REFERENCES

Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. *arXiv preprint arXiv:1806.05594*, 2018.

Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM (JACM)*, 58(2):1–34, 2011.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):1–11, 2022.

Irwan Bello, William Fedus, Xianzhi Du, Ekin D Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting resnets: Improved training and scaling strategies. *arXiv preprint arXiv:2103.07579*, 2021.

Diane Bouchacourt, Mark Ibrahim, and Ari Morcos. Grounding inductive biases in natural images: invariance stems from variations in data. *Advances in Neural Information Processing Systems*, 34: 19566–19579, 2021.

Andrew Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. *arXiv preprint arXiv:2102.06171*, 2021.

Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *arXiv preprint arXiv:2103.14899*, 2021.

Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks, 2017.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *arXiv preprint arXiv:2104.13840*, 2021.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.

Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021.

Stéphane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. *arXiv preprint arXiv:2103.10697*, 2021.

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*, pp. 248–255. Ieee, 2009.

Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13733–13742, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *arXiv preprint arXiv:2106.09681*, 2021.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations, 2018.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International conference on machine learning*, pp. 1802–1811. PMLR, 2019.

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pp. 3165–3176. PMLR, 2020.

Nathan Frey, Ryan Soklaski, Simon Axelrod, Siddharth Samsi, Rafael Gomez-Bombarelli, Connor Coley, and Vijay Gadepally. Neural scaling of deep chemical models, 2022.

Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Jian Guo, He He, Tong He, Leonard Lausen, Mu Li, Haibin Lin, Xingjian Shi, Chenguang Wang, Junyuan Xie, Sheng Zha, Aston Zhang, Hang Zhang, Zhi Zhang, Zhongyue Zhang, Shuai Zheng, and Yi Zhu. Gluoncv and gluonnlp: Deep learning in computer vision and natural language processing. *Journal of Machine Learning Research*, 21(23):1–7, 2020. URL http://jmlr.org/papers/v21/19-429.html.

Brian C Hall. Lie groups, lie algebras, and representations. In *Quantum Theory for Mathematicians*, pp. 333–366. Springer, 2013.

Dongyoon Han, Sangdoo Yun, Byeongho Heo, and YoungJoon Yoo. Rethinking channel dimensions for efficient model design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 732–741, 2021a.

Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. *arXiv preprint arXiv:1812.01187*, 2018.

Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. *arXiv preprint arXiv:2103.16302*, 2021.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Michael J Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. Lietransformer: Equivariant self-attention for lie groups. In *International Conference on Machine Learning*, pp. 4533–4543. PMLR, 2021.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kaggle and EyePacs. Kaggle diabetic retinopathy detection, July 2015. URL `https://www.kaggle.com/c/diabetic-retinopathy-detection/data`.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Z"ollner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6:27988, 2016.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Dmitry Laptev, Nikolay Savinov, Joachim M Buhmann, and Marc Pollefeys. Ti-pooling: transformation-invariant pooling for feature learning in convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 289–297, 2016.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 991–999, 2015.

Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 510–519, 2019.

Hanxiao Liu, Zihang Dai, David R So, and Quoc V Le. Pay attention to mlps. *arXiv preprint arXiv:2105.08050*, 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021b.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 181–196, 2018a.

Dhruv Kumar Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018b.

Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5048–5057, 2017.

Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect. *ACM Transactions on Graphics*, 39(4), Jul 2020. ISSN 1557-7368. doi: 10.1145/3386569.3392410. URL `http://dx.doi.org/10.1145/3386569.3392410`.

Chris Olah, Nick Cammarata, Chelsea Voss, Ludwig Schubert, and Gabriel Goh. Naturally occurring equivariance in neural networks. *Distill*, 5(12):e00024–004, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10428–10436, 2020.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks?, 2021.

Antônio H Ribeiro and Thomas B Schön. How convolutional neural networks deal with aliasing. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2755–2759. IEEE, 2021.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions, 2019.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning, 2016.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019a.

Mingxing Tan and Quoc V. Le. Mixconv: Mixed depthwise convolutional kernels, 2019b.

Mingxing Tan and Quoc V Le. Efficientnetv2: Smaller models and faster training. *arXiv preprint arXiv:2104.00298*, 2021.

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile, 2019.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021a.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021b.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *arXiv preprint arXiv:2103.17239*, 2021c.

Cristina Vasconcelos, Hugo Larochelle, Vincent Dumoulin, Rob Romijnders, Nicolas Le Roux, and Ross Goroshin. Impact of aliasing on generalization in deep convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10529–10538, 2021.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 210–218. Springer, 2018.

Chien-Yao Wang, Hong-Yuan Mark Liao, I-Hau Yeh, Yueh-Hua Wu, Ping-Yang Chen, and Jun-Wei Hsieh. Cspnet: A new backbone that can enhance learning capability of cnn, 2019.

Maurice Weiler and Gabriele Cesa. General E(2)-Equivariant Steerable CNNs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021.

Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5028–5037, 2017.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *CoRR*, abs/1905.00546, 2019. URL `http://arxiv.org/abs/1905.00546`.

Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation, 2019.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Muller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pp. 7324–7334. PMLR, 2019.

Zhi Zhang, Tong He, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of freebies for training object detection neural networks. *arXiv preprint arXiv:1902.04103*, 2019.

Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, Sercan Ö Arik, and Tomas Pfister. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3417–3425, 2022.

Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Oriented response networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 519–528, 2017.

Appendix

Table of Contents

A Aliasing Extended Discussion

When working with signal f we can write the fourier series

$$f(k) = \sum_{n=-\infty}^{\infty} F(n) e^{i2\pi kn/N}$$

where $F(n)$ are the frequency components and the values n are called *harmonics*. Take the f to be discretely sampled at uniformly spaced points in a bounded interval $[-1, -1+\Delta, ..., 1-\Delta, 1]$, with $\Delta = 1/K$. Because the sampling rate is limited, it is impossible to correctly measure components $F(n)$ where $n > K/2$. Evaluated only at the grid points, such content could have identical values to components with lower frequencies, causing fundamental ambiguities:

$$\sin(2\pi(k+nK)t+\phi) = \begin{cases} +\sin(2\pi(k+nK)t+\phi) & k+nK \geq 0 \\ -\sin(2\pi|k+nK|t+\phi) & k+nK < 0 \end{cases}$$

The default mechanism for resolving these ambiguities in the reconstruction is to choose the lowest frequency component for the corresponding observations, leading to the aliasing operation given in Equation 2.

This operation can also be considered a translation in the frequency domain. Crucially, operations in frequency domain have corresponding operations in the spatial domain, and thus aliasing can give rise to recognizable patterns in images with poorly chosen resolutions, for example moire patterns. This relationship also means aliasing's effects on translations in frequency space, for example, can effect translational spatial symmetries.

To make this relationship explicit, let us consider the translational symmetries of the first set of feature maps in a CNN in two scenarios. In both scenarios the transformation is downward translation of the input by 10% of its height. First, let us consider the case where this transformation happens to result in a translation by a discrete numbers of pixels, p in the feature maps. Obviously the pixels at the bottom of the image become lost to the boundary and thus cannot be recovered from the corresponding feature maps, as would be required for equivariance, as illustrated in Figure 1. As the amount of this translation gets smaller and smaller, however, the effect of the boundary should decrease, and yet the ability to recover the image can still be strongly affected by innate signal processing properties.

Consider the case where the CNN has a stride of 2. The feature maps will have half the width of the original image. Therefore the Nyquist frequency will also be half that of the Nyquist frequency of the image, and there will be aliasing of all the frequencies in between the original Nyquist frequency and the new value. When we try to reverse the transformation by translating p pixels upwards, the resulting translation will no longer be the inverse of the translation on the image. Therefore we cannot achieve perfect equivariance.

As another important subcase, let's also consider the non-linear activation in the CNN layer by itself. If we apply the non-linearity to a translated input, we can simply use the fact the result was a discrete translation in the output space to map the values at the grid points to values at different grid points under the reverse transformation. In this case there is clearly no issue introduced from the frequency domain properties of non-linearities on their own.

Now let's consider a translation of $1/p$ pixels. In this case, reconstructing the image after the translation is non-trivial, and we need to perform interpolation to calculate the values of the corresponding continuous image at the points that will become translated to the evaluation points. In order to perform this interpolation we must actually consider the full frequency spectrum of the image. Now the effects of pointwise non-linearities can become apparent. Because non-linearities can introduce high frequency content, these high frequencies become important when reconstructing the signal using interpolation. Aliasing makes this reconstruction fundamentally challenging and thus equivariance is impossible to achieve.

B Lie Groups, Lie Derivatives, and LEE

B.1 Lie Groups and Local/Global Notions of Equivariance

The key to understanding why the local - global equivalence holds is that $(\exp(X) - 1) = \sum_{k=1}^{\infty} X^k/k!$ has the same nullspace as X (here repeated application of X on a function f is just the repeated directional derivative, and this is the definition of a vector field used in differential geometry). Since they have the same nullspace, the space of functions for which $\exp(X)f = f$ is the same as the space $Xf = 0$. The same principle holds for $\rho(\exp(X))f = f$ and $d\rho(X)f = 0$ since $\rho(\exp(X)) = \exp(d\rho(X))$ (a basic result in representation theory, which can be found in (Hall, 2013)) where $d\rho$ is the corresponding Lie algebra representation of ρ, which for vector fields is the Lie derivative $d\rho(X) = L_X$. Hence carrying over the constraint for each element $\forall X \in \mathfrak{g} : L_X f = 0$ is equivalent to $\forall X \in \mathfrak{g} : \rho(\exp(X))f = f$ which is the same as $\forall g \in G : \rho(g)f = f$. Unpacking the representation ρ_{12} of f, this is just the global equivariance constraint $\forall g \in G : \rho_2(g)^{-1} f(\rho_1(g)x) = f(x)$.

B.2 Lie Derivative Chain Rule

Suppose we have two functions $h : V_1 \to V_2$ and $f : V_2 \to V_3$, and corresponding representations ρ_1, ρ_2, ρ_3 for the vector spaces V_1, V_2, V_3. Expanding out the definition of ρ_{31},

$$\begin{aligned}
\rho_{31}(g)[f \circ h](x) &= \rho_3(g)^{-1} f(h(\rho_1(g)x)) \\
&= \rho_3(g)^{-1} f(\rho_2(g)\rho_2(g)^{-1} h(\rho_1(g)x)) \\
&= \rho_{32}(g)[f] \circ \rho_{21}(g)[h](x).
\end{aligned}$$

From the definition of the Lie derivative, and using the chain rule that holds for the derivative with respect to the scalar t, and noting that $g_0 = \mathrm{Id}$ so $\rho(g_0) = \mathrm{Id}$, we have

$$\begin{aligned}
\mathcal{L}_X(f \circ h)(x) &= \frac{d}{dt}\left(\rho_{31}(g_t)[f \circ h](x)\right)\Big|_0 \\
&= \frac{d}{dt}\left(\rho_{32}(g_t)[f] \circ \rho_{21}(g_t)[h](x)\right)\Big|_0 \\
&= \left(\frac{d}{dt}\rho_{32}(g_t)[f]\big|_{t=0}\right) \circ \rho_{21}(g_0)[h](x) + \left[d(\rho_{32}(g_0)[f])\Big|_{h(x)}\right]\left(\frac{d}{dt}\rho_{21}(g_t)[h]\big|_{t=0}\right)(x) \\
&= \left(\frac{d}{dt}\rho_{32}(g_t)[f]\big|_{t=0}\right) \circ h(x) + df|_{h(x)}\left(\frac{d}{dt}\rho_{21}(g_t)[h]\big|_{t=0}\right)(x) \\
&= (\mathcal{L}_X f) \circ h(x) + df|_{h(x)}(\mathcal{L}_X h)(x),
\end{aligned}$$

where $df|_{h(x)}$ is the Jacobian of f at $h(x)$ and $df|_{h(x)}(\mathcal{L}_X h)(x)$ is understood to be the Jacobian vector product of $df|_{h(x)}$ with $(\mathcal{L}_X h)(x)$, equivalent to the directional derivative of f along $(\mathcal{L}_X h)(x)$. Therefore the Lie derivative satisfies a chain rule

B.3 Stochastic Trace Estimator for Layerwise Metric

Unrolling this chain rule for a sequence of layers $\mathrm{NN}(x) = f_{N:1}(x) := f_N(f_{N-1}(...(f_1(x))))$, or even an autograd DAG, we can identify the contribution that each layer f_i makes to the equivariance error of the whole as the sum of terms $C_i = df_{N:i+1}\mathcal{L}_X f_i$, $\mathcal{L}_X(\mathrm{NN}) = \sum_{i=1}^N C_i$.

Each of these C_i, like $\mathcal{L}_X(\mathrm{NN})$ measure the equivariance error for all of the outputs (which we define to be the softmax probabilities), and are hence vectors of size K where K is the number of classes. In order to summarize the C_i as a single number for plotting, we compute their norm $\|C_i\|$ which satisfy $\|\mathcal{L}_X(\mathrm{NN})\| \le \sum_i \|C_i\|$.

To compute $df_{N:i+1}\mathcal{L}_X f_i$, one can use autograd to perform Jacobian vector products (as opposed to typical vector Jacobian products) and build up $df_{N:i+1}$ in a backwards pass. Unfortunately doing so is quite cumbersome in the PyTorch framework where the large number of available models are implemented and pretrained. A trick which can be used to speed up this computation is to use stochastic trace estimation (Avron & Toledo, 2011). Since vector Jacobian products are cheap and easy, we can compute $\|C_i\|^2 = \mathbb{E}[\hat{A}]$ as the expectation of the estimator $\hat{A} = (1/N)\sum_n^N (z_n^\top C_i)^2 = (1/N)\sum_n^N (z_n^\top df_{N:i+1}\mathcal{L}_X f_i)^2$ with iid. Normal probe vectors $z_n \sim \mathcal{N}(0, I)$, and the quantity $z_n^\top df_{N:i+1}$ which is a standard vector Jacobian product.

One can see that $\mathbb{E}[\hat{A}] = C_i^\top \mathbb{E}[zz^\top]C_i = C_i^\top I C_i = \|C_i\|^2$. We can then measure the variance of this estimator to control for the error and increase N until this error is at an acceptable tolerance (we use $N = 100$ probes). The convergence of this trace estimator is shown in Figure 9 (right) for several different layers of a ResNet-50. In producing the final layerwise attribution plots, we average the computed quantity $\|C_i\|$ over 20 images from the ImageNet test set.

C LEE Theorems

C.1 LEE and consistency regularization

As shown in Athiwaratkun et al. (2018), consistency regularization with Gaussian input perturbations can be viewed as an estimator for the norm of the Jacobian of the network, but in fact when the perturbations are not Gaussian but from small spatial transformations, consistency regularization actually penalizes the Lie derivative norm. In the Π-model (Laine & Aila, 2016) (the most basic form of consistency regularization), the consistency regularization minimizes the norm of the difference of the outputs of the network when two randomly sampled transformations T^a and T^b are applied to the input,

$$\mathrm{L}_{\mathrm{cons}} = \|f(T^a(x)) - f(T^b(x))\|^2. \tag{7}$$

Suppose that the two transformations are representations of a given symmetry group and can be written as $T^a = \rho(g_a)$ and $T^b = \rho(g_b)$, and the group elements can be expressed as the flow generated by a linear combination of the vector fields which form the Lie Algebra: $g_a = \Phi_{\sum_i a_i X_i}$ for some coefficients $\{a_i\}_{i=1}^d$ and likewise for g_b. We can define the log map, mapping group elements top their generator values in this basis: $\log(g_a) = a$. Then, assuming a_i are small (and therefore the transformations are small), Taylor expansion yields $\mathrm{L}_{\text{cons}} = \|f(x) + \sum_i a_i \mathcal{L}_{X_i} f(x) + O(a^2) - [f(x) + \sum_j b_j \mathcal{L}_{X_j} f(x) + O(b^2)]\|^2$. Therefore, taking the expectation over the distribution which a and b are sampled over (which is assumed to be centered with $\mathbb{E}[a_i] = \mathbb{E}[b_i] = 0$ as well as the input distribution x, we get that

$$\mathbb{E}_{a,b,x}[\mathrm{L}_{\text{cons}}] = 2\mathbb{E}[\| \sum_i \mathcal{L}_{X_i} f(x) \|_\Sigma^2] + \text{higher order terms}, \tag{8}$$

where $\|\|_\Sigma^2$ denotes the norm with respect to the covariance matrix $\Sigma = \mathrm{Cov}(a) = \mathrm{Cov}(b)$.

When the transformations are not parameter space perturbations such as dropout, but input space perturbations like translations (which have been found to be far more important to the overall performance of the method (Athiwaratkun et al., 2018)), we can show that consistency regularization coincides with minimizing the expected Lie derivative norm. In this sense, consistency regularization can be viewed as an intervention for reducing the equivariance error on unlabeled data.

C.2 Translation LEE and aliasing

Below we show that spatial aliasing directly introduces translation equivariance error as measured by the Lie derivative, where the aliasing operation $A[\cdot]$ is given by Equation 2. The Fourier series representation of an image $h(x, y)$ with pixel locations (x, y) is H_{nm} with spatial frequencies (n, m), where the band limited reconstruction

$$h(x, y) = \tfrac{1}{2\pi} \sum_{nm} H_{nm} e^{2\pi i(xn+ym)} = F^{-1}[H]$$

and F^{-1} is the inverse Fourier transform, and the sums range over frequencies of $-M/2$ to $+M/2$ for both n and m where M is the image height and width (assumed to be square for convenience).

Applying a continuous translation by $t\mathbf{v}$ along vector $\mathbf{v} = (v_x, v_y)$ to the input means resampling the translated band limited continuous reconstruction $h(x, y)$ at the grid points.

$$T_{t\mathbf{v}}[h](x, y) = h(x - tv_x, y - tv_y) = \tfrac{1}{2\pi} \sum_{n,m=-M/2}^{M/2} H_{nm} e^{2\pi i[(x-tv_x)n+(y-tv_y)m]}$$

To simplify the notation, we will consider translations along only x and suppress the m index of H_{nm}, effectively deriving the result for the translations of a 1d sequence, but that extends straightforwardly to the 2 dimensional case.

$$T_{t\mathbf{v}}[h](x) = h(x - tv_x) = \tfrac{1}{2\pi} \sum_{n=-M/2}^{M/2} [H_n e^{-2\pi i t v_x n}] e^{2\pi i x n}$$

Applying the aliasing operation, sampling the image to a new size M' (with Nyquist frequency $M'/2$), we have

$$\begin{aligned} A[T_{t\mathbf{v}}[h]](x) &= \tfrac{1}{2\pi} \sum_{n=-M/2}^{M/2} [H_n e^{-2\pi i t v_x n}] e^{2\pi i x \mathrm{Alias}(n)} \\ &= \tfrac{1}{2\pi} \sum_{n'=-M'/2}^{M'/2} \left[\sum_{n=\mathrm{Alias}^{-1}(n')} H_n e^{-2\pi i t v_x n} \right] e^{2\pi i x n'} \end{aligned}$$

where the last line follows from applying a change of variables $n' = \mathrm{Alias}(n)$.

Applying the final inverse translation (which acts on the M' sampling rate band limited continuous reconstruction), we have

$$T_{-t\mathbf{v}}[A[T_{t\mathbf{v}}[h]]](x) = \tfrac{1}{2\pi} \sum_{n'=-M'/2}^{M'/2} \left[\sum_{n=\mathrm{Alias}^{-1}(n')} H_n e^{-2\pi i t v_x (n-n')} \right] e^{2\pi i x n'}.$$

Taking the derivative with respect to t, we have

$$\begin{aligned}\mathcal{L}_{\mathbf{v}}(A)(h) &= \tfrac{d}{dt}\Big|_0 T_{-t\mathbf{v}}[A[T_{t\mathbf{v}}[h]]] \\ &= \tfrac{1}{2\pi}\sum_{n'=-M'/2}^{M'/2}\Bigg[\sum_{n=\mathrm{Alias}^{-1}(n')} 2\pi i v_x(n'-n)H_n\Bigg]e^{2\pi i x n'}.\end{aligned}$$

Notably, for aliasing when the frequency is reduced by a factor of 2 from downsampling, there are only two values of n that satisfy $\mathrm{Alias}(n) = n'$: the value $n = n'$ and the one that gets aliased down, therefore when multiplied by $n - n'$ the sum

$$\Bigg[\sum_{n=\mathrm{Alias}^{-1}(n')} 2\pi i v_x(n'-n)H_n\Bigg]$$

consists only of a single term.

According to Parseval's theorem, the Fourier transform F is unitary, and therefore the norm of the function as a vector evaluated at the discrete sampling points $x = 1/M', 2/M', ...$ is the same as as the norm of the Fourier transform:

$$\begin{aligned}\|\mathcal{L}_{\mathbf{v}}(A)(h)\|^2 &= \|F[\mathcal{L}_{\mathbf{v}}(A)(h)]\|^2 \\ \|\mathcal{L}_{\mathbf{v}}(A)(h)\|^2 &= \sum_{n'=-M'/2}^{M'/2}\Bigg|\sum_{n=\mathrm{Alias}^{-1}(n')} 2\pi i v_x(n'-n)H_n\Bigg|^2 \\ \|\mathcal{L}_{\mathbf{v}}(A)(h)\|^2 &= \sum_{n=-M/2}^{M/2}(2\pi)^2 v_x^2(\mathrm{Alias}(n)-n)^2 H_n^2,\end{aligned}$$

using the fact that only one element is nonzero in the sum. Finally, generalizing to the 2d case, we have

$$\|\mathcal{L}_{\mathbf{v}}(A)(h)\|^2 = (2\pi)^2\sum_{nm} H_{nm}^2\big(v_x^2(\mathrm{Alias}(n)-n)^2 + v_y^2(\mathrm{Alias}(m)-m)^2\big), \tag{9}$$

showing how the translation Lie derivative norm is determined by the higher frequency components which are aliased down.

D LEARNED EQUIVARIANCE EXPERIMENTS

D.1 LAYER-WISE EQUIVARIANCE BASELINES

We use EQ-T and EQ-$\mathrm{T}_{\mathrm{frac}}$ (Karras et al., 2021) to calculate layer-wise equivariance by caching intermediate representations from the forward pass of the model. For image-shaped intermediate representations, EQ-T samples integer translations in pixels between -12.5% and 12.5% of the image dimensions in pixels. EQ-$\mathrm{T}_{\mathrm{frac}}$ is identical but with continuous translation vectors. The individual layer is applied to the transformed input and then the inverse group action is applied to the output, which is compared with the original cached output. Many different normalization could be chosen to compare equivariance errors across layers. The most obvious are $\frac{1}{N}$, $\frac{1}{\sqrt{N}}$, and $\frac{1}{1}$ (no normalization), where $N = C \times H \times W$. As we show in section 5, the normalization method can have a large effect of the relative contribution of a layer, despite the decision being relatively arbitrary (in contrast to LEE, which removes the need for doing so as the scale is automatically measured relative to the contribution to the output).

D.2 SUBNETWORK EQUIVARIANCE ANALYSIS

Another way one might use LEE to study the effects of different layers that make up a network is to break the network in question down into its constituent subnetworks (networks starting at the input and ending at every intermediate representation in the network) and calculate the LEE of the corresponding function. We show the result of this calculation on a ResNet50 in Figure 8.

As an alternative to our layerwise analysis, this method has a significant drawback that makes analysis challenging: the functions under consideration have different outputs. In our calculation, we applied batch normalization over the outputs in order to make their scales comparable. Despite this rescaling, comparing activations and preactivations, for example, remains challenging. By contrast, our layerwise breakdown specifically targets a layer's contribution to a shared output.

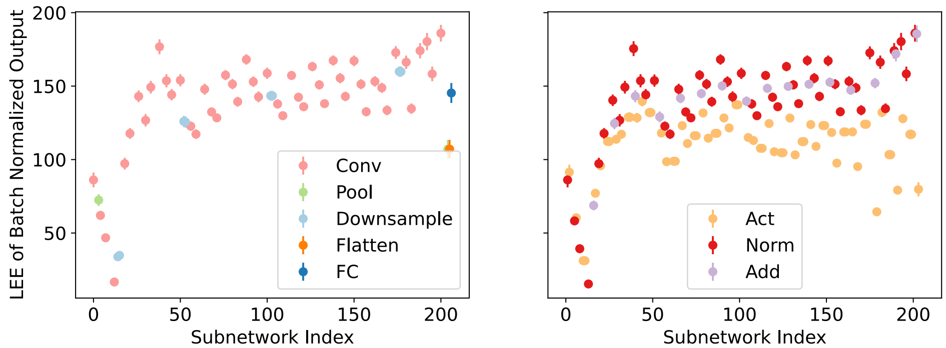

Figure 8: LEE calculated over the subnetworks of a ResNet50. Specifically a subnetwork is constructed between the input and every intermediate representation in the network's computation graph. We use batch normalization of the outputs to make the output scale of different subnetwork comparable. For visual clarity, layer types are broken across the left and right plots, which share the same axes. Similar to the pattern observed in Figure 4, we see a rapid increase in equivariance error in the early layers of the network, followed by many smaller increases later in the network. Unlike in our layerwise decomposition, comparison across layer types is challenging in this setting because layers have significantly different outputs. For example, comparing activations with preactivations is complicated by the ReLUs acting as contractions of the input and having potentially many zeroed values.

D.3 Model List

The models included in Figure 1 are

- Early CNNs: ResNets (He et al., 2015), ResNeXts (Xie et al., 2017), VGG (Simonyan & Zisserman, 2014), Inception (Szegedy et al., 2016), Xception (Chollet, 2017), DenseNet (Huang et al., 2017), MobileNet (Sandler et al., 2018), Blur-Pool Resnets and Densenets (Zhang, 2019), ResNeXt-IG (Mahajan et al., 2018a), SeResNe*ts (Hu et al., 2018), ResNet-D (He et al., 2018), Gluon ResNets (Guo et al., 2020; Zhang et al., 2019; 2020), SKResNets (Li et al., 2019), DPNs (Chen et al., 2017)

- Modern CNNs: EfficientNet (Tan & Le, 2019a; 2021), ConvMixer (?), RegNets (Radosavovic et al., 2020), ResNet-RS, (Bello et al., 2021), ResNets with new training recipes (Wightman et al., 2021), ResNeSts (Zhang et al., 2020), RexNet (Han et al., 2021a), Res2Net (Gao et al., 2019), RepVGG (Ding et al., 2021), NFNets (Brock et al., 2021), XNect (Mehta et al., 2020), MixNets (Tan & Le, 2019b), ResNeXts with SSL pretraining (Yalniz et al., 2019), DLA (Yu et al., 2019), CSPNets (Wang et al., 2019), ECA NFNets and ResNets (Brock et al., 2021), HRNet (Sun et al., 2019), MnasNet (Tan et al., 2019)

- Vision transformers: ViT (Dosovitskiy et al., 2020), CoaT (Dai et al., 2021), SwinViT (Liu et al., 2021b), (Bao et al., 2021), CaiT (Touvron et al., 2021c), ConViT (d'Ascoli et al., 2021), CrossViT (Chen et al., 2021), TwinsViT (Chu et al., 2021), TnT (Han et al., 2021b), XCiT (El-Nouby et al., 2021), PiT (Heo et al., 2021), Nested Transformers (Zhang et al., 2022)

- MLP-based architectures: MLPMixer (Touvron et al., 2021b), ResMLP (Touvron et al., 2021a), gMLP (Liu et al., 2021a), MLP-Mixers with (Si)GLU (Wightman, 2019)

D.4 Alternative End-to-End Equivariance Metrics

Discrete Consistency We adopt the consistency metric from Zhang (2019), which simply measures the fraction of top-1 predictions that match after applying an integer translation to the input (in our case by 10 pixels). Instead of reporting consistency numbers, we report $(1 - \%$ matching), so that consistency because a measure of equivariance error. Equivariant models should exhibit end-to-end invariance, high consistency, and low equivariance error.

Expected Group Sample Equivariance Inspired by work in equivariant architecture design (Finzi et al., 2020; Hutchinson et al., 2021), we provide an additional equivariance metric for comparison against the Lie derivative. Following (Hutchinson et al., 2021), we sample k group elements in the neighborhood of the identity group element, with sampling distribution $\mathcal{D}(G)$, and calculate the sample equivariance error for model f as $\frac{1}{k}||\rho_2^{-1}(g)f(\rho_1(g)x) - f(x)||$. For translations we take $\mathcal{D}(G)$ to be Uniform$(-5, 5)$ in pixels.

Versus LEE There are several reasons why the continuous lie derivative metric is preferable over discrete and group sample metrics. Firstly, it allows us to break down the equivariance error layerwise enabling more fine grained analysis in a way not possible with the discrete analog. Secondly, the metric is less dependent on architectural details like the input resolution of the network. For example, for discrete translations by 1 pixel, these translations have a different meaning depending on the resolution of the input, whereas our lie derivatives are defined as the derivative of translations as a fraction of the input size, which is consistently defined regardless of the resolution. Working with the vector space forming the Lie algebra rather than the group also removes some unnecessary freedom in how one constructs the metric. Rather than having to choose an arbitrary distribution over group elements, if we compute the Lie derivatives for a set of basis vectors of the lie algebra, we have completely characterized the space, and all lie derivatives are simply linear combinations of the computed values. Finally, paying attention to continuous transformations reveals the problems caused by aliasing which are far less apparent when considering discrete transformations, and ultimately the relevant transformations are continuous and we should study them directly.

D.5 LEE for Additional Transformations

Beyond the 3 continuous transformations that we study with Lie derivatives above, there are many more that might reveal important properties of the network. Here we include an three additional transformations–hyperbolic rotation, brightening, and stretch.

Figure 9 (left) shows that, perhaps surprisingly, models with high accuracy become more equivariant to hyperbolic rotations. We suspect this surprisingly general equivariance to diverse set of continuous transformations is probably the result of improved anti-aliasing learned implicitly by more accurate models. LEE does not identify any significant correlation between brightening or stretch transformations and generalization ability.

D.6 Rotated MNIST Finetuning

In order to test the ability of SOTA imagenet pre-trained models to learn equivariance competitive with specialized architectures, we adapted the example rotated MNIST notebook available in E2CNN repository (Weiler & Cesa, 2019). We use the base model and default finetuning procedure from (He et al., 2021), finetuning for 100 epochs, halving the learning rate on loss plateaus.

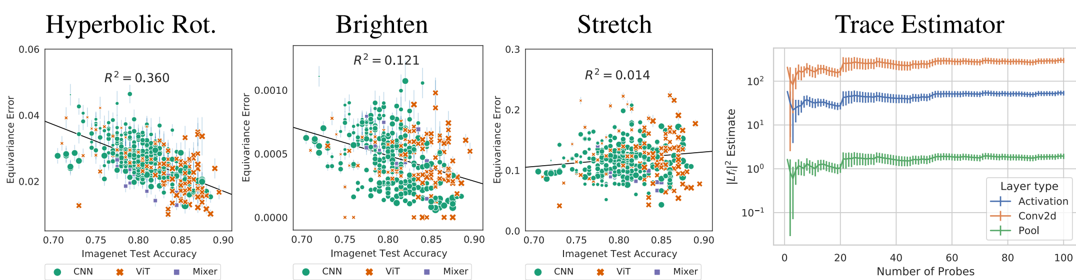

Figure 9: **(Left)**: Extending Figure 5 we show the Lie derivate norm for hyperbolic rotation, brightening, and stretch transformations. We observe that more accurate models are also more equivariant to hyperbolic rotations and to brighten transformation, to a more limited extent. In the case of hyperbolic rotations, this result is surprising, as nothing has directly encouraged this equivariance. One possible explanation is decreased aliasing in models with higher accuracy. Marker size indicates model size. Error bars show one standard error over the images use to evaluate the Lie derivative. **(Right)**: Cumulative mean and standard error of the estimator (computed for translations on a ResNet-50).