🔍 LAYER BY LAYER, MODULE BY MODULE: CHOOSE BOTH FOR OPTIMAL OOD PROBING OF VIT

Ambroise Odonnat* Noah's Ark Lab, Inria
Vasilii Feofanov 42.com
Laetitia Chapel L'Institut Agro, IRISA
Romain Tavenard Univ. Rennes 2, IRISA
Ievgen Redko Noah's Ark Lab

ABSTRACT

Recent studies have observed that intermediate layers of foundation models often yield more discriminative representations than the final layer. While initially attributed to autoregressive pretraining, this phenomenon has also been identified in models trained via supervised and discriminative self-supervised objectives. In this paper, we conduct a comprehensive study to analyze the behavior of intermediate layers in pretrained vision transformers. Through extensive linear probing experiments across a diverse set of image classification benchmarks, we find that distribution shift between pretraining and downstream data is the primary cause of performance degradation in deeper layers. Furthermore, we perform a fine-grained analysis at the module level. Our findings reveal that standard probing of transformer block outputs is suboptimal; instead, probing the activation within the feedforward network yields the best performance under significant distribution shift, whereas the normalized output of the multi-head self-attention module is optimal when the shift is weak.

vit-probing

1 INTRODUCTION

Foundation models, which rely primarily on the transformer architecture (Vaswani et al., 2017), have achieved impressive performance in a wide range of areas such as natural language processing (Brown et al., 2020; Touvron et al., 2023), computer vision (Siméoni et al., 2025), time series forecasting (Ilbert et al., 2024; Nie et al., 2023), and mathematical reasoning (Comanici et al., 2025). These models are typically pretrained on massive amount of diverse data (Du et al., 2025; Shukor et al., 2025) to encode general knowledge and then adapted to downstream tasks via finetuning (Hu et al., 2022; Lee et al., 2023) or used as frozen feature extractors (El-Nouby et al., 2024; Oquab et al., 2024) (note that the goal is different in tool-augmented large language models where models learn to use a tool instead of incorporating knowledge in their weights (Houliston et al., 2025; Lewis et al., 2020; Schick et al., 2023)). However, the reliability of foundation models remains a critical challenge in real environments. As they operate at scale, distribution shifts inevitably occur, and hidden representations remain effective only if the knowledge integrated during pretraining provides a robust prior that withstands drift at deployment.

Representations under drift. In practice, distribution shifts can severely degrade the performance by making the features extracted on out-of-distribution (OOD) data uninformative (Quionero-Candela et al., 2009). For foundation models, where pretraining data is often inaccessible, detecting and responding to this drift is a particularly challenging task. Notably, while some approaches assume the type of shift is known a priori (Lee et al., 2023; Uselis & Oh, 2025; Xie et al., 2024, 2025), this assumption rarely holds for general-purpose models. In this setting, Skean et al. (2025) challenged the conventional wisdom that final-layer representations are universally optimal, demonstrating

*Correspondence to ambroise.odonnat@gmail.com

that intermediate layers can yield superior performance. In particular, they argued that autoregressive vision models benefit from intermediate layers, whereas the final layer remains optimal for vision transformers (ViT, Dosovitskiy et al., 2021). Other recent works appear to contradict this conclusion (Bolya et al., 2025; Uselis & Oh, 2025), necessitating further investigation.

Our approach. In this paper, we study pretrained ViTs on out-of-distribution downstream image classification tasks and investigate when and why intermediate layers outperform the final layer in a linear probing setup (Section 3). We identify the distribution shift between pretraining and downstream data as the driving factor of this phenomenon, finding intermediate layers to be significantly more robust than the final ones. Motivated by prior works on component-wise adaptation (Odonnat et al., 2026; Zhao et al., 2024), we conduct a fine-grained study by probing each type of transformer module: normalization layers, multi-head attention modules, residual connections, and feedforward layers[1] (Section 4). We find that transformer modules are not equally resilient to the shift.

Takeaways. Our analysis reveals two *actionable* takeaways summarized below:

1. In ID settings, final layers always yield better performance than intermediate layers;
2. In OOD settings, probing inputs and activations of intermediate feedforwards is better.

2 Experimental setup

Vision transformer. A ViT takes as input 2D images, which are split into square patches and fed to a succession of transformer encoders. As shown in Fig. 2, each block consists of alternating multi-head attention modules (**MHA**) and feedforward networks. The latter combines two fully connected layers, **FC1** ($d \to 4d$) and **FC2** ($4d \to d$), separated by a GeLU activation (**Act**, Hendrycks & Gimpel, 2016). Two LayerNorms (Ba et al., 2016), **LN1** and **LN2**, precede the MHA and FFN, and two residual connections, **RC1** and **RC2**, follow them. In this work, we track the outputs of the 8 operations within each layer, denoting them by the name of their corresponding module. Note that the standard approach is to probe **RC2**, the output of the transformer block (El-Nouby et al., 2024; Oquab et al., 2024; Siméoni et al., 2025).

Implementation. All our experiments are conducted with an 86M-parameter ViT pretrained on ImageNet-21k (Deng et al., 2009). For a given hidden representation, linear probing is done by pooling the embeddings of the `CLS` token and applying a logistic regression with the `L-BFGS` solver. We consider a diverse set of 11 commonly used classification benchmarks: Cifar10, Cifar100 (Krizhevsky, 2009); 5 variants from Cifar10-C (Hendrycks & Dietterich, 2019): Contrast, Gaussian Noise, Motion Blur, Snow, Speckle Noise; 2 domains from DomainNet (Peng et al., 2019): Clipart, Sketch; Flowers102 (Nilsback & Zisserman, 2008) and Pets (Parkhi et al., 2012). The preprocessing protocol follows Dosovitskiy et al. (2021). The full implementation details are given in Appendix A.

3 Distribution shift degrades the performance of final layers

Skean et al. (2025) showed that intermediate layers of large language models consistently outperform the final ones. When conducting a similar analysis on vision transformers, they observe increasing downstream performance toward the final layers, except for the AIM model (El-Nouby et al., 2024), which is the only autoregressive vision model. It leads them to conclude that the benefit of intermediate layers is not modality-dependent but rather a byproduct of pretraining, with the autoregression as the driving factor.

Motivation. We notice that the vision experiments performed by Skean et al. (2025) are limited to ImageNet (Deng et al., 2009), which is included in the pretraining set of the models studied (Bao et al., 2022; Dosovitskiy et al., 2021; He et al., 2022; Oquab et al., 2024). We go beyond this in-distribution (ID) setting and perform linear probing on a diverse set of out-of-distribution (OOD) downstream data. As a sanity check, we conduct the same experiment in the ID scenario by finetuning

[1] In what follows, we use the terms module and component interchangeably, always referring to normalization layers, multi-head attention modules, residual connections, and feedforward layers.

the pretrained model on each dataset, respectively. The training protocol follows Dosovitskiy et al. (2021), see Appendix A for details.

Results. In Fig. 1, the linear probing performance across layers is displayed for the pretrained model (solid line) and for the finetuned model (dashed line). For each layer, probing is done on **RC2**, the output of the transformer block. As we have no information on the degree of distribution shift between pretraining and downstream data, we hypothesize that the stronger the shift, the larger the performance gap between frozen and finetuned encoders. We sort the plots in decreasing order of finetuning performance (see Table 3) from left to right: Flowers102, Cifar10, Contrast, and Speckle Noise. Our findings are twofold: **(1)** in the OOD scenario, we observe that the deeper representations become worse as the shift increases from left to right; **(2)** conversely, in the ID setting, the best visual embedding is at the end of the network. Similar patterns can be observed on all datasets in Fig. 5.

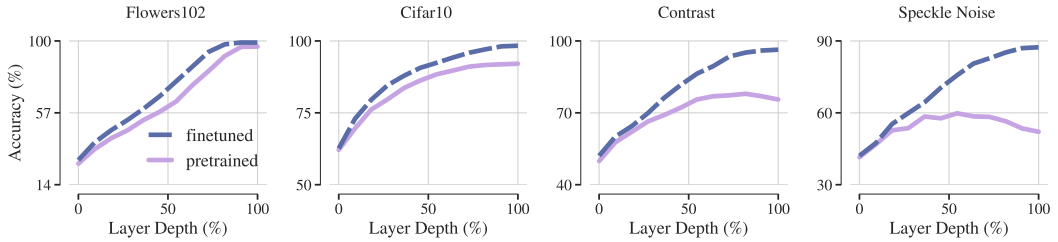

Figure 1: **Layer by layer.** Evolution of the linear probing performance across each layer of an 86M ViT pretrained on ImageNet (the x-axis is the depth percentage of the layer). The solid line denotes the model only pretrained, and the dashed line denotes the model finetuned on the dataset at hand. From **left** to **right**, the shift between the pretraining and the downstream data increases. The stronger the shift, the worse the final layers.

Distribution shift as the culprit. From Fig. 1, we conclude that the intermediate layers are more robust to distribution shifts than final layers. This can be intuitively understood by the fact that layers tend to specialize closer to the classification head. Our findings show that the benefit of intermediate representations is not merely a byproduct of the pretraining objective, as suggested by Skean et al. (2025), but also a consequence of the eventual presence of distribution shift. As such, when finetuning is prohibitive, being able to identify whether the setting is ID or OOD is crucial to know which accuracy profile to expect (from left to right in Fig. 1) and which layer to probe.

4 Not all transformer modules are worth probing on OOD data

In the previous section, we observed that the linear probing performance of the last transformer layers tends to degrade under distribution shifts. Inspired by Odonnat et al. (2026), where the authors demonstrated that the transformer modules do not adapt to downstream data equally, we investigate whether probing after a specific module within a transformer layer has an impact on the performance. We conduct a similar but finer-grained analysis to that in Section 3 on a broader set of downstream data across **both layers and modules** of the 86M ViT pretrained on ImageNet (Deng et al., 2009).

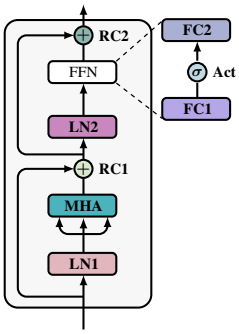

Figure 2: **Transformer block.**

Table 1: **Module by module.** For each module, we report the best linear probing accuracy over the layers. The best performance per dataset is in **bold** and the module with the highest win rate is in gray.

Dataset	LN1	MHA	RC1	LN2	FC1	Act	FC2	RC2
Cifar10	91.94	91.98	92.19	92.20	**92.28**	85.30	89.98	92.07
Cifar100	69.39	67.27	69.88	**69.97**	68.98	69.75	60.92	69.63
Contrast	78.05	74.55	78.30	79.05	78.15	**80.20**	70.85	77.95
Gaussian Noise	57.65	60.40	59.05	58.10	58.75	**61.85**	56.70	58.20
Motion Blur	66.85	64.30	68.50	67.75	66.80	**71.15**	60.90	67.75
Snow	67.30	66.15	68.10	67.30	67.35	**69.30**	61.50	67.70
Speckle Noise	59.75	61.85	60.25	59.95	59.90	**63.35**	58.00	59.80
Clipart	47.66	43.82	48.66	48.37	45.74	**49.34**	40.97	48.33
Sketch	32.36	30.95	33.32	33.07	31.11	**34.90**	28.45	32.99
Flowers102	96.58	96.34	96.58	**96.62**	96.44	91.64	95.23	**96.62**
Pet	88.36	88.33	89.48	**89.51**	88.47	83.46	85.80	89.18

Results. In Table 1, we report for each dataset-module pair the best linear probing accuracy achieved across the layers. Our findings are threefold: **(1)** the standard probing on transformer block outputs (**RC2**) is suboptimal on all datasets but one, Flowers102, for which most components perform equally well; **(2) FC2** is the worst module to probe, with the lowest accuracy on 10 out of 12 datasets, while **Act** is the best-performing one overall, with the highest win-rate over all datasets. We note that it outperforms other components by a large margin when the shift is strong, despite being less good on easier datasets such as Cifar10, Flowers102, or Pet. **(3)** Other modules yield comparable results, with **LN2** being slightly above. We further study these modules along the depth in Fig. 3. Akin to Fig. 1, the plots are ordered in terms of increasing distribution shift between pretraining and downstream data: Flowers102, Cifar10, Contrast, and Speckle Noise. We can see that **FC2** consistently yields the worst performance. When the shift increases, **Act** is the best module in intermediate layers while its performance in final layers plunges. On the contrary, **LN2** and **RC2** yield subpar but more stable results. Similar patterns can be observed on all datasets in Fig. 6.

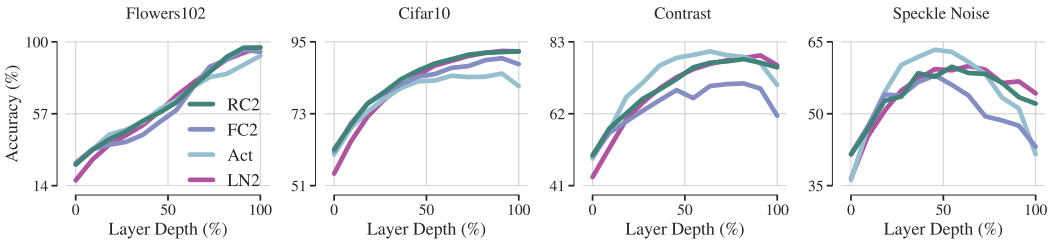

Figure 3: **Layer by layer, module by module.** Evolution of the linear probing performance of transformer modules across the layers of an 86M ViT pretrained on ImageNet. From **left** to **right**, the shift between the pretraining and the downstream data increases.

Signal propagation. We found in Section 3, intermediate representations are more robust to distribution shifts than final layers. As the residual stream of each encoder bears information from the previous layers, this could explain why the accuracy profiles of **LN2** and **RC2** are less concave than **Act** and **FC2**. Furthermore, all modules are maps from $(\mathbb{R}^d)^n$ to $(\mathbb{R}^d)^n$, except the feedforward network, where **FC1** increases the dimension of tokens to $4d$ and **FC2** decreases back to d. We hypothesize that by operating in a higher dimension, **FC1** and **Act** help promote feature disentanglement, which would benefit the probing. Since **Act** filters the potential noise induced by the projection, it may explain its higher accuracy. Conversely, **FC2** compresses the input, which may impact the linear separability of data. Another interesting perspective comes from seeing feedforward networks as key-value memory (Geva et al., 2021). The observed behavior can then be understood by the fact that **FC1** and **Act** capture semantic information in the inputs, which can be useful for linear probing, while **FC2** merely reflects a distribution over tokens. Our findings motivate further study of the hidden representations of transformers. A promising direction would be to extend the analysis of Skean et al. (2025) with information-theoretic, geometric, and invariance measures at the level of transformer modules. A key takeaway for practitioners is that probing after the activation might lead to the best performance, provided the correct layer is chosen. A safer approach, if the shift is difficult to detect, is to probe the **LN2** module, rather than the standard choice of **RC2**.

5 Discussion

We study linear probing on pretrained vision transformers across both layers and modules on a diverse set of classification benchmarks. We find that the discrepancy between pretraining and downstream data is at fault for the degradation of the final layers' performance, while intermediate representations are more robust. We further notice that the standard choice of probing the outputs of transformer blocks is not optimal. In comparison, the hidden representations after the feedforward activation in intermediate layers are the richest under significant distribution shift, whereas the outputs of the LayerNorm preceding the feedforward network are better when the shift is almost negligible. Our work provides a novel perspective on the role of vision transformer hidden representations. We hope it will help guide efficient methods towards detecting distribution shifts and identifying the layers and modules to probe.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=p-BhZSz59o4.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=INqBOmwIpG.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, and Ori Ram et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. IEEE, 2009. URL https://ieeexplore.ieee.org/abstract/document/5206848/.

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. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, and Flood Sung et al. Kimi-vl technical report, 2025. URL https://arxiv.org/abs/2504.07491.

Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Ángel Bautista, Vaishaal Shankar, Alexander T Toshev, Joshua M. Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 12371–12384, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Sam Houliston, Ambroise Odonnat, Charles Arnal, and Vivien Cabannes. Provable benefits of in-tool learning for large language models, 2025.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

HuggingFace. Transformers. https://github.com/huggingface/transformers, 2025. Accessed: 2025-09-21.

Romain Ilbert, Ambroise Odonnat, Vasilii Feofanov, Aladin Virmaux, Giuseppe Paolo, Themis Palpanas, and Ievgen Redko. SAMformer: Unlocking the potential of transformers in time series forecasting with sharpness-aware minimization and channel-wise attention. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=8kLzL5QBh2.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V*, pp. 491–507, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58557-0. doi: 10.1007/978-3-030-58558-7_29. URL https://doi.org/10.1007/978-3-030-58558-7_29.

Alex Krizhevsky. Learning multiple layers of features from tiny images. In *Technical report*, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=APuPRxjHvZ.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008. URL https://api.semanticscholar.org/CorpusID:15193013.

Ambroise Odonnat, Laetitia Chapel, Romain Tavenard, and Ievgen Redko. Vision transformer finetuning benefits from non-smooth components, 2026. URL https://arxiv.org/abs/2602.06883.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. URL https://api.semanticscholar.org/CorpusID:279027499.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009. ISBN 0262170051.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Yacmpz84TH.

Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures, 2025. URL https://arxiv.org/abs/2507.09404.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=WGXb7UdvTX.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Arnas Uselis and Seong Joon Oh. Intermediate layer classifiers for OOD generalization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ByCV9xWfNK.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Renchunzi Xie, Ambroise Odonnat, Vasilii Feofanov, Weijian Deng, Jianfeng Zhang, and Bo An. Mano: Exploiting matrix norm for unsupervised accuracy estimation under distribution shifts. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=mH1xtt2bJE.

Renchunzi Xie, Ambroise Odonnat, Vasilii Feofanov, Ievgen Redko, Jianfeng Zhang, and Bo An. Leveraging gradients for unsupervised accuracy estimation under distribution shift. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=FIWHRSuoos.

Bingchen Zhao, Haoqin Tu, Chen Wei, Jieru Mei, and Cihang Xie. Tuning layernorm in attention: Towards efficient multi-modal LLM finetuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YR3ETaElNK.

Appendix

A Implementation details

Vision transformers. In vision transformers (ViT, Dosovitskiy et al., 2021), input 2D images are split into square patches of size P, which are then flattened and linearly embedded into dimension d. A classification token (`CLS`) is prepended to the sequence of patch tokens before positional embeddings are added. The obtained sequence $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ is fed through a succession of transformer layers (Vaswani et al., 2017), where the output representation of the `CLS` token serves as the final encoder output. In our code, we follow the original ViT implementation from Dosovitskiy et al. (2021) and use a convolutional layer to embed images (see Dosovitskiy et al., 2021, §"Hybrid Architecture"). This is also the standard in the implementation from HuggingFace (2025). In Fig. 4, we display the implementation of the ViT-Base model with a classification head for 10 classes (we renamed our package "my_lib" to respect the anonymity).

```
# Python snippet to print the ViT architecture
from my_lib import ViT

model = ViT(name="base", n_classes=10)
print(model)

# Corresponding output
Transformer(
  (embedding): Embedding(
    (patching): PatchImages(
      (patching): Sequential(
        (0): Conv2d(3, 768, kernel_size=(16, 16), stride=(16, 16))
        (1): Flatten(start_dim=2, end_dim=-1)
      )
    )
  )
  (blocks): ModuleList(
    (0-11): 12 x TransformerBlock(
      (attn_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (attn): SelfAttention(
        (qkv_mat): Linear(in_features=768, out_features=2304, bias=True)
        (output): Linear(in_features=768, out_features=768, bias=True)
      )
      (ffn_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (ffn): FeedForward(
        (fc1): Linear(in_features=768, out_features=3072, bias=True)
        (fc2): Linear(in_features=3072, out_features=768, bias=True)
      )
    )
  )
  (output): Output(
    (output_layer): ClassificationLayer(
      (output_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (output): Linear(in_features=768, out_features=10, bias=True)
    )
  )
)
```

Figure 4: ViT-Base Implementation.

Data preprocessing. All our experiments are conducted on a varied collection of 11 classification benchmarks: Cifar10, Cifar100 (Krizhevsky, 2009); variants from Cifar10-C (Hendrycks & Dietterich, 2019) with severity 5: Contrast, Gaussian Noise, Motion Blur, Snow, Speckle Noise; 2 domains from

DomainNet (Peng et al., 2019), a challenging benchmark typically used for domain generalization: Clipart, Sketch; Flowers102 (Nilsback & Zisserman, 2008) and Pets (Parkhi et al., 2012). The preprocessing follows Dosovitskiy et al. (2021) and Kolesnikov et al. (2020): for training data, we apply random cropping, a 224×224 image resizing, and random horizontal flip for training images. For validation and test data, the 224×224 image resizing is applied before center cropping images. All images are normalized using the ImageNet (Deng et al., 2009) statistics. It ensures images with mean $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$. For datasets that do not have predefined training and test sets (i.e., datasets from Cifar10-C and DomainNet), we manually create *deterministic* training and test sets following a $80\% - 20\%$ split. The deterministic part is crucial to ensure no data contamination.

Finetuning setup. Our finetuning experiments follow the protocol from Dosovitskiy et al. (2021) with a resolution of 224×224. We optimize models with the Stochastic Gradient Descent (SGD), a momentum of 0.9, no weight decay, a cosine learning rate decay, a batch size of 512, and gradient clipping at norm 1. The finetuning resolution is of 224. For each pair of dataset - configuration, we perform a sweep over 4 learning rates, as summarized in Table 2, and conduct 3 runs with different seeds relative to network initialization and dataloaders. For each run, we monitor the training using a validation set (20% of the training set). The final performance is the test accuracy of the checkpoint that achieves the best validation accuracy.

Table 2: **Finetuning hyperparameters**. We report the choice of optimizer, batch size, training steps, and learning rates.

dataset	optimizer	batch size	training steps	learning rates η
Cifar10	SGD	512	10000	{1e−3, 3e−3, 1e−2, 3e−2}
Cifar100	SGD	512	10000	{1e−3, 3e−3, 1e−2, 3e−2}
Contrast	SGD	512	10000	{1e−3, 3e−3, 1e−2, 3e−2}
Gaussian Noise	SGD	512	10000	{1e−3, 3e−3, 1e−2, 3e−2}
Motion Blur	SGD	512	10000	{1e−3, 3e−3, 1e−2, 3e−2}
Snow	SGD	512	10000	{1e−3, 3e−3, 1e−2, 3e−2}
Speckle Noise	SGD	512	10000	{1e−3, 3e−3, 1e−2, 3e−2}
Clipart	SGD	512	20000	{3e−3, 1e−2, 3e−2, 6e−2}
Sketch	SGD	512	20000	{3e−3, 1e−2, 3e−2, 6e−2}
Flowers102	SGD	512	5000	{1e−3, 3e−3, 1e−2, 3e−2}
Pets	SGD	512	4000	{1e−3, 3e−3, 1e−2, 3e−2}

Table 3: **Full finetuning results.** We report the best top-1 accuracy (%) on the test set over the learning rate grid of each dataset (↑). Each entry shows the mean and standard deviation over three finetuning runs with different seeds.

Dataset	Cifar10	Cifar100	Contrast	Gaussian Noise	Motion Blur	Snow	Speckle Noise	Clipart	Sketch	Flowers102	Pets
full finetuning	$99.02_{\pm 0.02}$	$92.74_{\pm 0.05}$	$97.23_{\pm 0.18}$	$87.14_{\pm 1.16}$	$94.67_{\pm 0.14}$	$95.42_{\pm 0.13}$	$89.58_{\pm 0.43}$	$78.50_{\pm 0.49}$	$71.30_{\pm 0.26}$	$99.15_{\pm 0.05}$	$94.57_{\pm 0.29}$

B ADDITIONAL EXPERIMENTS

We display in Figs. 5 and 6 the additional results on all benchmarks related to the experiments of Sections 3 and 4, respectively.

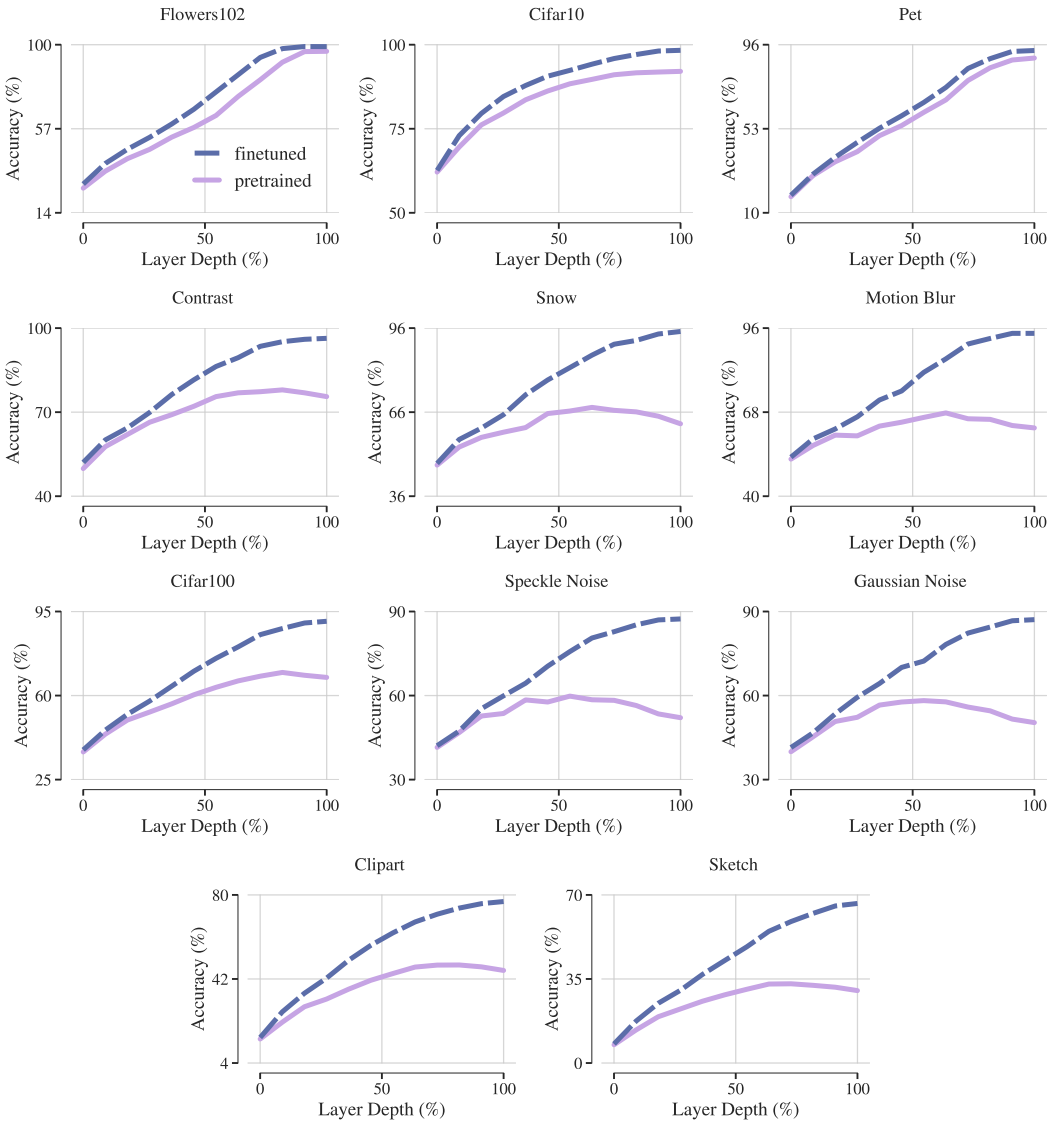

Figure 5: **Layer by layer.** Evolution of the linear probing performance across the layers of an 86M ViT pretrained on ImageNet. The solid line denotes the model only pretrained, and the dashed line denotes the model finetuned on the dataset at hand. From **left** to **right**, the shift between the pretraining and the downstream data increases. The stronger the shift, the worse the final layers.

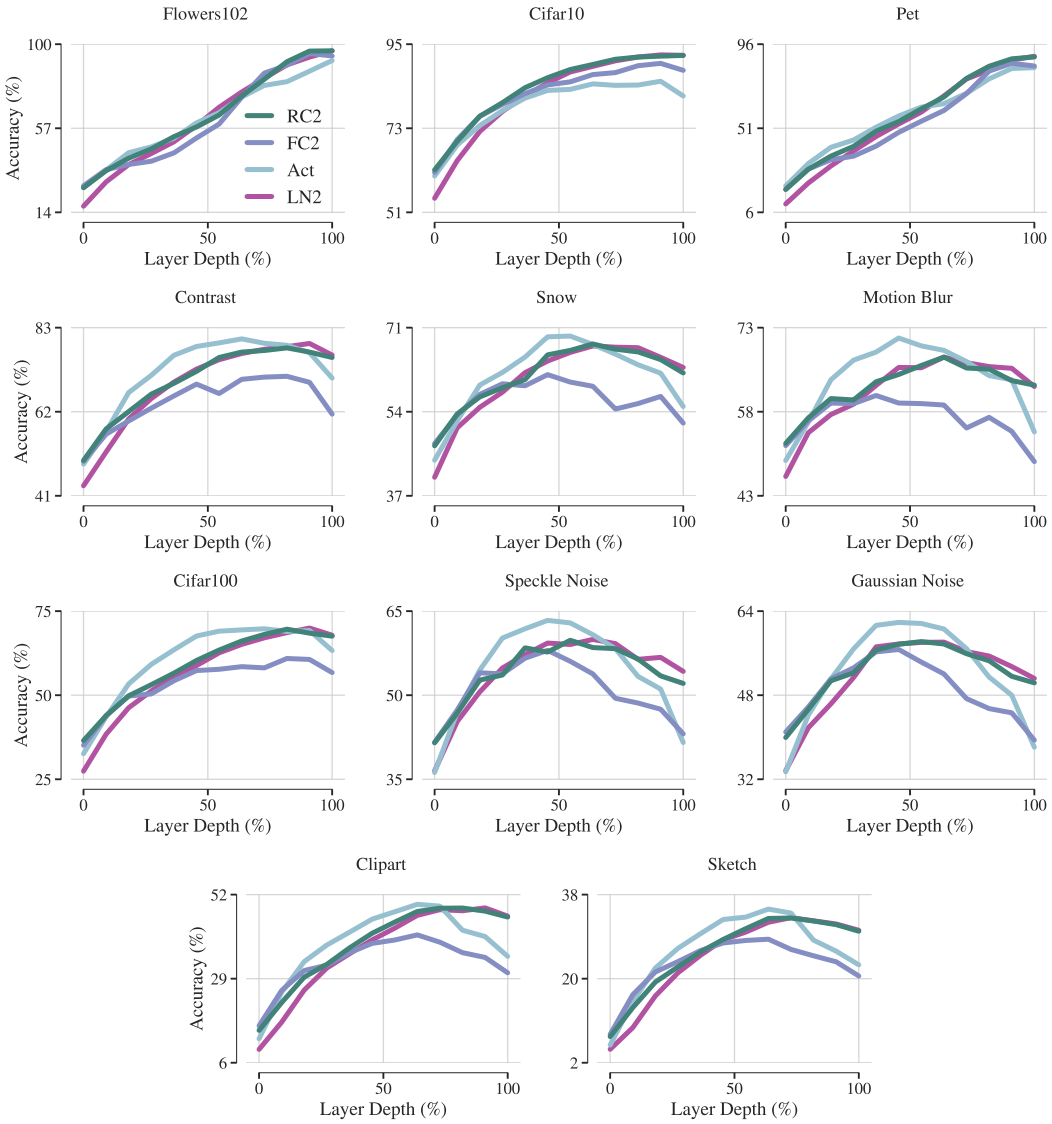

Figure 6: **Layer by layer, module by module.** Evolution of the linear probing performance of transformer modules across the layers of an 86M ViT pretrained on ImageNet. From **left** to **right**, the shift between the pretraining and the downstream data increases.