Supplementary Materials for
Generalization of CNNs on Relational Reasoning with Bar Charts

1 Overview

The following supplementary materials offer additional data that support or explain the main findings of our paper. In Section 2, we provide the performance of different network architectures examined in our first revisiting study trained with different hyper-parameters. Section 3 and 4 offer detailed experimental results of the IID and OOD tests, specifically, the mean and confidence intervals in figure 5 and 6 of the main paper. In Section 5, we report the generalization performance of CNNs that were guided by segmentation masks. Section 6 details the performance, measured by Intersection over Union (IoU), of mask-enhanced CNN under various levels of perturbations. Section 7 compares the completion time between CNNs and humans across various types of input data, demonstrating their relative efficiency. Section 8 shows the visualizations produced by four popular neural network interpretation methods. Finally, Section 9 provides the training and validation loss curves in our replication experiment of Haehn et al. [2] for investigating the training validity. These supplementary sections are designed to enhance the reader's understanding of our study by providing detailed data and in-depth analyses.

2 Performance of different network architectures in the revisiting experiment

In our study, we trained multiple networks, including a multi-layer perceptron (MLP), traditional convolutional neural network architectures like AlexNet [6], LeNet [7], VGG19 [10] and ResNet152 [3], more modern designs like DenseNet [4] and EfficientNet [11], as well as a Relation Network [8] designed which tailored for relational reasoning. Table 1 reports the performance of different network architectures with different hyper-parameters in type 1 of the position-length experiment. We used 5-fold cross-validation, dividing the dataset into five equally sized subsets. We then iteratively used four subsets to train the model and reserved the remaining subset for validation. The best performance for each network was highlighted in bold. The best performance for VGG19 and ResNet152 was below -2.4, while other networks were above -2.4. Specifically, the best performance of the Relation Network which is designed for relational reasoning was 0.11, indicating poorer performance compared to VGG19 and ResNet152. Therefore, we chose VGG19 and ResNet152 as the primary focus for analysis in this paper.

Table 1: Performance of different network Architectures with different hyper-parameters in type 1 of the position-length experiment, where the cell with ✗ indicates that the corresponding parameter is invalid.

Network	Optimizer	Learning rate	Momentum	Weight decay	Validation mean	Validation CI	Test mean	Test CI
MLP	SGDM	High	Classic	✗	0.65	0.15	0.65	0.17
			Nesterov	✗	0.79	0.23	0.76	0.24
		Low	Classic	✗	0.61	0.41	0.62	0.41
			Nesterov	✗	**0.35**	0.13	**0.36**	0.13
	AdamW	High	✗	High	4.09	0.08	4.07	0.10
			✗	Low	4.09	0.09	4.07	0.08
		Low	✗	High	1.85	0.08	1.78	0.09
			✗	Low	1.94	0.14	1.87	0.14
AlexNet	SGDM	High	Classic	✗	-0.30	0.19	-0.33	0.16
			Nesterov	✗	-0.20	0.15	-0.22	0.14
		Low	Classic	✗	1.85	1.80	1.80	1.82
			Nesterov	✗	2.51	1.88	2.52	1.85
	AdamW	High	✗	High	-1.80	0.05	-1.79	0.04

			✗	Low	-2.00	0.31	-2.00	0.30
		Low	✗	High	**-2.12**	0.09	**-2.11**	0.06
			✗	Low	-1.87	0.05	-1.85	0.05
LeNet	SGDM	High	Classic	✗	-1.34	0.10	-1.32	0.09
			Nesterov	✗	-1.41	0.10	-1.42	0.11
		Low	Classic	✗	-0.73	0.30	-0.73	0.30
			Nesterov	✗	-0.50	0.52	-0.51	0.50
	AdamW	High	✗	High	-1.90	0.06	-1.90	0.05
			✗	Low	-1.99	0.05	-1.94	0.08
		Low	✗	High	**-2.21**	0.27	**-2.21**	0.26
			✗	Low	-1.91	0.09	-1.90	0.08
VGG19	SGDM	High	Classic	✗	1.43	0.05	1.44	0.05
			Nesterov	✗	1.10	0.22	1.08	0.22
		Low	Classic	✗	1.41	0.15	1.40	0.16
			Nesterov	✗	1.04	0.27	1.05	0.27
	AdamW	High	✗	High	-2.32	0.11	-2.34	0.11
			✗	Low	**-2.41**	0.10	**-2.42**	0.09
		Low	✗	High	-1.89	0.35	-1.90	0.34
			✗	Low	-1.71	0.25	-1.72	0.25
ResNet152	SGDM	High	Classic	✗	-0.61	0.24	-0.64	0.26
			Nesterov	✗	-0.60	0.21	-0.60	0.21
		Low	Classic	✗	0.72	0.24	0.72	0.24
			Nesterov	✗	0.70	0.25	0.73	0.23
	AdamW	High	✗	High	**-2.76**	0.10	**-2.76**	0.10
			✗	Low	-2.42	0.17	-2.42	0.17
		Low	✗	High	-2.71	0.10	-2.71	0.11
			✗	Low	-2.53	0.24	-2.53	0.25
DenseNet	SGDM	High	Classic	✗	-0.34	0.25	-0.35	0.25
			Nesterov	✗	-0.43	0.28	-0.38	0.30
		Low	Classic	✗	0.71	0.28	0.71	0.30
			Nesterov	✗	0.85	0.20	0.83	0.23
	AdamW	High	✗	High	-2.17	0.49	-2.17	0.49
			✗	Low	-2.33	0.28	-2.33	0.27
		Low	✗	High	**-2.37**	0.21	**-2.36**	0.21
			✗	Low	-2.29	0.30	-2.28	0.30
EfficientNet	SGDM	High	Classic	✗	0.96	0.17	0.96	0.16
			Nesterov	✗	0.91	0.10	0.88	0.09
		Low	Classic	✗	1.82	0.22	1.80	0.20
			Nesterov	✗	1.34	0.27	1.34	0.27
	AdamW	High	✗	High	-0.01	1.81	-0.01	1.81
			✗	Low	-2.17	0.19	-2.17	0.18
		Low	✗	High	**-2.37**	0.14	**-2.37**	0.14
			✗	Low	-1.71	0.45	-1.69	0.44
Relation Network	SGDM	High	Classic	✗	0.57	0.07	0.57	0.08
			Nesterov	✗	0.53	0.08	0.49	0.06
		Low	Classic	✗	1.08	0.18	1.09	0.17
			Nesterov	✗	0.96	0.14	0.91	0.15
	AdamW	High	✗	High	0.15	0.01	0.05	0.02
			✗	Low	0.16	0.02	0.12	0.02
		Low	✗	High	0.14	0.03	0.06	0.02
			✗	Low	**0.11**	0.04	**0.09**	0.06

3 Performance of ResNet152 on IID and OOD tests

In this section, we present the experimental results of ResNet152 generalization on the GRAPE dataset for standard and perturbed chart visualizations. Table 2 shows that the mean and confidence intervals of MLAE values produced by CNNs on performing generalization tests of nine parameters on five types of bar charts. Perturbation levels represent varying degrees of perturbation. We find that the relational reasoning ability of CNNs is heavily influenced by most visual parameters.

Table 2: The means and CIs of MLAE values produced by CNNs on performing generalization tests of nine parameters on five types of bar charts.

Perturb. level		-45%		-30%		-15%		0%		15%		30%		45%	
		MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI
Title position	Type 1	-2.44	0.16	-2.51	0.14	-2.52	0.14	-2.52	0.14	-2.51	0.15	-2.51	0.15	-2.53	0.14
	Type 2	3.04	0.06	-2.50	0.16	-2.50	0.17	-2.40	0.21	-1.93	0.35	-2.01	0.24	-2.50	0.16
	Type 3	-2.04	0.25	-2.14	0.23	-2.05	0.24	-2.02	0.24	-1.97	0.25	-2.02	0.25	-2.12	0.23
	Type 4	-1.87	0.25	-1.86	0.22	-1.90	0.21	-1.91	0.23	-1.86	0.25	-1.85	0.24	-1.87	0.22
	Type 5	1.66	0.48	-1.98	0.30	-2.07	0.28	-2.06	0.29	-2.01	0.29	-1.94	0.31	-2.03	0.30

Perturb. level		-3		-2		-1		0		1		2		3	
		MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI
Title font size	Type 1	-2.52	0.14	-2.52	0.15	-2.52	0.14	-2.52	0.14	-2.51	0.14	-2.51	0.14	-2.49	0.14
	Type 2	-2.49	0.18	-2.50	0.17	-2.50	0.18	-2.50	0.17	-2.49	0.18	-2.50	0.18	-2.50	0.18
	Type 3	-2.05	0.23	-2.07	0.23	-2.05	0.23	-2.03	0.23	-2.05	0.23	-2.03	0.23	-2.08	0.23
	Type 4	-1.90	0.22	-1.91	0.22	-1.90	0.22	-1.91	0.22	-1.90	0.22	-1.90	0.22	-1.88	0.22
	Type 5	-2.07	0.29	-2.07	0.28	-2.07	0.28	-2.08	0.28	-2.07	0.28	-2.07	0.28	-2.00	0.30

Perturb. level		-25		-20		-15		-10		-5		0		5	
		MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI
Background color	Type 1	4.34	0.44	4.32	0.45	4.40	0.52	4.31	0.45	-2.16	0.31	-2.52	0.14	-2.52	0.14
	Type 2	4.31	0.45	4.32	0.45	4.32	0.44	4.34	0.43	4.40	0.57	-2.45	0.21	-2.50	0.17
	Type 3	4.34	0.43	4.44	0.58	4.07	0.66	-0.92	0.69	-1.98	0.24	-2.04	0.23	-2.03	0.23
	Type 4	4.80	0.55	5.17	0.42	5.30	0.43	4.00	1.23	-1.75	0.27	-1.91	0.22	-1.82	0.26
	Type 5	3.85	0.50	3.89	0.42	4.00	0.41	-0.28	0.50	-1.98	0.30	-2.08	0.28	-2.08	0.28

Perturb. level		-15		-10		-5		0		5		10		15	
		MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI
Bar color	Type 1	-2.08	0.27	-2.40	0.16	-2.50	0.14	-2.52	0.14	-2.50	0.14	-2.40	0.17	-2.27	0.23
	Type 2	-0.79	1.61	-2.46	0.20	-2.59	0.15	-2.50	0.17	-2.35	0.24	-2.08	0.33	-1.75	0.38
	Type 3	-2.09	0.22	-2.08	0.22	-2.06	0.23	-2.04	0.23	-1.99	0.24	-1.95	0.24	-1.91	0.24
	Type 4	3.69	1.09	-1.43	0.38	-1.89	0.24	-1.90	0.22	-1.89	0.22	-1.87	0.22	-1.85	0.22
	Type 5	1.31	0.91	-1.19	0.53	-1.97	0.31	-2.08	0.28	-2.06	0.28	-2.07	0.28	-2.05	0.29

Perturb. level		-5		0		5		10		15		20		25	
		MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI
Stroke color	Type 1	-2.51	0.15	-2.52	0.14	-2.51	0.14	-2.48	0.15	-2.45	0.17	-2.34	0.23	-2.15	0.34
	Type 2	-2.50	0.18	-2.50	0.17	-2.50	0.18	-2.45	0.19	-2.28	0.28	-1.84	0.69	-0.91	1.58
	Type 3	-2.03	0.23	-2.04	0.23	-2.04	0.23	-2.04	0.23	-2.02	0.23	-1.95	0.25	-1.83	0.30
	Type 4	-1.92	0.23	-1.90	0.22	-1.86	0.22	-1.79	0.25	-1.50	0.46	-0.44	1.57	0.56	2.10
	Type 5	-2.07	0.27	-2.08	0.28	-2.03	0.30	-1.89	0.34	-1.49	0.58	-0.25	1.93	1.04	2.56

Perturb. level		-3		-2		-1		0		1		2		3	
		MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI
Bar width	Type 1			-0.36	1.05	-1.71	0.43	-2.52	0.14	-2.08	0.30	-0.55	0.65		
	Type 2	0.90	0.72	-0.42	0.57	-2.09	0.22	-2.50	0.17	-0.10	0.29	1.00	0.35	1.39	0.53
	Type 3			-1.72	0.27	-1.96	0.23	-2.03	0.23	-1.82	0.24	-1.58	0.26		
	Type 4	5.72	0.29	5.61	0.31	4.22	0.74	-1.91	0.22	0.72	0.61	5.47	0.34	4.36	0.54
	Type 5	1.27	0.67	-0.03	0.47	-0.59	0.58	-2.08	0.28	-1.27	0.50	2.19	0.77	2.26	0.55

Perturb. level		-1		0		1	
		MLAE	CI	MLAE	CI	MLAE	CI
Stroke width	Type 1	-0.79	0.85	-2.52	0.14	-1.37	0.36
	Type 2	3.84	0.67	-2.50	0.17	3.67	0.86
	Type 3	-0.85	0.97	-2.03	0.23	-2.07	0.26
	Type 4	4.34	0.51	-1.91	0.22	0.63	0.49
	Type 5	5.60	0.41	-2.08	0.28	0.71	0.60

Perturb. level		-3		-2		-1		0		1		2		3	
		MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI	MLAE	CI
Dot position	Type 1	4.26	0.47	2.06	0.54	-2.10	0.21	-2.52	0.14	-1.27	0.30	3.92	0.54	4.26	0.47
	Type 2	-2.49	0.16	-2.50	0.17	-2.50	0.17	-2.50	0.17	-2.51	0.18	-2.52	0.19	-2.51	0.20
	Type 3	1.28	1.04	3.71	0.62	-1.79	0.25	-2.03	0.23	-1.69	0.28	3.78	0.66	4.18	0.49
	Type 4	2.76	0.72	0.55	0.82	-1.83	0.26	-1.91	0.22	-1.75	0.34	1.69	0.93	1.88	1.09
	Type 5	0.60	1.10	-1.06	0.66	-1.79	0.36	-2.08	0.28	-1.64	0.38	-0.43	0.99	1.29	1.04

Perturb. level		1-9|94-100		10-93	
		MLAE	CI	MLAE	CI
Bar length	Type 1	4.76	0.42	-2.52	0.14
	Type 2	1.69	1.41	-2.50	0.17
	Type 3	4.52	0.68	-2.03	0.23
	Type 4	1.94	0.61	-1.91	0.22
	Type 5	2.02	0.56	-2.08	0.28

4 Robustness comparison between CNNs and humans under the largest perturbation

In this section, we report the comparative data of performance and generalization in CNNs and human subjects. Table 3 presents the mean and CI of MLAE values produced by humans and CNNs on five types of bar charts without and with the largest level perturbations on eight parameters. We observe that humans are more robust than CNNs, with all tested levels of the most influential parameters. We surmise that while humans are primarily swayed by the length of the bars, CNNs are impacted by various other factors.

Table 3: The means and 95%CIs of MLAE values produced by CNNs and humans on five types of bar charts without and with the largest level perturbations on eight parameters.

Type	Parameter	CNN mean	CNN 95%CI	Human mean	Human 95%CI
Type 1	Standard	-2.47	0.02	1.40	0.38
	Title position	-2.41	0.02	1.44	0.37
	Stroke width	-1.18	0.07	1.40	0.50
	Bar width	0.18	0.15	1.28	0.44
	Bkgd color	3.97	0.09	1.42	0.46
	Stroke color	-1.75	0.09	1.15	0.40
	Bar color	-1.95	0.05	0.90	0.45
	Bar length	4.42	0.10	2.66	0.24
	Dot position	3.92	0.09	1.74	0.32
Type 2	Standard	-2.39	0.03	1.85	0.32
	Title position	2.95	0.03	1.88	0.31
	Stroke width	3.34	0.13	2.21	0.26
	Bar width	0.80	0.09	1.80	0.34
	Bkgd color	3.99	0.09	2.03	0.33
	Stroke color	-0.03	0.18	1.39	0.39
	Bar color	-0.10	0.17	1.74	0.30
	Bar length	1.61	0.17	2.63	0.26
	Dot position	-2.40	0.03	1.88	0.24
Type 3	Standard	-1.91	0.05	1.02	0.52
	Title position	-1.89	0.06	1.17	0.45
	Stroke width	-1.91	0.05	1.37	0.41
	Bar width	-1.67	0.05	1.60	0.40
	Bkgd color	4.03	0.09	1.35	0.39
	Stroke color	-1.47	0.10	0.75	0.49
	Bar color	-1.83	0.05	1.09	0.38
	Bar length	4.21	0.10	2.64	0.33
	Dot position	1.33	0.14	1.71	0.39
Type 4	Standard	-1.81	0.05	1.92	0.35
	Title position	-1.75	0.06	1.53	0.40
	Stroke width	0.53	0.08	2.36	0.30
	Bar width	5.55	0.05	2.21	0.30
	Bkgd color	4.49	0.10	1.78	0.47
	Stroke color	0.96	0.19	1.95	0.34
	Bar color	3.08	0.16	2.13	0.25
	Bar length	1.88	0.13	3.02	0.31
	Dot position	1.80	0.14	2.20	0.25
Type 5	Standard	-1.91	0.06	2.14	0.35
	Title position	1.41	0.08	2.22	0.34
	Stroke width	0.59	0.09	2.22	0.30
	Bar width	1.40	0.11	1.94	0.41
	Bkgd color	3.63	0.09	2.31	0.40
	Stroke color	1.42	0.22	1.97	0.44
	Bar color	1.59	0.15	2.16	0.29
	Bar length	1.69	0.15	2.70	0.31
	Dot position	1.42	0.14	2.20	0.29

5 Quantitative analysis of Grad-CAM map

In this section, we report quantitative evaluation of all test images by calculating the Intersection over Union (IoU) between the area of target bars and the high-intensity region of the Grad-CAM map. By incorporating IoU as a quantitative evaluation metric, we can gain deeper insights into the effectiveness of the Grad-CAM technique in highlighting relevant regions in the input images. Table 4 presents the IoU values of the type 1 over different levels of perturbations of one of eight parameters, as well as the IoU values after mask-enhanced. When without masks, this indicates that the CNN regions mainly used for relational inference on bar charts are rarely the target bars. After mask-enhanced, the segmentation masks significantly enhance the ability of CNNs to localize the target bars in bar charts.

Table 4: The Intersection over Union between the high-intensity regions in Grad-CAM map and target bars areas.

	Perturb. level	-3		-2		-1		0		1		2		3	
		Mean	CI	Mean	CI	Mean	CI	Mean	CI	Mean	CI	Mean	CI	Mean	CI
Mask	Title position	0.30	0.09	0.30	0.09	0.30	0.09	0.30	0.09	0.30	0.09	0.30	0.09	0.30	0.09
	Title font size	0.30	0.09	0.30	0.09	0.30	0.09	0.30	0.09	0.30	0.09	0.30	0.09	0.30	0.09
	Bkgd color	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09
	Bar color	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09
	Stroke color	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09	0.29	0.09
	Bar width			0.25	0.06	0.27	0.07	0.30	0.09	0.32	0.10	0.34	0.11		
	Stroke width					0.30	0.09	0.28	0.09	0.30	0.09				
	Dot position	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09	0.28	0.09
	Bar length							0.28	0.09	0.12	0.02				
Type 1	Title position	0.009	0.029	0.009	0.028	0.009	0.026	0.008	0.025	0.008	0.024	0.008	0.024	0.009	0.027
	Title font size	0.009	0.027	0.009	0.027	0.009	0.026	0.009	0.026	0.009	0.026	0.009	0.026	0.009	0.026
	Bkgd color	0.0	0.0	0.0	0.0	0.000	0.002	0.001	0.011	0.004	0.020	0.009	0.026	0.010	0.029
	Bar color	0.028	0.050	0.025	0.047	0.018	0.039	0.009	0.027	0.004	0.019	0.002	0.017	0.002	0.016
	Stroke color	0.008	0.025	0.009	0.026	0.010	0.028	0.011	0.029	0.011	0.030	0.011	0.030	0.011	0.029
	Bar width			0.001	0.000	0.000	0.006	0.009	0.026	0.013	0.036	0.017	0.043		
	Stroke width					0.013	0.033	0.009	0.026	0.003	0.016				
	Dot position	0.009	0.026	0.009	0.027	0.009	0.027	0.009	0.026	0.009	0.026	0.009	0.026	0.009	0.026
	Bar length							0.009	0.026	0.003	0.177				

6 Performance of segmentation mask-enhanced CNN

In this section, we present data supporting our approach to enhancing CNN generalization with segmentation masks, showing improved robustness and generalization. Table 5 shows how MLAE values change over different levels of perturbations of one of eight parameters in mask-enhanced CNN model. We find that the robustness of CNNs against perturbations in visual encodings of bar charts, such as title position, background color, and bar color, has significantly improved. While exhibiting improved robustness against various perturbations, the mask-enhanced CNN model remains susceptible to alterations in bar width, stroke width and bar length.

Table 5: Generalization performance of CNNs provided with segmentation masks.

Perturb. level	-3		-2		-1		0		1		2		3	
	Mean	CI	Mean	CI	Mean	CI	Mean	CI	Mean	CI	Mean	CI	Mean	CI
Title position	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.58	0.12
Title font size	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.58	0.12
Bkgd color	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.58	0.12
Bar color	-2.51	0.13	-2.55	0.12	-2.57	0.12	-2.58	0.12	-2.58	0.12	-2.57	0.12	-2.56	0.12
Stroke color	-2.58	0.11	-2.58	0.12	-2.57	0.12	-2.57	0.12	-2.56	0.12	-2.55	0.13	-2.53	0.13
Bar width			0.52	0.46	-1.03	0.43	-2.58	0.12	-1.43	0.41	-0.70	0.45		
Stroke width					-2.37	0.18	-2.58	0.12	-2.45	0.16				
Dot position	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.58	0.12	-2.59	0.12	-2.59	0.11	-2.60	0.11
Bar length							-2.58	0.12	2.28	1.09				

7 Comparison of completion time between CNNs and humans

We recorded the total time each participant took to complete the experiments and compared it with the ResNet152's inference time on the test set. The box plot in Figure 1 illustrates the distribution of completion time for humans and CNNs across different types. The results showed that human inference time was significantly longer than that of CNNs in all types.

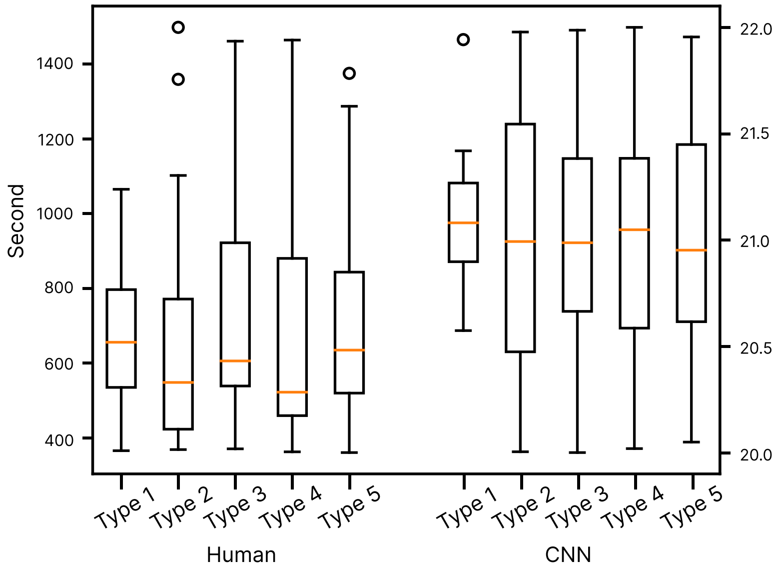

Figure 1: The completion time of humans and CNNs across different types of charts.

8 Visualizations from different neural network interpretation methods

In this section, we employ various neural network interpretation techniques, including Grad-CAM [9], LayerCAM [5], Score-CAM [12] and Deep Feature Factorization [1], to gain insights into the behavior of CNNs trained on our tasks. As shown in Figure 2, these methods can yield different results, reflecting the broader challenge in the XAI community of determining which techniques reliably capture the true decision-making process of neural networks.This variability indicates that different methods may emphasize distinct regions of importance, which suggests that caution is necessary when interpreting these saliency maps.

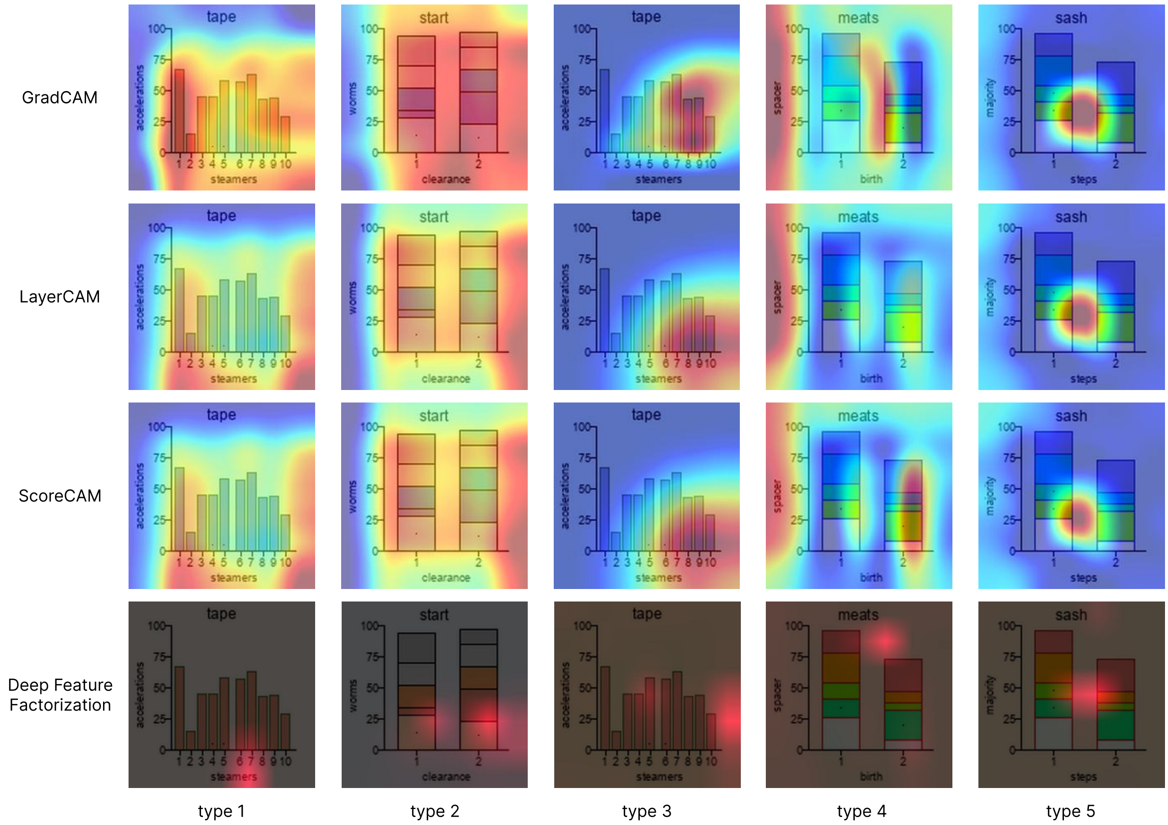

Figure 2: The example visualizations of four neural network interpretation methods.

9 Loss curves for training and validation

In our replication of Haehn et al. [2], all our models outperformed their best model (VGG19 with an MLAE of 3.51), including even the simple MLP (MLAE of 0.35). Given this unexpected outcome, we reviewed the training process by comparing the training and validation loss curves to rule out issues such as overfitting.

Figure 3 shows the training and validation loss curves of four example network architectures — ResNet152, VGG19, Relation Network, and MLP — each trained with hyperparameters that produce the best performance. These curves represent the Mean Squared Error (MSE) and Log Absolute Error (LAE) across training epochs. Although MSE is used as the training objective function, its values are too small to effectively reflect the differences in model performance. Therefore, we use LAE to better capture and signify these variations.

The training curves demonstrate a clear downward trend as the epochs advance, indicating effective learning from the training dataset. Similarly, the validation curves mirror this trend, suggesting that the models are generalizing well to the validation set. These observations imply that the models of different architectures are learning effectively without evident signs of overfitting.

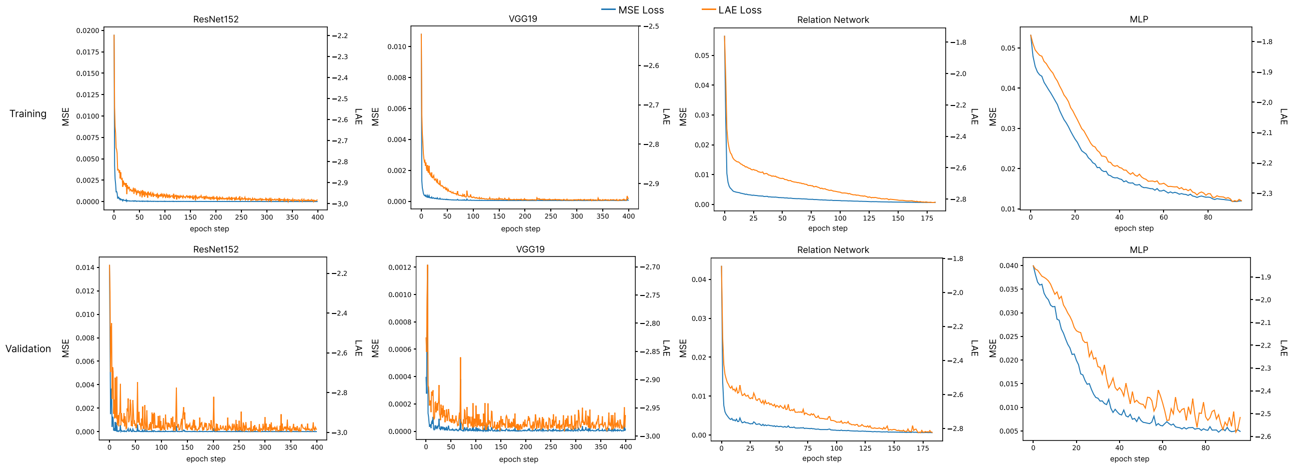

Figure 3: Loss curves for training and validation of four example neural networks.

References

[1] E. Collins, R. Achanta, and S. Susstrunk. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 336–352, 2018. DOI: 10.1007/978-3-030-01264-9_21

[2] D. Haehn, J. Tompkin, and H. Pfister. Evaluating 'graphical perception' with CNNs. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):641–650, 2019. DOI: 10.1109/TVCG.2018.2865138

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90

[4] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017. DOI: 10.1109/CVPR.2017.243

[5] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. DOI: 10.1109/TIP.2021.3089943

[6] A. Kirzhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. DOI: 10.1145/3065386

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. DOI: 10.1109/5.726791

[8] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. DOI: 10.5555/3295222.3295250

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017. DOI: 10.1109/ICCV.2017.74

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2015. DOI: 10.48550/arXiv.1409.1556

[11] M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 6105–6114. PMLR, 2019. DOI: 10.48550/arXiv.1905.11946

[12] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25, 2020. DOI: 10.1109/CVPRW50498.2020.00020