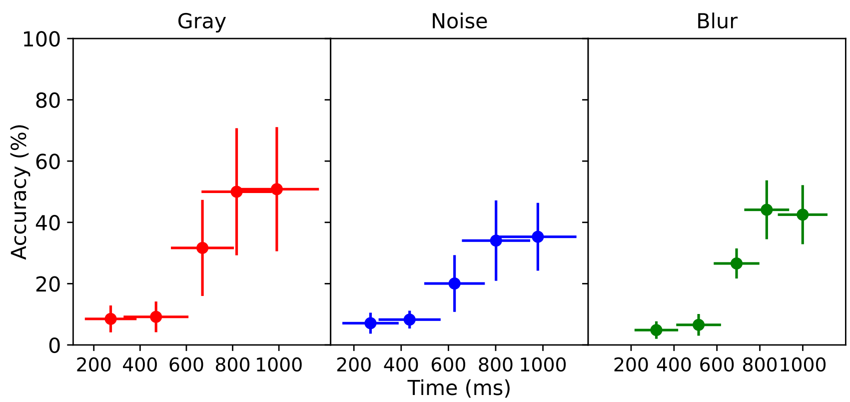

Figure 1: *Scatter plots showing mean and standard deviation of accuracy and reaction time across participants for different image perturbations*. Each point corresponds to a block of trials which required participants to respond within a specific duration in ms $\in \{200, 400, 600, 800, 1000\}$. 5 blocks of trials were run, and hence 5 points in each subplot.

We observe that participants in all experiments were highly accurate in their reaction times, showing means very close to the desired response time value and small variances (100 ms). Accuracies also showed the desired increasing trend, with small variance for short reaction times (200-600 ms) and large variance for longer reaction times(600-1000 ms).

B Correlations as a function of perturbation parameters

We report additional analysis of human and network (MSDNet [1], SCAN [2], ConvRNN [3]) results. In Table 1, we report the Pearson's r correlation coefficients of the human and network data at different levels of noise. As expected, human observers achieve the highest correlation to the average human results, followed closely by the MSDNet network.

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

Table 1: *Correlation (Pearson's r coefficient, ↑) of humans and each network with average human performance, evaluated across three noise (0.0, 0.04, 0.16)*. Each condition corresponds to a noise value used for human evaluation. For each network, noise value that shows the highest correlation with humans is found and shown. For each condition, the highest correlation is in bold, the second highest is underlined.

Observer	Noise		
	0	**0.04**	**0.16**
Human	**0.94**	**0.84**	**0.75**
MSDNet	0.93	**0.84**	0.74
ConvRNN	0.89	0.81	0.65
SCAN	0.74	0.66	0.51

We conduct a similar experiment with blur perturbations in Table 2. Interestingly, we find that there is a stronger correlation of human observers to the average human performance with blur than with noise. Similar to the previous result, humans have the highest correlation to humans, followed by MSDNet. For high blur values, MSDNet achieved a higher correlation to average human results than human-human correlation.

Table 2: *Correlation (Pearson's r coefficient, ↑) of humans and each network with average human performance, evaluated across three blur (0.0, 0.04, 0.16) conditions*. Each condition corresponds to the blur value used for human evaluation. For each network, the blur value that shows highest correlation with humans is found and shown. For each condition, the highest correlation is in bold, the second highest is underlined.

Observer	Blur		
	0	**1.0**	**3.0**
Human	**1.00**	**0.97**	0.70
MSDNet	0.96	0.95	**0.72**
ConvRNN	0.93	0.92	0.68
SCAN	0.94	0.92	0.68

C Performance range analysis

We additionally report performance ranges, i.e., the differences between the maximum and minimum accuracies, for human and neural networks. Table 3a reports results for noise and Table 3b reports results for blur. We find that the performance ranges for humans decrease as higher levels of either noise or blur is added. For networks the same phenomenon mostly holds for noise. For blur, we see that performance ranges actually increase for the SCAN and ConvRNN networks.

D Variation of performance range with training perturbations

In Figure 2, we study how the amount of training perturbation affects performance of networks when evaluated on perturbation-free images. We find that the performance ranges increase for all networks. For MSDNet, range increases from 13.87% to 19.24%; SCAN, from 4.34% to 6.58%; and ConvRNN from 9.03% to 14.18%.

E Network variants and parameter summary

MSDNet-L is the original MSDNet architecture, with FLOPs range from 15.13 MFLOPs to 75.86 MFLOPs. MSDNet-M refers to a model where we change filter dimensions in the classifiers from 128 to 32. MSDNet-M has fewer parameters and is 30% of the size of MSDNet-L. MSDNet-S is smaller in size and has early exits from 3.56 MFLOPs to 12.21 MFLOPs. It is 3.35% of the size of the MSDNet-L. It utilizes 1-1-1 setting for the bottleneck factor as compared to 1-2-4 setting in the original MSDNet. This is to add constraints to the original network and inhibit the ability of the initial layers to reach higher accuracy. The first early classifier is placed after 3 layers and the rest

Table 3: *Performance range (max accuracy minus min accuracy) of networks and human average reported evaluated across different noise/blur values*. For noise experiments, networks were trained with Gaussian noise with 0 mean and random batch standard deviation $\in [0.0, 0.05]$. For blur experiments, networks were trained with Gaussian blur with 0 mean and random batch standard deviation $\in [0.0, 0.9]$. The **No Perturbation** column reports the corresponding result with no noise used for training or testing, as a reference.

Observer	**No Perturbation**	**Testing Noise**											
	No Train/Test Noise	**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	**0.1**	**0.16**
Human	44.22	-	-	-	-	35.26	-	-	-	-	-	-	13.54
MSDNet	13.87	19.24	15.22	10.87	8.48	7.06	5.31	3.88	2.81	1.74	1.11	0.99	-
ConvRNN	9.03	14.18	5.63	3.89	2.90	2.89	4.04	6.38	6.56	5.03	3.02	-	-
SCAN	4.34	6.58	7.63	8.24	8.57	8.85	8.00	6.07	3.63	2.56	1.64	1.51	-

(a) Experiments with noise.

Observer	**No Perturbation**	**Testing Blur**									
	No Train/Test Noise	**0**	**0.25**	**0.5**	**0.75**	**1**	**1.25**	**1.5**	**1.75**	**2**	**3**
Human	44.22	-	-	-	-	42.67	-	-	-	-	15.86
MSDNet	13.87	18.27	18.29	22.08	5.16	7.65	7.69	7.06	6.52	6.13	-
ConvRNN	9.03	8.06	-	-	-	9.15	9.55	9.42	9.90	10.08	-
SCAN	4.34	4.28	-	-	-	5.37	16.90	17.26	6.00	4.20	-

(b) Experiments with blur.

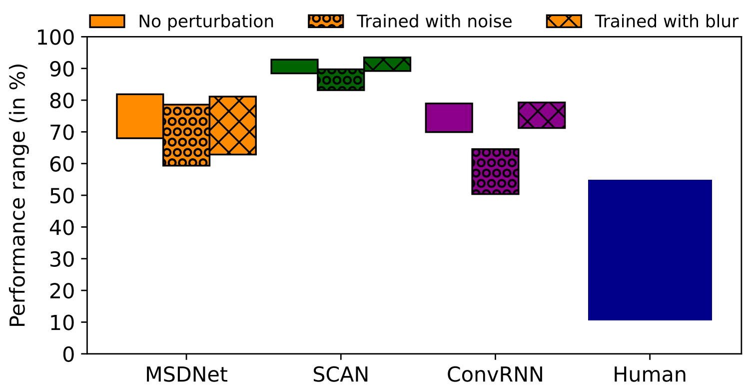

Figure 2: *Comparison between performance range of humans, and networks trained with different image perturbations and evaluated on grayscale CIFAR-10 images.* Training conditions considered are **a.** no perturbation (grayscale), **b.** Gaussian noise (0 mean, random batch standard deviation $\in [0.0, 0.05]$), **c.** Gaussian blur (0 mean, random batch standard deviation $\in [0.0, 0.9]$)

of classifiers are placed after every 2 layers. The first block contains scales of 8, 14 and 16 which sets up representations for the layers in the next blocks. We also change the filter dimensions in the classifiers from 128 to 32. Table 4 shows the comparison of MSDNet variants over the number of parameters and FLOPs range.

In Table 4, we summarize the number of parameters and FLOPs used by each network evaluated in the paper. Our experiments indicate that correlation to human performance does not necessarily increase with additional parameters.

F Additional image visualizations

We report visualizations of images with contrast adjustment and perturbations used for neural network experiments in Figure 3.

Table 4: *Number of parameters and range of MFLOPs for each network.* Correlation to human performance does not necessarily increase with additional parameters.

Observer	**# Params** ($\times 10^6$)	**MFLOPs**	
		Min.	**Max.**
MSDNet-S	0.10	3.56	12.21
MSDNet-M	0.90	12.36	54.21
MSDNet-L	2.98	15.13	75.86
ConvRNN	26.91	-	167060.12
SCAN-R9	8.71	76.76	173.14
SCAN-R18	14.98	190.86	627.72
SCAN-R34	25.09	266.94	1233.54

Figure 3: *Example images from the CIFAR-10 dataset [4] along with visualizations of image perturbations considered for neural network experiments.* Numbers in parentheses correspond to standard deviations for 0-mean Gaussian distributions in pixel units.

G Compute resources

In order to train and test models for all our experiments, we used resources from an internal cluster at New York University. All networks were trained using a single NVIDIA Tesla V100 GPU requiring < 32 GB of memory. Training time for all networks was under 8 hours. For each run of inference, we used a single NVIDIA GeForce GTX 1080 Ti GPU.

References

[1] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Weinberger. Multi-scale dense networks for resource efficient image classification. In *ICLR*, 2018.

[2] Linfeng Zhang, Zhanhong Tan, Jiebo Song, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. SCAN: A scalable neural networks framework towards compact and efficient models. *NeurIPS*, 2019.

[3] Courtney J Spoerer, Patrick McClure, and Nikolaus Kriegeskorte. Recurrent convolutional neural networks: a better model of biological object recognition. *Frontiers in psychology*, 8:1551, 2017.

[4] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.