# Do Gabor-filter constrained neural networks generalize better?

Sam Acquaviva and Reece Shuttleworth

Massachusetts Institute of Technology

May 17, 2023

**Abstract**

It is well known that humans have incredibly flexible and general object recognition capabilities, especially in comparison to machine vision systems. In an attempt to learn more about what might be causing this disparity in performance and to improve machine vision systems, we explore whether constraining a machine vision system to be more human-like improves the performance of the system in learning new tasks. Specifically, Gabor filters (shown in Figure 1) have been used to model simple cells in the visual cortex of mammals [5, 18], and previous work has shown that constraining the first layer of a convolutional neural network (CNN) to parameterize the Gabor function can lead to faster learning and higher performance [1].

We extend this work by testing if these Gabor-constrained models (GaborNets) *generalize* better to new datasets. We also run human experiments using adversarial examples from both datasets in order to analyze the disparity in performance between humans and both networks on these adversaries. Our experiments show that, while the GaborNet does learn more robust representations, it does not learn faster or converge to a higher accuracy. Additionally, our replication of prior work shows that previous results demonstrating performance improvements are limited to very specific models and datasets. Our human experiments validate these experimental results, showing that people's performance is not more similar to GaborNets than CNNs. These null results imply that even though something might model the human brain well, it will not necessarily improve performance of machine learning methods.[1]

## 1 Introduction

Humans have incredibly flexible and general object recognition capabilities. Much of the flexibility of human visual cognition is due to abilities down-stream of the visual system, as evidenced by learning new categories of objects with a single example [15], using compositionality to combine the meaning of two visual objects [20], and learning different levels of abstraction in visual categorization [2].

In comparison to machine vision systems, such as neural networks, humans learn much faster in object recognition tasks [24] and are also much better at generalizing from training

[1] The code to recreate all of our experiments is at `https://github.com/samacqua/gabor-constrained-nns`

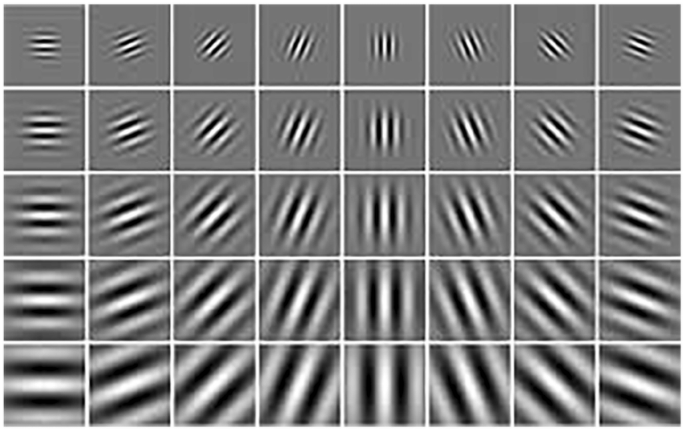


Figure 1: A visualization of Gabor filters with different parameterizations.[4]

data to related data [16]. It appears that this difference in learning speed and generalization ability is also not just due to the capacity of neural networks, as even large pre-trained vision models cannot one-shot [2] learn new objects with the same accuracy as humans [22].

Since people can also learn quickly outside of the vision domain, it is clear that not all of people's visual reasoning ability is due to the visual recognition system. However, perhaps current architectures can learn from biology. What aspects of the human visual system can improve a neural network's ability to learn new concepts? For a model to "learn a new concept" well in the visual domain, it should learn the new task quickly, maintain previous performance on old tasks, and be robust in its new ability.

Neural networks struggle with each of these. Neural networks, in all domains (including vision), require orders of magnitude more data to match human performance [16]. Neural networks, when trained on a new task, struggle to retain performance on previously trained tasks in what is known as "catastrophic forgetting" [12]. Neural networks, in classification tasks, are known to be sensitive to "adversarial examples", or images that "trick" the network [9].

So, we want to take one biologically inspired architectural modification and apply it to the problem of learning a new classification task. For reasons outlined in Section 2.2, we investigate how incorporating Gabor filters into the early layers of neural networks might address some of these problems. We hypothesize that doing so may help with

- Speed of convergence and final accuracy.
- Performance on original task.
- Adversarial robustness.

[2] one-shot learning refers to the use of one training example when training to recognize a new object or class.

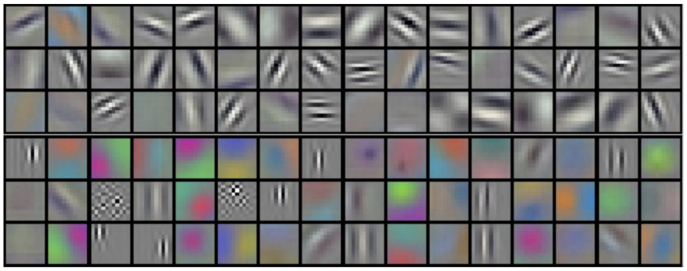


Figure 2: Convolutions in the first convolutional layer learned by AlexNet[14], a deep CNN trained on ImageNet. Without any constraint to do so, some of these convolutions look like Gabor filters.

# 2 Related Work

## 2.1 Gabor Filters

The Gabor function [10] is a 2-dimensional linear sinusoidal filter that can produce waves of various size, intensity, and orientation [4]. Gabor filters with different parameters are shown in figure 1, and defined below in Equation 1, where $g$ defines the brightness at position ($x$, $y$). The parameters $\omega, \theta, \phi$, and $\sigma$ control the size, intensity, and orientation of the produced waves.

$$g(x, y, \omega, \theta, \phi, \sigma) = exp(-\frac{x'^2 + y'^2}{2\sigma^2})exp(i(\omega x' + \phi)) \tag{1}$$

$$x' = x\cos\theta + y\sin\theta \tag{2}$$

$$y' = -x\cos\theta + y\cos\theta \tag{3}$$

Previous work has used the Gabor filter to model simple cells in the visual cortex of mammals [5, 18]. Also, a filter bank of Gabor filters is believed to be a good representation for V1, which is the first stage of visual processing in the brain [19]. In computer vision tasks, Gabor filters have been used to detect faces [3] and represent textures [11]. Together, these successful applications in modeling biology and in improving performance on classic computer vision tasks suggest that Gabor filters may be a good filter to use to extract robust features for machine vision tasks.

## 2.2 GaborNets

Interestingly, convolutional neural networks (CNNs) tend do learn filters that appear to look like Gabor filters, even without being explicitly constrained to the function class [14](see figure 2). In the CNN displayed in figure 2, the first layer is a set of completely unconstrained 11 by 11 filters that were trained using backpropagation. So, it is surprising that they seemingly mimic the same filter that appears in biological visual systems and performs so well on other vision benchmarks.

Inspired by this result, prior work has constrained the first convolutional layer to directly parameterize the Gabor function (specifically, they only parameterize the real part of Equation 1), instead of learning the convolution *de novo* [1]. Concretely, instead of learning the values of each convolutional filter value directly, the Gabor-constrained network (called GaborNet) backpropagates the error through the parameters $\omega$, $\theta$, $\phi$, and $\sigma$ at each ($x$, $y$) position in the filter. So, throughout training, the filters in the first layer are always Gabor functions.

Despite the fact that unconstrained CNNs achieve Gabor-like filters after training, GaborNets consistently outperform their unconstrained counterparts on multiple datasets, converging faster and to a higher accuracy. However, the authors of the GaborNet paper do not test their CNN-variant on any other task than image classification from scratch.

### 2.3 Fine-tuning, stability-plasticity, and adversarial examples

Fine-tuning is a common approach in deep learning, where one adapts a pre-trained neural network to address specific tasks or domains [8]. Fine-tuning exploits the shared structure between many tasks, allowing one to train a neural network without starting from scratch. Researchers often employ this technique to transfer knowledge from large-scale pre-trained models, such as BERT [6] and early versions of GPT [21], to target applications, capitalizing on the robust feature extraction capabilities of these architectures to enhance performance and reduce training time.

The stability-plasticity dilemma is a fundamental challenge in machine learning. This conundrum arises from the need to strike an optimal balance between the model's ability to retain previously acquired knowledge (stability) and its capacity to incorporate new information (plasticity). Overemphasizing stability may lead to a rigid model incapable of adapting to novel data, whereas excessive plasticity can result in *catastrophic forgetting*, where the network's performance on the original task is significantly degraded when it is trained on a new task [12].

Adversarial examples are a class of carefully crafted input examples that exploit the model's underlying architecture to "trick" the model into incorrectly classifying the image with minimal changes to the original image. One method for generating adversarial examples is the Fast Projected Gradient Descent (PGD) attack [17]. PGD works by iteratively applying a small perturbation in the direction of the gradient of the loss function, with respect to the input data, and subsequently projecting the perturbed input back into a valid input domain. This results in the generation of adversarial examples that are imperceptibly different from the original inputs but can cause misclassification in trained models. The PGD attack is often considered a stronger adversary due to its iterative nature and ability to exploit model vulnerabilities more effectively than single-step attacks. As such, PGD is frequently used as a benchmark for evaluating the robustness of machine learning models against adversarial attacks.

## 3 Methods

### 3.1 Machine Experiments

In our machine vision experiments, we investigate whether constraining the low-level features of a neural network to fit the Gabor function leads to better generalization. In CNNs, earlier layers are associated with lower-level filters, like edge detectors, and later layers are associated with higher-level features, like an entire face [14]. Therefore, in order to test the generalization ability of the learned low-level features, we focus on the first layer of the network.

In order to test whether the learned low-level features generalize well, we do the following, repeating both for a vanilla CNN and the same CNN that has had its first layer constrained to fit the Gabor function (the Gabornet): First, we train each model separately on either the CIFAR-10 [13] or Fashion-MNIST [23] dataset. We then freeze the first layer of weights, which represent the low-level features, so that they cannot be changed. We then fine-tune each network using the dataset that it was not trained with initially.

We test the generalization ability of the low-level features by looking at both the convergence speed and the accuracy of the networks during fine-tuning. We can also check the performance of the network on the original dataset during the fine-tuning phase to see if Gabor-constrained filters help prevent catastrophic forgetting. Finally, we can then create adversarial examples for the trained network to see if the newly learned classification is more robust when the first layer is Gabor-constrained.

Concretely, we do the following:

1. Train a classifier (GaborNet or vanilla CNN) on dataset A.
2. Freeze the first-layer weights of the classifier.
3. Train the rest of network weights on dataset B.
4. Analyze the convergence speed / accuracy on the new dataset, sustained performance on the original dataset, and robustness to adversarial examples.

### 3.1.1 Models

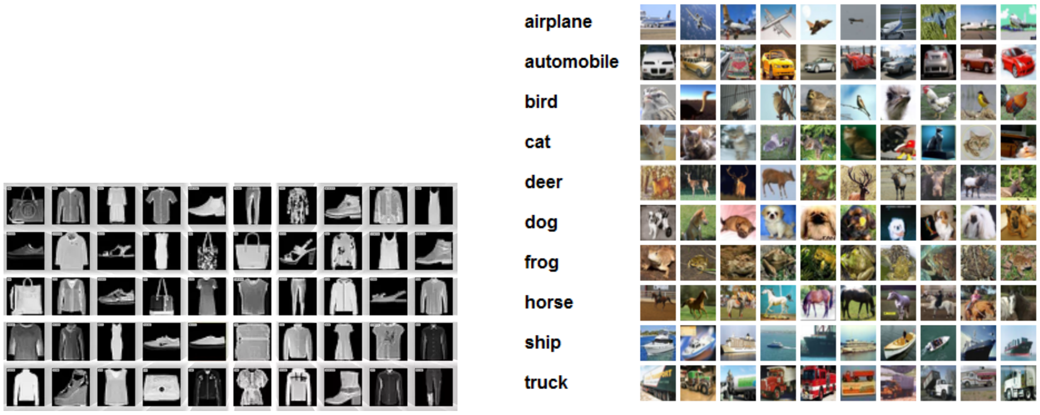
To ensure that any difference in performance is due to the Gabor-constrained layer, rather than due to a quirk of the specific model architecture, we will test 2 model types.

The first model is a *linear probe.* It is simply 2 convolutional layers (either a vanilla layer or a Gabor-constrained layer), followed a by a ReLU non-linearity, then a fully connected layer. For this model, the 2 layers of convolution are frozen during training. The second model is a small CNN which has another convolutional layer and another linear layer, which are not frozen during training. For more model details, visit the project codebase.

### 3.1.2 Datasets

Since we freeze the first layer before fine-tuning, our methodology tests whether the low-level features learned from the other dataset generalize to capture features of the new dataset. We intentionally chose two datasets that share the same number of classes (10), which enables us to swap datasets before fine-tuning without any architecture adaptation, since the final layer of the network depends on the number of classes. Additionally, we wanted the datasets to have very different objects so that new learning would be required to perform well on one dataset after being trained on the other.

For these reasons, we chose Fashion-MNIST [23], a dataset of full-color 28x28 images of various clothing items, and CIFAR-10 [13], a dataset of full-color 32x32 images of items like airplanes and frogs. Examples from both datasets are shown in Figure 3. Importantly, there are no identical classes between the two datasets. Since CIFAR-10 is in color (3 channels), and Fashion-MNIST is in black-and-white (1 channel), we ran preliminary experiments to



(a) Fashion-MNIST

(b) CIFAR-10

Figure 3: Examples from Fashion-MNIST [23] and from CIFAR-10 [13].

determine if casting one dataset to the other's color-space would impact performance and found that the results are qualitatively identical. For simplicity of experiments, we transform CIFAR-10 to black-and-white for the reported results and rescale Fashion-MNIST to 32x32.

## 3.2 Human Experiments

We conduct humans experiments in order to investigate the performance of humans in comparison to our machine vision models. To do this, we use adversarial examples from both networks to identify whether humans are more robust to adversarial examples than the networks that generated them. Interestingly, these results can also help us give evidence that Gabor filters may be closer to what is going on in the brain than vanilla CNNs are. This is because if there is a clear difference in performance on adversarial examples in which humans perform much worse on adversarial examples generated from the GaborNet in relation to the vanilla CNN, this would suggest that the adversarial examples from the GaborNet are better adversarial examples *for the brain as well*, rather than those from the vanilla CNN. This would lead us to believe that the Gabor filters therefore could be a better representation of what is occurring in the brain in comparison to the vanilla CNN.

### 3.2.1 Using Adversarial Examples

In this experiment, we test humans with adversarial examples from both networks. This enables us to investigate if humans are more robust to these adversarial examples than both the vanilla CNN and the GaborNet. Adversarial examples are examples that are similar to examples in the training data but differ due to small perturbations that lead to incorrect classifications. In this context, successful adversarial examples are images that are based on images from either Fashion-MNIST or CIFAR-10 but have been changed slightly such that the machine vision model incorrectly classifies it, even though it correctly classified the image that the adversarial example was based on. Adversarial examples are found by methods like the Fast Gradient Signed Method (FSGM)[9]. For these methods, an $\epsilon$ value, a number between 0 and 1, is specified that determines the amount and magnitude of changes allowed to the real example. This impacts how similar the adversary is to the real image.

Given this context, we create the adversarial examples in the following manner, repeating for both a low and a high $\epsilon$:[3]

1. Randomly sample 5 images from both CIFAR-10 and Fashion-MNIST, resulting in 10 images total.

2. For each of the 10 images, run the Fast Gradient Signed Method (FSGM)[9] on both GaborNet and the Vanilla CNN. This results in 20 adversarial examples in total (each of the 10 images gets an adversary based on GaborNet or the Vanilla CNN.)

Note that we do not replace an image after it has been sampled, so we end up with 40 adversarial examples based on 20 unique images. As a baseline to compare with human performance, we take the 40 adversarial examples and run each through the network that generated it and determine if it is correctly classified.

We test humans on a computer in the following manner until they have classified all 40 adversarial examples:

1. Randomly sample one of the images.

2. Present it to the participant for 1500 milliseconds.

3. Present the 10 possible classes of the image's respective dataset and prompt the participant to classify the image as one of the 10 possible classes.

We present the adversarial example to the participant for a limited time in order to more fairly compare the machine vision models to humans. We want to test the humans snap judgement of what the image is, not their ability to sit there and deduce what it is based on clues. This test is designed to take 5 to 10 minutes, meaning fatigue will not play a significant role in the experiment. We analyze which images are incorrectly and correctly classified and run Student's t-test to see if there is a meaningful difference between the performance of adversaries generated from the GaborNet and adversaries generated from the vanilla CNN.

# 4 Results

## 4.1 Machine Experiments

### 4.1.1 Reproduction of Previous Results

Before running our experiments, we wanted to reproduce the results claimed in the original GaborNet paper to ensure that our implementation is correct. Specifically, we want to recreate the performance on the Cats versus Dogs dataset [7], where the original paper shows impressive improvements upon a CNN baseline, reaching 70% accuracy in less than half the training time and reaching a performance 6% higher [1]. The paper showed slight performance improvements on other datasets, but none was as extreme as on the Cats versus Dogs dataset.

While we were able to reproduce this performance improvement (see Figure 4), we ran two baselines which show that the GaborNet architecture might not universally offer performance improvements. First, we found that we could get the majority of the performance

[3] We heuristically choose a low $\epsilon$ to be 0.05 and a high $\epsilon$ to be 0.3.

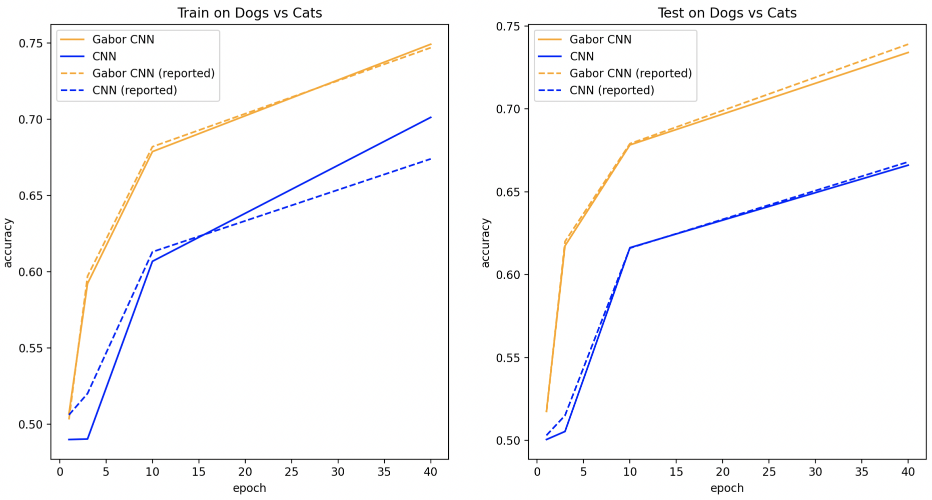


Figure 4: Replication of previous results.

improvement over CNNs without any learning in the Gabor-constrained layers. This means that, while the Gabor-filters do improve performance on this dataset, learning through the filters is not necessary. Second, we found that downscaling either the dataset image size (from 256x256 to 64x64) and model size (removing several layers), or the filter size (from 15x15 to 5x5) completely removed any performance improvement. These results are summarized in Table 1.

These experiments show that our implementation of the GaborNet is correct (because we can reproduce previous results), but that the performance improvements offered by the architecture in the image classification task are not universal.

### 4.1.2 Generalization Hypothesis

We found that, in all 5 runs, the full models that use the first layer from training on dataset A consistently outperform the *random* baseline, where the first layer is randomly initialized

| | | replication | | | |
|---|---|---|---|---|---|
| epoch | reported | original | no learning | 64x64 images & small CNN | 5x5 kernel |
| 1 | 0.014 | **0.017** | 0.009 | -0.073 | -0.051 |
| 3 | 0.105 | **0.112** | 0.110 | -0.034 | 0.021 |
| 10 | **0.063** | 0.062 | 0.056 | 0.011 | -0.019 |
| 20 | **0.071** | 0.068 | 0.063 | 0.002 | 0.008 |

Table 1: Difference in accuracy between the test set performance of the GaborNet and CNN ($acc_{\text{GaborNet}} - acc_{\text{CNN}}$) on the Cats versus Dogs dataset. The highest accuracy per epoch is bolded. Note that the performance improvement of the GaborNet is reduced or completely gone for the down-scaled image and kernel variants, but it is very similar for the variant where there is no learning in the first layer.

|  | random | Linear probe |  | full CNN |  |
|---|---|---|---|---|---|
| dataset | baseline | GaborNet | CNN | GaborNet | CNN |
| CIFAR-10 | 0.406 | 0.447 | 0.418 | 0.440 | 0.493 |
| Fashion-MNIST | 0.829 | 0.756 | 0.774 | 0.844 | 0.881 |

Table 2: Accuracy on Dataset B after 10 epochs, averaged across 5 runs.

and frozen before training on dataset B, which in turn outperforms the linear probe models.

We also found that, for both the linear probe class and the full CNN model class, the models with a first layer that is not constrained to be a Gabor filter consistently outperform their Gabor-constrained counterpart. This ordering of performance was consistent across all 5 trials and for both orderings of datasets.

So, we conclude based on the data presented in Table 2 that Gabor filters learned via backpropagation do not provide an advantage over traditional convolutional filters in creating low-level features that generalize to new datasets.

### 4.1.3 Plasticity Hypothesis

We found that GaborNet has a statistically significant higher accuracy than the CNN and the frozen baseline on Dataset A at each stage of training on Dataset B (see Figure 5, left). When first trained on Fashion-MNIST, the unconstrained CNN seems to be outperformed by the frozen baseline model.

However, when one looks at the original task accuracy as a function of the *accuracy* on Dataset B, rather than the training iteration, neither the GaborNet nor the baseline have a significant advantage over the CNN (see Figure 5, right). For a given performance accuracy score on Dataset B, the unconstrained CNN actually has the higher accuracy on Dataset A, followed by the GaborNet, then by the baseline. So, the GaborNet's higher accuracy on Dataset A shown in Figure 5 is not due to an improved management of both classification task skills, but rather because the CNN learns the new task more quickly.

Although the GaborNet has a statistically significant higher accuracy on dataset A as a function of the training iteration on dataset B, we conclude based that Gabor filters learned via backpropagation do not provide an advantage over traditional convolutional filters in preventing catastrophic forgetting.

### 4.1.4 Adversarial Robustness Hypothesis

To test adversarial robustness, we used the Projected Gradient Descent (PGD) attack [17] to generate adversarial examples from the entire test set. The strength of the perturbation to generate the adversarial example in PGD is controlled by a hyperparameter $\epsilon$. For two values, $\epsilon = 0.3$ and $\epsilon = 0.05$, we tested robustness by measuring the accuracy after perturbing all images in the test set.

We used the student's t-test and found that the GaborNet has a statistically significant ($p < 0.05$) higher accuracy on the adversarial test set than the unconstrained counterpart, for both the linear probe and full CNN, both adversarial strengths, and both orderings of datasets (see Table 3).

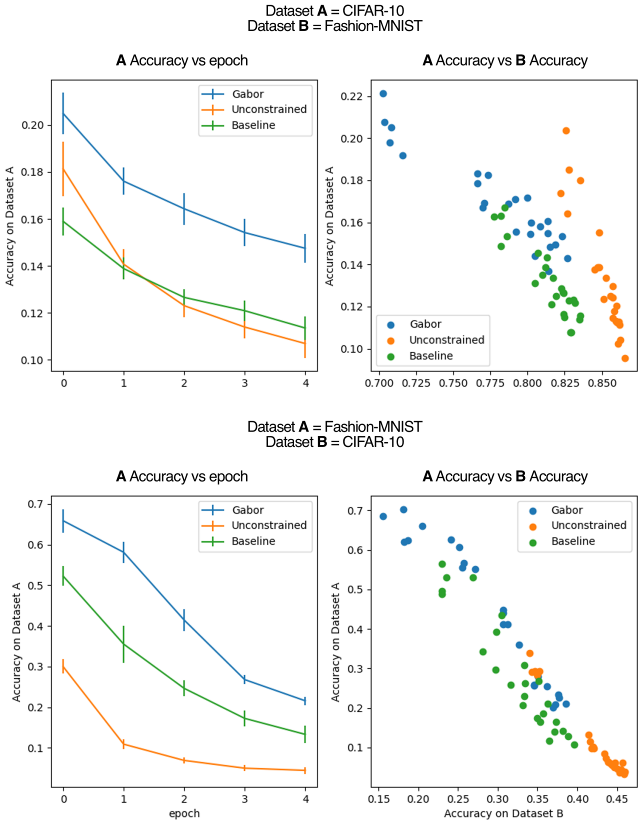


Figure 5: Accuracy of the models at each stage of convergence. **Top:** The Accuracy of the full CNN models when first trained on CIFAR-10, then finetuned on Fashion-MNIST. **Bottom:** The Accuracy of the full CNN models when first trained on Fashion-MNIST, then finetuned on CIFAR-10. **Left:** The accuracy on Dataset A versus the training epoch on Dataset B. Bars are 95% confidence intervals across 5 runs. **Right:** Scatter plot of the accuracy on Dataset A versus the accuracy of that model checkpoint on dataset B.

| | Linear probe | | full CNN | |
|---|---|---|---|---|
| dataset & $\epsilon$ | GaborNet | CNN | GaborNet | CNN |
| CIFAR-10 $\epsilon = 0.05$ | **0.0160 ± 0.004** | 0.0029 ± 0.001 | **0.0603 ± 0.009** | 0.0384 ± 0.009 |
| Fashion-MNIST $\epsilon = 0.05$ | **0.4503 ± 0.007** | 0.2415 ± 0.020 | **0.4643 ± 0.049** | 0.4093 ± 0.005 |
| CIFAR-10 $\epsilon = 0.3$ | **0.0004 ± 0.000** | 0.0000 ± 0.000 | **0.0133 ± 0.004** | 0.0070 ± 0.003 |
| Fashion-MNIST $\epsilon = 0.3$ | **0.1442 ± 0.010** | 0.0007 ± 0.000 | **0.1880 ± 0.049** | 0.1189 ± 0.0119 |

Table 3: Average accuracy of the model on adversaries generated with strength $\epsilon$. 95% confidence intervals based on 5 runs. Bold indicates statistically significant higher value when comparing within the model family (linear probe or full CNN).

| | GaborNet | CNN |
|---|---|---|
| Fashion-MNIST | 0.781 ± 0.076 | 0.852 ± 0.055 |
| CIFAR-10 | 0.795 ± 0.058 | 0.781 ± 0.077 |

Table 4: Average accuracy on adversarial examples based on each dataset and model. 95% confidence intervals are also indicated.

## 4.2 Human Experiments

### 4.2.1 Performance on Adversarial Examples

We test 20 subjects, all of which are MIT students, on the adversarial examples using jsPsych. As we can see in Table 4, we find that there are similar mean accuracies across both datasets and models. This suggests that there is likely no difference in human performance on these adversaries. In order to ensure the validity of our conclusion, we run Student's t-test on the two different populations of results (GaborNet performance versus vanilla CNN performance) for a p-value of 0.05. We calculate a statistic of 0.7228. Looking at the table for a p-value of 0.05 and 18 (20-2) degrees of freedom and the calculated statistic, we find a value of 1.732. Since 0.7228 is less than 1.732, we fail to reject the null hypothesis at the p-value of 0.05. Therefore, we conclude that there is no difference between the performance of humans on adversaries generated for the Gabornet and adversaries generated for the vanilla CNN.

# 5 Conclusion

We get several null results from our experiments. First, we fail to replicate the results of previous work[1] that influenced the direction of this project. We also report worse performance using Gabor filters in comparison to a vanilla CNN baseline for both Fashion-MNIST and CIFAR-10. Lastly, we fail to reject our null hypothesis in our human experiments and conclude that there is no difference between the performance of humans on adversaries generated for the GaborNet and adversaries generated for the vanilla CNN.

These results suggest a lesson regarding the connections between human cognition and machine learning methods. Frequently, inspiration is sought from the brain when building new machine models. For example, basic neural networks are inspired by neurons and their connection to each other in the brain. While this paradigm has led to good results in many ways, it is important not to get attached to certain methods thought to emulate cognition in the brain when designing machine models. It could be the case that a more naive or different method works better. That was the case with our results, which failed to replicate prior results which suggested that the use of human vision inspired Gabor filters outperformed naive backpropagation. It may be the case that using biologically inspired methods sometimes works better, while sometimes cold, unfeeling methods like backpropagation succeed. This tells us that machine vision models do not have to be similar to humans in order to work well.

There is also a lesson in our application of the Gabor filter itself. We assumed that since previous work showed that the GaborNet worked well on certain datasets in comparison to naive methods, that it would also outperform on most datasets and would maintain better performance while scaling up models. However, this was not the case. This reiterates the wisdom to not take things for granted, and that just because something works on a certain dataset and architecture, does not mean that it will also work on other datasets or architectures. This is especially true in brittle machine models.

# References

[1] Andrey Alekseev and Anatoly Bobe. "GaborNet: Gabor filters with learnable parameters in deep convolutional neural networks". In: *CoRR* (2019). URL: `http://arxiv.org/abs/1904.13204`.

[2] Hans Op de Beeck and Johan Wagemans. "Visual object categorisation at distinct levels of abstraction: A new stimulus set". In: *Perception* 30.11 (2001), pp. 1337–1361. DOI: `10.1068/p3120`.

[3] Jie Chen et al. "Novel Face Detection Method Based on Gabor Features". In: (2005).

[4] Issam Dagher, Sandy Mikhael, and Oubaida Al-Khalil. "Gabor face clustering using affinity propagation and structural similarity index". In: *Multimedia Tools and Applications* 80 (2021).

[5] John G. Daugman. "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters". In: *J. Opt. Soc. Am. A* (1985).

[6] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* 2019. arXiv: `1810.04805 [cs.CL]`.

[7] *Dogs vs. Cats.* URL: `https://www.kaggle.com/c/dogs-vs-cats`.

[8] Zihao Fu et al. *On the Effectiveness of Parameter-Efficient Fine-Tuning.* 2022. arXiv: `2211.15583 [cs.CL]`.

[9] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples.* 2015. arXiv: `1412.6572 [stat.ML]`.

[10] G H Granlund. "In search of a general picture processing operator". In: *Computer Graphics and Image Processing* 8 (1978), pp. 155–173.

[11] Anil Jain, Nalini Ratha, and Sridhar Lakshmanan. "Object detection using Gabor filters". In: *Pattern Recognition* 30 (1997).

[12] James Kirkpatrick et al. "Overcoming catastrophic forgetting in neural networks". In: *Proceedings of the National Academy of Sciences* 114.13 (Mar. 2017), pp. 3521–3526. DOI: `10.1073/pnas.1611835114`. URL: `https://doi.org/10.1073%2Fpnas.1611835114`.

[13] Alex Krizhevsky. "Learning Multiple Layers of Features from Tiny Images". In: (2009).

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: (2012).

[15] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. "Human-level concept learning through probabilistic program induction". In: *Science* (2015). DOI: `10.1126/science.aab3050`. URL: `https://www.science.org/doi/abs/10.1126/science.aab3050`.

[16] Brenden M. Lake et al. "Building machines that learn and think like people". In: *Behavioral and Brain Sciences* 40 (2017).

[17] Aleksander Madry et al. *Towards Deep Learning Models Resistant to Adversarial Attacks*. 2019. arXiv: `1706.06083 [stat.ML]`.

[18] S. Marĉelja. "Mathematical description of the responses of simple cortical cells∗". In: *J. Opt. Soc. Am.* 70 (1980).

[19] Bruno A. Olshausen and David J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision Research* (1997). DOI: `https://doi.org/10.1016/S0042-6989(97)00169-7`. URL: `https://www.sciencedirect.com/science/article/pii/S0042698997001697`.

[20] Steven T. Piantadosi and Richard N. Aslin. "Compositional Reasoning in Early Childhood". In: *PLoS ONE* 11 (2016).

[21] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2019).

[22] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *CoRR* (2021).

[23] Han Xiao, Kashif Rasul, and Roland Vollgraf. "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms". In: *CoRR* (2017). URL: `http://arxiv.org/abs/1708.07747`.

[24] Bo Zhang. "Computer vision vs. human vision". In: *9th IEEE International Conference on Cognitive Informatics (ICCI'10)* (2010), pp. 3–3.