Evaluating 'Graphical Perception' with CNNs

Daniel Haehn, James Tompkin, and Hanspeter Pfister

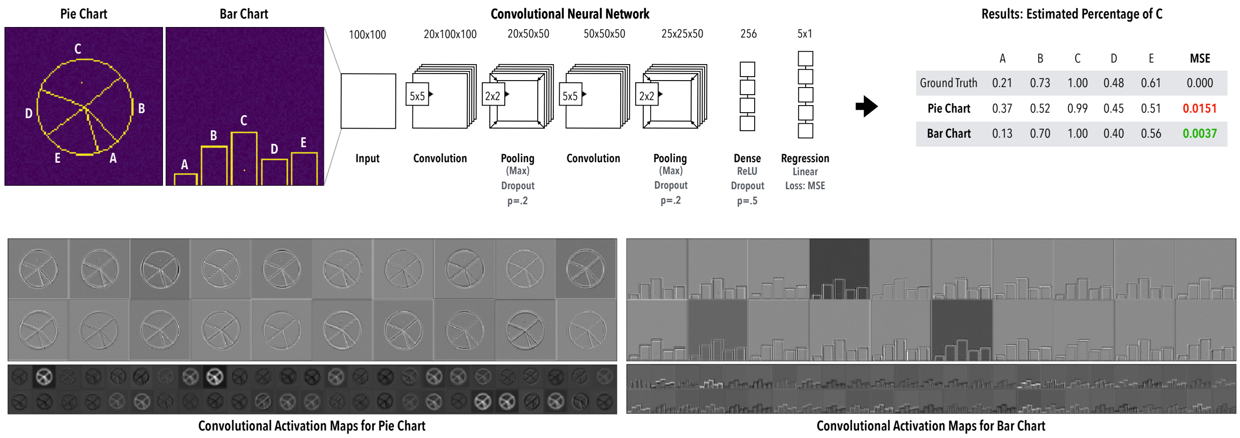

Fig. 1. **Computing Cleveland and McGill's Position-Angle Experiment using Convolutional Neural Networks.** We replicate the original experiment by asking visual cortex inspired machine learning classifiers to assess the relationship between values encoded in pie charts and bar charts. Similar to the findings of Clevenland and McGill [3], our experiments show that CNNs read quantities more accurately from bar charts (mean squared error, MSE in green).

Abstract—Convolutional neural networks can successfully perform many computer vision tasks on images, and their learned representations are often said to mimic the early layers of the visual cortex. But can CNNs understand graphical perception for visualization? We investigate this question by reproducing Cleveland and McGill's seminal 1984 experiments, which measured human perception efficiency of different visual encodings and defined elementary perceptual tasks for visualization. We measure the graphical perceptual capabilities of four classifiers on a) elementary perceptual tasks with increasing parametric complexity, b) the position-angle experiment that compares pie charts to bar charts, c) the position-length experiment that compares grouped and divided bar charts, and d) the bars and framed rectangles experiment where visual cues aid perception. We also study how feed-forward neural networks obey Weber's law, which defines the proportional relation between perceivable information and distribution density. We present the results of these experiments to foster the understanding of how CNN classifiers succeed and fail when applied to data visualizations.

Index Terms—Machine Perception, Deep Learning

1 Introduction

Artificial intelligence has taken the world of technology by storm. Deep multilayer neural networks are being successfully applied in a wide range of applications that are regularly outperforming humans in object recognition [13, 20, 21]. Originally inspired by neuroscientific discoveries, the recent advances in deep learning have been the direct results of engineering efforts, more specifically in convolutional neural networks (CNNs). While there has been significant advancement, this does not mean we understand what CNNs are doing and we are in fact treating them as blackboxes without detracting from their success [7, 19]. Our current knowledge of biological vision suggests that modern machine learning models indeed mimic the underlying biology by abstracting the many details of biological neural networks [9, 24].

Despite tremendous research efforts and generating massive datasets, we are far from fully understanding biological vision. Similarly, we are constantly developing inventive features for deep artificial networks without truly understanding it in its entirety. As a result, a double discrepancy is observed. Advances in neuroscience can revolutionize machine learning by reverse-engineering neural circuits, yielding new classifiers which, in turn, can help process the massive biological data. There are many existing questions which must be answered in order to fill this gap of our understanding.

We focus on perception... experiments from cleveland mcgill..

1.1 Biological Vision

Biological vision is an extremely powerful system which allows humans the ability, and seemingly without effort, to recognize an enormous amount of distinct objects in the world. Object detection is extremely difficult and therefore is especially impressive as light intensities can change by levels of magnitude and contrast between foreground and background is so often low. In addition, the visual scene changes every time the human body or human eyes move. This visual system exhibits a very noisy structure but because it is organized by layers it has inspired the mathematical theory of multilayer neural networks. What is remarkable is that even though current machine learning models do not resemble the complexity of its biological pendant, they inherently generalize extremely well. Neural networks trained on one specific task can be used to perform detection or segmentation of, seemingly, unrelated objects with relatively minor retraining. The reported classification performance is superior to that of humans and the question in

- *Daniel Haehn, and Hanspeter Pfister are with the Paulson School of Engineering and Applied Sciences at Harvard University. E-mail: {haehn,pfister}@seas.harvard.edu.*
- *James Tompkin is with the Thomas J. Watson Sr. Center for Information Technology at Brown University. E-mail: james_tompkin@brown.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

regards to their functionality opens an interesting research topic.

In 1962 Hubel and Wiesel were the first to begin studying the visual cortex from the standpoint of a neuroscientist. Their experimental findings on cats and macaque monkeys suggested a hierarchy of cells with increasing complexity which was then later transferred to the hierarchical model of different layers. Twenty years later, this insight was translated to the Neocognitron quantitative model, by Fukushima and Miyake, which ultimately led to the important work of Hinton, Bengio, and LeCun in the 1980s. Their work on stochastic gradient descent approximation, and the availability of faster computer hardware then led to todays breakthrough of deep learning networks. In the last decade, this field has exhibited rapid growth, constant evolution, and new applications in various domains.

GOALS

- reduce the gap between neurobiology and data science to advance the understanding of visual cortex inspired machine learning
- insights for infovis for machines

CONTRIBUTIONS

- experiments of cleveland mcgill with systematic parametrization and evaluation for computational perception
- ranking of elementary perceptual tasks for machine perception
- validation of weber's law for feedforward neural networks
- many insights for infovis for machines

2 Previous Work

Graphical Perception. Cleveland and McGill [3] introduced the concept of *graphical perception* and investigated how different visual attributes and encodings are perceivable by humans. They define *elementary perceptual tasks* as mental-visual stimuli to understand encodings in visualizations.

[10] Heer and Bostock, CHI 2010 — Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design

[22] Comparing linegraph vs. scatterplot

[4] [6] [14] 2D vs. 3D Evaluation (vector fields)

[12] [17] [23] [18] color maps

2d vs. 3d + color [1]

black hat vis [5]

timeseries [11]

munzner [15]

open vs. closed shaped [2]

visualization ranking [8]

Visual Cortex Inspired Machine Learning. MLP, LeNet, VGG, ImageNet, XCeption, ResNets etc and work from THomas Serre

Computing Perception.

Pineo et al. [16] present a method to automatically evaluate and optimize visualizations using a computational model of human vision, based on a neural network simulation of the early perceptual processing in the retina and primary visual cortex. [JT] Copied from their abstract.

3 Experimental Setup

The experiments shown in this paper are either supervised regression or classification tasks. We formulate any estimation of quantities (e.g. angles, positions, lengths etc.) as a regression problem between 0 and 1. The output indicates the percentage in regards to the degrees of freedom of the individual experiment. If the experiment involves a choice, we formulate it as a classification problem.

Table 1. We use different feature generators as input to a multilayer perceptron which performs linear regression or the classification task. This yields different sets of trainable parameters. We also train the MLP directly on the visualizations without any additional feature generation.

Classifier	Trainable Parameters	Optimization
MLP	2,560,513	SGD (Nesterov momentum)
LeNet + MLP	8,026,083	Learning rate: 0.0001
VGG19 + MLP	21,204,545	Momentum: 0.9
Xception + MLP	25,580,585	Batchsize: 32
		Epochs: 1000 (Early Stopping)

3.1 Measures

Accuracy. We use the same metric as Cleveland and McGill to measure accuracy.

$$\log_2(|\text{predicted percent} - \text{true percent}| + .125) \quad (1)$$

Confidence Intervals. We follow the notion of Cleveland and McGill to compute the confidence intervals.

Efficiency. We use the convergence rate based on the decrease of loss per training epoch as an indicator for the efficiency of the classifier in combination with a visual encoding. For regression tasks the loss is defined as mean squared error (MSE) and for classification tasks the loss is categorical cross-entropy.

3.2 Classifiers

Our classifiers are built upon a multilayer perceptron (MLP) which is a feedforward artificial neural network. We combine this MLP with different convolutional neural networks (CNNs) for preprocessing and feature generation. These include the traditional LeNet trained from scratch, as well as VGG19 and Xception trained using ImageNet.

Multilayer Perceptron. The multilayer perceptron in this paper has 256 neurons which are activated as rectified linear units (Fig. 2). We then add a dropout layer to prevent overfitting and compute linear regression or classification (softmax).

Convolutional Neural Networks. We use CNNs to generate additional features as input to the MLP. We train the *LeNet* classifier with tune it specifically towards each visualization. For *VGG19* and *Xception*, we generate features using previously trained weights on ImageNet.

Optimization. All networks are optimized using stochastic gradient descent with Nesterov momentum using fixed parameters (Table 1). We train for 1000 epochs but stop early if the loss does not decrease for ten epochs.

Environment. We run all experiments on an NVIDIA DGX1 machine with Tesla V100 graphical processing units. We use the KERAS framework with tensorflow.

3.3 Data

We create all visualizations as parametrized rasterized images without interpolation. The number of parameters differs per experiment as summarized in Table 2 and section 4.1. We add subtle random noise (0.05) to each pixel to introduce additional variation.

Training/Validation/Test Splits. We specify the size of each split set as follows: 60,000 training images, 20,000 validation images, and 20,000 test images. We then randomly add parameterized visualizations to the sets while guaranteeing that each set is disjunct from each other in terms of encoded variables. This eliminates leakage during training and evaluation. We also scale each set independently:

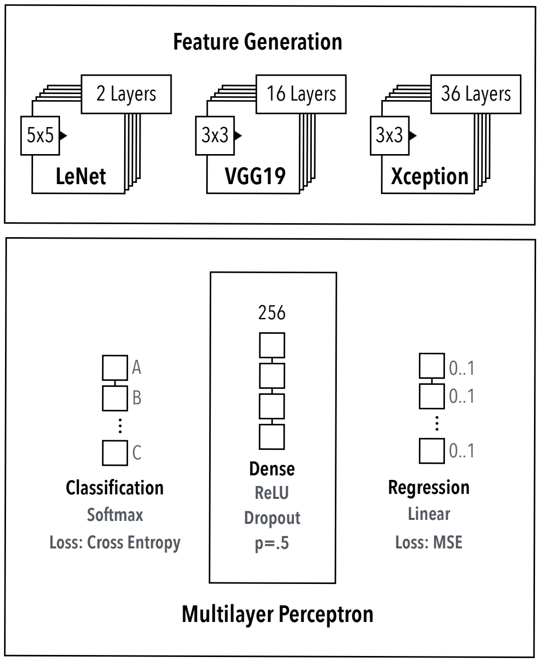

Fig. 2. The multilayer perceptron (MLP) in our experiments has 256 neurons which are activated as rectified linear units (ReLU). We use Dropout regularization to prevent overfitting. We learn categorical and unordered dependent variables using the softmax function and perform linear regression for continuous variables. The MLP can learn the visualizations directly but we also learn features generated by LeNet (2 conv. layers, filter size 5), VGG19 trained on ImageNet (16 conv. layers, filter size 3×3), or Xception trained on ImageNet (36 conv. layers, filter size 3×3) to increase the number of trainable parameters.

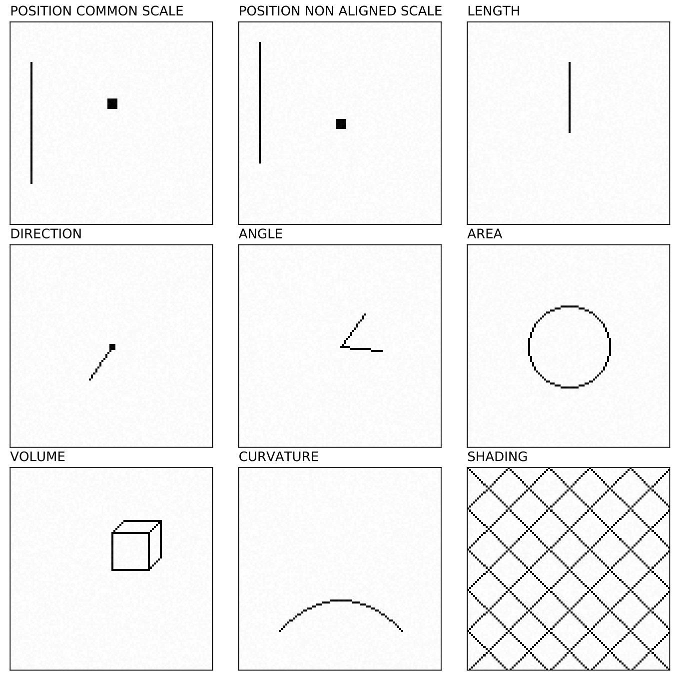

Fig. 3. **Elementary Perceptual Tasks.** Rasterized visualizations of the elementary perceptual tasks as defined by Cleveland and McGill [3] (color saturation excluded). We vary the parameters of each perceptual task and then assess the interpretability of feed-forward neural networks.

images to the range of $-.5$ to .5 and labels to the range of 0.0 to 1.0.

Cross-classifier variability. We also evaluate classifiers previously trained with one visualization on the same type of visualizations with different parameters by decreasing and increasing the variability of the generated images.

4 Elementary Perceptual Tasks

Cleveland and McGill describe the mapping of graphical elements to quantitative variables as *elementary perceptual tasks* and introduce a list of ten different encodings in their paper [3]. We create visualizations of these tasks as rasterized images (Fig. 3).

4.1 Parametrizations

We generate multiple parameterizations for each elementary perceptual task (Fig. 3) and sequentially increase the number of parameters. For instance, for *position non-aligned scale* we first only vary the origin of the coordinate system which yields just 10 different parameters. We then include translation along the y-axis with a significant increase in variability. We then also add x-movement and a variable spot size. This results in more complex datasets depending on the variability setting. Table 2 shows the different settings. It is important to consider this variability when evaluating different classifiers with individual trainable parameters (Table 1).

4.2 Hypotheses

We proposed four hypotheses entering the elementary perceptual task experiment:

- **H1.1 Visual cortex inspired classifiers are able to connect graphical elements to their quantitative variables.** While much simpler models than their biological pendant, convolutional neural networks are heavily influenced by our biological knowledge of the visual system. Such classifiers therefor follow the same principles as human perception.
- **H1.2 Computed perceptual performance is dependent on classifier complexity.** We evaluate multiple classifiers with different numbers of trainable parameters. A more complex classifier (with higher number of parameters) will perform better on elementary perceptual tasks.
- **H1.3 Some visual encodings are better than others for computations.** Cleveland and McGill order the elementary perceptual tasks by accuracy. We investigate whether this order is also relevant for computing graphical perception.
- **H1.4 Classifiers trained on perceptual tasks can generalize to more or less complex variations of the same task.** Recent research suggests that convolutional neural networks generalize extremely well. While the underlying reasons are mainly yet unknown, this property allows them to perform on variations of a similar perceptual task.

5 Position-Angle Experiment

The position-angle experiment was originally performed by Cleveland and McGill to measure whether humans can better perceive quantities encoded as positions or as angles [3]. The actual experiment then compares pie charts versus bar charts since these map down to elementary position and angle judgement. We create rasterized images mimicking Cleveland and McGill's proposed encoding and investigate computational perception of our four classifiers.

5.1 Hypotheses

We proposed four hypotheses entering the elementary perceptual task experiment:

- **H2.1 Computed perceptual performance is better using bar charts than pie charts.** Cleveland and McGill report that position judgements are almost twice as accurate as angle judgements.

Table 2. **Variability of Elementary Perceptual Tasks.** We sequentially increase the number of parameters for every visual encoding of the elementary perceptual tasks. This introduces variability and increasingly more complex datasets.

Elementary Perceptual Task	Variability	Parameters
Position Common Scale	Position Y	60
	+ Position X	3600
	+ Spot Size	21600
Position Non-Aligned Scale	Position Y	600
	+ Position X	36000
	+ Spot Size	216000
Length	Length	60
	+ Position Y	2400
	+ Position X	144000
	+ Width	864000
Direction	Angle	360
	+ Position Y	21600
	+ Position X	1296000
Angle	Angle	90
	+ Position Y	5400
	+ Position X	324000
Area	Radius	40
	+ Position Y	800
	+ Position X	16000
Volume	Cube Sidelength	20
	+ Position Y	400
	+ Position X	8000
Curvature	Midpoint Curvature	80
	+ Position Y	1600
	+ Position X	64000
Shading	Density	100
	+ Position Y	2000
	+ Position X	40000

This renders bar charts superior to pie charts and should also be the case for convolutional neural networks.

- **H2.2 Classifiers can learn position faster than angles.** We assume that understanding bar charts is easier than understanding pie charts. We suspect that our classifiers learn encodings of positions faster than of angles resulting in more efficient training and faster convergence.

6 Position-Length Experiment

This is the one where we estimate two selected bars compared to the longest one - very similar to the previous one but, not yet done. We basically test divided versus grouped barchart and we estimate one relation between two marked quanitities: what percent the smaller is of the larger.

There are five types: type 1-3 this is a position judgement along a common scale. (btw all classifiers seem to do that extremely well in the elementary tasks so we assume this will work well here too). Types 4-5 are length judgements and we know that the classifiers struggle with that quite a bit.

The setup from Cleveland McGill is first a classification task: which one is smaller? and then a regression task: how much smaller. So we have to see how to encode this.

6.1 Hypotheses

We proposed two hypotheses entering the elementary perceptual task experiment:

- **H3.1 Grouped bar charts are better computational perceivable than divided bar charts.** A grouped bar chart involves judging a position while a divided bar chart most likely (if not the bottom is looked at) requires length judgements. Classifiers are better at judging position than at judging length so grouped bar charts are easier to grasp in terms of computational perception.

- **H3.2 not yet** Any ideas?

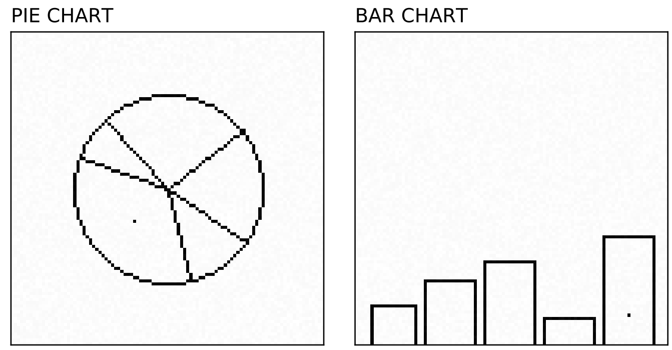

Fig. 4. **Position-Angle Experiment.** We create rasterized visualizations of pie charts and bar charts to follow Cleveland and McGill's position-angle experiment. The experimental task involves the judgement of different encoded values in comparison to the largest encoded values. The pie chart (left) and the bar chart (right) visualize the same data point. In their paper, Cleveland and McGill report less errors using bar charts.

6.2 Discussion

JT: Look at the relative difficulty of the tasks. In Cleveland and McGill, types 1-5 were post-ordered by their log error such that type 1 was easiest and type 5 was hardest. Is this still the case with our CNNs?

7 Bars and Framed Rectangles Experiment

Visual cues can help converting graphical elements back to their real world variables. Cleveland and McGill introduced the bars and framed rectangles experiment which judges the elementary perceptual task of position along non-aligned scales [3].

7.1 Hypotheses

We proposed two hypotheses entering the elementary perceptual task experiment:

- **H4.1 Classifiers can leverage additional visual cues.** The original bar and framed rectangle experiment shows how visual cues aid humans in mapping graphical elements to quantitative variables. This should be the same for feed-forward neural networks since they are inspired by the visual system.

- **H4.2 Weber's law can be transferred to computational perception.** Cleveland and McGill confirmed Weber's law based on the bar and framed rectangle experiment. For humans, the ability to perceive change within a distribution is proportional to the size of the initial distribution.

7.2 Weber-Fechner's Law

As identified by Cleveland and McGill, the bar and framed rectangle experiment is closely related to Weber's law. This psychophysics law states that perceivable difference within a distribution is proportional to the initial size of the distribution. Weber's law goes hand-in-hand with Fechner's law. We conduct an additional experiment based on the original illustrations of the Weber-Fechner law to investigate whether this law can be applied to computational perception of our classifiers (Fig. 7).

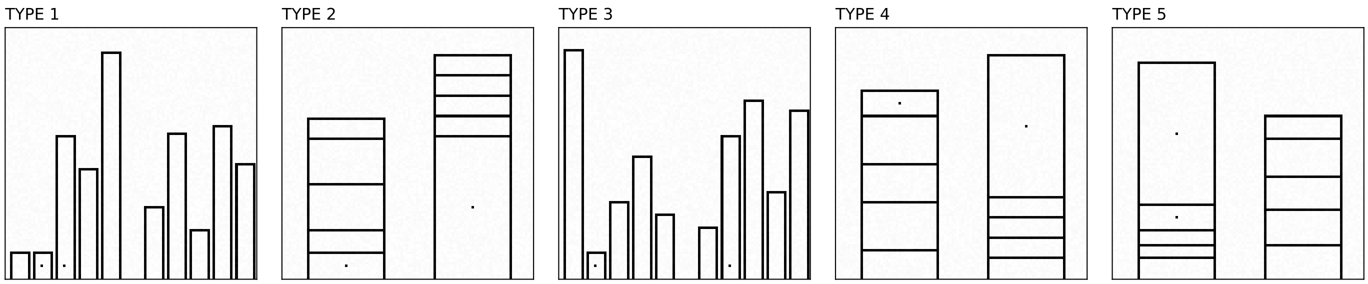

Fig. 5. **Position-Length Experiment.** (Not yet) Rasterized versions of the graphs of Cleveland and McGill's position-length experiment. The perceptual task involves comparing. the two dot-marked quantities across five different visual encodings of either grouped or divided bar charts. We evaluate which type of bar chart performs better with our neural networks as a combined classification and regression problem. The first task is to select which of the marked quantities is smaller (classification) and the second task is to specify how much smaller it is (regression).

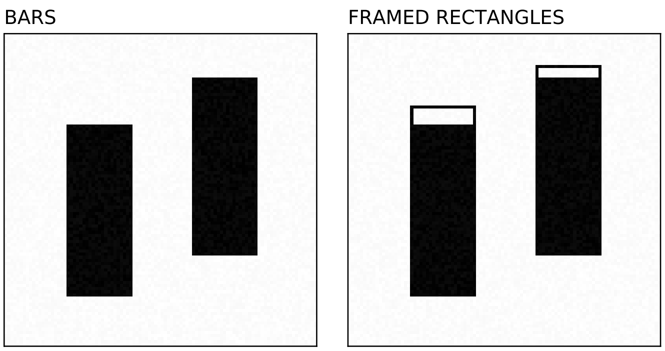

Fig. 6. **Bars and Framed Rectangles Experiment.** Cleveland and McGill introduce the bars and framed rectangles experiment which measures the perceptual task of judging position along non-aligned scales. For humans, it is easier to decide which of two bars represent a larger height if a scale is introduced by adding framed rectangles (right). In this case, the right bar is heigher as visible with less free space when adding the frame. We evaluate whether such a visual aid also helps machines to perceive visually encoded quantities.

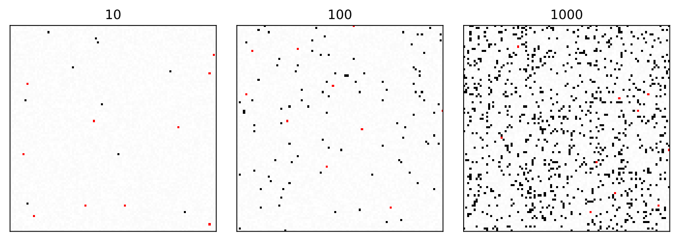

Fig. 7. **Weber-Fechner Law.** The Weber-Fechner law states that the perceivable differences within a distribution is proportional to the initial size of the distribution. The lower square contains 10 more dots than the upper one on both sides. However, the difference is easily perceivable on the left while the squares on the right almost look the same. We generate rasterized visualizations similar to this setup and evaluate our classifiers.

8 Results and Discussion

8.1 Elementary Perceptual Tasks

some are good and some are bad.. why?

Computational Perception Ranking.

Cleveland McGills Ranking - can we observe something similar?

1. Position along a common scale e.g. scatter plot
2. Position on identical but nonaligned scales e.g. multiple scatter plots
3. Length e.g. bar chart
4. Angle & Slope (tie) e.g. pie chart
5. Area e.g. bubbles
6. Volume, density, and color saturation (tie) e.g. heatmap
7. Color hue e.g. newsmap

Cross-classifier variability.

Can a neural network generalize on simple perceptual tasks?

8.2 Position-Angle Experiment

Bar charts are more accurate (Fig. 11) and networks converge faster (Fig. 10). This is great.

8.3 Position-Length Experiment

8.4 Bars and Framed Rectangles Experiment

First run indicates that framed rectangles perform better but we dont really know it yet.

9 Conclusions

Future work: allow insights for infovis for machines

References

[1] M. Borkin, K. Gajos, A. Peters, D. Mitsouras, S. Melchionna, F. Rybicki, C. Feldman, and H. Pfister. Evaluation of artery visualizations for heart disease diagnosis. *IEEE transactions on visualization and computer graphics*, 17(12):2479–2488, 2011.

[2] D. Burlinson, K. Subramanian, and P. Goolkasian. Open vs. closed shapes: New perceptual categories? *IEEE Transactions on Visualization and Computer Graphics*, 24(1):574–583, Jan 2018. doi: 10.1109/TVCG.2017.2745086

[3] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[4] A. Cockburn and B. McKenzie. Evaluating the effectiveness of spatial memory in 2d and 3d physical and virtual environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pp. 203–210. ACM, New York, NY, USA, 2002. doi: 10.1145/503376.503413

[5] M. Correll and J. Heer. Black hat visualization. In *Workshop on Dealing with Cognitive Biases in Visualisations (DECISIVe), IEEE VIS*, 2017.

[6] A. Forsberg, J. Chen, and D. Laidlaw. Comparing 3d vector field visualization methods: A user study. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), 2009.

[7] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[8] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and*

Computer Graphics, 20(12):1943–1952, Dec 2014. doi: 10.1109/TVCG.2014.2346979

[9] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

[10] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 203–212. ACM, 2010.

[11] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1303–1312. ACM, New York, NY, USA, 2009. doi: 10.1145/1518701.1518897

[12] G. Kindlmann, E. Reinhard, and S. Creem. Face-based luminance matching for perceptual colormap generation. In *Proceedings of the Conference on Visualization '02*, VIS '02, pp. 299–306. IEEE Computer Society, Washington, DC, USA, 2002.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012.

[14] D. H. Laidlaw, R. M. Kirby, C. D. Jackson, J. S. Davidson, T. S. Miller, M. da Silva, W. H. Warren, and M. J. Tarr. Comparing 2d vector field visualization methods: a user study. *IEEE Transactions on Visualization and Computer Graphics*, 11(1):59–70, Jan 2005. doi: 10.1109/TVCG.2005.4

[15] T. Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press, 2015.

[16] D. Pineo and C. Ware. Data visualization optimization via computational modeling of perception. *IEEE Transactions on Visualization and Computer Graphics*, 18(2):309–320, Feb 2012. doi: 10.1109/TVCG.2011.52

[17] P. Rheingans. Color, change, and control of quantitative data display. In *Visualization, 1992. Visualization'92, Proceedings., IEEE Conference on*, pp. 252–259. IEEE, 1992.

[18] B. E. Rogowitz and A. D. Kalvin. The "which blair project": a quick visual method for evaluating perceptual color maps. *Proceedings Visualization, 2001. VIS '01.*, pp. 183–556, 2001.

[19] R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[22] Y. Wang, F. Han, L. Zhu, O. Deussen, and B. Chen. Line graph or scatter plot? automatic selection of methods for visualizing trends in time series. *IEEE Transactions on Visualization and Computer Graphics*, 24(2):1141–1154, Feb 2018. doi: 10.1109/TVCG.2017.2653106

[23] C. Ware. Color sequences for univariate maps: Theory, experiments and principles. *IEEE Computer Graphics and Applications*, 8(5):41–49, 1988.

[24] D. L. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356, 2016.

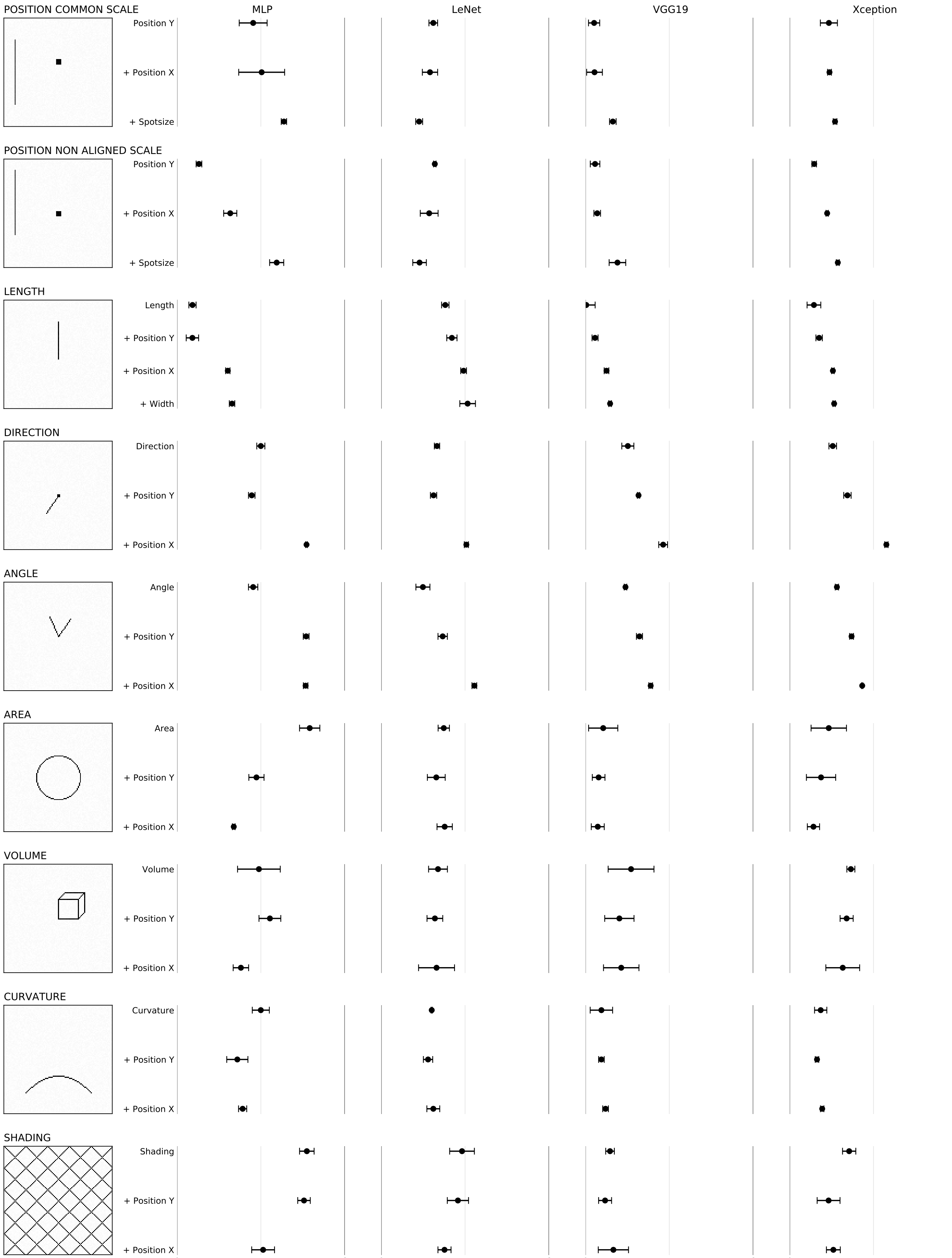

Fig. 8. **Computational results of Elementary Perceptual Tasks experiment.** Log absolute error means and 95% confidence intervals for computed perception of different classifiers on the *elementary perceptual tasks* introduced by Cleveland and McGill 1984 [3]. We test the performance of a Multi-layer Perceptron (MLP), the LeNet Convolutional Neural Network, as well as feature generation using the VGG19 and Xception networks trained on ImageNet.

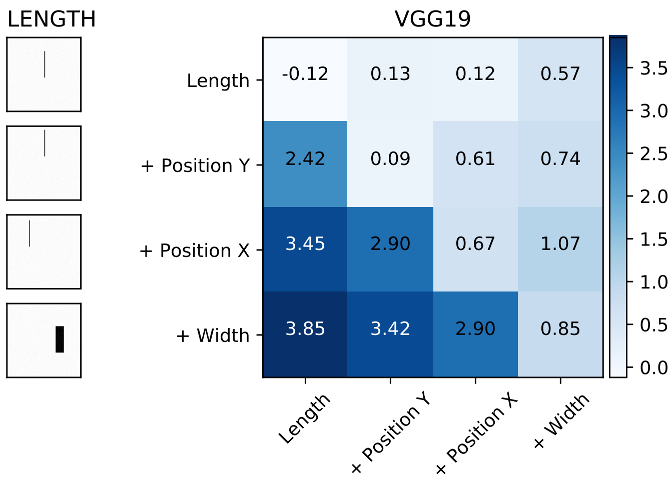

Fig. 9. **Cross-classifier variability for the perceptual task of measuring length.** We use predictions of LeNet classifiers trained on different parametrizations of the *curvature* elementary perceptual task and measure the mean logistic absolute error (MLAE). The lower score, the better. Classifiers trained on curves with variable position can generalize even if the axis of translation varies. However, classifiers trained on fixed positions of curves are not able to measure translated curves.

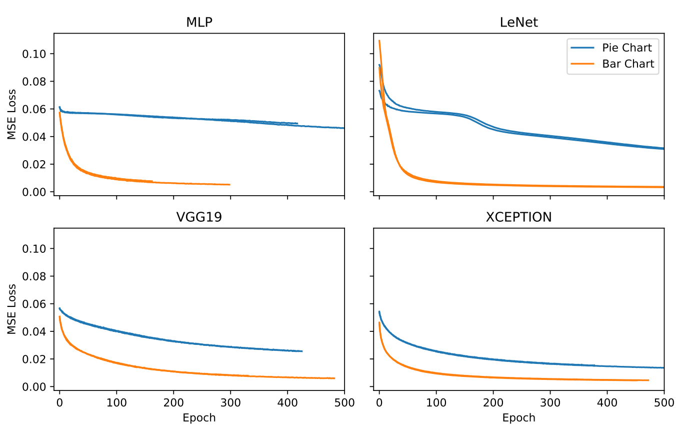

Fig. 10. **Classifier Efficiency of the Position-Angle experiment.** Mean Square Error (MSE) loss for the *position-angle experiment* as described by Cleveland and McGill [3] which compares the visualization of pie charts and bar charts. We report the MSE measure for both encodings of four different classifier on previously unseen validation data.

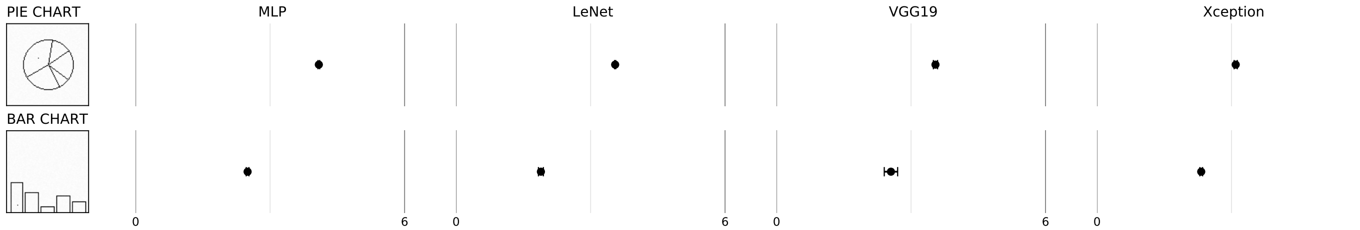

Fig. 11. **Computational results of the Position-Angle experiment.** Log absolute error means and 95% confidence intervals for the *position-angle experiment* as described by Cleveland and McGill [3]. We test the performance of a Multi-layer Perceptron (MLP), the LeNet Convolutional Neural Network, as well as feature generation using the VGG19 and Xception networks trained on ImageNet.

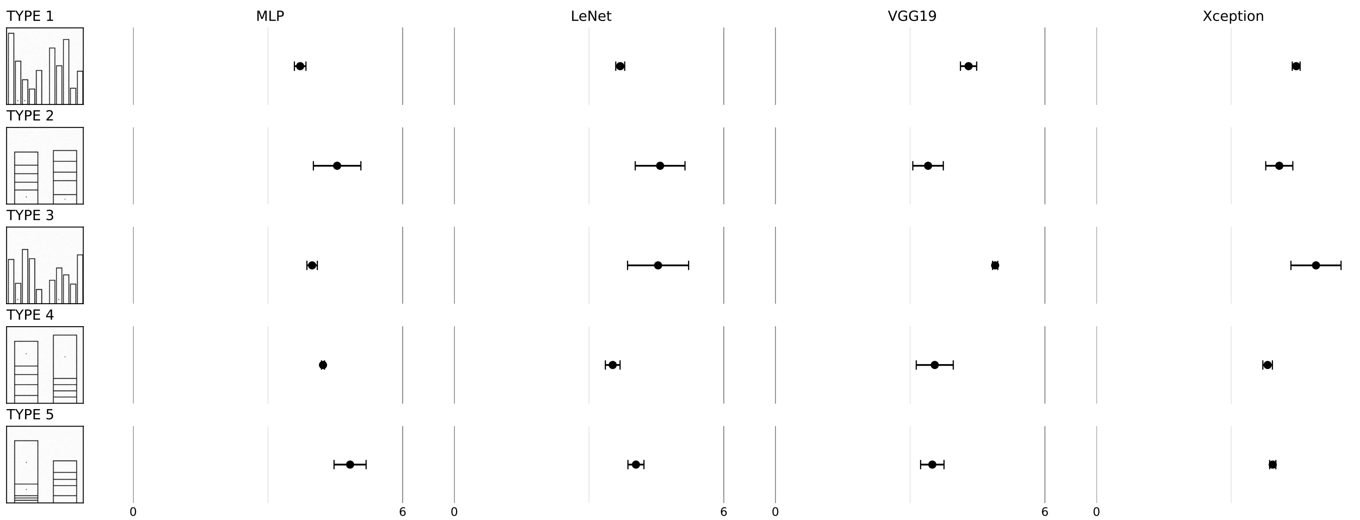

Fig. 12. **Computational results of the Position-Length experiment.** Log absolute error means and 95% confidence intervals for the *position-length experiment* as described by Cleveland and McGill [3]. We test the performance of a Multi-layer Perceptron (MLP), the LeNet Convolutional Neural Network, as well as feature generation using the VGG19 and Xception networks trained on ImageNet.

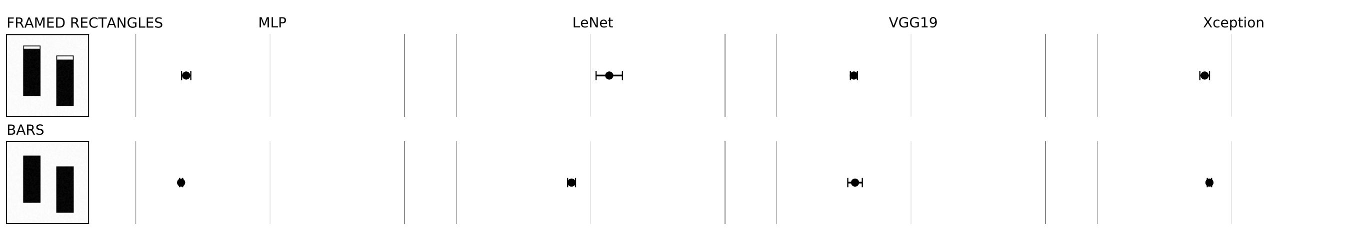

Fig. 13. **Computational results of the Bars-and-Framed-Rectangles experiment.** Log absolute error means and 95% confidence intervals for the *bars-and-framed-rectangles experiment* as described by Cleveland and McGill [3]. We test the performance of a Multi-layer Perceptron (MLP), the LeNet Convolutional Neural Network, as well as feature generation using the VGG19 and Xception networks trained on ImageNet.

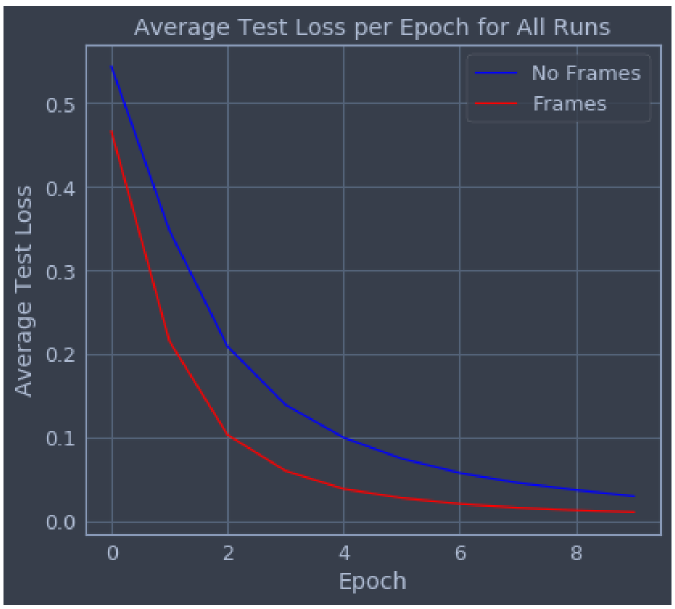

Fig. 14. **Classifier Efficiency of the Bars and Framed Rectangles experiment.** Categorical Cross-Entropy loss for the *bars and framed rectangles experiment* as described by Cleveland and McGill [3]. The frame around the bars adds an additional visual cue enables faster network convergence. This is not yet reproducible!