CONTROLLABLE CONCEPT TRANSFER OF INTERMEDIATE REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

With the proliferation of large pre-trained models in various domains, transfer learning has gained prominence where intermediate representations from these models can be leveraged to train better (target) task-specific models, with possibly limited labeled data. Although transfer learning can be beneficial in many cases, it can also transfer undesirable information to target tasks that may severely curtail its performance in the target domain or raise ethical concerns related to privacy and/or fairness. In this paper, we propose a novel approach for controlling the transfer of user-determined semantic concepts (viz. color, glasses, etc.) in intermediate source representations to target tasks without the need to retrain the source model which can otherwise be expensive or even infeasible. Notably, this is also a bigger challenge than blocking concepts in the input representation as a given intermediate source representation is biased towards the source task it was originally trained to solve, thus possibly further entangling the desired concepts. We qualitatively and quantitatively evaluate our approach in the visual domain showcasing its efficacy for classification and generative source models.

1 INTRODUCTION

Deep neural networks (DNN) have achieved unprecedented performance in various computer vision, natural language (NLP) problems such as image classification (Sun et al., 2017; Mahajan et al., 2018), object detection (Girshick, 2015; Ren et al., 2015), segmentation (Long et al., 2015; He et al., 2017), question answering (Min et al., 2017; Chung et al., 2017), and machine translation (Zoph et al., 2016; Wang et al., 2018) etc. One of their strengths is the ability to learn task-specific hidden representations rather than relying on predefined image features. In an ideal scenario, there is an abundance of labeled training samples to learn a good hidden representation. However, it is often expensive, time-consuming, or unrealistic to collect sufficient training data. In such scenarios, transfer learning (Pan & Yang, 2009) has emerged as one of the promising learning paradigms. Transfer learning utilizes knowledge from information-rich source tasks to learn a specific (often information-poor) target task.

One of the most widely used approaches for transfer learning is *fine-tuning* (Sharif Razavian et al., 2014) where the target DNN being trained is initialized with the weights of a source DNN that has been pre-trained on a large dataset from a related task. Another popular approach involves matching/combining the hidden representation or the gradient of the output of the target model with that of the source model (Jang et al., 2019; Li et al., 2018; Murugesan et al., 2022). These approaches are extensively used in improving prediction performance and robustness of many vision and NLP tasks (Hendrycks et al., 2019; Devlin et al., 2018), while reducing training time and resources. However, transferring from large pre-trained models (fine-tuning or representation transfer) could propagate undesirable *concepts* encoded in source models to downstream tasks. For example, a source model, trained to classify cats vs dogs, with most cat images in gray-scale and dog images in color, could incorrectly associate the concept of *color* to the images of the dog and pass this biased knowledge to downstream tasks. In real-world applications, this could have serious consequences. Among several examples, (Steed & Caliskan, 2021) showed that embeddings extracted from pre-trained image models exhibit racial and gender bias that they learn from training datasets. Similarly, (Kennedy et al., 2020) demonstrated that hate speech classifiers finetuned from BERT (Devlin et al., 2018) resulted in frequent false positives when certain group identifiers (e.g., *Muslim*, *black*) were mentioned in the text.

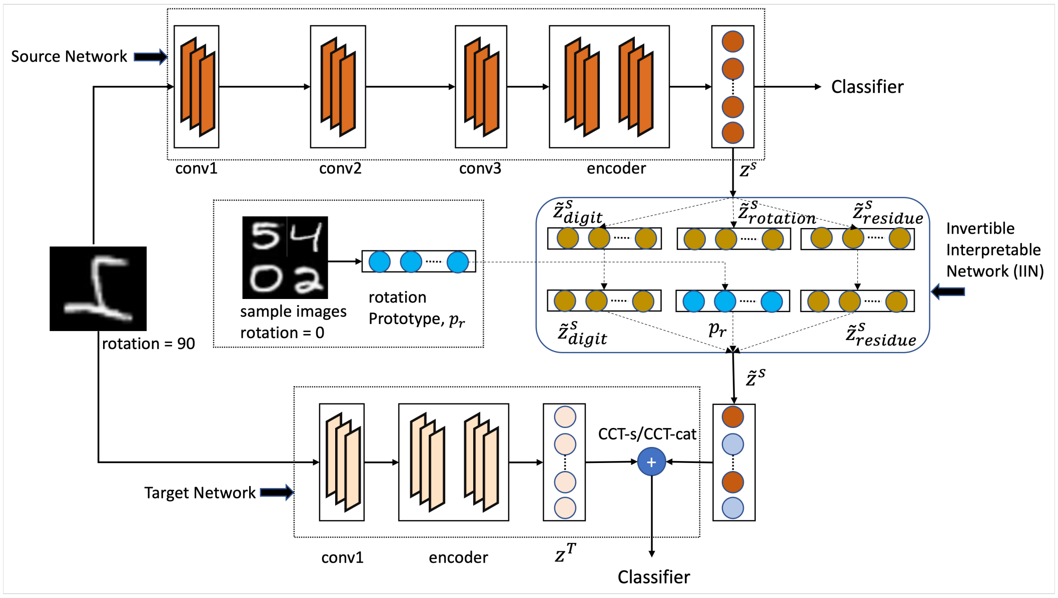

Figure 1: Illustration of our proposed approach on rotated-MNIST dataset. We show how the $rotation$ concept is blocked from the image before transferring it to the target task. First, hidden representation z^s from a pre-trained source network is factorized using IIN into $z^s_k, k \in K = \{digit, rotation, residue\}$. To block the $rotation$ concept, the factor for $z^s_{rotation}$ is set to a rotation $prototype, p_r$ generated using a few sample images with no rotation. Now the IIN is used to invert the modified factors to $\tilde{z}^s$ which can either be directly used *Controllable Concept Transfer-source (CCT-s)* or concatenated with the target representation z^T *Controllable Concept Transfer-concatenate (CCT-cat)* to train the target task.

While there are several approaches to *mitigate* the impact of unintended knowledge transfer in target models, ranging from data augmentation that balances target datasets (Park et al., 2018; Dixon et al., 2018) to adversarial training for generating robust hidden representations against certain spurious concepts (Zhang et al., 2018; Wang et al., 2020; Fan et al., 2021), the controllable transfer using the intermediate representation of the source model has been largely unexplored. Typically, large pretrained (source) models are learned with imbalanced/biased data, and retraining these models to remove undesirable concepts might not be ideal. Our work takes a novel concept-based knowledge transfer approach to address this problem where we address the following question:

How can we most effectively ***control*** *the intermediate representation of a source model by blocking a specific concept while keeping other concepts (largely) intact before transferring to downstream tasks?*

Towards this goal, we propose a transfer learning method, Controllable Concept Transfer (CCT), to block the undesirable concepts in the hidden representation of the source model before transferring to a downstream task. Note that this is a challenging problem than simply blocking concepts directly in the input representation since the intermediate source representation and the concepts extracted from it could be biased towards the source task. To address this problem, we propose a novel approach using Invertible Interpretable Network (IIN) (Esser et al., 2020) to disentangle the concepts in the source representation and adapt them to the target task, which to the best of our knowledge, has not been previously done in transfer learning. We propose two transfer learning settings and demonstrate our approach to image classification tasks using two real-world datasets. In both these settings, we find that our approach successfully blocks concepts from the source intermediate representation.

Figure 1 illustrates our approach to block the concept of *rotation* from the source model using IIN explained later. We evaluate the performance of controllable concept transfer both qualitatively and quantitatively. In addition to the accuracy for evaluation of target task performance, we adopt a mutual information-based metric based on MINE (Belghazi et al., 2018) to quantify the measure of concept removal. Our qualitative analysis presents decoded images with different blocked concepts.

2 RELATED WORK

Transfer Learning from a large pretrained source model is a well-known approach to learning target tasks with limited labels (Pan & Yang, 2009). One of the most common transfer learning techniques is fine-tuning a pretrained source model (Sharif Razavian et al., 2014), where network layers from the source model are frozen, and a new classifier head is trained for the target task. Recent works align the source and target features to transfer relevant knowledge - either by matching network weights (Xuhong et al., 2018; Jang et al., 2019), attention maps (Li et al., 2018; Zagoruyko & Komodakis, 2016), Jacobians (Srinivas & Fleuret, 2018) or model reprogramming (Chen, 2022). Another line of work uses the source model to better guide the target network by transferring feature maps automatically to improve the target task performance (Murugesan et al., 2022). In all the above methods, the transferred knowledge is typically not interpretable. To understand what knowledge is being transferred from source and target networks, a few methods use attention maps to visualize the key features from the source model useful for the target task (Murugesan et al., 2022; Jang et al., 2019) or a series of experimental analyses on the finetuned target model to study the importance of transferred knowledge (Neyshabur et al., 2020). In this paper, we take a different approach to transfer learning and propose a principled way of controlling what semantically meaningful *concepts* can be transferred from the source model to the target task.

With an increasing interest in **Model Interpretability**, several approaches have been proposed to understand the inner workings of deep neural classifiers, specifically through human understandable high-level concepts as activation vectors (Kim et al., 2018; Zhou et al., 2018; Chen et al., 2020), or individual neurons (Erhan et al., 2009; Olah et al., 2017; Zeiler & Fergus, 2014; Bau et al., 2017). However, the representation of the semantic concepts is distributed across the hidden layers of the network (Fong & Vedaldi, 2018) and none of these methods can (confidently) claim that the features (i.e. neurons) identified from intermediate representations are associated *only* with the specific concept and are largely independent of other concepts (Montavon et al., 2017; Yosinski et al., 2015). A related line of work trains the models that explicitly encode concepts in their intermediate representations (Koh et al., 2020; Chen et al., 2020; Losch et al., 2019). However, this approach alters the network architecture and typically deteriorates overall performance (Zhou et al., 2016). Unlike these works, we propose a novel approach to transfer learning by blocking or allowing the relevant concepts transferred from the source model to the target network for better interpretability.

Unlike DNNs, **Generative Models** are trained with the explicit goal to produce images from samples of a specific distribution. Variational auto-encoders (Kumar et al., 2018; Higgins et al., 2017) reconstruct images from a representation whose marginal distribution is matched to a standard normal distribution. Generative Adversarial Networks (GAN) (Goodfellow et al., 2020; Hoang et al., 2018) map samples from a standard normal distribution to realistic images as judged by a discriminator. While these approaches are invertible, they are not interpretable, limited to representations with a linear structure, and cannot be applied to arbitrary representations from a source network. This motivates our choice of Invertible Neural Networks (Dinh et al., 2014; Jacobsen et al., 2018; Kingma & Dhariwal, 2018; Esser et al., 2020) for our transfer learning problem setup as they can identify disentangled concepts for interpretability, invert them and map them back to relevant features in the intermediate representations of the source model.

3 CONTROLLABLE CONCEPT TRANSFER METHOD

It has been widely observed that machine learning models learn context-specific correlations in datasets. For example, a model trained to classify different indoor scenes would learn to associate the presence of a *bed* to the output class of “bedroom” vs *couch* to the output class of “living-room”. It is postulated that such high-level semantic concepts or representations are useful to differentiate them (Neyshabur et al., 2020), and it is further possible to reuse these learned patterns to generalize to new (related) tasks by transferring representations to downstream tasks. However, spurious associations in the transferred knowledge could hinder the performance of a target task. For instance, an *accent chair* could be exclusively associated with “living-room” and when the model encounters a novel environment where it is present in “bedroom”, the source representation could be biased to classify the input as “living-room”. The transfer learning method to modify the source intermediate representation to block a certain *concept* would prove useful in such situations. The canonical way to block a certain concept from hidden representation would be to retrain the model with new images that satisfy the desired constraint. However, it is hard to predict how the retraining affects other related concepts. The goal of our work is to tackle this problem *directly* and modify the hidden representation of the source model in a targeted manner. For instance, in our earlier example, we would ideally want to block the concept of *accent chair* without affecting other concepts in the hidden representation of the source model before transferring it to the downstream task. Next, we provide a brief overview of the Invertible Neural Network and a recent work by Esser et al. (2020) which motivates our framework, after which we detail our work. Throughout this paper, we use the terms "intermediate representation" and "hidden representation" interchangeably.

3.1 BACKGROUND: DISENTANGLING SEMANTIC CONCEPTS IN INTERMEDIATE REPRESENTATION

Let f be the given neural network with L layers that maps the input image $x \in \mathbb{R}^{h \times w \times c1}$ through a series of hidden layers to the final output $f(x)$. Often, intermediate representation $E(x) \in \mathbb{R}^{H \times W \times C2}$ at a hidden layer does not convey any semantic meaning knowledge. Often the mapping from the intermediate representation to semantically meaningful representation is well-defined, whereas the inverse is not straightforward. In this paper, we are interested in invertible representation learning that maps intermediate representation to a semantically meaningful concepts and vice verse. Esser et al. (Esser et al., 2020) developed an approach titled Invertible Interpretable Network (IIN) to factorize the hidden representation from a model into user-defined semantic concepts. Specifically, they map arbitrary representations into a space of interpretable representations – a non-linear mapping between the two domains. This mapping is invertible, i.e., any modification in the domain of semantic concepts concurrently alters the original representation. In this scenario, it takes the flattened version of $E(x)$, denoted $z \in \mathbb{R}^N$ (where $N = H \cdot W \cdot C$), and factorize it into $\tilde{z} = (\tilde{z}_k)_{k=0}^{K} \in \mathbb{R}^N$, where each of the $K + 1$ factors of $\tilde{z}_k \in \mathbb{R}^{N_k}$ with $\sum_{k=0}^{K} N_k = N$ represents an interpretable concept that is normally distributed $\mathcal{N}(\tilde{z}_k|0, \mathbf{1})$. Calling this transformation I, we have $\tilde{z} = I(z)$.

To encode semantic representation into each factor $\tilde{z}_k$, they constrain (i) each factor $\tilde{z}_k$ to vary with exactly one interpretable concept and (ii) $\tilde{z}_k$ to be invariant to all other variations. This is ensured through training pairs (x^a, x^b), which specify semantics through similarity, i.e. image pairs that both have a semantic concept of *accent chair* in them. Each semantic concept, indexed by $F \in \{1, ..., K\}$, has image pairs $(x^a, x^b) \sim p(x^a, x^b|F)$ to the corresponding factor $\tilde{z}_F$. To capture the remaining variability that is not captured by the K concepts, a residual concept $\tilde{z}_0$ is introduced. This ensures that any change made to the factorized semantic concept $\tilde{z}_k$ is reflected in the original representation space. Calling this transformation I^{-1}, we have $I^{-1}(\tilde{z}) = z$. Intuitively, the goal is to have a bijective mapping so that modifications of the disentangled semantic factors correctly translate back to the original representation. Please refer to Algorithm 2 and (Esser et al., 2020) for further details.

3.2 CONCEPT BLOCKING AND TRANSFER

In this section, we focus our attention on controlling the concept transfer by blocking undesirable concepts in the hidden representation of the source classifier and transferring relevant concepts to the downstream task. To describe our approach, let's take a simple example of *rotation* concept added to MNIST images (LeCun et al., 1998). At the high level, the goal is to take the hidden representation at a layer L of the pre-trained model and block the *rotation* concept without affecting other concepts.

How to block a concept? Let us assume that we have a pre-trained source network f^s that takes input images x from a target task and produces hidden representation at layer $L-1$, $f_{L-1}^s(x) = z^s$, i.e., the layer before classifier head c_L^s. Let us define two semantic concepts specific to the target task as *digit* and *rotation*. We first train the IIN I to take the hidden representation z^s and factorize it according to concepts such that $\tilde{z}_k^s = (I(z^s))_k$, where $k \in \{digit, rotation, residue\}$. Training is done using pairs of images that contain a common concept, i.e., the same *digit* to map $\tilde{z}_{digit}^s$ or the same *rotation* for $\tilde{z}_{rotation}^s$. In addition, there is a *residue* factor $\tilde{z}_{residue}^s$ that encodes all other variations unaccounted by these concepts. Since the IIN imposes a one-to-one mapping from the

[1] Where h, w, c are height, width and channel dimensions of input image

[2] Where H, W, C are height, width, and channel dimensions of intermediate representation

original representation space (z^s) to a factorized space ($\tilde{z}^s$), we can edit the factorized representation $\tilde{z}^s_{rotation}$ without affecting the other factors. Given the IIN I, suppose one wants to block the *rotation* concept. We sample a few example images $\{r_1, ..., r_n\}$ which are not *rotated* and pass them through our pre-trained source network and IIN to obtain their *rotation* embedding and take the mean to create a $prototype$ embedding, $p_r = \frac{1}{n}\sum_{i=1}^{n}(I(f^s_{L-1}(r_i))_{rotation})$, which is indicative of absence of the rotation concept.

How to transfer? Next, we proceed to training the target model f^t that takes as input images x and maps to a hidden representation at layer $L{-}1$, $f^t_{L-1}(x) = z^t$, the layer before classifier head c^t_L. The input image is also passed through the source model to get the source intermediate representation, which is then fed to IIN. The *rotation* concept is blocked by replacing the corresponding factor $\tilde{z}^s_{rotation}$ with the $prototype, p_r$. The updated hidden representation $I^{-1}(\tilde{z}^s)$ is then transferred to the target classifier c^t_L. We consider two variations of transferring knowledge from the source model to the target task,

1. **Controllable Concept Transfer - source** (CCT-s) where we freeze the layers up to $L-1$ of source network, attach a classifier head for target task c^t_L and train the classifier head for the target task with updated source representation, $c^t_L(I^{-1}(\tilde{z}^s))$.
2. **Controllable Concept Transfer - concatenate** (CCT-cat) where we concatenate the updated source representation with that of the pre-trained target network before passing it through the target classifier, $c^t_L([I^{-1}(\tilde{z}^s) \oplus z^t])$ where $\oplus$ represents concatenation operation.

The entire pipeline is presented in Figure 1 and Algorithm 1 as CCT-cat (CCT-s follows similarly but would remove the target network f^t and combination operation in Step 15). Note that this approach works for multiple concepts by simply generating each $prototype$, p_i and editing the corresponding factorized representation $\tilde{z}^s_i$ for concept i.

Algorithm 1 Controllable Concept Transfer - concatenate (CCT-cat) method

```
 1: Inputs: Target training dataset $D_T$; Target classifier loss $\mathcal{L}_{c^t}(\cdot)$; Combination operation $\bigoplus$; Seed weight
    parameters: $\mathcal{W}_{c^t}[0]$; Source pre-trained network $f^s$, Target pre-trained network $f^t$; Number of Epochs $E$,
    Layer $L$, Concepts $K$ to block.
 2: $I \leftarrow$ TRAIN-IIN
 3: for $concept \in [1:K]$ do
 4:     Randomly sample $n$ images $\{r_1, r_2, ...r_n\}$ without concept.
 5:     $p_{concept} \leftarrow \frac{1}{n}\sum_{j=1}^{n}(I(f^s_L(r_j))_{concept})$
 6: end for
 7: Randomly shuffle $D_T$.
 8: for $epoch \in [1:E]$ do
 9:     for $batch \in D_T$ do
10:         $x \leftarrow D_T[batch]$.
11:         $\tilde{z}^s \leftarrow I(f^s_L(x))$
12:         for $concept \in [1:K]$ do
13:             $\tilde{z}^s_{concept} \leftarrow p_{concept}$
14:         end for
15:         $\mathcal{W}_{c^t}[batch] \leftarrow \mathcal{W}_{c^t}[batch-1] - \eta_{batch}\nabla_{\mathcal{W}_{c^t}}\mathcal{L}_{c^t}(f^t_L(x) \bigoplus I^{-1}(\tilde{z}^s))$
16:     end for
17: end for
18: Output: Trained model $c^t_L$ with last iterate of $\mathcal{W}_{c^t}$
```

4 Evaluation on Classification tasks

In this section, we consider how concept blocking affects the performance of a target model. In particular, we present two scenarios of image classification tasks: (i) Transfer from the rotated-EMNIST trained source model to the rotated-MNIST classification task, and (ii) Transfer between CelebFaces attribute classifiers. For both experiments, we use a deeper 6-layer convolutional neural network (CNN) for the source model and a 3-layer CNN for the target model. Additional details about the experimental setup and datasets can be found in Appendix A.3. For each experiment, we consider two variants of transfer: Controllable Concept Transfer - source (CCT-s) and Controllable

Table 1: Mean accuracy (over three runs) for rotated-EMNIST to rotated-MNIST transfer task. Experiments are conducted by varying the proportion (%) of rotated samples {90, 180, 270} in the training dataset from 1% to 75%. We compare the performance of three models: Target only (TG), CCT-s and CCT-cat. For CCT-s and CCT-cat we conduct experiments without blocking any concept (noedit) vs blocking rotation (edit).

Method	Fraction (%) of rotated images in target task dataset												
	1	2	3	4	5	6	7	8	9	10	25	50	75
TG	0.52	0.58	0.62	0.65	0.68	0.70	0.71	0.73	0.74	0.75	0.84	0.88	0.90
CCT-s(noedit)	0.57	0.59	0.61	0.62	0.63	0.64	0.65	0.66	0.67	0.68	0.75	0.80	0.81
CCT-cat(noedit)	0.61	0.65	0.68	0.70	0.72	0.73	0.75	0.76	0.77	0.78	**0.86**	**0.90**	**0.91**
CCT-s(edit)	**0.68**	**0.68**	0.69	0.69	0.69	0.69	0.69	0.69	0.70	0.70	0.73	0.75	0.77
CCT-cat(edit)	0.64	0.68	**0.71**	**0.73**	**0.74**	**0.76**	**0.76**	**0.77**	**0.78**	**0.79**	0.85	0.89	0.90

Concept Transfer - concatenate (CCT-cat) and compare the performance against training the target model independently (TG).

4.1 Transfer between heterogeneous datasets

In this experiment, we study the effect of blocking the concept of $rotation$ in a rotated-EMNIST trained source model to a rotated-MNIST transfer task. First, we pre-train the source model with the rotated-EMNIST dataset, where the goal is to classify 26 different English letters. Next, we train the IIN to factorize the concepts of $digit$ and $rotation$ using layer $L-1$ representation of the pre-trained source model by probing it with rotated-MNIST images as inputs. Finally, we train the target task of classifying 10 digits without blocking any concepts, termed CCT-s(noedit) and CCT-cat(noedit), and compare the performance to $rotation$ blocked transfer, termed CCT-s(edit) and CCT-cat(edit). To force the target model to rely on the source model for the $rotation$ concept, we vary the amount of rotated training samples in the target dataset from 1% to 75%. We assume that, at lower percentages of rotated training samples in the target dataset, the target model relies on the source model for a good representation of the $rotation$ concept. Top-1 accuracy for these experiments is presented in Table 1 with maximum accuracy in bold for each column. Each experiment is repeated thrice and the mean accuracy is presented. As expected, at lower percentages of rotated samples ($\leq$ 10), we see that blocking the concept of $rotation$ boosts the performance of transfer with CCT-s(edit) and CCT-cat(edit) performing best. As the percentage of rotated samples in the training data increases, the target model learns a better representation of the $rotation$ concept and relies less on the source model. This is evidenced by TG and CCT-cat(noedit) having comparable performance.

4.2 Evaluation on CelebFaces Attributes

In this experiment, we consider a homogeneous setup where the transfer is done between two CelebA tasks. We train the source model to identify 4 different concepts, *Smiling*, *Wearing_Lipstick*, *Heavy_Makeup*, and *High_Cheekbones*, based on a multi-label classifier. We then train an IIN to factorize these concepts using layer $L-1$ representation of the source network. Finally, we train the target binary classification task to identify if the given image is labelled $Male$ or Not_Male. Specifically, we train several models with/without blocking concepts, CCT-s(noedit), CCT-cat(noedit), CCT-s(edit), and CCT-cat(edit), and compare against an independently trained target model (TG).

	Blocked concept				
Method	Smiling	Wearing Lipstick	Heavy Makeup	High Cheekbones	Both Makeup
TG	0.9432				
CCT-s (noedit)	0.9313				
CCT-cat (noedit)	**0.9466**				
CCT-s (edit)	0.9210	0.6113	0.8761	0.9210	0.5984
CCT-cat (edit)	0.9462	0.9401	0.9444	0.9461	0.9401

Table 2: Mean accuracy (over three runs) for CelebA transfer task. Experiments are conducted by blocking concepts *Smiling*, *Wearing_Lipstick*, *Heavy_Makeup*, *High_Cheekbones* one at a time and blocking both *Wearing_Lipstick*, *Heavy_Makeup* at once (*Both_Makeup*). As before we compare the performance of three models: TG, CCT-s, and CCT-cat. For CCT-s and CCT-cat we perform experiments with (edit) and without (noedit) concept blocking.

As seen in Table 2, just using the source model CCT-s(noedit) gives a similar performance to the independently trained target model (TG). Blocking the concepts of $Smiling$ and $High\_Cheekbones$ do not have a big impact on CCT-s(edit) performance suggesting that these concepts are less relevant to the target task. However, blocking $Wearing\_Lipstick$ and $Heavy\_Makeup$ individually causes a drop in CCT-s(edit) performance. Next, we blocked both concepts simultaneously and found that the performance dropped further. This suggests that the target classifier is relying on concepts such as $Wearing\_Lipstick$ and $Heavy\_Makeup$ for classifying $Male$ vs Not_Male, and for fairness purposes, we need to block these concepts from being transferred to the target model.

5 How well does Concept Blocking work?

In the previous section, we showed that blocking concepts influence the performance of the target task. Here, we explore how well the concepts are blocked from source hidden representation. In particular, we consider two sets of experiments. (i) In the quantitative experiments, we adopt mutual information (MI) based metric (Belghazi et al., 2018) to measure the MI between concept and hidden representation before and after blocking a concept. In both cases, we edit the model using a single $prototype$ to block one concept at a time. (ii) In the qualitative experiments, we use an autoencoder architecture as a source model, which facilitates visualization of concept blocking by decoding IIN edited hidden representations to human understandable images.

5.1 Quantitative experiments using mutual information

The goal of this experiment is to quantitatively assess whether information about undesired concepts is contained or 'hiding' in transferred representations. In order to formally evaluate the performance of concept blocking in this scenario, we employ an information-theoretic analysis of the transferred representations. Specifically, we measure the mutual information between a specific concept and the hidden representations. To compute mutual information between concepts and neural network representations, we adapt the *mutual information based neural estimator* (MINE) proposed by Belghazi et al. (2018), where the authors present a way to estimate mutual information between high dimensional random variables using a trainable neural network that they term a $statistics\ network$. In simple words, given two random variable X and Z, the authors propose a neural information measure defined as,

$$I_\Theta(X,Z) = \sup_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}_{\mathbb{XZ}}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}_X \bigotimes \mathbb{P}_Z}[e^{T_\theta}])$$

where the expectations are estimated using samples drawn from $\mathbb{P}_{XZ}$ and $\mathbb{P}_X \bigotimes \mathbb{P}_Z$ while the objective is maximized by gradient ascent. Please refer to (Belghazi et al., 2018) for more details.

In the original paper, Belghazi et al. (2018) used X to represent the input image and Z to represent the latent representation. For our purposes, we are interested in estimating the mutual information (MI) between the concept of a random variable and the hidden representation. In particular, we estimate the MI between the concept C and original hidden representation z^s, $I(C, z^s)$, and compare with the MI between concept C and edited hidden representation $\tilde{z}^s$, $I(C, \tilde{z}^s)$. We start with a heterogeneous setup where the source model is trained on the rotated-colored-EMNIST dataset (Cohen et al., 2017), where each colored image is $rotated$ by a random angle drawn from $r \in \{90, 180, 270\}$. However, we probe the source model with the rotated-colored-MNIST dataset for MI estimation. Here, MI is measured for $color$, $rotation$ and $digit$ with respect to hidden representations – before and after blocking $color$ and $rotation$ concepts. To estimate MI, we use the color RGB vector, and angle of rotation as proxy concept random variables C. These results are presented in Figure 2(a). We see that the MI for $digit$ does not change much when blocking the other two concepts. However, the MI for $color$ and $rotation$ drop significantly when the respective concepts are blocked.

We proceed to conduct experiments on the CelebA (Liu et al., 2015) dataset, where the source model is trained on a multi-label classification task to identify 4 different concepts $\{Smiling, Wearing\_Lipstick, Heavy\_Makeup, High\_Cheekbones\}$. Next, each concept is blocked and MI is measured for every concept. To estimate MI, we use the binary value of the presence/absence of each concept as a proxy random variable. Results are presented in Figure 2(b). As demonstrated by the plots, the MI for the blocked concept reduces (almost) to zero in most cases. These findings suggest that the selected concept is blocked from the updated hidden representation without significantly affecting the other concepts. Additionally, our choice of blocking concepts by editing the final layers in the source is also justified by looking at Figure 3, where we see that more complex concepts appear only in the later layers of the source network.

5.2 Comparing against concept activation based blocking (CAB)

To test the need for IINs to block concepts in hidden representations, we conduct an experiment with concept activation-based blocking (CAB) motivated by previous works (Kim et al., 2018; Zhou

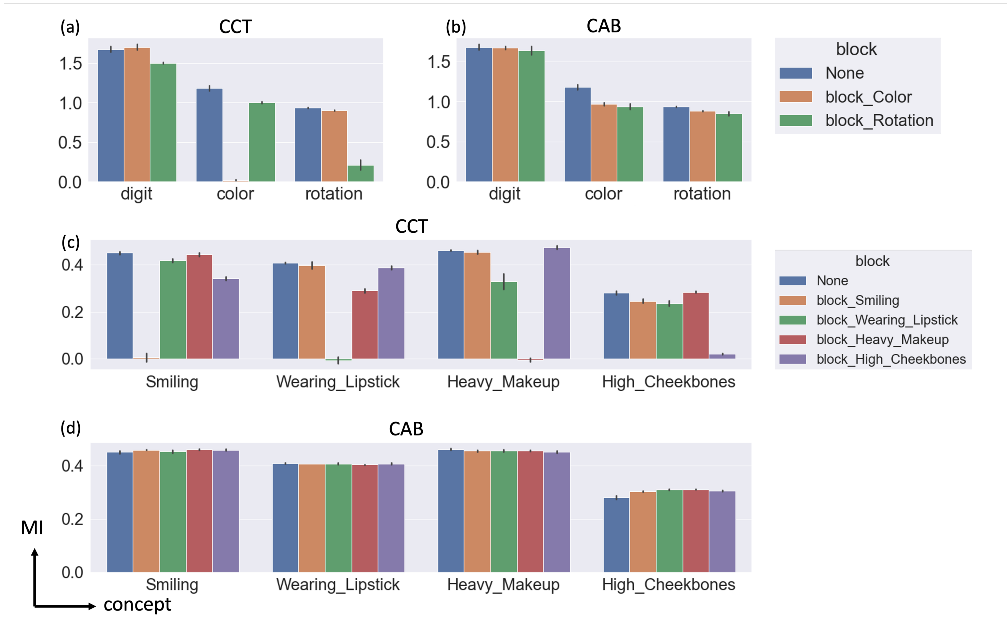

Figure 2: Experiments on mutual information (MI) between concepts and intermediate representation. Each group of bars represents a concept and the color of the bars represents the blocked concept. MI for rotated-colored-MNIST dataset $digit$, $color$, and $rotation$ concepts when $color$ and $rotation$ are blocked using (a) CCT method and (b) CAB method. MI for CelebA dataset where source model is trained on a multi-label classification task to identify 4 different concepts $\{Smiling, Wearing\_Lipstick, Heavy\_Makeup, High\_Cheekbones\}$. Each concept is then blocked individually and MI is reported for all concepts using the (c) CCT method and (d) CAB method.

et al., 2018; Chen et al., 2020). In this experiment, for each concept, we train a linear classifier based on logistic regression where the hidden representation from the source network is used as input and each model predicts if the concept is present/absent. We then identify the top neurons for each concept based on model coefficients and use them to create the $prototypes, p_i$, for each concept. To block a concept i, we just set those candidate neurons to the corresponding p_i during transfer. We tested this approach in rotated-colored-MNIST and CelebA by measuring MI values similar to previous experiments. Results are presented in Figure 2(b, d). As demonstrated, the MI for all concepts remains high even after using CAB to block them. This suggests that the information about the concept is still transferred after directly blocking in the hidden representation space. On the other hand, blocking through the factorized representation of IIN leads to much less information transfer.

5.3 Qualitative experiments using autoencoder architecture

To visually evaluate concept blocking, we replace the source model with an autoencoder. Consider that f^s takes the input image $x \in \mathbb{R}^{h\times w\times c}$ and encodes it into the hidden representation $z = f^s(x) \in \mathbb{R}^N, N = H \cdot W \cdot C$. A decoder d^s is then used to map the hidden representation back to original image space $\tilde{x} = d^s(z) \in \mathbb{R}^{h\times w\times c}$. First we consider a simple dataset where we add the concept of *color* to the MNIST dataset (colored-MNIST) (LeCun et al., 1998). This is done by multiplying RGB values in gray-scale MNIST images. A source model is trained to reconstruct the input images and IIN is trained to disentangle the concepts of $digit$ and $color$. We then edit the color factor of a few randomly drawn sample images with the $prototype$ p_c of $color$ as depicted in Figure 4(a). Next, we consider a more realistic dataset, CelebA (Liu et al., 2015), which is a large-scale dataset of celebrity faces with attributes. We train the source model to reconstruct CelebA images and then train the IIN to disentangle 3 concepts – *Eyeglasses*, *No_Beard*, and *Smiling*. We then proceed to edit these concepts by using corresponding $prototypes, p_i$. Results for blocking the $Eyeglasses$ concept in a few randomly drawn sample images are presented in Figure 4(b). Additional figures for blocking concept of $Smiling$ and No_Beard are presented in Figure 8. As demonstrated in Figure 4(a), our method is able to block the concept of *color* without affecting the *digit* concept. In Figure 4(b), we see that the concept of $Eyeglasses$ is successfully blocked from sample images, demonstrating that this approach can be used to block complicated concepts.

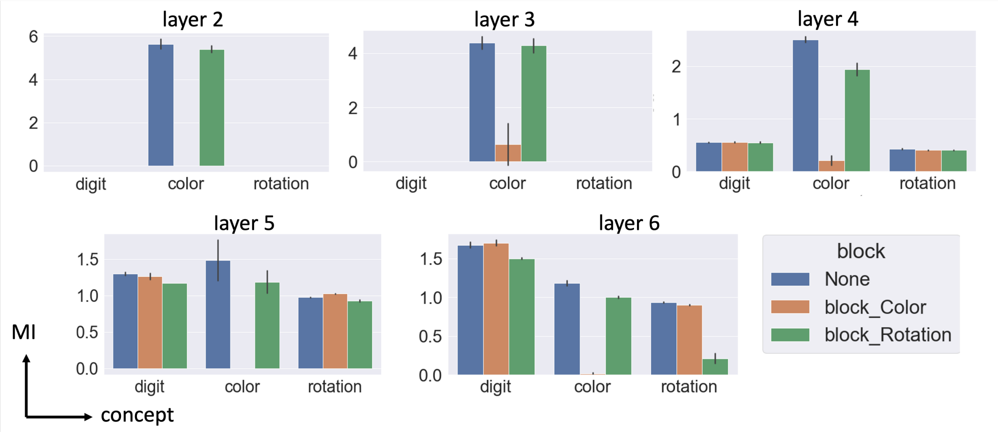

Figure 3: Experiment comparing concept blocking using hidden representations from different layers for the rotated-colored-MNIST dataset. As can be seen the $color$ concept is learned first followed by $digit$ and $rotation$. It is thus best to intervene (i.e. block) in the final layer (i.e. layer 6) where the model has learned the critical concepts such as $digit$ which we want to transfer.

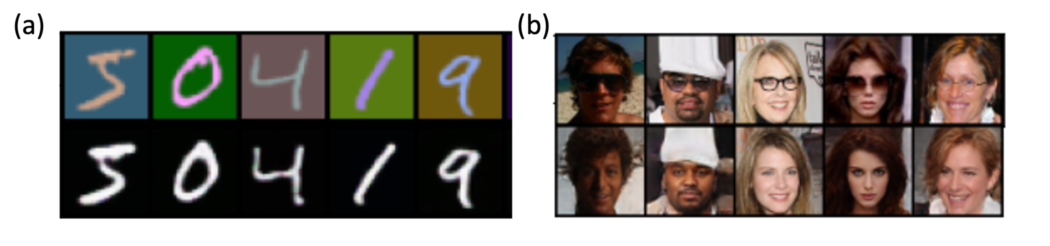

Figure 4: Visualization of concept blocking using an autoencoder as source network. (a) In the first row, we present randomly drawn colored-MNIST images. We then proceed to block the concept of $color$ from these images and show them in the second row. (b) For CelebA dataset, we consider randomly drawn samples with the $Eyeglasses$ attribute and proceed to block them. Examples of other concepts being blocked are in the appendix.

6 Discussion

We have seen in this paper how one can effectively block certain semantic concepts from being transferred from a source model to a target model while allowing other concepts and information to be transferred. While we were (largely) successful in this endeavor, there may be situations where it is difficult to block a certain concept, while allowing others to pass. The reason for this is that concepts can be (statistically or causally) correlated and so blocking one will lead to also (at least partially) blocking the other. For instance, it may be impossible to block the *Smiling* concept while allowing *High Cheekbones* to be transferred. In the future, it would be interesting to consider such cases where the user not only determines which concepts to block but also which to specifically pass and to arrive at a strategy that best satisfies these requirements. The strategy may also involve informing the user that the constraints are impossible to satisfy. We blocked concepts by setting the corresponding latent vector in the IIN disentangled representation to mean/median values based on images lacking that concept and then inverting back to the source intermediate representation. However, there may be other ways to set these values, possibly taking inspiration from the explainable AI and fairness literature (Došilović et al., 2018; Mehrabi et al., 2021), where for methods such as SHAP (Lundberg & Lee, 2017) and MACEM (Dhurandhar et al., 2019), there are different ways to determine null/base values indicative of no information. To summarize, we have provided a novel approach to block desired semantically meaningful concepts in the transfer learning setting with applications to interpretability, fairness, and privacy. We have evaluated our approach both qualitatively, through intuitive visual examples, and quantitatively, based on an (adapted) mutual information metric, highlighting our method's efficacy.

Ethics and Reproducibility Statement

Transfer learning is a very well-studied research area and yet the clear understanding of what, where, and how the knowledge is transferred from a source task to a target task is still largely unexplored. With the wide adoption of deep learning technologies, explaining or understanding the reasons behind their decisions has become extremely important in many critical applications (Arya et al., 2019). This work sheds some light on explaining what knowledge is transferred from the source network to the target task by controlling it via human-understandable concepts. Of course, our method may not be perfect in blocking these concepts and some information leakage is possible. Nonetheless, we provide a human-in-the-loop avenue for controlling what semantically meaningful information may get transferred.

Experimental details are provided in Sections 4 and 5 of the main paper and Appendix A.3. All datasets are public. Code will be provided during the discussion phase through an anonymized link.

References

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Pin-Yu Chen. Model reprogramming: Resource-efficient cross-domain machine learning. *arXiv preprint arXiv:2202.10629*, 2022.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Yu-An Chung, Hung-Yi Lee, and James Glass. Supervised and unsupervised transfer learning for question answering. *arXiv preprint arXiv:1711.05345*, 2017.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Amit Dhurandhar, Tejaswini Pedapati, Avinash Balakrishnan, Kartik Ahuja Pin-Yu Chen, Karthikeyan Shanmugam, and Ruchir Puri. Model agnostic contrastive explanations for structured data. https://arxiv.org/abs/1906.00117, 2019.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, 2018.

Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 0210–0215, 2018. doi: 10.23919/MIPRO.2018.8400040.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

Patrick Esser, Robin Rombach, and Bjorn Ommer. A disentangling invertible interpretation network for explaining latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9223–9232, 2020.

Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, and Chuang Gan. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? *Advances in Neural Information Processing Systems*, 34:21480–21492, 2021.

Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8730–8738, 2018.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2712–2721, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR 2018*, 2017.

Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. Mgan: Training generative adversarial nets with multiple generators. In *International conference on learning representations*, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-revnet: Deep invertible networks. *arXiv preprint arXiv:1802.07088*, 2018.

Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Learning what and where to transfer. In *International Conference on Machine Learning*, pp. 3030–3039. PMLR, 2019.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing hate speech classifiers with post-hoc explanation. *arXiv preprint arXiv:2005.02439*, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

A. Kumar, P. Sattigeri, and A. Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *ICLR 2018*, 2018.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, and Jun Huan. Delta: Deep learning transfer using feature map with attention for convolutional networks. In *International Conference on Learning Representations*, 2018.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Max Losch, Mario Fritz, and Bernt Schiele. Interpretability beyond classification output: Semantic bottleneck networks. *arXiv preprint arXiv:1907.10882*, 2019.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31, 2018.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

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

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 2021.

Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. Question answering through transfer learning from large fine-grained supervision data. *arXiv preprint arXiv:1702.02171*, 2017.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017.

Keerthiram Murugesan, Vijay Sadashivaiah, Ronny Luss, Karthikeyan Shanmugam, Pin-Yu Chen, and Amit Dhurandhar. Auto-transfer: Learning to route transferrable representations. *arXiv preprint arXiv:2202.01011*, 2022.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *arXiv preprint arXiv:2008.11687*, 2020.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99, 2015.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014.

Suraj Srinivas and François Fleuret. Knowledge transfer with jacobian matching. In *International Conference on Machine Learning*, pp. 4723–4731. PMLR, 2018.

Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 701–713, 2021.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

Yijun Wang, Yingce Xia, Li Zhao, Jiang Bian, Tao Qin, Guiquan Liu, and Tie-Yan Liu. Dual transfer learning for neural machine translation with marginal distribution regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8919–8928, 2020.

LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*, pp. 2825–2834. PMLR, 2018.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 119–134, 2018.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*, 2016.