Probing the Equivariance of Image Embeddings

Cyrus Rashtchian[*]
Google Research
cyroid@google.com

Charles Herrmann[*]
Google Research
irwinherrmann@google.com

Chun-Sung Ferng[*]
Google Research
csferng@google.com

Ayan Chakrabarti
Google Research
ayanchakrab@google.com

Dilip Krishnan
Google Research
dilipkay@google.com

Deqing Sun
Google Research
deqingsun@google.com

Da-Cheng Juan
Google Research
dacheng@google.com

Andrew Tomkins
Google Research
tomkins@google.com

Abstract

Probes are small networks that predict properties of underlying data from embeddings, and they provide a targeted way to illuminate the information in embeddings. While analysis with probes has become standard in NLP, there has been less exploration in vision. Our goal is to understand the invariance vs. equivariance of popular image embeddings (e.g., MAE, SimCLR, or CLIP) under certain distribution shifts. By doing so, we investigate what visual aspects from the raw images are encoded into the embeddings by these foundation models. Our probing is based on a systematic transformation prediction task that measures the visual content of embeddings along many axes, including neural style transfer, recoloring, icon/text overlays, noising, and blurring. Surprisingly, six embeddings (including SimCLR) encode enough non-semantic information to identify dozens of transformations. We also consider a generalization task, where we group similar transformations and hold out several for testing. Image-text models (CLIP, ALIGN) are better at recognizing new examples of style transfer than masking-based models (CAN, MAE). Our results show that embeddings from foundation models are equivariant and encode more non-semantic features than a supervised baseline. Hence, their OOD generalization abilities are not due to invariance to such distribution shifts.

1 Introduction

Large pre-trained networks, sometimes known as *foundation models*, provide a 'general-purpose' embedding for multiple data modalities (Bommasani et al., 2021; Zhou et al., 2023). The models perform very well on several downstream tasks and exhibit robustness to dataset shift. In large ML systems, raw data is often pre-processed using embeddings, and training a small network on top of a frozen embedding is a scalable and desirable solution. A central research direction is to develop foundation models that are easily adapted for current and future applications. Hence, it is important to be able to evaluate what information these embeddings capture and what aspects they ignore.

Probes provide a way to determine what information can be extracted about data after computing an embedding. The idea is to train a small network (a.k.a., probe) to predict certain properties about the underlying data using only an embedding of the data (Alain & Bengio, 2017). In addition to helping

[*]Equal contribution.

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

explain embedding models, probes also inform whether embeddings can be used for downstream tasks that rely on certain information about the data. In NLP, researchers have probed text embeddings for information about syntax trees, sentence length, word order, and so on (Belinkov, 2021; Conneau et al., 2018; Hewitt & Manning, 2019; Li et al., 2022). For image-text models, it is also possible to probe the visual information through text prompts or questions that describe visual attributes, such as color, shape, and material (Liu et al., 2022; Paik et al., 2021; Zhang et al., 2022).

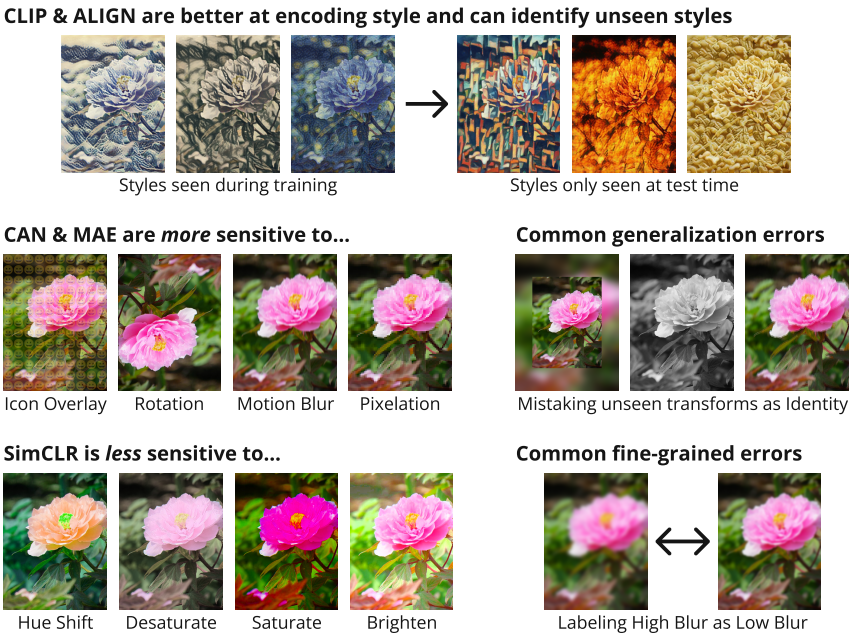


Figure 1: Main takeaways from analyzing the performance of frozen image embeddings on our transformation prediction tasks. We draw conclusions about (in)sensitivity by looking the accuracy when detecting whether a particular transformation has been applied to an image. We evaluate both a fine-grained version (31 classes, same train and test transforms) and a generalization version (10 classes, with 28 training and 15 additional test transforms grouped into categories). Both versions contain the 'Identity' class as the original image.

Unfortunately, probes using text prompts are limited to concepts describable in short text snippets. This overlooks visual attributes that are important for certain tasks. For example, a core part of trust and safety involves filtering content for policy violations, which is challenging when adversaries manipulate images to evade filters (Cao et al., 2022; Stimberg et al., 2023; Yuan et al., 2019). These manipulations include noise, recoloring, overlays, and neural style transfer based on a reference image. One solution is to develop filtering models that are highly invariant. However, it is also critical to produce manipulation *detectors*, to identify malicious users who upload such images. Unlike semantic classification, the new task of manipulation detection must instead use embeddings that encode information about the transformations that we wish to detect.

Can a single embedding be robust against dataset shift, and yet, sensitive to many transformations? And even if it is possible, are today's popular pre-trained embeddings successful at doing so? These are the questions we study in this paper. Through image probing experiments, we evaluate the equivariance of several popular embeddings to dozens of transformations (see Figure 1 for examples) and offer insights into the information encoded by image-only and image-text foundation models.

1.1 Visual probing using image transformations

Our new approach to visual probing centers around predicting how an image has been transformed. The core idea of the experiment is to modify an image and then see if this change is detectable after computing the image's embedding. For example, consider two images: one that is a sample from ImageNet, and another where the same image has been slightly blurred. Then, compute embeddings for both images and throw out the original pixels. Assume that in both cases a linear probe will

predict the correct ImageNet class for the image. The next question is: does the embedding contain enough information to determine which image was blurred and which was unaltered?

If the embedding contains sufficient information to detect blurring, then it should be possible to train a network to perform well on a 'blurry or not' classification task. Specifically, we can apply Gaussian blur to all images in ImageNet, and we can train a probing network to predict whether the transformation has been applied given access only to the image embeddings. Foundation models that capture more of the transformation information will perform better on this task, whereas models that perform poorly must be insensitive to the transformation. Note that freezing the embedding model is crucial for this analysis. If we fine-tuned on the transformation prediction task, then we would not know whether the original model extracted the transformation information or not.

One of our goals is to interpret the information that is kept or lost when using popular vision embeddings. Another goal is to evaluate the effectiveness of embeddings for the task of predicting transformations. Our motivation comes from using embeddings for tasks such as detecting manipulation, predicting style, or data cleaning. We carefully design the set of transformations, ensuring enough variety to elicit whether embeddings capture different types of visual content. For example, we include image filtering (e.g, hue shift, saturate), occlusion (e.g., icon or text overlay), corruption (e.g., noise, pixelate), natural domain shift (e.g., motion blur, brighten, crop), and neural style transfer. Probing visual embeddings complements the prior work on training models to be more invariant or equivariant (Dangovski et al., 2021; Dubois et al., 2022; Von Kügelgen et al., 2021; Xiao et al., 2020).

Using our set of transformations, we propose two transformation prediction tasks. The first is a *fine-grained* variation. Here, the probe sees all 30 transformations during training and the goal is to classify them. We design the fine-grained transformations so that we expect trained humans to achieve 100% accuracy. Any errors from the probe indicate a lack of information in the embedding.

Our second task focuses on *generalization*. We group the transformations into 9 categories, and we only train the probe using some in each category. During test time, the probe should recognize the category of the transformation. This task measures two things: (1) whether an embedding organizes transformation information in a generalizable way, and (2) whether we can use the embedding for more realistic prediction with held-out transformations. Specifically, for (2), we may want to detect domain shifts, such as image manipulations or natural perturbations. However, we cannot train with all relevant transformations. Instead, we desire a model that can correctly classify unseen, but similar, transformations. For example, we have a category based on neural style transfer, but the probe only sees some styles during training. Nonetheless, transferring styles from other references produces visually analogous images (e.g., Figure 1, first row). Our generalization task evaluates whether the features in the embedding suffice to also detect new examples in the same transformation category.

We evaluate a representative sample of vision foundation models based on their pre-training algorithms: (1) a masked autoencoder (MAE) (He et al., 2022), a canonical masking-based method that fills image portions during pre-training, (2) SimCLR (Chen et al., 2020), an image-only contrastive loss that encourages invariance to a few transformations, (3) CAN (Mishra et al., 2022), which combines masking, contrastive, and noise prediction, (4) CLIP (Radford et al., 2021), a standard image-text self-supervised method, (5) ALIGN (Jia et al., 2021), another contrastive image-text model, and (6) a supervised method that trains with the ImageNet-1k semantic class labels.

2 Probing Embeddings by Predicting Transformations

Evaluating only the typical semantic accuracy on class labels leaves open questions regarding what information from the raw data is retained or lost in the embedding. Instead, we probe the embeddings, measuring the ability of a network to determine how an image has been transformed. Assume we have T image transformations, such as style transfer, recoloring, overlays, noising, or blurring. Here, for *transformation*, we take a broad definition. One option is a well-defined function, such as adding Gaussian noise independently to each pixel. Another possibility is to have random parameters, such as uniformly choosing a value and increasing the image's saturation by this much. Finally, we can have transformation families, containing several *sub-transformations*. For example, the "color quantizing" transformation contains sub-transformations that modify hue, invert colors, or solarize the image. We apply the T transformations to images in the train/test sets. This generates $T + 1$ copies of the dataset, including the original images. This process defines a $(T + 1)$-way classification problem, labeling each image either with 'Identity' or one of the T transformations.

Table 1: Transformation prediction accuracies for six embeddings on our transformed version of ImageNet-1k. Fine-grained has 31 classes (30 transforms), and generalization has 10 classes (28 training and 15 held-out test transforms). MLP, one hidden layer, width 2048. Averaged over 5 runs, all std. dev. below 0.19. Right columns present transformation accuracies on the subset of test data with unseen sub-transformations for two categories of the generalization task, noise and style transfer.

EMBEDDING	FINE-GRAINED	GENERALIZATION	HELD-OUT NOISE	HELD-OUT STYLE
CAN	**98.27**	88.12	86.33	49.29
MAE	97.67	86.79	**94.67**	28.92
SIMCLR	93.05	87.32	59.55	54.27
CLIP	96.45	**90.99**	76.58	**86.24**
ALIGN	96.66	89.22	85.69	69.61
SUPERVISED	94.12	79.11	61.06	41.90

2.1 Experimental set-up

Datasets. We evaluate with transformed versions of ImageNet-1k (Russakovsky et al., 2015). In addition to the original image (Identity), we apply 30 transformations to each train/test image. This leads to 31 classes for the fine-grained transformation prediction task. We also construct a generalization dataset with 10 categories, where each category contains one or more transformations along with a range of parameters (e.g., noise level or type of style transfer). The test set transformations form a superset of those applied to the training images.

Metrics. The *transformation prediction accuracy* is the fraction of images receiving the correct transformation label. In the fine-grained case, the model predicts one of 31 transformation classes; in the generalization case, it predicts one of 10. For both cases, we average over a test set with size being (# classes) times (# original images), i.e., (# classes) $\times$ 50k for ImageNet-1k.

Embeddings. While it is hard to control for all aspects of the foundation models, we enable a fair analysis for three image-only embeddings. We train the CAN, MAE, and SimCLR algorithms all on JFT-300M (Sun et al., 2017); they output a 1024-dim. embedding from a Vision Transformer (ViT) L/16. Compared to CLIP/ALIGN, the number of training images is also similar (300M vs 400M). The SimCLR model also contains a projection to a 128-dim. embedding that we use for one comparison. CLIP uses ViT L/14 for a 768-dim. image embedding. ALIGN uses EfficientNet-L2 for the image encoder and outputs a 1376-dim. embedding. We were given access to the ALIGN weights. Our baseline is a 1024-dim. embedding from a supervised ViT L/16 trained on ImageNet-1k.

2.2 What have we learned about embeddings?

We compare transformation prediction for six embeddings in Table 1. All embeddings perform well on this probing task: over 93% accuracy for the fine-grained task and over 79% accuracy for the generalization task. These embeddings preserve fairly detailed information about the input image that can be extracted with minimal post-processing (2-layer MLP). In the **fine-grained task** (a test of which embedding has the most detailed information about the image), the CAN probe performs the best with MAE being a close second. In the **generalization task** (a test of how well an embedding's information about transformations can generalize), the two text-image embeddings (CLIP and ALIGN) perform better than all other methods.

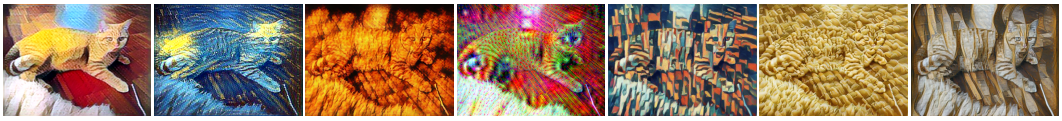
Robustness to domain shifts is not due to invariance. By analyzing transformation prediction, we have concluded that several embeddings are equivariant along many axes. Figure 1 has summarized these insights, which help inform a choice between competing models. All of the models capture a lot of transformation information, which is quite unexpected given their robustness to OOD data like ImageNet variants. Thus, a lack of invariance does *not* imply poor generalization on OOD data.

Modern embeddings suffice to classify maliciously transformed data. Prior work shows that classifiers are susceptible to transformation-based attacks, such as style transfer, Gaussian noise, or recoloring (Cao et al., 2022; Goodman & Wei, 2019; Hao et al., 2021; Hosseini et al., 2017; Li et al., 2019; Yuan et al., 2019). Classifying malicious transformations is an important direction for content safety, beyond OOD and anomaly detection (Salehi et al., 2021). Our work shows that we can solve this task using MLPs on top of pre-computed embeddings instead of custom pixel-based classifiers.

2.3 Generalization analysis: held-out styles

Table 2 zooms in the on the style transfer accuracies, showing the fraction of correct prediction for each of the thirteen styles that are displayed in Figure 2. All styles have the 'Style Transfer' label for the generalization task. When the styles are seen in the train set, then the validation accuracy is nearly perfect and often 100% for these six styles. For the held-out styles, we see that the models may struggle to recognize the unseen types of style transfer. Overall, CLIP performs the best, sometimes by a large margin. Interestingly, SimCLR also performs well, often better than CAN, MAE, and Supervised. In several cases, the models perform worse than chance level (below 10%).

(a) Identity (b) Art. Mosaic (c) Art. Udnie (d) Great Wave (e) The Scream (f) Starry Night (g) Taligas1920



(h) Art. Candy (i) Art. S. Night (j) Bonfire (k) DeepDream (l) Black Figure (m) Pasta (n) Violon

Figure 2: Style sub-transformations from the 'Style Transfer' category of the generalization dataset. We also include the 'Identity' as the original image for reference. The six styles in the top row are in both the train and test sets, while the bottom seven styles only appear in the test set. All styles have the 'Style Transfer' label. Table 2 has accuracies for the different embeddings for each of these styles.

Table 2: Accuracies for the style transfer sub-transformations in the generalization dataset. We report the fraction of style-transferred images for which the model predicts the 'Style Transfer' label correctly (out of 10 possible labels). The top six styles are in the train and test sets; bottom seven only appear in the test set. Numbers in red are below chance level (< 10% correct).

STYLE	CAN	MAE	SIMCLR	CLIP	ALIGN	SUPERVISED
ARTISTIC MOSAIC	**100**	**100**	**100**	**100**	**100**	**100**
ARTISTIC UDNIE	**100**	**100**	**100**	**100**	99.99	99.99
GREAT WAVE OFF KANAGAWA	**100**	**100**	**100**	**100**	**100**	**100**
THE SCREAM	**100**	**100**	99.97	99.97	99.95	99.97
STARRY NIGHT	99.98	**100**	99.98	99.98	99.98	99.97
TALIGAS 1920	99.99	**100**	99.99	99.99	99.99	99.99
ARTISTIC STARRY NIGHT	53.66	47.13	5.91	**91.33**	70.91	90.75
ARTISTIC CANDY	55.09	37.09	79.00	**95.09**	77.48	6.47
BONFIRE	46.33	0.00	2.33	**92.12**	69.24	9.93
DEEP DREAM	3.13	0.12	26.34	**52.10**	11.84	3.40
LANDSCAPE BLACK FIGURE	99.92	98.94	99.75	**99.99**	99.97	79.43
PASTA	3.73	0.46	70.07	**73.39**	58.87	39.84
VIOLON	83.15	18.67	96.46	**99.69**	98.98	63.49

3 Conclusion

We investigated a new probing task to shed new light on image embeddings. We showed that popular models capture enough information to distinguish dozens of transformations. Our experiments uncovered ways in which SimCLR is more invariant than CAN and MAE, and the types of transformations that are captured by self-supervised vision models vs. image-text models, such as CLIP and ALIGN. We also found that the self-supervised models perform better than a supervised baseline, suggesting that optimizing an embedding directly for semantic information (i.e., ImageNet-1k classes) does not by default retain as much transformation information. For more experiments and further discussion, see the full version of our paper at `https://arxiv.org/abs/2307.05610`.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. In *ICLR*, 2017.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and alternatives. *arXiv preprint arXiv:2102.12452*, 2021.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., and Brunskill, E. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Cao, Y., Xiao, X., Sun, R., Wang, D., Xue, M., and Wen, S. Stylefool: Fooling video classification systems via style transfer. *arXiv preprint arXiv:2203.16000*, 2022.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. What you can cram into a single &!#* vector: Probing sentence embeddings for linguistic properties. In *ACL*, volume 1, pp. 2126–2136, 2018.

Dangovski, R., Jing, L., Loh, C., Han, S., Srivastava, A., Cheung, B., Agrawal, P., and Soljacic, M. Equivariant self-supervised learning: Encouraging equivariance in representations. In *International Conference on Learning Representations*, 2021.

Dubois, Y., Ermon, S., Hashimoto, T., and Liang, P. Improving self-supervised learning by characterizing idealized representations. In *NeurIPS*, 2022.

Goodman, D. and Wei, T. Cloud-based image classification service is not robust to simple transformations: A forgotten battlefield. *arXiv preprint arXiv:1906.07997*, 2019.

Hao, Q., Luo, L., Jan, S. T., and Wang, G. It's not what it looks like: Manipulating perceptual hashing based applications. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 69–85, 2021.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, 2019.

Hosseini, H., Xiao, B., and Poovendran, R. Google's cloud vision api is not robust to noise. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pp. 101–105. IEEE, 2017.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

Li, X., Ji, S., Han, M., Ji, J., Ren, Z., Liu, Y., and Wu, C. Adversarial examples versus cloud-based detectors: A black-box empirical study. *IEEE Transactions on Dependable and Secure Computing*, 18(4):1933–1949, 2019.

Liu, X., Yin, D., Feng, Y., and Zhao, D. Things not written in text: Exploring spatial commonsense from visual signals. *arXiv preprint arXiv:2203.08075*, 2022.

Mishra, S., Robinson, J., Chang, H., Jacobs, D., Sarna, A., Maschinot, A., and Krishnan, D. A simple, efficient and scalable contrastive masked autoencoder for learning visual representations. *arXiv preprint arXiv:2210.16870*, 2022.

Paik, C., Aroca-Ouellette, S., Roncone, A., and Kann, K. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In *EMNLP*, pp. 823–835, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.63. URL https://aclanthology.org/2021.emnlp-main.63.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., and Clark, J. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M. H., and Sabokrou, M. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021.

Stimberg, F., Chakrabarti, A., Lu, C.-T., Hazimeh, H., Stretcu, O., Qiao, W., Liu, Y., Kaya, M., Rashtchian, C., Fuxman, A., et al. Benchmarking robustness to adversarial image obfuscations. *arXiv preprint arXiv:2301.12993*, 2023.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. *arXiv preprint arXiv:1707.02968*, 2017.

Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Xiao, T., Wang, X., Efros, A. A., and Darrell, T. What Should Not Be Contrastive in Contrastive Learning. *arXiv preprint arXiv:2008.05659*, 2020.

Yuan, K., Tang, D., Liao, X., Wang, X., Feng, X., Chen, Y., Sun, M., Lu, H., and Zhang, K. Stealthy porn: Understanding real-world adversarial images for illicit online promotion. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 952–966, 2019. doi: 10.1109/SP.2019.00032.

Zhang, C., Van Durme, B., Li, Z., and Stengel-Eskin, E. Visual commonsense in pretrained unimodal and multimodal models. In *NAACL*, pp. 5321–5335, 2022.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023.