Sparse Mixture-of-Experts are Domain Generalizable Learners

Bo Li[1]* **Yifei Shen**[2]* **Jingkang Yang**[1] **Yezhen Wang**[3] **Jiawei Ren**[1]
Tong Che[3,4] **Jun Zhang**[2] **Ziwei Liu**[1]
[1]S-Lab, Nanyang Technological University
[2]The Hong Kong University of Science and Technology
[3]Mila-Quebec AI Institute [4]Nvidia Research
{libo0013,ziwei.liu}@ntu.edu.sg

Abstract

In domain generalization (DG), most existing methods focused on the loss function design. This paper proposes to explore an orthogonal direction, i.e., the design of the backbone architecture. It is motivated by an empirical finding that transformer-based models trained with empirical risk minimization (ERM) outperform CNN-based models employing state-of-the-art (SOTA) DG algorithms on multiple DG datasets. We develop a formal framework to characterize a network's robustness to distribution shifts by studying its architecture's alignment with the correlations in the dataset. This analysis guides us to propose a novel DG model built upon vision transformers, namely *Generalizable Mixture-of-Experts (GMoE)*. Experiments on DomainBed demonstrate that GMoE trained with ERM outperforms SOTA DG baselines by a large margin.

1 Introduction

Generalizing to out-of-distribution (OOD) data is an innate ability for human vision, but highly challenging for machine learning models [23, 11, 19]. Domain generalization (DG) is one approach to address this problem, which encourages models to be resilient under various distribution shifts such as background, lighting, texture, shape, and geographic/demographic attributes.

From the perspective of representation learning, there are several paradigms towards this goal, including domain alignment [10, 15], invariant causality prediction [1, 17], meta-learning [2, 32], ensemble learning [20, 3], and feature disentanglement [28, 31]. The most popular approach to implementing these ideas is to design a specific loss function (e.g., [10] for domain alignment, [1, 17] for invariant causal prediction, [2, 32] for meta-learning). Recent studies have shown that these approaches improve ERM and achieve promising results on large-scale DG datasets [29].

Meanwhile, in various computer vision tasks, the innovations in backbone architectures play a pivotal role in performance boost and have attracted much attention [14, 16, 18]. Inspired by these empirical successes, we conjecture that *backbone architecture design would be promising for DG*. In this paper, we formally investigate the impact of the backbone architecture on DG and propose to develop effective DG methods by backbone architecture design. Specifically, our main contributions are summarized as follows:

A Novel View of DG: In contrast to previous works, this paper initiates an exploration of the backbone architecture in DG. Based on algorithmic alignment [30], we prove that a network is more robust to distribution shifts if its architecture aligns with the invariant correlation.

*Equal contribution

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

A Novel Model for DG: Based on our theoretical analysis, we propose Generalizable Mixture-of-Experts (GMoE) and prove that it enjoys a better alignment than vision transformers. GMoE is built upon sparse mixture-of-experts [25] and vision transformer [7], with a theory-guided performance enhancement for DG.

2 On the Importance of Neural Architecture for Domain Generalization

In this section, we investigate the impact of the backbone architecture on DG by using an algorithmic alignment framework [30]. We first define the alignment.

Definition 1. *(Alignment; [30]) Let $\mathcal{N}$ denote a neural network with n modules $\{\mathcal{N}_i\}_{i=1}^n$ and assume that a target function for learning $y = g(\boldsymbol{x})$ can be decomposed into n functions $f_1, \cdots, f_n$. The network $\mathcal{N}$ aligns with the target function if replacing $\mathcal{N}_i$ with f_i, it outputs the same value as algorithm g. The alignment value between $\mathcal{N}$ and f is defined as*

$$\textit{Alignment}(\mathcal{N}, f, \epsilon, \delta) := n \cdot \max_i \mathcal{M}(f_i, \mathcal{N}_i, \epsilon, \delta), \tag{1}$$

where $\mathcal{M}(f_i, \mathcal{N}_i, \epsilon, \delta)$ denotes the sample complexity measure for $\mathcal{N}_i$ to learn f_i with ϵ precision at failure probability δ under a learning algorithm when the training distribution is the same as the test distribution.

To have a tractable analysis for nonlinear function approximation, we first make an assumption on the distribution shift.

Assumption 1. *Denote $\mathcal{N}_1$ as the first module of the network (including one or multiple layers) of the network. Let $p_{train,\mathcal{N}_1}(\boldsymbol{s})$ and $p_{test,\mathcal{N}_1}(\boldsymbol{s})$ denote the probability density functions of features after $\mathcal{N}_1$. Assume that the support of the training feature distribution covers that of the test feature distribution, i.e., $\max_{\boldsymbol{s}} \frac{p_{test,\mathcal{N}_1}(\boldsymbol{s})}{p_{train,\mathcal{N}_1}(\boldsymbol{s})} \leq C$, where C is a constant independent of the number of training samples.*

In DG, the target function is an invariant correlation across the training and test datasets. For simplicity, we assume that the labels are noise-free.

Assumption 2. *(Invariant correlation) Assume there exists a function g_c such that for training data, we have $g_c(\mathcal{N}_1(\boldsymbol{x})) = y, \forall x \in \mathcal{E}_{tr}$, and for test data, we have $\mathbb{P}_{D_{te}}[\|g_c(\mathcal{N}_1(\boldsymbol{x})) - y\| \leq \epsilon] > 1 - \delta$.*

The following theorem shows that if the neural architecture aligns with the invariant correlation, ERM is sufficient to achieve a good performance.

Theorem 1. *(Impact of Backbone Architecture in Domain Generalization) Denote $\mathcal{N}' = \{\mathcal{N}_2, \cdots, \mathcal{N}_n\}$. Assuming we train the neural network with ERM, and Assumption 1, 2, hold. If Alignment$(\mathcal{N}', g_c, \epsilon, \delta) \leq |\mathcal{E}_{tr}|$, we have $\mathbb{P}_{D_{te}}[\|\mathcal{N}(\boldsymbol{x}) - y\| \leq O(\epsilon)] > 1 - O(\delta)$.*

ViT with ERM versus ResNet50 with DG algorithms The existing DG methods often adopt ResNet50 as the backbone [13] and we compare ViT trained with ERM with ResNet50 trained with SOTA DG algorithms in Table 1. It is shown that ViT outperforms ResNet50 on three datasets while underperforms ResNet50 on TerraInc. We will use Theorem 1 to explain this experiment. According to the analysis in [21], multi-head attentions (MHA) are low-pass filters with a *shape* bias while convolutions are high-pass filters with a *texture* bias. In VLCS, OfficeHome, and DomainNet, the shape is invariant across some domains (e.g., among natural, sketch, and paint of DomainNet). On the contrary, the object to recognize is often in a narrowed local region in TerraInc, which corresponds to the local bias of CNNs. As a result, a ViT simply trained with ERM can outperform CNNs trained with SOTA DG algorithms on the four datasets while the ViT's performance deteriorates on TerraInc.

To improve ViT's performance, Theorem 1 suggests that we should exploit the properties of invariant correlations. In image recognition, objects are described by functional parts (e.g., visual attributes), with words associated with them [33]. The configuration of the objects has a large degree of freedom, resulting in different shapes among one category. Therefore, functional parts are more fundamental than shape in image recognition and we will develop backbone architectures to capture them in the next section.

3 Generalizable Mixture-of-Experts for Domain Generalization

In this section, we propose Generalizable Mixture-of-Experts (GMoE) for domain generalization, supported by effective neural architecture design and theoretical analysis.

3.1 Mixture-of-Experts Layer

In this subsection, we introduce the mixture-of-experts (MoE) layer, which is an essential component of GMoE. One ViT layer is composed of an MHA and an FFN. In the MoE layer, the FFN is replaced by mixture-of-experts and each expert is implemented by an FFN [25]. Denoting the output of the MHA as $\boldsymbol{x}$, the output of the MoE layer with N experts is given by

$$f_{\text{MoE}}(\boldsymbol{x}) = \sum_{i=1}^{N} G(\boldsymbol{x})_i \cdot E_i(\boldsymbol{x}) = \sum_{i=1}^{N} \text{TOP}_k(\text{Softmax}(\boldsymbol{W}\boldsymbol{x})) \cdot \boldsymbol{W}^2_{\text{FFN}_i} \phi(\boldsymbol{W}^1_{\text{FFN}_i} \boldsymbol{x}), \tag{2}$$

where $\boldsymbol{W}$ is the learnable parameter for the gate, $\boldsymbol{W}^1_{\text{FFN}_i}$ and $\boldsymbol{W}^2_{\text{FFN}_i}$ are learnable parameters for the i-th expert, $\phi(\cdot)$ is a nonlinear activation function, and $\text{TOP}_k(\cdot)$ operation is a one-hot embedding that sets all other elements in the output vector as zero except for the elements with the largest k values where k is a hyperparameter. Given $\boldsymbol{x}_{\text{in}}$ as the input of the MoE layer, the update is given by

$$\boldsymbol{x} = f_{\text{MHA}}(\text{LN}(\boldsymbol{x}_{\text{in}})) + \boldsymbol{x}_{\text{in}}, \quad \boldsymbol{x}_{\text{out}} = f_{\text{MoE}}(\text{LN}(\boldsymbol{x})) + \boldsymbol{x},$$

where f_{MHA} is the MHA layer, LN represents layer normalization, and $\boldsymbol{x}_{\text{out}}$ is the output of the MoE layer.

3.2 Visual Attributes and Sparse MoEs

In real world image data, the label depends on multiple attributes. Capturing *diverse* visual attributes is especially important for DG. For example, the definition of an *elephant* in the Oxford dictionary is "a very large animal with thick grey skin, large ears, two curved outer teeth called tusks, and a long nose called a trunk". The definition involves three shape attributes (i.e., large ears, curved outer teeth, and a long nose) and one texture attribute (i.e., thick grey skin). In the IID ImageNet task, using the most discriminative attribute, i.e., the thick grey skin [12], is sufficient to achieve high accuracy. However, in DomainNet, elephants no longer have grey skins while the long nose and big ears are preserved and the network relying on grey skins will fail to generalize.

Algorithm 1: `Conditional Statements`

Define intervals
$I_i \subset \mathbb{R}, i = 1, \cdots, M$
Define functions
$h_i, , i = 1, \cdots, M+1$
switch $h_1(\boldsymbol{x})$ **do**
  case I_i **do**
    apply h_{i+1} to $\boldsymbol{x}$

To efficiently capture the visual attributes and combine them for DG, conditional statements (e.g., IF/ELSE in programming) are needed. First, to obtain these diverse attributes, different filters should be applied to different regions of the image. For example, the operation for *curved outer teeth* is that if the patches belong to the teeth, we apply a filter to obtain its shape. Second, for image recognition, the network should leverage conditional statements to integrate multiple visual attributes for learning the definition of this class. For example, the recognition of an elephant should check if the key attributes are placed in the proper position. In literature, the MoE layer is considered as an effective approach to implement conditional computations [25, 24]. We formalize this intuition in the next theorem.

Theorem 2. *An MoE module in equation 2 with N experts and $k = 1$ aligns with the conditional statements in Algorithm 1 with*

$$\textit{Alignment} = \begin{cases} (N+1) \cdot \max\left(\mathcal{M}^*_{\mathscr{P}}, \mathcal{M}(G, h_1, \epsilon, \delta)\right), & \textit{if } N < M, \\ (N+1) \cdot \max\left(\max\limits_{i \in \{1, \cdots, M\}} \mathcal{M}(f_{FFN_i}, h_{i+1}, \epsilon, \delta), \mathcal{M}(G, h_1, \epsilon, \delta)\right), & \textit{if } N \geq M, \end{cases} \tag{3}$$

where $\mathcal{M}(\cdot, \cdot, \cdot, \cdot)$ is defined in Definition 1, and $\mathcal{M}^*_{\mathscr{P}}$ is the optimal objective value of the following optimization problem:

$$\begin{aligned} \mathscr{P}: & \underset{\mathcal{I}_1, \cdots \mathcal{I}_N}{\textit{minimize}} \quad \max_{i \in \{1, \cdots, N\}} \quad \mathcal{M}(f_{FFN_i}, ([1_{I_j}]_{j \in \mathcal{I}_i} \circ h_1)^T \cdot [h_j]_{j \in \mathcal{I}_i}, \epsilon, \delta) \\ & \textit{subject to} \quad \cup_{i=1}^N \mathcal{I}_i = \{2, 3, \cdots, M+1\}, \end{aligned}, \tag{4}$$

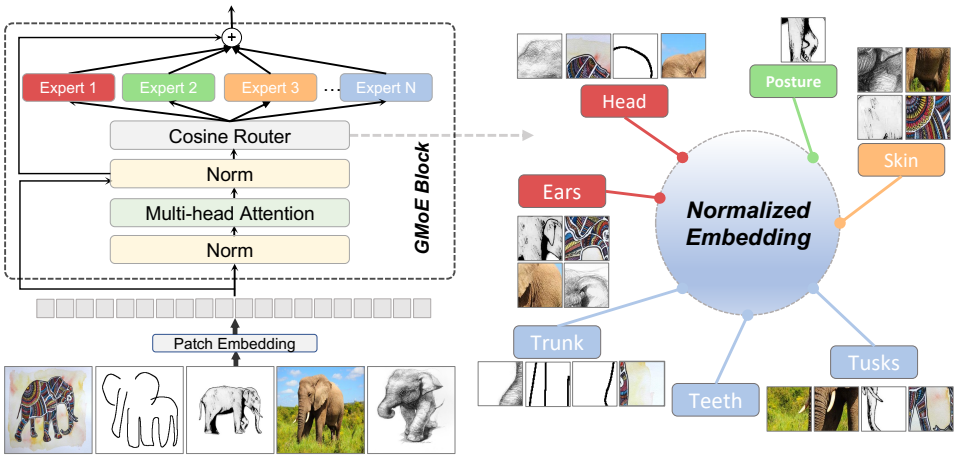

Figure 1: Overview architecture of GMoE. The cosine router distributes normalized image patches of different visual attributes to corresponding experts.

where 1_{I_j} *is the indicator function on interval* I_j.

Remark 1. (Interpretations of Theorem 2) In algorithmic alignment, the network better aligns with the algorithm if the alignment value in equation 1 is lower. The alignment value between MoE and conditional statements depends on the product of $N+1$ and a sample complexity term. When we increase the number of experts N, the alignment value first decreases as multiple experts decompose the original conditional statements into several simpler tasks. As we further increase N, the alignment value increases because of the factor $N+1$ in the product. Therefore, the MoE aligns better with conditional statements than with the original FFN (i.e., $N=1$).

3.3 Adapting MoE to Domain Generalization

In literature, there are several variants of MoE architectures, e.g., [24, 8], and we should identify one for DG. By algorithmic alignment, in order to achieve a better generalization, the architecture of sparse MoEs should be designed to effectively handle visual attributes. In the following, we discuss our architecture design for this purpose.

For the routing scheme, linear routers (i.e., equation 2) are often adopted in MoEs for vision tasks [24] while recent studies in NLP show that the cosine router achieves better performance in cross-lingual language tasks [6]. For the cosine router, given input $\boldsymbol{x} \in \mathbb{R}^d$, the embedding $\boldsymbol{W}\boldsymbol{x} \in \mathbb{R}^{d_e}$ is first projected onto a hypersphere, followed by multiplying a learned embedding $\boldsymbol{E} \in \mathbb{R}^{d_e \times N}$. Specifically, the expression for the gate is given by $G(\boldsymbol{x}) = \text{TOP}_k\left(\text{Softmax}\left(\frac{\boldsymbol{E}^T \boldsymbol{W}\boldsymbol{x}}{\tau \|\boldsymbol{W}\boldsymbol{x}\| \|\boldsymbol{E}\|}\right)\right)$, where τ is a hyper-parameter. We opine that cosine routers are more powerful in DG as $\boldsymbol{E}$ can be interpreted as the codebook for for visual attributes [9, 33] and the dot product between $\boldsymbol{E}$ and $\boldsymbol{W}\boldsymbol{x}$ with ℓ_2 normalization is a matched filter.

As for the number of MoE layers, *Every-two* and *last-two* are two commonly adopted placement methods in existing MoE studies [24]. Specifically, every-two refers to replacing the even layer's FFN with MoE, and last-two refers to placing MoE at the last two even layers. For IID generalization, every-two often outperforms last-two [24]. We argue that last-two is more suitable for DG as the conditional sentences for processing visual attributes are high-level.

4 Experimental Results

We evaluate GMoE on DomainBed [13]. We present results in Table 1 with **train-validation selection**, which include baseline methods and recent SOTA DG algorithms and GMoE trained with ERM. The results demonstrate that GMoE without DG algorithms already outperforms counterparts on almost all the datasets.

Table 1: Overall out-of-domain accuracies with train-validation selection criterion.

Algorithm	PACS	VLCS	OfficeHome	TerraInc	DomainNet
ERM (ResNet50) [27]	85.7 ± 0.5	77.4 ± 0.3	67.5 ± 0.5	47.2 ± 0.4	41.2 ± 0.2
IRM [ArXiv 20] [1]	83.5 ± 0.8	78.5 ± 0.5	64.3 ± 2.2	47.6 ± 0.8	33.9 ± 2.8
FISH [ICLR 22] [26]	85.5 ± 0.3	77.8 ± 0.3	68.6 ± 0.4	45.1 ± 1.3	42.7 ± 0.2
SWAD [NeurIPS 21] [4]	88.1 ± 0.1	79.1 ± 0.1	70.6 ± 0.2	50.0 ± 0.3	46.5 ± 0.1
Fishr [ICML 22] [22]	85.5 ± 0.2	77.8 ± 0.2	68.6 ± 0.2	47.4 ± 1.6	41.7 ± 0.0
MIRO [ECCV 22] [5]	85.4 ± 0.4	79.0 ± 0.0	70.5 ± 0.4	**50.4 ± 1.1**	44.3 ± 0.2
ERM (ViT-S/16) [ICLR 21] [7]	86.2 ± 0.1	79.7 ± 0.0	72.2 ± 0.4	42.0 ± 0.8	47.3 ± 0.2
GMoE-S/16 (Ours)	**88.1 ± 0.1**	**80.2 ± 0.2**	**74.2 ± 0.4**	48.5 ± 0.4	**48.7 ± 0.2**

References

[1] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, 2019. 1, 5

[2] Manh-Ha Bui, Toan Tran, Anh Tuan Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *CoRR*, 2021. 1

[3] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 2021. 1

[4] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, 2021. 5

[5] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. *CoRR*, 2022. 5

[6] Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. On the representation collapse of sparse mixture of experts. *CoRR*, abs/2204.09179, 2022. 4

[7] 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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 2, 5

[8] William Fedus, Jeff Dean, and Barret Zoph. A review of sparse expert models in deep learning. *arXiv preprint arXiv:2209.01667*, 2022. 4

[9] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. *Advances in neural information processing systems*, 20, 2007. 4

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 2016. 1

[11] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems 34*, 2021. 1

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 3

[13] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *9th International Conference on Learning Representations, ICLR*, 2021. 2, 4

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 1

[15] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8256–8266, 2018. 1

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1

[17] David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron C. Courville. Out-of-distribution generalization via risk extrapolation (rex). In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR, 2021. 1

[18] 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. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV*, 2021. 1

[19] Yi Ma, Doris Tsao, and Heung-Yeung Shum. On the principles of parsimony and self-consistency for the emergence of intelligence. *Frontiers of Information Technology & Electronic Engineering*, pages 1–26, 2022. 1

[20] Massimiliano Mancini, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Best sources forward: Domain generalization through source-specific nets. In *2018 IEEE International Conference on Image Processing*, 2018. 1

[21] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 2

[22] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. *arXiv preprint arXiv:2109.02934*, 2021. 5

[23] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 1

[24] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems, NeurIPS 2021, December 6-14, 2021*, 2021. 3, 4

[25] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR*, 2017. 2, 3

[26] Yuge Shi, Jeffrey Seely, Philip H. S. Torr, N. Siddharth, Awni Y. Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *CoRR*, 2021. 5

[27] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991. 5

[28] Yufei Wang, Haoliang Li, Lap-Pui Chau, and Alex C. Kot. Variational disentanglement for domain generalization. *CoRR*, 2021. 1

[29] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Taylan Cemgil, et al. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021. 1

[30] Keyulu Xu, Jingling Li, Mozhi Zhang, Simon Du, Kenichi Kawarabayashi, and Stefanie Jegelka. What can neural networks reason about? In *Proceedings of International Conference on Learning Representation*, Apr. 2020. 1, 2

[31] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P. Xing. Towards principled disentanglement for domain generalization. *CoRR*, 2021. 1

[32] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021. 1

[33] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014. 2, 4