A Mechanistic Analysis of Adversarial Fine-tuning of Vision Transformers

Hannah Gao*
MIT
Cambridge, MA
hanngao@mit.edu

Isha Agarwal*
MIT
Cambridge, MA
agarwali@mit.edu

Dylan Hadfield-Menell
MIT
Cambridge, MA
dylanhm@mit.edu

Rachel Ma
MIT
Cambridge, MA
rachelm8@mit.edu

Abstract

The widespread use of image classification models in high-risk, real-world situations necessitates making these models robust to slight disturbances or perturbations, such as blurring or sharpening, in the input images. While vision transformers (ViTs) play an integral role in many modern-day multi-modal models like Vision-Language-Models (VLMs) and Vision-Language-Action (VLA) models, they have received a lack of attention in the setting of robustness. In this work, we analyze the effects of adversarial fine-tuning, a popular method for improving model robustness to image perturbations, on a ViT's performance on perturbed and regular images through a mechanistic lens. We adversarially train a ViT on low-frequency and high-frequency image corruptions, and attempt to explain changes in downstream model performance through an examination of the model's attention mechanisms, internal representations, and knowledge evolution. Overall, our results suggest that, while fine-tuning on inputs with common corruptions improves model performance and certainty on new instances of corrupted data, these improvements do not transfer to other classes of corruptions not seen in the training. Additionally, despite observing changes in visual attention and knowledge evolution across layers, we found that adversarial training did not lead to fundamental changes in the sparse representations learned by ViTs.

1. Introduction

Image classification models have wide-ranging applications, from medical image analysis to self-driving cars [1, 26]. It is important to develop classification models invariant to image perturbations, including minor modifications such as blurs, that commonly occur in the real world. For example, security cameras should continue to work to classify threats, even when there is fog or other weather elements that may interfere with the lens's ability to get a clear photo. Additionally, given the growing popularity of vision-language models (VLMs) and vision-language-action (VLA) models, it is more important than ever to study the robustness of vision models, like Vision Transformers (ViTs), which are directly being augmented into such models for perception [12].

One way of making models more robust is training on more adversarial examples. We look at the effect of adversarial training on simple and common image corruptions. It is important to understand exactly how adversarial training is influencing the model, mechanistically, as fine-tuning has been shown to have inadvertent side-effects on learning and performance, such as catastrophic forgetting [27].

Our project investigates how robust latent representations of images are to image perturbations in the form of common image corruptions. We examine how the internal activations of a Vision Transformer (ViT) image classification model change when given an image versus its corrupted counterpart. We also explore whether fine-tuning the ViT on perturbed images improves robustness of the model's learned features.

We explore the effects of adversarially training a ViT on corrupted images through two main questions: 1) What is the impact of adversarial training on downstream ViT performance? and 2) How can we explain the changes in performance mechanistically?

To answer these questions, we compare how the way models process perturbed and original images differ before and after adversarial fine-tuning on different image corruptions. To this end, we examine how different the model attentions are for blurred versus original images in each model. Then, we train a Sparse Autoencoder (SAE), which provides us with interpretable feature representations, on each model to understand how internal representations of blurred versus original images are affected by adversarial training.

Our main contributions can be summarized as follows:

1. We fine-tune ViT models on two different families of common corruptions and compare downstream performance and find that adversarially fine-tuning on low fre-

*Equal contribution.

quency perturbations does not translate to improved performance on high-frequency perturbations.
2. We apply interpretability techniques to understand how the adversarial fine-tuning changes the model both in its attention mechanism and the latent representations.
3. We observe that adversarially fine-tuned models exhibit greater confidence in correct answers and the emergence of correct classifications earlier in during input processing.

2. Related Works

Robustness of Vision Models. While in recent years vision models have demonstrated impressive capabilities across tasks ranging from object-identification to spatial analysis, they still suffer performance degradations when given images with some form of perturbation.

Perturbations consist of adversarial attacks, which are perturbations specially engineered to fool the model, or common corruptions [9] such as blurs or noise, which resemble more natural perturbations that may occur in the real world.

While most works have focused on adversarial attacks on vision models [8, 13, 17], several studies have also examined the the effects of common image corruptions and have found vision models to be vulnerable to various degrees to several common corruption types, particularly to high-frequency perturbations [22, 24]. However, there has been relatively little work to analyze the robustness of ViTs to image perturbations [14] and particularly to these common corruptions. Furthermore, to our knowledge, such works only focus simply on downstream performance impacts of common corruptions and propose mitigations like adversarial training [7, 22] without going any further to mechanistically understand the underlying reason for observed performance changes.

Mechanistic Interpretability on Vision Models. While most prior mechanistic interpretability works have focused on language models [3, 15, 23], more recent literature has started to adapt techniques for language models such as examining activations, logit attribution and training SAEs for vision models [10, 11, 20]. There have also been works that have come up with or applied vision-specific interpretability techniques [19, 21].

However, the internal workings of vision models remain understudied, especially in the context of fine-tuning and robustness. Given that robustness to perturbations is a crucial property for any vision model deployed in the real world where images may naturally be perturbed, we are interested in studying and quantifying this phenomenon by studying the model's internal processing of perturbed images.

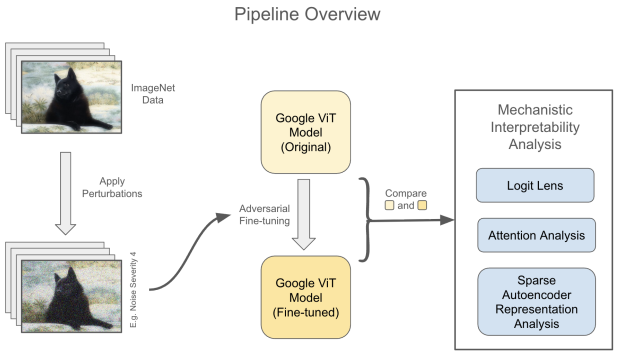

Figure 1. Overview of methods. ViT is adversarially trained on perturbed ImageNet images, and the fine-tuned and original base ViT are analyzed and compared using various mechanistic interpretability techniques. Images sourced (and modified) from ImageNet ILSVRC2012 dataset.

3. Methods

We apply common corruptions to an image dataset and use these perturbed images to adversarially fine-tune a ViT. In addition to comparing the base and fine-tuned models on downstream performance on new perturbed images, we additionally compare the two models using a suite of mechanistic interpretability techniques (see Fig. 1)

3.1. Adversarial Training

We perform adversarial training by fine-tuning a ViT on image corruptions on the ImageNet ILSVRC2012 dataset [4]. We broadly explore two main categories of image corruptions: low-frequency and high-frequency corruptions. For our low-frequency corruption, we apply the common Gaussian blur, and for our high-frequency corruption, we apply a Gaussian noise filter. We apply these filters each at severity levels 1, 2, and 4, where for the Gaussian blur the level is used as the standard deviation of the blur kernel, while for Gaussian noise we follow the implementation of the severity levels used in ImageNetC [9]. In both cases, a higher severity level indicates a higher degree of image corruption.

3.2. Class Prediction Progression Across Layers

To understand how the model's class prediction evolves across layers, we apply the logit lens technique [16]. For LLMs, this techniques applies the unembedding matrix matrix $W_U \in \mathbb{R}^{d_{model} \times |V|}$ to project an intermediate hidden state $h^{(l)} \in \mathbb{R}^{d_{model}}$, producing logits$^{(l)}$ over the vocabulary V. Taking the token with the maximum logit in logits$^{(l)}$ then allows us to determine what the model "predicts" the next token to be at that layer.

We modify the technique for a ViT classifier by examining the classifier head output when applied to $c^{(l)}$, the CLS token representation at an intermediate layer l, to produce

logits$^{(l)}$. We then apply a softmax to get the model's probability distribution over the possible classes for that layer.

For inputs on which the model considers the correct classification at some point in its processing, we quantify the layer at which the correct answer first emerges as a prediction. Similar to logit lens for LLMs, we take the model's class prediction at layer l to be the maximum logit after applying the classifier head to $c^{(l)}$.

3.3. Attention Analysis

We consider how the model's underlying attention structure changes on perturbed inputs. Let $A^{(l)}_{h,i,j}$ denote the attention output of head h in layer l for query token i attending to token j.

We measure how close attentions for the original versus perturbed input are at each layer by examining squared difference and cosine similarity. To quantify how perturbation impacts attention distribution over the token, we also compute the entropy $H^{(l)}_{h,i}$:

$$H^{(l)}_{h,i} = -\sum_{j=1}^{T} A^{(l)}_{h,i,j} \log(A^{(l)}_{h,i,j}). \quad (1)$$

We then take an average over all heads and query tokens to produce average entropy $\overline{H}^{(l)}$ for each layer. We compare $\overline{H}^{(l)}$ for an original image x and a perturbed image $\tilde{x}$ to see how perturbation affects attention spread over the image patches.

3.4. Representational Analysis

To understand how learned representations in the ViT are affected by adversarial training, we explore the internal activations of the ViT. To disentangle the internal representations of the model into a dictionary of sparse concepts, we train Sparse Autoencoders (SAEs) on the outputs of the penultimate ViT transformer layer, following guidance from previous works such as [6] which suggest placing SAEs close to the end of the model.

An SAE encodes representations into a higher-dimensional representation space and then decodes back to the original dimension, with the objective of reconstructing representations as closely as possible while simultaneously enforcing sparsity in the training loss to ensure that the features in the SAE are sufficiently disentangled, i.e. that very few SAE neurons activate strongly on a given input.

We train both a Vanilla ReLU SAE [3], on the loss:

$$||x - \text{Dec}(\text{Enc}(x))||_2^2 + \lambda||\text{Enc}(x)||_1 \quad (2)$$

and a BatchTopK SAE [2], using the loss:

$$||x - \text{Dec}(z_{topk})||_2^2 + \lambda||z_{topk}||_1 + \alpha L_{aux} \quad (3)$$

Model	Acc.	Top 5 Acc.	Top 10 Acc.	ECE
Base	0.649	0.860	0.903	0.054
Blur 1 tuned	0.725	0.905	0.941	0.046
Blur 2 tuned	0.744	0.922	0.953	0.054
Blur 4 tuned	0.753	0.929	0.954	0.055

Table 1. Comparison of accuracies and expected calibration error (ECE) for models fine-tuned on images with Gaussian blur severity levels 1, 2, and 4.

where $z_{topk} = \text{BatchTopK}(\text{Enc}(x))$, and Enc and Dec are the typical encoding and decoding functions consisting of an encoding/decoding matrix respectively and a bias term. We use the BatchTopK function provided by [2].

4. Experiments

All experiments are performed on Google's ViT-B/16 model [4, 5, 25] which is a vision classifier pre-trained on ImageNet21 (containing 4 million images and 21k target classes) and fine-tuned on ImageNet ILSVRC2012 (containing 1 million images and 1k target classes) [4, 18].

We apply each of our 6 perturbations (3 levels of Gaussian blur and 3 levels of Gaussian noise) on 10k images from held-out ImageNet validation data, and fine-tune the ViT separately on each of these datasets. We fine-tune for up to 10 epochs, using early-stopping, learning rate 5e-5, AdamW optimizer, and a linear learning-rate scheduler.

4.1. Adversarial Finetuning Results

We evaluate all the adversarially trained models on 1k new blurred and noised image samples, and compare their performance with the original base ViT.

We notice that the ViTs fine-tuned on Gaussian blurs resulted in increased downstream accuracy (i.e., top-1-accuracy), top-5 accuracy, and top-10 accuracy on classifying new blurred images, and the performance gain was monotonically increasing in the level of blur used in the adversarial training data (i.e., blur-level-4-tuned ViTs performed the best). (See Tab. 1) We noticed the same trend when evaluating the various levels of Gaussian-noise-tuned models on new noised image samples.

Interestingly, we notice that across the blur-tuned models (i.e., models fine-tuned on Gaussian blur datasets), the metric that saw the greatest improvement on novel blurred data was top-1-accuracy, followed by top-5-accuracy, and finally top-10-accuracy; the same pattern was observed for noise-tuned models. (See Fig. 2 and Fig. 3).

These results suggest that adversarial training on higher severity of common corruptions like Gaussian blurs and Gaussian noise lead to increased model certainty in the correct classification on future similar corruptions: while all

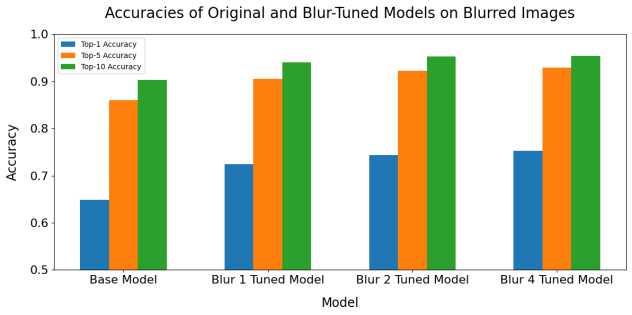

Figure 2. Top-1-accuracy, top-5-accuracy, and top-10 accuracy of the blur-tuned models on blur-4 test images.

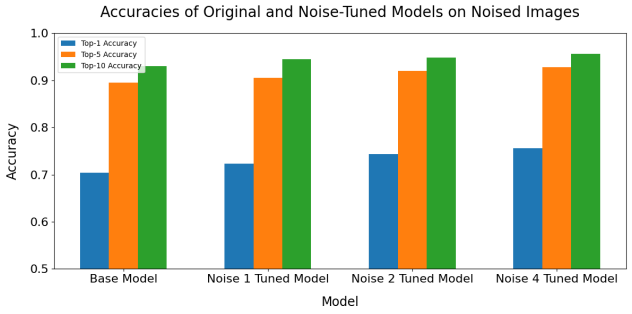

Figure 3. Top-1-accuracy, top-5-accuracy, and top-10 accuracy of the noise-tuned models on noise-4 test images.

models seem to have similar rates of predicting the correct answer within their top-10 (and even top-5) answers, the models fine-tuned on higher levels of blur and noise filters had notably higher top-1 accuracies.

We observed minimal to no changes in the top-1, top-5, and top-10 accuracies for blur-tuned models evaluated on the noise-4 test data and for noise-tuned models evaluated on the blur-4 test data. This indicates that robustness gained from adversarial training on one family of corruptions is not transferrable, though not harmful, to robustness in other families of corruptions.

We also observe minimal to no changes in the top-1, top-5, and top-10 accuracies of the adversarially-trained models on uncorrupted ImageNet images, suggesting that moderate amounts of fine-tuning on corruptions may confer neither general performance benefits nor harm to ViTs.

4.2. Class Prediction Progression Across Layers

We look at the average probability of predicting the correct class throughout the layers of each fine-tuned model. The results for both the Gaussian blur and noise are included in Fig. 4.

The results suggest that while all models start out with about the same probability of predicting the correct class, models fine-tuned on the highest filter severity (which also exhibited the best performance on these datasets) tend to

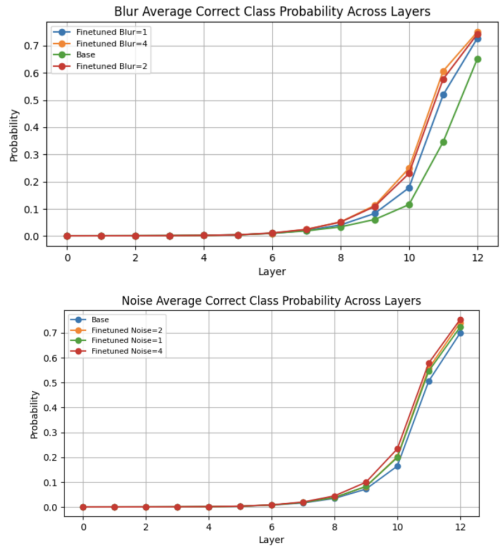

Figure 4. Probability of predicting the correct class for base and finetuned models on dataset with Gaussian blur (top) and Gaussian noise (bottom) of severity 4 applied.

overtake other models by having a higher probability of predicting the correct class in a mid-to-late layer. This suggests the mechanism for improved accuracy may be localized to those layers, or that these layers may be responsible for some difference in processing corrupted images, though further investigation is required to determine whether such a mechanism exists and may be controlled.

When restricting to inputs for which model predicts the correct class at some layer, we find that the first layer in which the fine-tuned models learns (i.e. has the highest logit for) the correct class is lower than the original model on the adversarial datasets, as shown in Fig. 5 and Fig. 6. The effect is more pronounced for Gaussian blur than for Gaussian noise. Overall, this suggests that after fine-tuning, ViTs are able to tell what the correct image classification is earlier on in processing for adversarial perturbations.

4.3. Attention Analysis

We look at the results of the difference in attention entropy between the original and blurred images for each model in Fig. 7 and Fig. 8. For both perturbations, the difference starts out negative, suggesting that the perturbed image attentions are more centralized as opposed to spread out over the whole image than the original image attentions. However, as the model processes the image, the difference becomes positive.

For the blur dataset in Fig. 7, we observe that the difference is highest for the original model, suggesting that its blurred image attention entropy is lower and possibly more skewed towards certain image patches as opposed to being

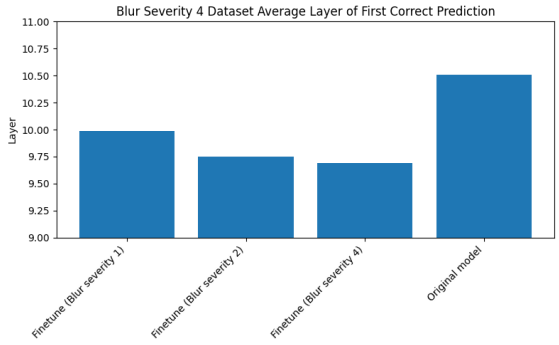

Figure 5. The average first layer the model predicts the correct class (restricted to only consider inputs on which the models predict the correct class at some layer) on the blur severity four dataset.

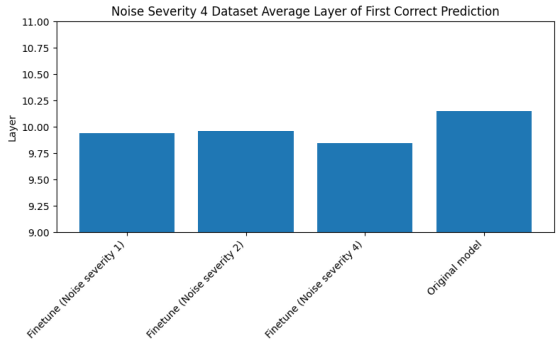

Figure 6. The average first layer the model predicts the correct class (restricted to only consider inputs on which the models predict the correct class at some layer) on the noise severity four dataset.

more evenly spread out. This suggests that robustness fine-tuning for Gaussian blurs may enable models to focus on more specific patches of the image as opposed to the overall image more broadly.

On the other hand, for Gaussian noise, we do see the original model having a higher entropy difference around layers 2 through 5, but this effect is mitigated in later layers. This suggests that fine-tuning on high frequency noise may not change the attention distribution as much.

4.4. SAE Representation Analysis

We quantify the similarity between a ViT representations of unseen ImageNet images and their corrupted counterpart by calculating the cosine between the extracted SAE activations when the original and corrupted images are fed into the model.

For both BatchTopK and Vanilla SAE, we use an SAE expansion factor of 32 and λ. The BatchTopK SAE additionally uses $\alpha = \frac{1}{32}$ and k=32. Both types of SAEs

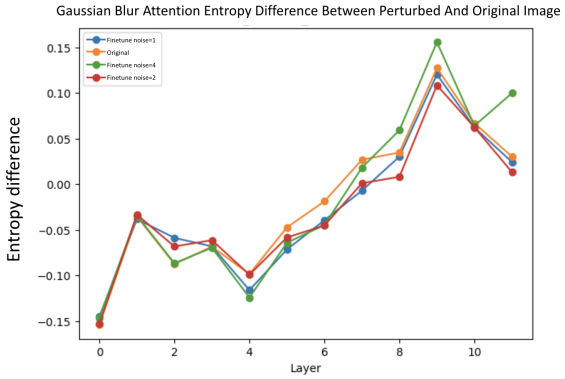

Figure 7. The average difference in attention entropy between an original image and blurred image across the layers of the model.

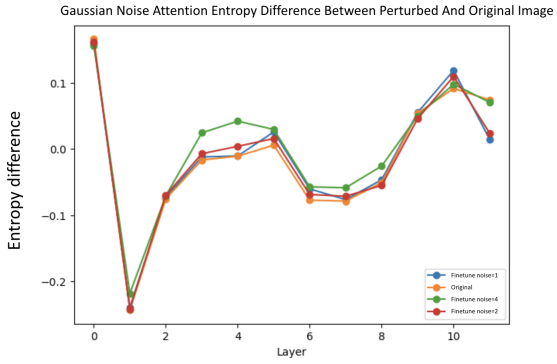

Figure 8. The average difference in attention entropy between an original image and noised image across the layers of the model.

are trained for 15 epochs with early stopping and an initial learning rate from 1e-3 paired with a Cosine Annealing scheduler. We train the SAEs on a mix of uncorrupted and corrupted images to learn a comprehensive feature dictionary.

In general, we observe that the distributions are similar between the models further fine-tuned on corrupted data and the base model. Fig. 9 and Fig. 10 show the distribution of the cosine similarities between SAE activations for corresponding patches in images and their blurred counterpart in a blur-4-tuned ViT compared to the base ViT.

The nearly-identical distributions of cosine similarities between the two models suggest that the improvement in model robustness gained from adversarial training may not be closely tied to fundamental differences in learned representations of the model.

5. Discussion and Conclusions

In this study, we analyze how ViTs process perturbed images. We fine-tune ViT classifiers on common perturbations and compare how the internal processing of these models

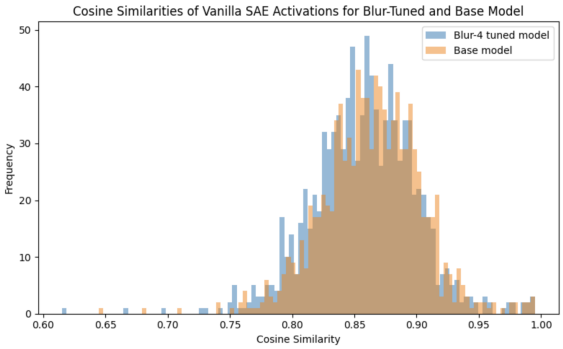

Figure 9. Distribution of cosine similarities of vanilla SAE activations between pairs of original and corrupted images for a blur-4-tuned ViT and the base ViT.

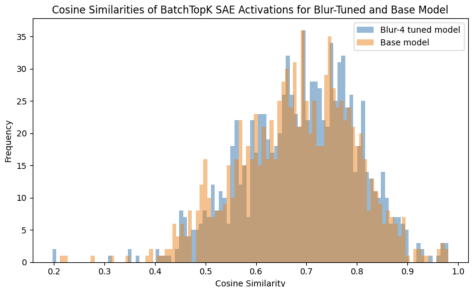

Figure 10. Distribution of cosine similarities of BatchTopK SAE activations between pairs of original and corrupted images for a blur-4-tuned ViT and the base ViT.

differs to understand the impacts of adversarial fine-tuning by looking at the class prediction progression across layers, the model's latent representation space, and its attention mechanism.

We find that adversarial training improves performance and certainty on the adversarial dataset. However, we find that improved performance does not transfer to other families of perturbations, underscoring the importance of comprehensive adversarial training data. Mechanistically, we observe that models fine-tuned on perturbations are more likely to predict the correct answer at an earlier layer than the original model and have more certainty in their answer.

Future work could investigate how training transfers between different filters of the same family. Future work for the mechanistic analysis could include trying to isolate and control the key differences between the base and the fine-tuned model.

References

[1] Abu Shad Ahammed, Md Shahi Amran Hossain, and Roman Obermaisser. A computer vision approach for autonomous cars to drive safe at construction zone, 2024. 1

[2] Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders, 2024. 3

[3] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. 2, 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2, 3

[5] 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, 2021. 3

[6] Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[7] Suklav Ghosh, Sonal Kumar, and Arijit Sur. C-lead: Contrastive learning for enhanced adversarial defense, 2025. 2

[8] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. 2

[9] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. 2

[10] Nick Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don't need trained registers, 2025. 2

[11] Sonia Joseph and Neel Nanda. Laying the foundations for vision and multimodal mechanistic interpretability & open problems. https://www.alignmentforum.org/posts/kobJymvvcvhbjWFKe/laying-the-foundations-for-vision-and-multimodal-mechanistic, 2024. AI Alignment Forum post. 2

[12] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model, 2024. 1

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. 2

[14] Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples, 2021. 2

[15] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023. 2

[16] nostalgebraist. interpreting gpt: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020. LessWrong post, accessed 2026. 2

[17] Luke Rowe, Benjamin Thérien, Krzysztof Czarnecki, and Hongyang Zhang. A closer look at robustness to l-infinity and spatial perturbations and their composition, 2022. 2

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 3

[19] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019. 2

[20] Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Interpretable and testable vision features via sparse autoencoders, 2025. 2

[21] Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. Diffusion lens: Interpreting text encoders in text-to-image pipelines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 9713–9728. Association for Computational Linguistics, 2024. 2

[22] Muhammad Usama, Syeda Aishah Asim, Syed Bilal Ali, Syed Talal Wasim, and Umair Bin Mansoor. Analysing the robustness of vision-language-models to common corruptions, 2025. 2

[23] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022. 2

[24] Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. A survey on the robustness of computer vision models against common corruptions, 2024. 2

[25] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020. 3

[26] Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3dvlm: An efficient vision-language model for 3d medical image analysis, 2025. 1

[27] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models, 2023. 1