Test accuracy is not all you need: Examining neural network behavior at the classification boundary

Anonymous Author(s)
Affiliation
Address
email

Abstract

We evaluate the classification of both human volunteers and various neural network models on a set of GAN-generated images that reflect the transition from one MNIST class to another. We find that models that obtain the same test accuracy on the standard MNIST test data set exhibit different behavior on these images. Further, we find that although the number of misclassified images decreases with test accuracy, the spread in predictions over multiple runs on images that are difficult to classify (for humans) also decreases with test accuracy. We conclude that test accuracy is an insufficient metric to capture the behavior of a network on images that lie along the boundary between classes.

1 Introduction

Two people manually classifying images may easily obtain the same accuracy on a test data set. How do we determine which of the two is a better classifier? One option is to extend the challenge to a range of different data sets and tasks and evaluate their performance on each of these. Examples of such benchmarks are the Visual Task Adaptation Benchmark (VTAB) [8] for vision and SuperGLUE [7] for natural language processing models. However, the choice of these benchmark data sets may skew the direction of machine learning research as discussed in a recent preprint on "The Benchmark Lottery" by Dehghani et al. [1]. A further limitation that we want to discuss in this work is that across all these data sets the metric that we use to evaluate classifiers remains the same: the performance on a held out test set.

To be accepted in any of the leading machine learning journals, a paper that proposes a new neural network optimization scheme is generally expected to prove that its test accuracy on certain popular data sets, e.g., ImageNet for vision, matches (ideally exceeds) the state of the art. This effectively constrains the development of optimization schemes to a class of algorithms that will perform well using the metric and benchmark data sets generally accepted by the community. If we consider our search for new optimization schemes as being analogous to traversing a loss landscape, then overfitting on specific metric and benchmark choices will result in getting stuck in a local minimum, which is undesirable. Although expanding our range of data sets and metrics does not entirely resolve the problem, we argue that more research on this topic will lead to better and more robust optimizers.

In this work we study whether a model's accuracy on the standard test data set is an accurate indicator of its performance on a set of images that lie along the classification boundary between two known classes. We do this by comparing the classifications of different network architectures with identical test accuracies to the classifications provided by a group of human volunteers. Our experiment raises the question of how we want networks to behave on images that could belong to either of two known classes and independently illustrates the limitations of using test accuracy as the sole metric for model evaluation.

Under review at ICBINB Workshop @ NeurIPS 2021. Do not distribute.

2 Experiment design

In this experiment we consider the behavior of different neural network models with identical test accuracies on images that capture the transition from one MNIST class to another. We measure their performance and compare their predictions with the labels assigned by a set of human volunteers.

Creating the data. We trained a Deep Convolutional Generative Adversarial Network (DCGAN) [6] on MNIST data [4]. We then used the trained DCGAN to generate images, selected ten images depicting the handwritten numbers 0 to 9, and stored the corresponding ten 100-dimensional vectors $z_0, \ldots, z_9$ which formed the input to the generator. To linearly interpolate between any given z_i and z_j (for $i \neq j$) we used the following equation:

$$z_\alpha = (1-\alpha)z_i + \alpha z_j, \quad \alpha \in [0,1], i,j \in \{0,1,\ldots,9\}. \tag{1}$$

To create our data set we considered 15 examples consisting of 15 pairs of (z_i, z_j), where we set $i < j$ and had each handwritten number represented three times, i.e., each z_k for $k \in \{0,1,\ldots,9\}$ appears three times across the 15 pairs. We then linearly interpolated between each pair using Eq. (1) with 11 values of α equally spaced between 0 and 1 (inclusive). This resulted in a total of 165 images (Appx. B). For illustration we have provided the linear interpolation between a (z_i, z_j) pair in Fig. 1.

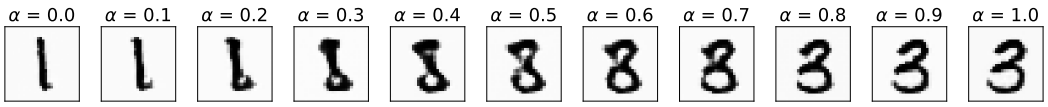

Figure 1: Images generated using Eq. (1), where $i = 1$ and $j = 3$.

Obtaining the neural network baseline. We trained a multilayer perceptron (MLP), convolutional neural network (CNN) [4], and Vision Transformer (ViT) [2] architecture using cross entropy loss and the Adam [3] optimizer with different hyperparameter choices (more details can be found in Appendix A) on the normalized MNIST training data set. We stopped training once a model reached a certain target accuracy $\pm 0.05\%$ on the test data set. We considered a maximum target test accuracy of 99.5% (for the CNN architecture) and 98.5% (for the MLP and ViT architectures). The resulting models were then fed the DCGAN-generated images.

Obtaining the human baseline. To obtain a human baseline for the DCGAN-generated images, we recruited 10 volunteers to manually label the images. The images were shuffled and then distributed to each of the labelers. The volunteers were not informed of the goal of the task, but were simply asked to label the images with numbers between 0 and 9 to the best of their ability, without consulting others. We then anonymously processed the data and re-sorted the manual labels.

3 Results

We consider the classification provided by the human volunteers to be the ground truth. This means that if all human classifiers agree on a specific label for an image we consider this to be the true label. On the other hand, if one or more of the human classifiers disagree on the label for a specific image, we consider this image to be a 'boundary case' example.[1] We will contrast the (dis-)agreement between the human classifiers with the (dis-)agreement between an equally large number of neural network (NN) models. As summarized in Table 1 the resulting classifications by the neural network models on the DCGAN-generated images can then be grouped into three categories:

- Type A: 'Misclassification': Images on which neural network models disagree, whereas human classifiers agree.
- Type B: Matching 'ground truth': Images on which neural network models (dis)agree and human classifiers also (dis)agree.[2]
- Type C: Images on which neural network models agree, whereas human classifiers disagree.

[1] With a larger number of human classifiers the number of boundary case examples would likely increase, but the number of human classifiers is kept constant throughout this experiment.

[2] An exception is the (rare) images on which both neural network models agree and human classifiers agree but the label assigned is different for the two groups: these images are misclassified by the neural networks.

Table 1: The different types that images belong to: type A are misclassified images on which networks disagree; for type B images the networks' predictions match the ground truth (indicated with a ✓); and finally type C is indicated with a "?" as it is unclear how networks should behave in this setting.

	NNs disagree	NNs agree
Human classifiers disagree	✓	?
Human classifiers agree	Misclassification	Human label $=$ NN label: ✓ Human label $\neq$ NN label: Misclassification

To compare the disagreement amongst a group of neural network models we need to decide which models to group together. We focus on the two extreme cases: the exact same model trained with different random seeds and a diverse set of models (MLP, CNN, and ViT) trained with different optimization schemes (full details in Appx. A). The former experiment illustrates the differences between model architectures, whereas the latter is particularly useful for comparing neural networks with humans. Each human is unique and hence we want to contrast the disagreements within the human classifiers with the disagreements amongst an equally diverse set of neural network models.

Do models with the same test accuracy misclassify the same images? We consider the diverse set of neural network models with equivalent test accuracies. Table 2 shows that the average % of images on which 2 models disagree is much lower than the % of images on which all 9 models disagree. Further, when focusing on the images on which the human classifiers agree, we see that the actual % of misclassified images, i.e., on which the label predicted by the neural network is different from the unanimous label assigned by human classifiers, is also much smaller per model than when measured across all 9 models. This illustrates that models that were trained to the same test accuracy behave differently on the DCGAN-generated data set and misclassify (and disagree on) different images.

Table 2: Three MLPs, CNNs, and ViTs were trained until a set test accuracy using Adam (Appx. A).

	% of misclassified images		% of images where	
Test accuracy	per model	across 9 models	2 models disagree	9 models disagree
97.5%	5.9 $\pm$1.7%	15.2%	11.3 $\pm$2.9%	28.5%
98%	4.3 $\pm$1.1%	11.5%	8.5 $\pm$2.3%	21.8%
98.5%	3.5 $\pm$0.8%	7.3%	7.4 $\pm$1.8%	18.2%

Do neural networks disagree on the same images as human classifiers? A human classifier disagrees with another human classifier on average on 9.3 $\pm$2.2% of the data set. As shown in Table 2 this falls within the same range as the number of disagreements between two neural network models (4th column), which raises the question of whether the neural networks disagree on the same images as the humans, namely the complicated 'boundary-case' examples. In Table 3 we compare the disagreements among the 9 neural network models with the disagreements between 9 humans. There exist a significant number of images on which either humans agree but neural network models do not (type A) or on which humans disagree whereas neural network models agree (type C). It is surprising that such a diverse set of neural network models disagree on a large set of images, but do not disagree on the difficult to classify 'boundary-case' examples.

Table 3: Same setting as in Table 2, where type A and type C are as detailed in Table 1.

	% of data set		
Test accuracy	on which NN models disagree	that is type A	that is type C
97.5%	28.5%	15.2%	6.1%
98%	21.8%	11.5%	9.1%
98.5%	18.2%	7.3%	8.5%

How does the maximum softmax value across models vary with the class? In Table 4 we show that the maximum softmax value across models is much lower for type A images than for type C images. Further, the maximum softmax value for images on which both humans and models disagree is higher than for images on which models disagree but humans do not (type A). Instead we would expect the 'boundary case' examples on which the humans also disagree to have similar maximum softmax values as the other examples on which the models disagree.

Table 4: Maximum softmax values across all neural network predictions on different image classes.

		Maximum softmax for test accuracy $\mathcal{A}$		
		$\mathcal{A}$ = 97.5%	$\mathcal{A}$ = 98%	$\mathcal{A}$ = 98.5%
(type A)	NNs disagree & Humans agree	0.83 ±0.178	0.85 ±0.172	0.83 ±0.187
(type B)	NNs disagree & Humans disagree	0.87 ±0.149	0.86 ±0.152	0.89 ±0.138
	NNs agree & Humans agree	0.99 ±0.050	0.99 ±0.051	0.99 ±0.038
(type C)	NNs agree & Humans disagree	0.97 ±0.067	0.96 ±0.091	0.97 ±0.082

How does the image type vary with respect to test accuracy of the considered models? In Figure 2 we consider the (dis)agreements between the predictions of a network across 10 random seeds that was trained each time to the same test accuracy. Standard deviations are obtained by repeating this 10 times (i.e., over 100 different seeds in total). Type A images form a clear representation of a specific handwritten digit as all the human volunteers agreed upon their label. We consider type A images to be misclassified by the neural network models and hence we expect the number of type A images to decrease with test accuracy of the considered models. For all three model architectures considered (MLPs, CNNs, and ViTs) we indeed observe that the percentage of type A images decreases with test accuracy (Figure 2, left).

Although our expectation for images on which human classifiers agree is clear, it is debatable what the desired behavior of a set of neural networks is for images on which human classifiers disagree. However, if a set of models is initially split across multiple labels for a (for humans) difficult to classify image, but then transitions to predict a unanimous label, we consider this to be an undesirable change as there does not exist a single true answer in this setting. In Figure 2 (right) we study how the number of type C images (on which humans disagree, but neural networks agree) changes with the test accuracy of the considered models. We find that the number of difficult to classify images that neural networks unanimously agree on increases with test accuracy.

Finally, Figure 2 clearly shows the variations across different network architectures, which were trained using the same optimizer until the same test accuracy was reached. An MLP architecture (blue) misclassifies the lowest number of images (type A) for any given test accuracy, but also predicts 'boundary-case' images with the largest 'certainty' (type C). Test accuracy is therefore not an accurate indicator of the performance of these models on the considered data set.

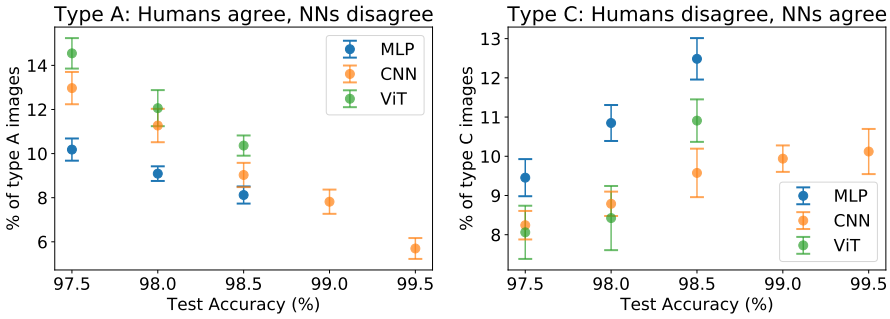

Figure 2: The % of type A (left) and type C (right) images with respect to the test accuracy of the neural network models considered. We measure the disagreements between the same neural network architectures trained with 10 different seeds. All models were trained using Adam with $h = 5\text{e-}4$.

4 Conclusion

To show the limitations of using test accuracy as the only metric for model evaluation we created a set of DCGAN-generated images that capture the transition from one MNIST class to another. We found that models with the same test accuracy misclassify (and disagree on) different sets of images. Further, we find that although a higher test accuracy corresponds with neural network models becoming more unanimous in their label prediction, this may be undesirable for those images on which humans disagree as they do not have a clear class label associated with them. Our experiments also highlight the differences between human and model predictions, as neural networks disagree amongst themselves on a different set of examples than where human classifiers disagree and when in disagreement have higher maximum softmax values on those images on which humans disagree.

References

[1] Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The Benchmark Lottery. *arXiv preprint arXiv:2107.07002*, 2021.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[3] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[4] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998.

[5] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. 2017.

[6] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016.

[7] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *NeurIPS*, 2019.

[8] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

A Experimental details

Generating the data. As the generator of the DCGAN architecture [6] we use five convolutional transpose layers with feature map size 64 with ReLU activation and batch normalization. For the discriminator we use five convolutional layers with Leaky ReLU activation and batch normalization. The weights are randomly initialized from $\mathcal{N}$(0,0.02) and both models are trained using Adam [3] with h = 2e-4 and $\beta_1 = 0.5$ as suggested by Radford et al. [6]. All other details on the data generation are provided in Section 2.

Model architectures. We consider three different neural network architectures: a multilayer perceptron (MLP), a convolutional neural network (CNN) [4], and a Vision Transformer (ViT) [2]. As MLP architecture we used a hidden layer with 500 input and output nodes and ReLU activation. Our CNN architecture has 4 convolutional layers and 3 fully connected layers, ReLU activation, batch normalization, max pooling, and dropout ($p = 0.4$) on the middle fully connected layer. Finally, the Vision Transformer (ViT) architecture is adapted from Dosovitskiy et al. [2], where we use 6 transformer blocks, 8 heads in the multi-head attention layer, and a MLP size of 128.

Training. We train the models using PyTorch [5] on the normalized MNIST [4] training data set with the Adam optimizer. Training is stopped when a model has reached the target test accuracy $\pm 0.05\%$ on the MNIST test data set. To create the diverse set of models we train three models of each architecture type: for MLP1, CNN1, and ViT1 we use Adam with step size $h =$ 5e-4, batch size (BS) = 128 and no weight decay (WD). For MLP2, CNN2, and ViT2 we use AdamW with $h =$ 5e-4, BS = 128, and WD = 1e-05. To introduce additional variation we train MLP3 using $h =$ 1e-3, BS = 256, WD = 1e-05; CNN3 without weight decay but using additional data augmentation; and ViT3 using WD = 5e-05. To create Figure 2 we train the same network across 10 random seeds to the same test accuracy using a batch size of 128 and the Adam optimizer with $h =$ 5e-4. We repeat this process 10 times (so using a total of 100 random seeds) to obtain standard deviations.

B Dataset

In Figure 3 we provide the full 165 DCGAN-generated images considered in this work. These images were shuffled before being given to the human volunteers. On top of each image we provide a ✓ if all 10 human classifiers agreed on the label for this image. If for an image there was disagreement between the human classifiers, we state the different labels proposed.

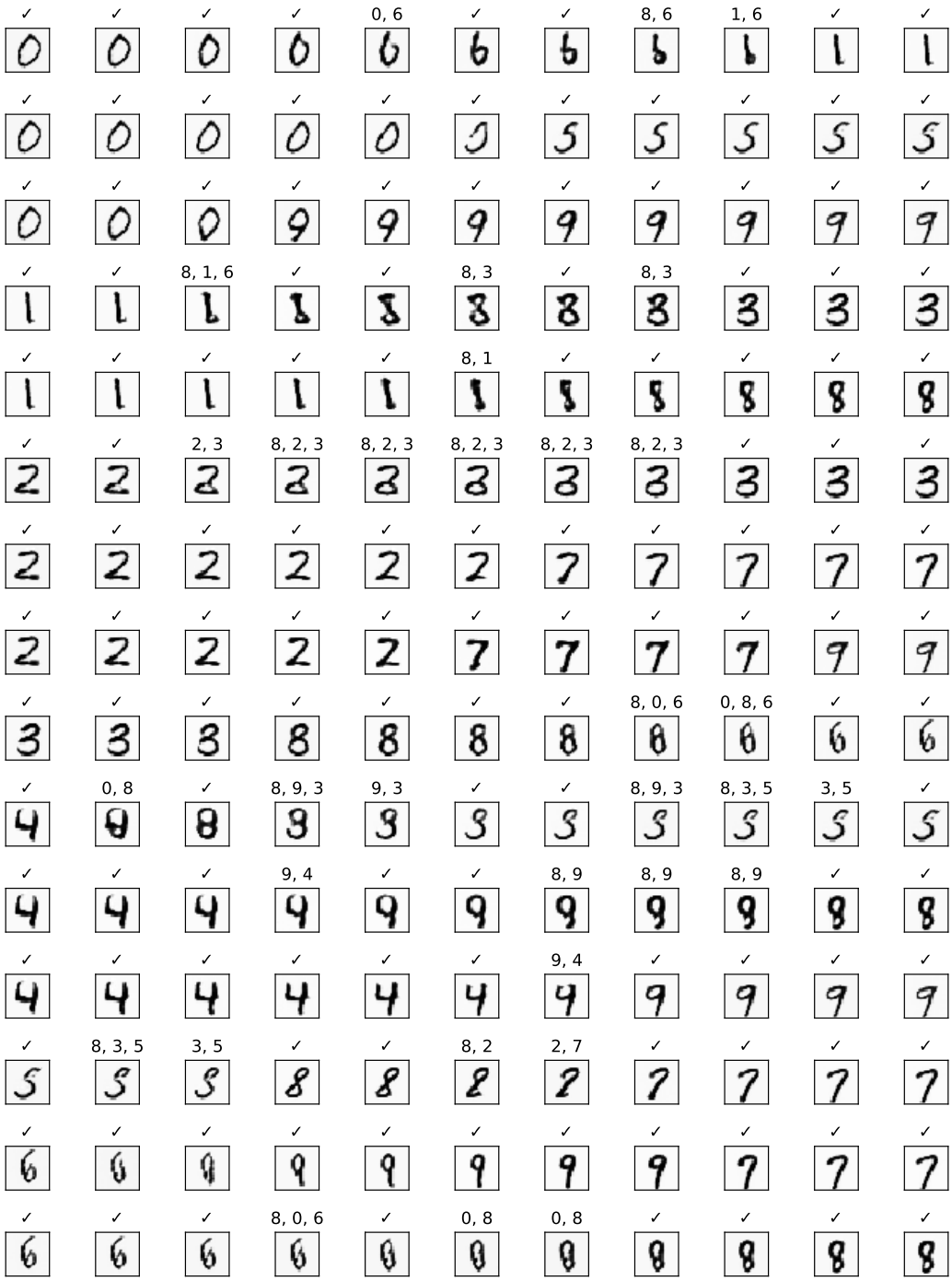

Figure 3: The dataset used in this paper. The data generation is described in Section 2. The labels provided by the human volunteers are listed on top of each image (✓ if there was no disagreement).

C Further results

In the paper we focused on showing the limitations of test accuracy as the sole metric for model evaluation. The results in this section further support this statement and try to answer some follow-up questions the reader may have.

Is an increased disagreement among human classifiers reflected by the behavior across neural network models? Throughout this work we have categorized the human classifications into two groups: images on which the humans agree and images on which they disagree. We now want to take a closer look at the latter category. We consider the number of humans disagreeing with the majority vote as a measure of human certainty, i.e., if 2 humans disagree with the majority vote the humans are 80% certain. Correspondingly, we can measure the level of 'certainty' of neural networks by studying how many networks disagree with the most prevalent label assigned to a specific image across all 10 networks. The results are summarized in Table 5. It is clear that there is a lot of variation across models that have the same test accuracy. Further, there is not a clear correlation between the human certainty on specific images and the model certainty. CNNs seem to have the lowest overall certainty (but still much higher than humans) on images where humans are $\leq$80% certain on. Since we only have 10 human volunteers in this study, it may be worth repeating this experiment with a much larger sample size to give a more accurate representation of human certainty. Further, the number of images in each "human certainty" class is rather limited and so scaling up along this axis is also worth exploring in future experiments.

Table 5: We compute the model certainty of a model across 10 random seeds on images that humans are hc% certain on. The number of images in each class are: 3 (hc = 90%), 9 (hc = 80%), 7 (hc = 70%), 6 (hc = 60%), and 6 (hc = 50%). We exclude the class on which humans are 40% certain as there only exists a single sample point. The standard deviations for the models are obtained by repeating the experiment 10 times (so using 100 random seeds in total).

		Model certainty on images that humans are hc% certain on				
Test acc	Model	hc = 90%	hc = 80%	hc = 70%	hc = 60%	hc = 50%
97.5%	MLP	81.7 ±5.6%	86.6 ±3.0%	90.9 ±2.4%	80.0 ±3.8%	94.7 ±1.5%
	CNN	96.3 ±2.8%	83.3 ±2.1%	81.4 ±2.6%	79.2 ±4.2%	87.3 ±2.9%
	ViT	85.0 ±5.8%	87.2 ±2.8%	82.7 ±3.1%	80.5 ±3.8%	93.5 ±3.1%
98%	MLP	84.0 ±2.0%	86.8 ±3.7%	94.9 ±1.6%	86.5 ±4.7%	93.5 ±1.7%
	CNN	98.0 ±1.6%	83.3 ±2.9%	83.3 ±2.5%	79.2 ±3.8%	88.0 ±2.2%
	ViT	89.0 ±5.4%	85.6 ±2.5%	85.9 ±3.6%	79.3 ±4.5%	91.5 ±1.9%
98.5%	MLP	86.7 ±3.7%	87.2 ±2.4%	94.3 ±2.6%	88.0 ±3.6%	94.2 ±2.0%
	CNN	99.0 ±2.1%	84.6 ±3.5%	81.6 ±3.5%	81.2 ±6.6%	87.8 ±2.0%
	ViT	88.3 ±3.7%	88.8 ±2.2%	90.3 ±2.5%	81.7 ±5.0%	92.7 ±2.6%

Can we visualize how the human and network (dis-)agreements are distributed across the dataset? In Figure 4 we visualize using a heatmap on which images humans disagree amongst themselves, but CNNs do not (in red) and on which images CNNs disagree amongst themselves, but humans do not (in blue). The images on which both networks and humans agree are indicated in gray, while images on which both networks and humans disagree are indicated in white. The significant number of red and blue squares clearly shows that CNNs disagree on a different set of images than the human classifiers. Further, we observe that both the number of blue squares (on which CNNs disagree, but humans do not) and the number of white squares (on which both CNNs and humans disagree) decrease with test accuracy of the considered models, which corroborates the results shown in Figure 2.

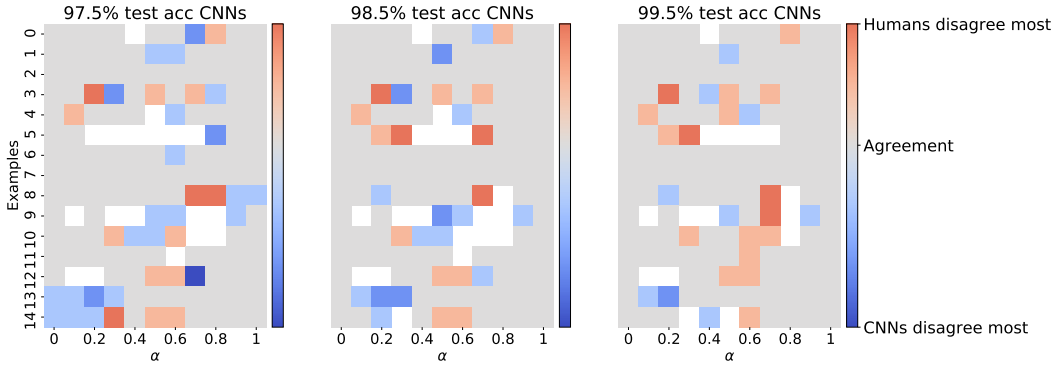

Figure 4: The disagreement between 10 human classifiers on the data set (Figure 3) is indicated in red, where the humans are either split over two classes (light red) or three classes (dark red). The disagreement between 10 CNN models with identical test accuracies (obtained using 10 different random seeds) is indicated in blue, where a darker color again indicates that the models were split over more classes. In gray we indicate images on which all human classifiers and all networks agree (although they may be misclassified by the networks). Finally, in white we indicate images on which both the humans and networks disagree. We observe that both the number of blue squares and (to a smaller extent) the number of white squares decreases with test accuracy, thus collaborating the results from Figure 2 that CNNs agree more as test accuracy increases.