FEATURE DISENTANGLEMENT FOR COVARIATE SHIFT ADAPTATION IN FEDERATED LEARNING

Le Peng, Gaoxiang Luo, Andrew Walker *
Department of Computer Science and Engineering
University of Minnesota, Twin Cities
Minneapolis, MN 55455, USA
{peng0347,luo00042,walk0655}@umn.edu

ABSTRACT

We aim to pioneer disentangled representation learning in building robust federated learning (FL) models in various computer vision (CV) problems. Disentangled feature representations are essential for the global model of FL to generalize well on the unseen domains (i.e., out-of-distribution (OOD) generalization). This research is societally important to quicken the deployment of most state-of-the-art CV models toward covariant-shift inputs, such as medical images with different modality parameter setups (e.g., brightness, Hounsfield Units). We proposed a content-style-based FL scheme for the image classification task in this work. Our results show that the proposed FL scheme outperforms baseline FedAvg in OOD generalization across multiple domains, as well as image reconstruction for successful disentanglement assurance of domain-invariant features and domain-specific features.

1 INTRODUCTION

Many models, especially deep models, are data-hungry. These data requirements increase as the test set becomes more diverse and unpredictable. A natural way to improve performance on a diverse test set is to collect a larger volume of diverse training data. However, in some machine learning (ML) domains, such as healthcare, annotated data is expensive, and a single institution may not hold enough data. Furthermore, there may be some concern about privacy regarding data sharing. Federated learning (FL) has arisen in recent years to address the need for more training data by pooling data from different sources while also meeting data privacy.

FL is the process of training a model on decentralized edge devices, each holding its local data and without sharing it. FL allows the global model to be updated with more data, theoretically improving model performance. However, the training data is often statistically heterogeneous in such a setting and manifests various distribution patterns. Clients may have distinct data acquisition processes, leading to differences in $\mathbf{X}$ data while the y labels may remain the same. For example, in x-ray imaging, different hospitals may use other scanners that are not calibrated in the same way but still want to diagnose the same disease. This study will focus on improving FL performance for situations where the marginal input distribution differs across clients, namely covariate shift.

One line of FL research is to build a global model that generalizes well across different data domains, which is our focus in this paper. A recent survey paper by Liu et al. (2021) points out the family of content-style models that enforce disentanglement. We're motivated by the content encoder that learns domain-invariant features such as shapes and blobs. We hypothesize that domain invariant features are essential in the classification performance of the global model across different clients, assuming the domain-specific features discovered by the style encoder primarily add biases toward specific clients.

We investigate the use of feature disentanglement to improve global model performance in an FL scheme with a considerable covariate shift across each client. We use feature disentanglement to

*These authors contributed equally.

isolate the invariant features across the different domains, enabling FL on very disparate client data and generalization on new test domains that are unlike the client data.

2 RELATED WORK

2.1 FEDERATED LEARNING ALGORITHMS

We focus on feature disentanglement approaches in this project but briefly introduce FL algorithms' problems with heterogeneous data to motivate our work. The most widely known FL algorithm, FedAvg, suffers when dealing with heterogeneous data. Because it does a naive averaging of each parameter in each round of communication, it is negatively affected by the misaligned loss gradients that may arise from different data distributions. More recently, various algorithms have been proposed, such as FedProx by Li et al. (2018) and FedBN by Li et al. (2021). FedProx adds a proximal term to the loss to enforce the soft constraint. However, the performance drops dramatically as the number of clients increases, as shown in Wang et al. (2020). FedBN solves the problem of feature shift by proposing a framework built upon FedAvg that allows local batch normalization. As more and more recent work (e.g., Kolesnikov et al. (2020)) points out that batch normalization is detrimental to model generalization and fails when batch size is small, it is questionable whether FedBN is a good solution to the problem.

2.2 FEATURE DISENTANGLEMENT

Feature disentanglement supposes that features can be learned in a guided way to have interpretable semantic meaning. Generally, it decorates the learned set of features into distinct subsets related to a meaningful physical property. The reasons for decoupling learned latent features might be for interpretability or easily distinguished semantical meaning. In our case, we are isolating features that are invariant across different domains. We now discuss the properties desired of our disentangled representation and how to achieve them during training.

2.2.1 DESIRED PROPERTIES OF A DISENTANGLED REPRESENTATION

Prior to Higgins et al. (2018), there was no formal mathematical definition of disentangled features. The paper formulates disentanglement of a given representation in relation to symmetries on "the world state". If we assume the world state has a number of symmetries that each change some physical properties and leave others invariant, we hope that a disentangled representation reflects these in an equivariant and decomposable manner.

Now, we represent this definition mathematically to emphasize the desired properties. Let the set of world states be W. The authors suppose observations O can be generated from the world state via a method $b : W \to O$. Representations Z can be generated from observations via a method $h : O \to Z$. Then, they define the composition $f : W \to Z, f = h \circ b$. Next, we make use of their definition for disentangled group action.

Definition 1 ***Disentangled Group Action:*** *A disentangled group action is defined for a given group action $\cdot : G \times W \to W$ if the group can be decomposed as $G = G_1 \times G_2$, there is a decomposition $W = W_1 \times W_2$ for some subspaces W_1 and W_2, and there are actions $\cdot_i : G_i \times W_i \to W_i, i\epsilon\{1, 2\}$ such that*

$$(g_1, g_2) \cdot (v_1, v_2) = (g_1 \cdot_1 v_1, g_2 \cdot_2 v_2)\,.$$

In other words, a group action is disentangled if can be decomposed into group actions that act independently on different subspaces of the physical representation while leaving the others invariant.

Then, suppose there is a group G of symmetries acting on W via a disentangled action $\cdot : G \times W \to W$. The paper shows that we can preserve the disentangled group action structure from W on Z to produce a *disentangled representation* if f is equivariant.

This definition yields two properties for disentangled representations (which previous literature had not agreed on): modularity, and (not-necessarily-)compactness. *Modularity* refers to the idea that a single latent dimension should only encode one action of the symmetry group. A latent dimension should not encode actions on two different symmetry groups. *Compactness* refers to the desire for

each group action to be encoded by a single latent dimension. This is *not* something the proposed definition requires; each disentangled subspace is allowed to be multidimensional. For our purposes, we hope to achieve a representation that is modular, so that we can use just the domain invariant dimensions. Similar to the authors, we do not care if our representation is compact.

Achille & Soatto (2018) describe other desirable properties: sufficiency, invariance to nuisances, and minimality. *Sufficiency* refers to the desire for the representation to be as informative as the original data. *Invariance to Nuisances* refers to the desire for the representation to be independent of any irrelevant nuisance variables. *Minimality* refers to the desire for the information contained in the network parameters to be minimized.

This paper uses a definition of *disentanglement* that is related to both the modularity and compactness properties described by Higgins et al. (2018). They use the word "disentangled" to mean the total correlation of the representation is minimized, that is, a maximally disentangled representation is one whose components are independent. Because of this discrepancy, we use the terminology "componentwise-independent" where appropriate, however, componentwise-independence and disentanglement are connected in that the former arises from the latter if compactness is satisfied (which we do not require). For a given disentangled representation, if the symmetry group actions are disentangled and thus independent, modularity implies that the components that encode an action are independent from the others. Achieving complete componentwise-independence is possible if each symmetry group action is controlled by just a single dimension, which is the case when the representation is compact. Future work should more rigorously connect these concepts.

Once again, mathematically represented: let x be the input data, z be the representation of x, y be our task target which we infer from z, and n be a nuisance variable which affects the observed x but is not informative to our task (i.e. $y \perp\!\!\!\perp n$).

A representation z is...		...if...
sufficient	:=	$I(y;z) = I(y;x)$
invariant	:=	$n \perp\!\!\!\perp y \implies I(n;z) = 0$
minimal	:=	$I(x;z)$ minimized
componentwise-independent	:=	$\text{TC}(z) = \text{KL}\left(p(z) \| \prod_i p(z_i)\right)$ minimized

Achille & Soatto (2018) show that, given sufficiency, minimality can be used to induce invariance and componentwise-independence, or given sufficiency, invariance can be used to induce minimality and componentwise-independence.

2.2.2 How Feature Disentanglement is accomplished in the literature

Variational autoencoders (VAEs) decompose classification-related factor through image reconstruction (Siddharth et al. (2017)). More specifically, they decouple factors of variation by forcing independence between different dimensions of latent features, which are factorized representations from the input. Among different variants of VAEs, β-VAE by Higgins et al. (2017) encourages disentanglement by forcing less information about the reconstruction by increasing the weight of the KL divergence term. Different from VAEs, recent generative adversarial networks (GANs) (Goodfellow et al. (2014)) manage to learn disentangled features by adding regularization terms. A typical example of such a regularization approach is InfoGAN by Chen et al. (2016), which enforces the learning of unstructured noise and structured features of data distribution, respectively. In FL, FedDis by Bercea et al. (2021) proposes using feature disentanglement to adapt to covariate-shifted domains. They assume some features are domain-invariant (content) while some are domain-variant (style). For example, different clients may have different scanners, which provide different styles, while the structural anatomy or content information is assumed to be the same across all clients. They separate content features from style features with contrastive learning, which enforces content consistency across style augmentations (e.g., brightness/contrast shift) and latent orthogonality (i.e., latent representations for content and style should be different). With these disentangled features, they create an FL model that only aggregates the domain-invariant features. This construction of FL allows the global model to generalize well despite different styles for each client. They improved segmentation performance by 11% over SOTA FL methods.

3 METHODS

We implement feature disentanglement to allow the model to learn the domain-invariant features exclusively and build a global classifier upon them. As shown in Bercea et al. (2021), disentanglement learning can improve federated learning when there is covariate shift in the data across clients.

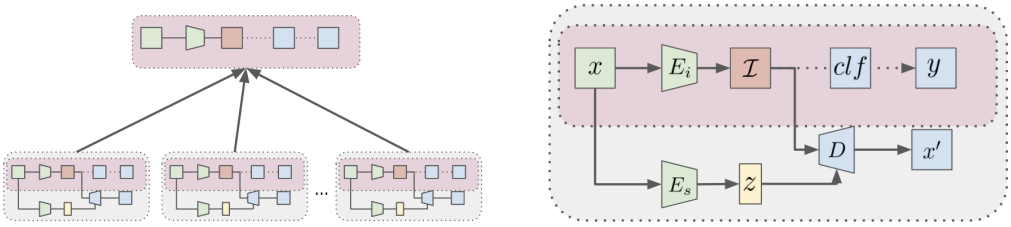

Figure 1: Basic architecture for disentanglement. $\boldsymbol{x}$ and $\boldsymbol{x'}$ are input and reconstructed images. $\mathcal{I}$, $\boldsymbol{z}$ are the latent tensor and vector. E_i and E_s are encoders of domain invariant features and domain specific features. clf stands for classifier.

However, we also address some limitations of the FedDis paper. They use inductive bias to assume the features beforehand (content vs. style) and separate contents from styles with data augmentations only that are invariant to content; however, we hope to develop a more generalizable approach that uses no inductive biases. Lee et al. (2021) and Nam et al. (2020) uses generalized cross-entropy loss by Zhang & Sabuncu (2018) to first learn easy-to-learn features (i.e., domain-specific feature in their statement), as well as data augmentation with diversified domain-specific samples to enhance feature disentanglement. However, there is no guarantee that the features are properly decoupled besides empirical success and image reconstruction. As a preliminary study, in this project, we start from unenforced disentanglement learning with inductive bias from archicture building blocks in CNN and present it in a FL scheme.

3.1 EXPERIMENTAL DESIGN

3.1.1 MODEL DESIGN

We adopt the architecture design from Huang et al. (2018) which consists of a content encoder, a style encoder, a joint decoder, and a classifier. When deployed in an FL manner, each client owns an independent model with the four modules. We emphasize that the style encoder is client-independent which means it does not join the aggregation step when performing FL.

Content Encoder The content encoder composes several convolutional layers followed by a few residual blocks. Inspired by Huang & Belongie (2017), we add instance normalization between convolutional layers for style removal. Actually, in our experiment, this significantly helps suppress style-related information.

Style Encoder The style encoder is similar to the content encoder with a few differences: 1) We replace the instance normalization with group normalization. 2) We add a global pooling layer at the end of the convolutional layers. Intuitively, the average pool operation is not sensitive to the structured content information, thus removing the spatial correlation and retraining the style data. 3) We put a fully connected layer at the end for dimension reduction.

Deocoder The deocoder receives inputs from both content encoder and style encoder. However, the style features is not directly pass into the network, instead we use a MLP to map the style features to a set of AdaIN Huang & Belongie (2017) parameters. As shown in eq. (1), AdaIN receives a content input x and a style input y. The goal is to align the channel-wise mean and variance of x to match those of y. As a result, AdaIN can embed the style features into the decoder.

$$\text{AdaIN}(x, y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y) \tag{1}$$

Classifier The classifier composes of two fully connected layers concatenating with the content encoder. We believe that domain invariant features, as extracted from the content encoder, are robust features that can be used to generalize well, even on out-of-distribution (OOD) samples. The classifier, therefore, only receives content features as input.

The final objective function is eq. (2) where $\mathcal{L}_{clf}$ measures the classification loss, $\mathcal{L}_{recon}$ is the reconstruction loss between input image and reconstructed image output by decoder.

$$\mathcal{L} = \mathcal{L}_{clf} + \alpha\mathcal{L}_{recon} \tag{2}$$

4 Experimental Results

4.1 Dataset and Experiment Setups

We conduct experiments on a digit dataset containing multiple sources with non-iid distribution images. Specifically, the five datasets are: MNIST LeCun et al. (1998), MNIST_M Ganin & Lempitsky (2015), SVHN Netzer et al. (2011), SythDigits Ganin & Lempitsky (2015) and USPS Hull (1994). We select MNIST and SythDigits for FL and others servers as OOD dataset for testing only. For MNIST and SythDigits, we randomly split 5% data for training to simulate the limited training data common in FL.

For training, we use ADAM optimizer with a learning rate of 0.001. We also use MSE and CrossEntropy as the reconstruction and classification loss function. We set a maximum of 300 epochs for model training and checkpoint the best model with the smallest loss on the holdout validation set. All the experiments are repeated three times to report the mean and standard deviation.

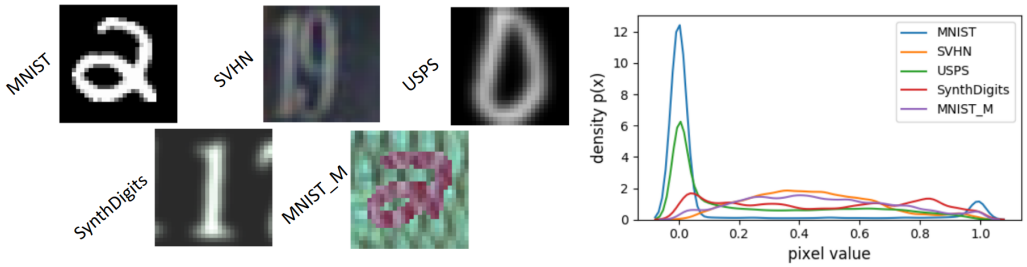

Figure 2: (a) sample from each dataset (b) pixel distribution of five dataset (credit toLi et al. (2021))

4.2 Result

We compare the disentanglement FL with the baseline FedAvg method on the digits dataset to show that our method is more generalizable on the OOD dataset. As shown in table 1, our proposed method performs better on the three OOD datasets by slightly sacrificing the performance on the training data domain. Although the two methods perform comparable on the USPS dataset, it can be explained by the fact that the MNIST and USPS have very similar pixel distribution patterns as shown in section 4.1. As for MNIST_M and SVHN, they are colored datasets and exhibit very different visual patterns compared with MNIST and SYtheDigits. Therefore, domain-invariant features are essential to ensure transferability, which we confirm by the result that our method tends to perform better on the MNIST_M and SVHN.

To better understand how the proposed method can be beneficial when covariate shift exists, we generate several images by swapping the style and content features. Intuitively speaking, we can think of the digit itself as the content in the digit dataset while the background and the colors represent

Table 1: Classification performance evaluated on five dataset based on accuracy (mean(std)). dataset colored in blue indicate out of distribution data.

	MNIST_M	MNIST	SVHN	USPS	SytheticDigit
baseline	62.51 (1.75)	97.56 (0.04)	65.13 (2.13)	94.00 (0.54)	92.40 (1.54)
ours	63.59(0.76)	97.24(0.14)	68.43(1.89)	94.90 (0.14)	90.24 (1.49)

the styles. As can be seen in fig. 3, the synthetic data inherit the style from other domains while the content keeps unchanged. This observation supports our conjecture that the content encoder and style encoder can automatically learn the corresponding features without interfering with each other. The result of this work can be very beneficial from the robustness point of view. We can explain the benefit with a simple case where color and labels are correlated. For example, when all digits "1" are in the red background in the training domain while others are not, it is very likely the model will learn the color information to distinguish "1" from other digits. However, if we test this model on a new domain where all the background is green, including digits "1", then the model's output on digits "1" with green background would be unpredictable.

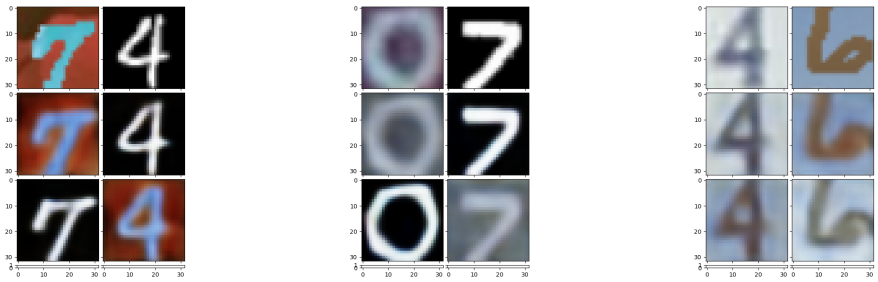

Figure 3: From top to bottom: origional image paris, reconstructed images and the synthetic images by swapping the feature vector; From left to right: image from MNIST_M and MNIST, SVHN and MNIST, SVHN and MNIST_M

5 Theoretical Analysis

5.1 Setup

We begin with an easier to analyze setting, using a simple toy dataset. Later this is generalized to a case compatible with the experiments in the main text.

Imagine that on our three clients, we have three datasets whose domains are explicitly factorable into content and color attributes (e.g. maybe the dataset was generated by beginning with grayscale MNIST images and artificially coloring each with a different distribution for each client). This factorization looks like this $W = W_{content} \times W_{color}$. The symmetry group we hope to be invariant over in the global model is color permutations. For simplicity, to define this group, we can imagine the set of all possible colors and let our transformation be the mapping of a color to any other color. This yields the permutation group S_n where n is the number of colors in the world, which we can rename G_{color}. Let $G_{content}$ represent the group of transformations on everything except color[1]. Then, the group of all possible transformations can be factorized as $G = G_{color} \times G_{content}$. This provides us the disentangled group actions $\cdot_{color} : G_{color} \times W_{color} \to W_{color}$ and $\cdot_{content} : G_{content} \times W_{content} \to W_{content}$.

[1]This is a slight simplification; not all transformations that appear in the world are part of a group. They must satisfy the properties of a group, e.g., invertibility. Identifying how this theoretical framework would accommodate other transformations could be future work.

If we manage to train a representation generating function $f = h \circ b : W \to Z$ that is equivariant to these group actions, we see our representation Z is also disentangled with the same group actions (result from Higgins et al. (2018)).

5.2 WHY DISENTANGLEMENT HELPS

Again, we hope to update the global model with local models that are trained on the same training distribution. Because the content distributions across all three clients are the same, while the color distributions are artificially generated to be distinct across all three clients, we would like to train on just $W_{content}$. We do this by training just on $Z_{content}$, which allows our model to be invariant to the group action $\cdot_{color} : G_{color} \times W \to W$. Thus, given the assumption that some content component of our data is disentangled and aligned across all clients, we can think of our method (at optimality) as perfectly aligning our domain space across clients.

For the disentanglement experiments in the main text, we have datasets from the real world. Because of this, it is hard to identify symmetries that were artificially baked into easier synthetic datasets. We can generalize the earlier results by replacing G_{color} with the analogous $G_{nuisance}$. This is also a simplification, but the existence of a $G_{nuisance}$ in the main text experiment is empirically supported by the successful disentanglement shown in Figure 3. Future work may include stronger theoretical support.

This can also be rephrased in the framework of Huang et al. (2018) to show that this disentanglement is successful and aligns the distributions of each client. We run a separate instantiation of their method on each client. Our data on the decomposed domain $W = W_{content} \times W_{nuisance}$ is drawn from distributions $p_j(x_i) = p_j(c) * p_j(n_i)$ where j represents each client and i represents each datapoint within the client. They assume the content distribution $p_j(c)$ is shared across datapoints within each client (*i.e.* $p_j(c_i) = p_j(c)$ for all i) whereas nuisance features are presumed to be drawn from a different distribution for each datapoint (*i.e.* $p_j(n_i)$). In their Proposition 2, they show that at optimality, their disentangling framework is able to properly learn that $p_j(c_i) = p_j(c)$ for all data on a single client. When paired with our assumption that each client has the same underlying content distribution (*i.e.* $p_j(c) = p(c)$), we see that the training distributions of each client have been aligned.

When viewed from these two perspectives, we see that disentangling our data representations and discarding the nuisance information aligns the clients' training distributions. This allows us to use traditional federated algorithms, even in the case of highly heterogeneous training distributions.

6 CONCLUSION

This paper investigates disentanglement learning in FL and proposes a content-style-based FL scheme for image classification. We compare our proposed method with the baseline FedAvg algorithm on a mixed digit dataset. Our experiment results show that FL with disentanglement learning can improve generalization performance on OOD data. Our visualization of the feature disentanglement suggests that some building blocks such as average pooling layers and instance normalization layers can encourage feature disentanglement.

As a preliminary study, there are some unsolved problems we would like to put them as our future work: 1) Our current approach is un-enforced disentanglement approach which means that strict disentanglement are not always guaranteed. We plan to explore the explicit supervision approach to complete our study in the future. 2) Currently, we only experiment on a simple dataset. However, in the real-world setting (i.e., natural image), features are highly entangled, and more challenging to disentangle them. We are interested in a more realistic setup and plan to test our approach to more complex datasets such as DomainNet.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations, 2018.

Cosmin I. Bercea, Benedikt Wiestler, Daniel Rueckert, and Shadi Albarqouni. Feddis: Disentangled federated learning for unsupervised brain pathology segmentation, 2021.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, 2016.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

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

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations, 2018.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–189, 2018.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 491–507. Springer, 2020.

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.

Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation, 2021.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Learning disentangled representations in the imaging domain, 2021.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. This work is unpublished., 2011.

N. Siddharth, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah D. Goodman, Pushmeet Kohli, Frank Wood, and Philip H. S. Torr. Learning disentangled representations with semi-supervised deep generative models, 2017.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels, 2018.