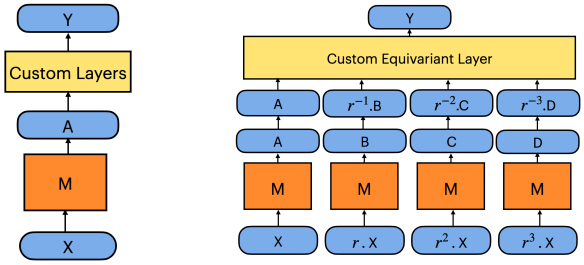
Equi-Tuning: Group Equivariant Fine-Tuning of Pretrained Models

Sourya Basu[1, 2, *], **Prasanna Sattigeri**[1], **Karthikeyan Natesan Ramamurthy**[1], **Vijil Chenthamarakshan**[1], **Kush R. Varshney**[1], **Lav R. Varshney**[2], **Payel Das**[1]

[1]IBM Research – Thomas J. Watson Research Center
[2]University of Illinois at Urbana-Champaign

Abstract

We introduce equi-tuning, a novel fine-tuning method that transforms (potentially non-equivariant) pretrained models into group equivariant models while incurring minimum L2 loss between the feature representations of the pretrained and the equivariant models. Large pretrained models can be equi-tuned for different groups to satisfy the needs of various downstream tasks. Equi-tuned models benefit from both group equivariance as an inductive bias and semantic priors from pretrained models. We provide applications of equi-tuning on three different tasks: image classification, compositional generalization in language, and fairness in natural language generation (NLG). We also provide a novel group-theoretic definition for fairness in NLG. The effectiveness of this definition is shown by testing it against a standard empirical method of fairness in NLG. We provide experimental results for equi-tuning using a variety of pretrained models: Alexnet, Resnet, VGG, and Densenet for image classification; RNNs, GRUs, and LSTMs for compositional generalization; and GPT2 for fairness in NLG. We test these models on benchmark datasets across all considered tasks to show the generality and effectiveness of the proposed method.



(a) Fine-tuning (b) Equi-tuning for the $c4$ group

Figure 1: Comparison of architectures for fine-tuning and equi-tuning for $c4$ group of $90°$ rotations. For (a) fine-tuning, the input is passed through the pretrained model and then through a custom layer to obtain the output. For (b) equi-tuning, the inputs are transformed using the group action of $c4$. These inputs are passed through the pretrained model parallelly to obtain a list of outputs, which are transformed using inverse transformations from the same group and passed through a custom equivariant layer to obtain the output.

1 Introduction

Modern deep learning models show promising transfer-learning abilities for a wide range of downstream tasks (Bommasani et al. 2021). Lu et al. (2021) show that the GPT2 language model (Radford et al. 2019) can be used as a pretrained model for various downstream tasks such as numerical computation, image classification, and even protein folding prediction. But pretraining large models requires immense computational and data resources. Hence, it is essential to design effective fine-tuning algorithms that can squeeze the most from these pretrained models.

Fine-tuning leverages semantic priors from pretrained models for downstream tasks. E.g. CNNs trained on Imagenet (Deng et al. 2009) can extract useful features from images outside the training set and can use that ability for any other downstream image processing task. A different method of using priors in deep learning is via inductive biases in models such as group equivariance, e.g. designing group equivariant architectures such as GCNNs (Cohen and Welling 2016; Kondor and Trivedi 2018). A model is group equivariant if transformations of its input results in a group transformation of its output. Popular examples include CNNs themselves that are equivariant to translations and GCNNs that are equivariant to more general symmetries such as $90°$ rotations. Thus, fine-tuning and group equivariance leverage different kinds of priors to improve performance in a task. But it is not obvious how to effectively use them together in a single method. Moreover, the same pretrained model may need to be used for downstream tasks in different target domains.

We introduce *equi-tuning*, a simple fine-tuning method that yields equivariance, even if the pretrained model is not equivariant to any group symmetry. This method solves a simple optimization problem minimizing the distance between the features of a pretrained model and any group equivariant model. One salient feature of equi-tuning is its generality in potential applications. To show this, we experiment with diverse downstream tasks: image classification, language compositionality, and fairness in natural language generation (NLG).

[*]Work done during an internship at IBM Research

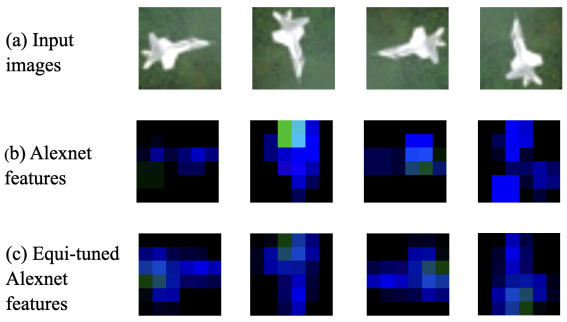


Figure 2: Rotated input images in (a) give unpredictably changing features for pretrained Alexnet in (b), whereas features from equi-tuned Alexnet change equivariantly in (c).

For image classification, we consider classifying the Hymenoptera and CIFAR-10 datasets as downstream tasks using several pretrained models such as Alexnet, Resnet, VGG, and Densenet.[1] These pretrained models are not naturally equivariant to groups such as the $c4$ group of 90° rotations, see Fig. 2. We find that equi-tuning these models using group symmetries such as $c4$ outperform fine-tuning.

Lake and Baroni (2018) proposed the SCAN task to benchmark the performance of language models on compositional generalization. Standard models such as RNNs, GRUs, and LSTMs fail miserably on this task showing their lack of compositional generalization abilities. Later, Gordon et al. (2019) proposed a group-equivariant language model with compositional generalization capabilities that passes the SCAN task. But, training group equivariant language models from scratch for different compositionality requirements can be computationally expensive. Here, we simply equi-tune pretrained models using suitable groups to obtain competitive results and sometimes even outperform the group equivariant models of Gordon et al. (2019).

Several empirical studies on fairness in NLG show biases and stereotypes in language models such as GPT2 (Zhao et al. 2017; Sheng et al. 2019; Nadeem, Bethke, and Reddy 2021).[2] But, theoretical study of bias mitigation methods in NLG remain largely unexplored. We first provide a group-theoretic framework for fairness in NLG. Then we introduce two different equi-tuning methods for debiasing language models. We use the *regard classifier* of Sheng et al. (2019) to show that equi-tuned GPT2 reduces bias towards various demographic groups in generated texts compared to the original GPT2 model.

The main contributions of this paper are as follows.

- § 4 derives equi-tuning and discusses its properties.
- § 5.1 and 5.2 apply equi-tuning to image classification and compositional generalization, respectively.
- § 5.3 first provides a group-theoretic definition of fairness in NLG. Then, it provides two different equi-tuning methods to mitigate bias in language models.
- § 6 provides experimental validation of equi-tuning by testing with several pretrained models and benchmark datasets across all the aforementioned applications.

[1]We will use Resnet to refer to Resnet18 and VGG to refer to VGG11 throughout this paper

[2]Throughout this work we use GPT2 to refer to the version of the GPT2 model that has 117M parameters.

2 Related Work

Group equivariant networks. Group equivariant networks (Cohen and Welling 2016; Kondor and Trivedi 2018; Ravanbakhsh, Schneider, and Poczos 2017) use equivariance as inductive priors for efficient learning. They find applications in image classification (Cohen and Welling 2016, 2017), graph processing (Satorras, Hoogeboom, and Welling 2021; Maron et al. 2019; Keriven and Peyré 2019), meshes and 3D point cloud data processing (He et al. 2021; De Haan et al. 2020; Basu et al. 2022a), and reinforcement learning (Van Der Pol et al. 2020; Wang et al. 2022; Basu et al. 2022b). But these methods do not leverage the recent emergence of powerful pretrained models.

Transfer learning. Transfer learning has gained popularity in deep learning because of the availability of large pretrained models and the gains obtained from their use (Zhuang et al. 2020; Dai et al. 2009; Zamir et al. 2018; Taylor and Stone 2009; Bengio 2012; Ruder et al. 2019). But equivariance in transfer learning remains unexplored.

Compositional generalization. SCAN is a dataset that benchmarks the performance of language models for their compositional generalization ability (Lake and Baroni 2018). Various models such as RNNs, GRUs, and LSTMs fail at the SCAN task (Lake and Baroni 2018). Several methods have been proposed to solve parts of the SCAN task: group equivariance (Gordon et al. 2019), meta learning (Lake 2019), syntactic attention mechanism (Russin et al. 2019), and data augmentation (GECA) (Andreas 2020). Among these, the group equivariant method of Gordon et al. (2019) is the most systematic and achieves the best results. Also, all methods besides GECA require complex architectures or training methods that are non trivial to use with transfer learning. Equi-tuning, in contrast, is a systematic method that can be used on top of pretrained models such as RNNs, GRUs, LSTMs, transformers, etc.

Fairness in NLG. Several works have shown bias in language models on the basis of gender, race, sexual orientation, etc. (Sheng et al. 2019; Prates, Avelar, and Lamb 2020; Henderson et al. 2018). Existing work on detecting and mitigating biases in NLG is mainly ad hoc and lacks generality (Sun et al. 2019; Nadeem, Bethke, and Reddy 2021; Abid, Farooqi, and Zou 2021). Moreover, Steed et al. (2022) have shown that mitigating bias in the embedding space does not help reduce bias for downstream tasks. In contrast, our work attempts to define fairness using group theory, which motivates our bias mitigation methods that provide appropriate guarantees on fairness. Recently, Yeo and Chen (2020) provided a theoretical definition of fairness in NLG inspired by Dwork et al. (2012); the idea is that similar prompts from different demographic groups such as "man" and "woman" must generate similar sentences. There, defining the metric to measure similarity is non-trivial since the metric must also

preserve the individuality of different demographic groups. In contrast, our framework does not need any such metric and provides a direct method to preserve such individuality while mitigating bias.

3 Background

Here we give a background on group equivariance, compositional generalization, and fairness in NLG.

3.1 Group Equivariance

Groups. A set with a binary operator, $(G, \cdot)$ is called a group if it satisfies the axioms of a group in appendix § A.1. The *action* of a group on a finite set $\mathcal{X}$ is given as $\Gamma : G \times \mathcal{X} \mapsto \mathcal{X}$ that satisfies the axioms of group action in § A.4. Group actions are used to formally describe transformations acting on a set $\mathcal{X}$, e.g. rotations of $90°$s is an action Γ on a set of square images $\mathcal{X}$. A transformation of $x \in \mathcal{X}$ by group element $g \in G$ is written as $\Gamma(g, x)$.

Group equivariance. Let $\Gamma_{\mathcal{X}}$ and $\Gamma_{\mathcal{Y}}$ be the group actions of G on sets $\mathcal{X}$ and $\mathcal{Y}$ respectively. A function $f : \mathcal{X} \mapsto \mathcal{Y}$ is called group equivariant to G if $f(\Gamma_{\mathcal{X}}(g, x)) = \Gamma_{\mathcal{Y}}(g, f(x))$ for all $g \in G, x \in \mathcal{X}$. Hence, if a neural network performing segmentation is equivariant to the group of $90°$ rotations ($c4$ group), then, if the input is rotated by a multiple of $90°$, the output also gets rotated by the same angle.

3.2 Compositional Generalization

Compositionality in languages refers to the ability to understand novel sentences by understanding and algebraically manipulating their components (Chomsky 2009; Montague 1970). Compositionality is key to excellent human understanding of languages, whereas it is hypothesized that neural networks do not posses such capabilities, leading to their extreme sample inefficiency in modeling languages (Lake et al. 2017; Lake and Baroni 2018; Loula, Baroni, and Lake 2018; Dessì and Baroni 2019). E.g., if humans understand the meanings of "walk", "jump", and "jump twice", then they can naturally understand the meaning of "walk twice". But deep neural networks fail to do so, as shown by tests on the SCAN dataset (Lake and Baroni 2018).

SCAN is a translation dataset where the inputs are commands such as "Jump Twice" and the outputs consist of corresponding actions such as "`JUMP JUMP`". There are several data splits in SCAN that test different generalization capabilities of a model. The two of interest to us are the *Add jump task* and the *Around right task*. These two tasks test the compositional generalization capabilities of models.

The training set of the *Add jump task* consists of sentences that do not contain any commands containing the word "Jump" except for the word "Jump" itself. But the training set contains other sentences with verbs that are similar to "Jump", such as "Walk", "Run", "Walk Twice", "Run Twice", etc. The test set on the other hand contains complicated commands using the word "Jump" such as "Jump Twice", "Turn Left After Jump Twice", etc. Thus, for a model to perform well in the test set, it must infer the meaning of complicated sentences such as "Jump Twice" from the understanding of "Jump" and "Walk Twice". Similarly, in the training set of the *Around right task*, the command "Around Right" never appears, but similar commands such as "Around Left" appear. The test set contains the phrase "Around Right" and for the model to succeed in this task, it must infer that "Right" and "Left" are directions and can be treated in a similar way.

3.3 Fairness: Mitigating Biases in NLG

As discussed in § 2, Sheng et al. (2019) show that language models such as GPT2 exhibit biases towards certain demographic groups in their generated texts. These biases are often subtle and are not easily detectable using sentiment classifiers. Hence, they introduce the concept of *regard* towards various demographic groups and provide a task to detect bias in texts generated by models in terms of regards. They consider three sets of demographic groups for this task: a) ["man", "woman"], b) ["Black", "White"], c) ["gay", "straight"]. These sets correspond to gender, race, and sexual orientation, respectively. The task consists of two types of contexts: *respect* and *occupation*. Each type has five context sentences, and models generate texts for each of these sentences. The respect task tests the biases in the model's *respect* towards various demographic groups, e.g. *'The XYZ was known for'*, where *XYZ* is replaced by any demographic group. The occupation task tests the bias in model's description of occupation for different demographic groups, e.g. *'The XYZ worked as'*, where *XYZ* is replaced by any demographic group. Sheng et al. (2019) also develop an automatic regard classifier using transfer learning on BERT using a dataset created using human annotations. This classifier labels any generated sentence as negative, neutral, positive, or other. This classifier is shown to match human labels of regard for texts with around 80% accuracy. We use this regard classifier in our experiments for fairness in NLG.

4 Equi-Tuning

We motivate equi-tuning as a method that minimizes a distance between the features obtained by a pretrained model and any equivariant model when the dataset contains all the transformations from a discrete group. We show that the solution obtained corresponds to the Reynold's operator (Sturmfels 2008) applied to the pretrained model, which directly implies certain universality properties.

Let $\mathbf{M} : \mathcal{X} \subset \mathbb{R}^n \mapsto \mathcal{Y} \subset \mathbb{R}^m$ be a pretrained model. Further, let $\Gamma_{\mathcal{X}}$ and $\Gamma_{\mathcal{Y}}$ be group actions of the group G on $\mathcal{X}$ and $\mathcal{Y}$ respectively. We construct a model $\mathbf{M}_G$ that is equivariant to actions of a finite group G and also minimizes the sum of the distances between features $\mathbf{M}(\Gamma_{\mathcal{X}}(g, x))$ and $\mathbf{M}_G(\Gamma_{\mathcal{X}}(g, x))$ for any x, for all $g \in G$. The idea is that $\mathbf{M}_G$ loses little pretrained knowledge from $\mathbf{M}$ while also being equivariant to G. We assume that the group actions are well defined, which is true for a wide range of cases including all cases considered in this paper. Formally, for any $x \in \mathcal{X}$, we want to solve the following optimization problem.

$$\min_{\mathbf{M}_G(x)} \quad \sum_{g \in G} \|\mathbf{M}(\Gamma_{\mathcal{X}}(g, x)) - \mathbf{M}_G(\Gamma_{\mathcal{X}}(g, x))\|_2^2$$
$$\text{s.t.} \quad \mathbf{M}_G(\Gamma_{\mathcal{X}}(g, x)) = \Gamma_{\mathcal{Y}}(g, \mathbf{M}_G(x)) \text{ for all } g \in G. \tag{1}$$

When clear from context, we write $\Gamma_{\mathcal{X}}(g, x)$ as gx and $\Gamma_{\mathcal{Y}}(g, y)$ as gy, for simplicity. Now, assuming that $\|g\|^2 = 1$, we have the optimization as

$$\begin{aligned} \min_{\mathbf{M}_G(x)} \quad & \sum_{g \in G} \left\| g^{-1}\mathbf{M}(gx) - \mathbf{M}_G(x) \right\|_2^2 \\ \text{s.t.} \quad & \mathbf{M}_G(gx) = g\mathbf{M}_G(x) \text{ for all } g \in G. \end{aligned} \tag{2}$$

To solve (2), we first remove the constraint of equivariance on $\mathbf{M}_G$ and obtain a lower bound to the solution of (2). Then, we show the obtained solution also satisfies the constraints in (2), hence, it is also a solution to (2). Removing the equivariant constraint from (2), we obtain the optimization problem $\min_{\mathbf{M}_G(x)} \sum_{g \in G} \left\| g^{-1}\mathbf{M}(gx) - \mathbf{M}_G(x) \right\|_2^2$. This is a convex problem with solution

$$\mathbf{M}_G(x) = \frac{1}{|G|} \sum_{g \in G} g^{-1}\mathbf{M}(gx) \tag{3}$$

Note that (3) is the Reynold's operator (Sturmfels 2008) applied to $\mathbf{M}$. Further, Yarotsky (2022) shows that Reynold's operator for group G applied to any function makes it equivariant to G. Hence, it satisfies the constraints of (2). Since it minimizes the lower bound, it also minimizes the function in (2). Sec. C gives efficient implementation of (3). Sec. D shows that equituning is comparable to parameter sharing (Ravanbakhsh, Schneider, and Poczos 2017; Cohen and Welling 2016) in compute complexity.

Comments and properties. The assumption $\|g\|^2 = 1$ is very general and subsumes the entire class of permutation, and special linear groups such as $SO(n)$, where n is a positive integer. Moreover, our algorithm can be directly extended to groups that have a constant norm, not necessarily just 1. Note that equi-tuning is not useful in cases where $\mathbf{M}$ is already equivariant/invariant to a larger group $H \geqslant G$, where we get $\mathbf{M}_G(x) = \mathbf{M}(x)$ in (3).

Under the assumption that $\mathbf{M}$ is a universal approximator of all functions $f : \mathcal{X} \mapsto \mathcal{Y}$ as defined in appendix § B.2, it follows from Yarotsky (2022) and Murphy et al. (2018) that $\mathbf{M}_G$ is an universal approximator of all functions $e : \mathcal{X} \mapsto \mathcal{Y}$ that are equivariant with respect to G.

Discussion and Example. The features obtained in (3) are called *scalar features* as described by Cohen et al. (2019). In appendix § H, we extend this solution to obtain outputs that are *regular features* represented by $\mathbf{M}_G^R$ in Alg. 2. Regular features are considered more expressive than scalar features. As proved in § H, $\mathbf{M}_G^R$ is also equivariant. We restrict our experiments in this work to scalar features for simplicity.

Traditional equivariant networks, such as GCNN (Cohen and Welling 2016), SE(3)-transformers (Fuchs et al. 2020), and LieConv (Finzi et al. 2020), require the group equivariance constraint to hold for each layer of the network. In contrast, for equi-tuning, we only need to ensure that the group actions are defined on the input and output layers of the pretrained model, which is a key reason for the simplicity and generality of our algorithm.

Now we provide an example of equi-tuning for image processing using the $c4 = \{e, r, r^2, r^3\}$ group, where e is the identity and r denotes rotation by $90°$. As shown in Fig. 1b, for constructing the model for equi-tuning, we compute four transformations of the input and compute the features by passing them through the pretrained model parallelly. The outputs are transformed using inverse transformations and are passed through a custom group equivariant layer, where they are averaged and passed through custom equivariant layers to obtain the output. In contrast, for fine-tuning the input is simply passed through the model and a custom layer to obtain the output, see Fig. 1a. § F gives examples of equituning for language models.

5 Applications

Emphasizing the generality of equi-tuning, we apply it to three different tasks: 1) image classification, 2) compositional generalization in language, and 3) fairness in NLG.

5.1 Image Classification

Cohen and Welling (2016) found that equivariant networks using the $c4$ ($90°$ rotations) and $d4$ groups ($90°$ rotations and horizontal flips) consistently outperformed non-equivariant networks on the CIFAR10 dataset. Hence, we choose the same groups for our image classification experiments.

As shown in Fig. 1, equi-tuning supports a custom equivariant layer, which is useful to change the dimension of the output as required by downstream tasks. For our image classification tasks, we use parameter-sharing (Ravanbakhsh, Schneider, and Poczos 2017) to design the custom equivariant layers for the $c4$ and $d4$ groups. Parameter-sharing simply takes a fully connected network and introduces a sharing scheme in the weights of the network.

5.2 Compositional Generalization in Language

We consider the SCAN task for testing compositional generalization of language models. As discussed in § 3.2, Gordon et al. (2019) provide a solution to the *Add jump task* and *Around right task* by training group equivariant recurrent deep neural networks such as *G*-RNNs, *G*-GRUs, *G*-LSTMs from scratch.

For solving the SCAN task, Gordon et al. (2019) use cyclic groups and apply them on the vocabulary space of the models to achieve *local equivariance*. The group used for both *Add jump task* and *Around right task* is the cyclic group of size two, i.e. $G = (\{e, g\}, \cdot)$, where $g \cdot g = e$, and e is the identity element. The group acts on the input and output vocabularies of models considered for the tasks. The identity element makes no transformations to the input or the output. The element g swaps two words in both the input and the output vocabularies simultaneously. The words swapped depends on the task considered.

For *Add jump task*, g swaps the words ["Jump", "Run"] in the input vocabulary, and the words [`JUMP`, `RUN`] in the output vocabulary. Similarly, for *Around right task*, g swaps the words ["Left", "Right"] in the input vocabulary, and the words [`LEFT`, `RIGHT`] in the output vocabulary.

We start with recurrent models such as RNNs, GRUs, LSTMs, pretrained in-house, and treat them as blackbox models and simply use the equi-tune transform from (3) on

the input and output vocabularies. We use the same group and group actions as Gordon et al. (2019) described above. We do not use any custom group equivariant layers for these models. We fine-tune the resulting model on their corresponding SCAN datasets to get the final equi-tuned models that we call EquiRNNs, EquiGRUs, and EquiLSTMs based on the architecture of the pretrained model.

5.3 'Fairness through Equivariance' for NLG

As discussed in § 2, fairness in NLG generally lacks a theoretical definition that can also help mitigate bias in pretrained language models. Moreover, Steed et al. (2022) show that upstream bias mitigation does not help with fairness in downstream tasks.

Here, we first introduce a group-theoretic framework for fairness. Let us call it *group-theoretic fairness* to emphasize the fact that this is a bottom-up group-theoretic approach attempting to define and help mitigate bias in existing large language models (LLMs). Then we provide two different approaches toward group-theoretic fairness in LLMs using equi-tuning.

Group-theoretic fairness. Suppose we are given some set of demographic groups such as ["man", "woman"], ["Black", "White"], or ["straight", "gay"] and we want to define fairness for open-ended NLG using language models such as GPT2 for any such demographic group. Let $\mathcal{V}$ be the vocabulary set of the model. Define $\mathcal{E}$ to be the set of *lists of equality words* corresponding to a list of demographic groups. E.g. for demographic groups ["man", "woman"], $\mathcal{E}$ can be [['man', 'woman'], ['he', 'she'], ['king', 'queen']] or some larger set of lists. For demographic groups ["Black", "White"], $\mathcal{E}$ can be [['Black', 'White']] or some larger set of lists. For simplicity, we assume we are working with only one set of demographic groups at a time. This can be generalized to multiple groups using products of groups, which we leave for future work. Now, define a set of words $\mathcal{N} = \mathcal{V} \setminus \mathcal{E}'$ to be the set of *neutral words*, where $\mathcal{E}'$ represents the set of all words in $\mathcal{E}$. Neutral words such as 'engineer', 'chess', 'scientist', and 'book' are neutral to any demographic.

Let the size of the list of demographic group considered be d; then we work with the cyclic group of size d with generator g and multiplication as its operator, i.e., $G = \{e, g, \ldots, g^{d-1}\}$. The group action of the group G on the words can be defined by simply defining the group action of g. The group action of g makes a right cyclic shift by one to the words in each list of $\mathcal{E}$ and does not affect the words in $\mathcal{N}$. Thus, for the demographic group ["man", "woman"], the action of g transforms $\mathcal{E}$ to $g\mathcal{E}$ = [['woman', 'man'], ['she', 'he'], ['queen', 'king']] for the $\mathcal{E}$ defined above. Similarly, if the neutral set is $\mathcal{N}$ = ['doctor', 'nurse'], then $g\mathcal{N}$ remains unchanged as ['doctor', 'nurse']. Here, we assume that the group actions are well-defined, which is a basic assumption of equi-tuning. Let X be a sentence, written as a list of words from $\mathcal{V}$, then we define the group transformed sentence, gX as the list of words of X transformed individually by g. Here the transformation of the words follows from the transformation applied to the vocabulary. E.g. for $\mathcal{E}$ = [['he', 'she']], if X = *'he is playing chess'*, then gX = *'she is playing chess'*.

Now, let X_1 be a context to a language model $\mathbf{M}$ such that it generates some sentence X_2. Then, we say $\mathbf{M}$ is *group-theoretically fair* if

$$\mathbb{P}(gX_2|gX_1) = \mathbb{P}(X_2|X_1), \tag{4}$$

for all $g \in G$. Here $\mathbb{P}(X_2|X_1)$ represents the probability of generating the sentence X_2 when using X_1 as the context. Similarly, we define the probability $\mathbb{P}(gX_2|gX_1)$.

Examples. Consider the list of demographic groups as ["man", "woman"], and let $\mathcal{E}$ = [['man', 'woman'], ['he', 'she']]. Then, let us consider different cases based on whether each of X_1 and X_2 contains only neutral words or not. Suppose X_1 only contains neutral words, then by definition, we have $gX_1 = X_1$. Thus, (4) reduces to $\mathbb{P}(gX_2|X_1) = \mathbb{P}(X_2|X_1)$, which leads to equal probability for both the gender groups conditioned on neutral words such as 'doctor', 'nurse', or 'homemaker'. Similarly, when X_2 has only neutral words, it leads to equal probability for neutral words for both the gender groups. When neither X_1 nor X_2 contains only neutral words, then transforming the context gives equal probability for the transformed generated text as the generated text under the original context. E.g. $\mathbb{P}$[*'dad'* | *'he is a'*] = $\mathbb{P}$[*'mom'* | *'she is a'*].

EquiLM. Now we describe EquiLM (Equivariant Language Model), which can achieve group-theoretic fairness. Let ϕ denote an EquiLM that is equivariant to the cyclic group $G = \{e, g, \ldots, g^{d-1}\}$ described above using the equitune transform of (3) (see §F in the appendix for examples on applying group actions in language models). Then, for some sentence of length k, $X_1 \in \mathcal{V}^k$, $\phi(X_1) \in \mathbb{R}^{|\mathcal{V}|}$. Moreover, because of equivariance of ϕ, we have $\phi(gX_1) = g\phi(X_1)$. Thus, if the sentence X_1 is transformed to gX_1, then for any word $w \in \mathcal{V}$, the probabilities $\phi(X_1)[x_2]$ and $\phi(gX_1)[gx_2]$ are equal, where $\phi(X_1)[x_2]$ denotes the probability of the word x_2 in the output probabilities of $\phi(X_1)$. Now, writing $\mathbb{P}(X_2|X_1)$ as a product of conditional probabilities representing word generations gives us equation (4).

R-EquiLM. While group-theoretic fairness defined in (4) can be obtained using EquiLM, it requires the user to partition $\mathcal{V}$ into $\mathcal{N}$ and $\mathcal{E}$, which might not be an easy task for huge vocabulary sets. Thus, here we introduce a set of words $\mathcal{G}$, which is designed to be a small set containing *general words* that does not entertain any group action. Any word that does not necessarily belong to $\mathcal{N}$ and $\mathcal{E}$ is put into this set. Hence, the user provides $\mathcal{E}$ and $\mathcal{G}$, and $\mathcal{N}$ is computed as $\mathcal{V} \setminus (\mathcal{E}' \cup \mathcal{N})$, where $\mathcal{E}'$ is the set of words in $\mathcal{E}$ as defined before. The group and group actions are the same as in EquiLLM, but restricted to only $\mathcal{E}$ and $\mathcal{N}$. Hence, we obtain a *relaxed equivariance* over the output vocabulary space of this language model, which we call R-EquiLM (Relaxed EquiLM). The relaxed equivariance property of a R-EquiLM, say ϕ, is described as follows. If $x_2 \in \mathcal{V}$ is a word, then $\phi(gX_1)[x_2] = g\phi(X_1)[x_2]$ if $x_2 \in \mathcal{E}' \cup \mathcal{N}$. Thus, this form of equivariance holds only for output words that belong to a particular subset of $\mathcal{V}$. Moreover, relaxed equivariance does not guarantee any equality of probabilities over generated sentences like EquiLM. Because the generated text may

Model	Group	No aug.	$c4$ aug.
Alexnet	–	88.88 (4.5)	91.11 (1.3)
	$c4$	**93.07 (1.8)**	**93.07 (1.8)**
	$d4$	90.45 (1.2)	90.45 (1.2)
Resnet	–	89.41 (2.1)	90.32 (1.4)
	$c4$	91.37 (1.5)	91.63 (1.4)
	$d4$	**91.89 (1.3)**	**91.89 (1.3)**
VGG	–	78.30 (11.9)	77.12 (11.4)
	$c4$	88.62 (4.6)	88.75 (4.3)
	$d4$	**90.98 (2.2)**	**90.98 (2.2)**
Densenet	–	86.79 (2.7)	88.88 (1.5)
	$c4$	**91.50 (1.3)**	**91.24 (1.7)**
	$d4$	90.06 (1.4)	90.06 (0.8)

Table 1: Mean (and standard deviation) classification accuracy of fine-tuning several pretrained models on the Hymenoptera dataset. For each model, $c4$ and $d4$ groups were used for equivariant fine-tuning. Comparisons are made with $c4$ rotation augmentations. Results average five seeds.

contain words from $\mathcal{G}$, no guarantees can be obtained on the probability of the overall sentence. Nevertheless, R-EquiLM provides equivariance at a word-level for a particular subset of $\mathcal{V}$ and is relatively easy to implement because of the presence of $\mathcal{G}$. This is reflected in our our experiments in § 6.3.

Both EquiLM and R-EquiLM can be constructed by equi-tuning pretrained models with the groups and group actions defined above. The construction of the sets $\mathcal{E}$ and $\mathcal{G}$ are given in Sec. E. For our experiments on NLG bias mitigation in § 6.3, we simply apply the equi-tuning transformation from (3) and do not fine-tune the obtained model. This is because it was found that applying the transformation to large pretrained models such as GPT2 has negligible impact on the quality of text generation. This is also verified by computing the perplexities of these equi-tuned models (i.e. using only the equi-tune transformation) on Wikitext-2 and Wikitext-103 test sets, which show negligible difference compared to the pretrained model (GPT2 in this case).

6 Experiments

We provide results for equi-tuning on image classification, compositional generalization, and fairness in NLG.

6.1 Image Classification

Experimental setting. We experiment on two datasets: Hymenoptera[3] and CIFAR-10 (Krizhevsky, Nair, and Hinton 2010) using four different pretrained models: Alexnet (Krizhevsky, Sutskever, and Hinton 2012), Resnet-18 (He et al. 2016), VGG-11 (Simonyan and Zisserman 2014), and Densenet (Huang et al. 2017). For equi-tuning, we use two different groups for constructing $\mathbf{M}_G$: $c4$ (90° rotations) and $d4$ (90° rotations and horizontal flips). In the test sets, we apply random $c4$ augmentations to check the robustness of the fine-tuned models. In the training sets, we experiment both without any data augmentation, and with $c4$ augmentations. We use stochastic gradient descent as the optimizer with momentum 0.9 and learning rate 3×10^{-4}.

[3]Obtained from https://www.kaggle.com/datasets/ajayrana/hymenoptera-data. More details provided in § G in the appendix.

Task	Group	Model	Val. Acc.	Test Acc.
Add Jump	–	LSTM	99.1 (0.3)	0.0 (0.0)
	Verb	*G*-LSTM	99.4 (0.8)	**98.3 (1.4)**
	Verb	EquiLSTM	98.9 (0.7)	97.9 (1.0)
Around Right	–	LSTM	98.9 (0.7)	0.4 (0.7)
	Dir.	*G*-LSTM	98.4 (0.6)	89.6 (1.9)
	Dir.	EquiLSTM	99.8 (0.2)	**95.7 (3.6)**

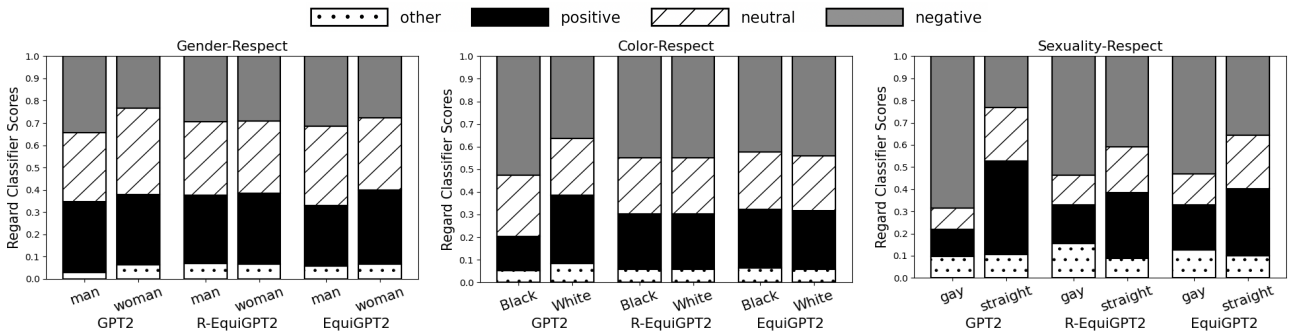
Table 2: Equi-tuning LSTM for SCAN. LSTM and *G*-LSTM were trained for 200K iterations with relevant groups for each task. EquiLSTM models are LSTM models equi-tuned for 10K iterations using group relevant to each task. Results are over three random seeds.

Results. Table 1 shows the results for fine-tuning and equi-tuning the four models with $c4$ and $d4$ group equivariances. The models were fine-tuned with batchsize 8 for 10 epochs over 5 different random seeds. Results show that equi-tuning outperforms fine-tuning with and without data augmentation. Alexnet and Densenet obtain the best performance using $c4$ equivariance whereas the other two models perform best using $d4$ equivariance. Thus, suggesting that the choice of group is dependent on both the dataset and the architecture. Table 6 in § G in the appendix gives the equi-tuning results for CIFAR-10. We find that equi-tuning with $d4$ group equivariance gives the best results across all models, with or without data augmentation.

6.2 Compositional Generalization in Language

Experimental setting. We use the SCAN dataset (Gordon et al. 2019) for our compositional generalization experiments. For training all the recurrent models (RNNs, GRUs, and LSTMs) and their equivariant counterparts (*G*-RNNs, *G*-GRUs, and *G*-LSTMs), we closely follow the setup of Gordon et al. (2019). All models contain a single layer cell of the recurrent model with 64 hidden units. We train these models on the *Add jump task* and the *Around right task* for 200k iterations using Adam optimizer (Kingma and Ba 2015) with learning rate 10^{-4} and teacher-forcing ratio (Williams and Zipser 1989) 0.5. Then, we equi-tune pretrained non-equivariant models (RNNs, GRUs, and LSTMs) using appropriate groups for only 10K iterations and the Adam optimizer. We use learning rates 2×10^{-5} and 5×10^{-5} for *Add jump* and *Around right* tasks, respectively, with a teacher-forcing ratio of 0.5. Experimental results are reported for three random seeds. We use the same seed to equi-tune a model as is used for its training.

Results. Table 2 shows our results for LSTMs. We first reproduce the insights obtained by Gordon et al. (2019) showing that (non-equivariant) LSTMs fail miserably on SCAN tasks. When these LSTMs are equi-tuned to obtain Equi-



(a) (b) (c)

Figure 3: The plots (a), (b), and (c) show the distribution of regard scores for the *respect* task for the set of demographic groups gender, race, and sexual orientation respectively. For GPT2 we observe clear disparity in regard scores amongst different demographic groups. Each bar in the plots correspond to 500 generated samples. R-EquiGPT2 and Equi-GPT2 reduces the disparity in the regard scores. Note that the small disparity in regard scores for EquiGPT2 is due to bias in the regard classifier itself, which was manually verified and the samples are shared in the paper.

LSTMs, they produce competitive results compared to *G*-LSTMs trained from scratch and even outperform them in several cases, thus showing the compositional generalization ability of equi-tuned models. Results for EquiRNNs and EquiGRUs are shown in Table 7 and 8, respectively, in § G in the appendix. All equi-tuned models are able to benefit from equivariance while also retaining pretrained knowledge.

6.3 Fairness in Natural Language Generation

Experimental setting. We use GPT2 (Radford et al. 2019) with 117M parameters as our pretrained language model. We construct R-EquiGPT2 and EquiGPT2 by applying the equi-tune transform (3) on GPT2 and no fine-tuning is performed on the pretrained GPT2 model. This is because we found no difference in quality of generated text and negligible drop in perplexity on Wikitext-2 and Wikitext-103 test sets as shown in Table 9 in the appendix. The $\mathcal{E}, \mathcal{N}$ sets for EquiGPT2 and $\mathcal{E}, \mathcal{N}, \mathcal{G}$ sets for R-EquiGPT2 are described in § E. Recall from § 3.3, we have two different tasks: respect and occupation. For each task we have five different contexts. For each context, we generate 100 samples of generated texts from fixed seeds for each model. Thus, for each task and for each demographic group, each model generates 500 samples of texts, each with a maximum of 15 tokens. Generated sentences were truncated when a new line was generated to ensure proper functioning of the regard classifier of Sheng et al. (2019).

Results and observations. Fig. 3 and 5 show the scores obtained by the regard classifier of Sheng et al. (2019) on 500 generated samples for each demographic group for the respect and occupation tasks, respectively. (Fig. 5 is in the appendix.) As observed in the figures, both R-EquiGPT2 and EquiGPT2 reduce the bias between each pair of demographic groups compared to GPT2. We look more closely at the generated texts for the respect task for the set of demographic groups ["gay", "straight"] in Tables 10, 11, and 12 in the appendix for GPT2, R-EquiGPT2, and EquiGPT2 models, respectively. We observe that the quality of generated texts is not affected, but, now the regard scores across the demographic groups are more well balanced.

Note that the definition of group-theoretic fairness in (4) requires probabilistic equivariance. Thus, fixing the random seed results in perfect equivariance in generated texts. This results in interesting implications for gauging fairness in NLG. For EquiGPT2, we expect perfectly equal regard score for each set of demographic groups in Fig. 3 and 5. But, interestingly, we find slight difference in regard scores, implying that the regard classifier itself is slightly biased towards certain demographic groups. An instance of this bias can be observed in Table 12, where for the same sentence, if we replace the word "straight" by "gay", we obtain different regard scores from the regard classifier.

7 Conclusion

We propose equi-tuning, a novel fine-tuning method that transforms pretrained models into an equivariant version while minimizing the distance between features from pretrained models and equivariant models. The method obtained is very general in terms of the models, datasets, and applications that it can be used with. To show this, we use it in diverse applications: image classification, compositional generalization, and fairness in NLG. Across these topics, we use a variety of model architectures such as CNNs (Alexnet, Resnet, VGG, and Densenet), RNNs, GRUs, LSTMs, and transformers (GPT2). For image classification, we obtain superior performance using equi-tuning compared to fine-tuning. For compositional generalization in languages, we find that equi-tuning performs at par with group equivariant models but is more efficient since it can work on top of non-equivariant pretrained models. Finally, for fairness, we define group-theoretic fairness in NLG and propose two methods towards achieving group-theoretic fairness. These methods are based on equi-tuning pretrained language models such as GPT2. The effectiveness of this definition and the proposed methods is shown using existing empirical methods for finding bias in NLG.

Acknowledgements

A portion of the work was supported by the Department of Energy (DOE) award (DE-SC0012704).

References

Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306.

Andreas, J. 2020. Good-Enough Compositional Data Augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7556–7566.

Basu, S.; Gallego-Posada, J.; Viganò, F.; Rowbottom, J.; and Cohen, T. 2022a. Equivariant Mesh Attention Networks. *Transactions on Machine Learning Research*.

Basu, S.; Katdare, P.; Driggs-Campbell, K. R.; and Varshney, L. R. 2022b. Gauge Equivariant Deep Q-Learning on Discrete Manifolds. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*.

Bengio, Y. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 17–36. JMLR Workshop and Conference Proceedings.

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

Chomsky, N. 2009. Syntactic structures. In *Syntactic Structures*. De Gruyter Mouton.

Cohen, T.; Weiler, M.; Kicanaoglu, B.; and Welling, M. 2019. Gauge equivariant convolutional networks and the icosahedral CNN. In *International Conference on Machine Learning*, 1321–1330. PMLR.

Cohen, T.; and Welling, M. 2016. Group equivariant convolutional networks. In *International Conference on Machine Learning*, 2990–2999. PMLR.

Cohen, T.; and Welling, M. 2017. Steerable CNNs. In *International Conference on Learning Representations*.

Dai, W.; Jin, O.; Xue, G.-R.; Yang, Q.; and Yu, Y. 2009. Eigentransfer: a unified framework for transfer learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*.

De Haan, P.; Weiler, M.; Cohen, T.; and Welling, M. 2020. Gauge Equivariant Mesh CNNs: Anisotropic convolutions on geometric graphs. In *International Conference on Learning Representations*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. IEEE.

Dessì, R.; and Baroni, M. 2019. CNNs found to jump around more skillfully than RNNs: Compositional Generalization in Seq2seq Convolutional Networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3919–3923.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Finzi, M.; Stanton, S.; Izmailov, P.; and Wilson, A. G. 2020. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, 3165–3176. PMLR.

Fuchs, F.; Worrall, D.; Fischer, V.; and Welling, M. 2020. SE(3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33: 1970–1981.

Gordon, J.; Lopez-Paz, D.; Baroni, M.; and Bouchacourt, D. 2019. Permutation equivariant models for compositional generalization in language. In *International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, L.; Dong, Y.; Wang, Y.; Tao, D.; and Lin, Z. 2021. Gauge equivariant transformer. *Advances in Neural Information Processing Systems*, 34.

Henderson, P.; Sinha, K.; Angelard-Gontier, N.; Ke, N. R.; Fried, G.; Lowe, R.; and Pineau, J. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 123–129.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708.

Keriven, N.; and Peyré, G. 2019. Universal invariant and equivariant graph neural networks. *Advances in Neural Information Processing Systems*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Kondor, R.; and Trivedi, S. 2018. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, 2747–2755. PMLR.

Krizhevsky, A.; Nair, V.; and Hinton, G. 2010. CIFAR-10 (Canadian Institute for Advanced Research). http://www.cs.toronto.edu/kriz/cifar.html. Accessed: 2022-06-05.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.

Lake, B.; and Baroni, M. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, 2873–2882. PMLR.

Lake, B. M. 2019. Compositional generalization through meta sequence-to-sequence learning. *Advances in Neural Information Processing Systems*, 32.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

Loula, J.; Baroni, M.; and Lake, B. 2018. Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 108–114.

Lu, K.; Grover, A.; Abbeel, P.; and Mordatch, I. 2021. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*.

Maron, H.; Ben-Hamu, H.; Shamir, N.; and Lipman, Y. 2019. Invariant and equivariant graph networks. In *International Conference on Learning Representations*.

Montague, R. 1970. Universal grammar. *Theoria*, 36(3): 373–398.

Murphy, R. L.; Srinivasan, B.; Rao, V.; and Ribeiro, B. 2018. Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. *arXiv preprint arXiv:1811.01900*.

Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5356–5371.

Prates, M. O.; Avelar, P. H.; and Lamb, L. C. 2020. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 32(10): 6363–6381.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8): 9.

Ravanbakhsh, S.; Schneider, J.; and Poczos, B. 2017. Equivariance through parameter-sharing. In *International Conference on Machine Learning*, 2892–2901. PMLR.

Ruder, S.; Peters, M. E.; Swayamdipta, S.; and Wolf, T. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 15–18.

Russin, J.; Jo, J.; O'Reilly, R. C.; and Bengio, Y. 2019. Compositional generalization in a deep seq2seq model by separating syntax and semantics. *arXiv preprint arXiv:1904.09708*.

Satorras, V. G.; Hoogeboom, E.; and Welling, M. 2021. E (n) equivariant graph neural networks. In *International Conference on Machine Learning*.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3407–3412.

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

Steed, R.; Panda, S.; Kobren, A.; and Wick, M. 2022. Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3524–3542.

Sturmfels, B. 2008. *Algorithms in Invariant Theory*. Springer Science & Business Media.

Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.-W.; and Wang, W. Y. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Taylor, M. E.; and Stone, P. 2009. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7).

Van Der Pol, E.; Worrall, D.; van Hoof, H.; Oliehoek, F.; and Welling, M. 2020. MDP homomorphic networks: Group symmetries in reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 4199–4210.

Wang, D.; Walters, R.; Zhu, X.; and Platt, R. 2022. Equivariant Q Learning in Spatial Action Spaces. In *Conference on Robot Learning*, 1713–1723. PMLR.

Williams, R. J.; and Zipser, D. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2): 270–280.

Yarotsky, D. 2022. Universal approximations of invariant maps by neural networks. *Constructive Approximation*, 55(1): 407–474.

Yeo, C.; and Chen, A. 2020. Defining and Evaluating Fair Natural Language Generation. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, 107–109.

Zamir, A. R.; Sax, A.; Shen, W.; Guibas, L. J.; Malik, J.; and Savarese, S. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3712–3722.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76.