HIERARCHICAL TRAINING FOR DISTRIBUTED DEEP LEARNING BASED ON MULTIMEDIA DATA OVER BAND-LIMITED NETWORKS

Siyu Qi, Lahiru D. Chamain, Zhi Ding

Department of Electrical and Computer Engineering
University of California Davis
Davis, CA, USA
{syqi, hdchamain, zding}@ucdavis.edu

ABSTRACT

Distributed deep learning (DL) plays a critical role in many wireless Internet of Things (IoT) applications including remote camera deployment. This work addresses three practical challenges in cyber-deployment of distributed DL over band-limited channels. Specifically, many IoT systems consist of sensor nodes for raw data collection and encoding, and servers for learning and inference tasks. Adaptation of DL over band-limited network data links has only been scantly addressed. A second challenge is the need for pre-deployed encoders being compatible with flexible decoders that can be upgraded or retrained. The third challenge is the robustness against erroneous training labels. Addressing these three challenges, we develop a hierarchical learning strategy to improve image classification accuracy over band-limited links between sensor nodes and servers. Experimental results show that our hierarchically-trained models can improve link spectrum efficiency without performance loss, reduce storage and computational complexity, and achieve robustness against training label corruption.

***Index Terms*—** Hierarchical training, image compression and classification, auto-encoders, information theory.

1. INTRODUCTION

Deep learning (DL) has become an increasingly important tool for multimedia processing particularly in IoT systems. Within such networked/distributed learning framework, low-cost encoders are deployed to compress and transmit data on low-power sensing devices to cloud/server nodes to carry out major learning tasks. Images can be efficiently compressed to lower-dimensional latent representations using auto-encoders (AEs) [1, 2, 3, 4], where an encoder at sensor nodes and a decoder/classifier at servers are optimized in an end-to-end (E2E) manner. In networked learning, bandwidth efficiency can be just as important as the overall accuracy. It is therefore vital to lower the coding rate of latent representations without severely compromising the task accuracy. Moreover, after deployment of encoders on source devices, the server nodes may directly channel the data obtained by encoders to separately-trained decoders [5, 6, 7]. Consequently, the encoders that are jointly optimized with one decoder may exhibit degraded performance with another reconfigured decoder. Hence, the reconfiguration flexibility with decoders is an essential characteristic of these encoders embedded on source devices. Another major issue in supervised learning is its reliance on labeled training data. In practice, however, errors in data annotation, inaccuracy in automatic label extraction process, or data poisoning attacks [8, 9] are commonplace that can lead to erroneous data labels [10]. For this reason, achieving robustness to certain level of corrupted training labels is critical to reliable supervised learning models.

In training DL models for image classification, cross-entropy (CE) loss function has been particularly effective. Despite its successes, CE-based training does not overcome the three obstacles above. In this work, we address these aforementioned practical considerations by integrating a newly-proposed information-theoretic learning principle of Maximal Coding Rate Reduction (MCR^2) [11] in training. Importantly, the principle of MCR^2 is capable of projecting input data to latent representations in low-dimensional subspaces that are inter-class discriminative and in-class compressive. In addition to offering better interpretability, MCR^2-trained classifiers have demonstrated stronger robustness against label noise. The lower-dimensional latent representations of MCR^2 can potentially provide valuable insight on deriving DL models subject to bandwidth constraints.

To this end, we present an MCR^2-guided AE architecture for cloud-based image classification. Leveraging the MCR^2 principle, we introduce a novel multi-phase hierarchical training (DuPHiL) strategy via a side channel. Moving beyond the traditional E2E training based on loss function superposition, we guide the encoder training with MCR^2 loss function from an auxiliary path, to acquire diverse and discriminant latent representations. Meanwhile, we train the decoder with CE loss function. This DuPHiL strategy separates encoder and decoder training, achieves the dual objectives of efficient compression and accurate classification, and gains stronger robustness to label corruption as well as better decoder reconfigurability.

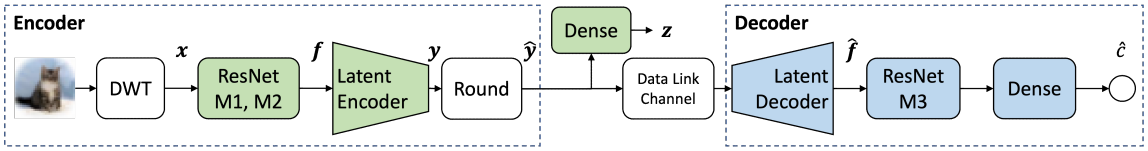


Fig. 1: Architecture of proposed wavelet-domain auto-encoder. "M" is short for "module". In Phase 1 of proposed DuPHiL strategy, green blocks are guided by MCR^2 loss; in Phase 2, blue blocks are guided by CE loss.

2. DISTRIBUTED LEARNING IN IOT

We consider a deployment scenario involving distributed encoder/classifier, where a source device is responsible for encoding image data for transmission whereas an edge server is responsible for decoding and classification. As standardized commercial image compression algorithm, such as JPEG 2000 [12], which is based on multi-level 2-dimensional Discrete Wavelet Transform (2D DWT), is commonly embedded on source nodes, bypassing reconstruction when processing JPEG-2000-encoded images on source device can save the decoding computation [13] and improve inference speed. Therefore, we adopt the modified ResNet proposed in [14] as the backbone for DWT-domain image classification, as shown in Fig. 1. The term "ResNet module" refers to two or more stacked ResNet blocks with same numbers of filters, each containing two convolutional layers and a shortcut.

At the source node, we incorporate a latent encoder consisting of a pooling layer and two dense layers for compression of the intermediate feature $\boldsymbol{f}$. At the edge server, the receiver begins with a latent decoder consisting of three transposed convolutional layers for latent recovery. The latent representations $\boldsymbol{y}$ are mapped (via rounding) to codeword $\hat{\boldsymbol{y}}$ before transmission to the receiver node over a communication data link. During training, as the rounding quantizer has zero derivative almost everywhere, we adopt the method proposed in [15] to solve the zero gradient problem by adding a random uniform noise in $(-1,\, 1)$ to $\boldsymbol{y}$ as a rounding relaxation. After performing DuPHiL, this cloud-based classification model can be distributively deployed as an Encoder on remote source node and a Decoder on the edge server.

3. PROPOSED HIERARCHICAL LEARNING

3.1. DuPHiL: Hierarchical Training

In the basic classification problem, consider a set of N samples $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\} \in \mathcal{R}^{D_{\rm in}\times N}$ and their class labels $\{c_1, \ldots, c_N\} \in [K]$, where $D_{\rm in}$ is the input data (image) size, N is number of samples in the dataset, and K is the number of classes. According to experience and empirical tests, a deep classifier is typically trained for a direct mapping from an input $\boldsymbol{x} \in \mathcal{R}^{D_{\rm in}}$ to its class label c.

Based on information theoretic foundation, the recent work of [11] suggested the MCR^2 loss that drives a DL model to extract more diverse and discriminant lower-dimensional latent representations $\boldsymbol{z} \in \mathcal{R}^{D_z}$, with $D_z < D_{in}$, from input before classification. Define the coding-rate-reduction loss

$$\mathcal{L}_{\rm MCR2} = -\Delta R(\boldsymbol{Z}) \doteq -R(\boldsymbol{Z}) + R_c(\boldsymbol{Z}, \boldsymbol{\Pi}) \qquad (1)$$

where, according to [16],

$$R(\boldsymbol{Z}) = \frac{1}{2}\log\det\left(\boldsymbol{I} + \frac{D_{\rm in}}{\epsilon^2 N}\boldsymbol{Z}\boldsymbol{Z}^T\right) \qquad (2)$$

is the average number of bits required to encode a learned representation $\boldsymbol{z}_i$ from $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N\} \in \mathcal{R}^{D_z\times N}$ up to an precision bound of ϵ. When a known partition $\boldsymbol{\Pi} = \{\boldsymbol{\Pi}_j\}_1^K$ that groups samples into classes, we can write the "group-wise average bit rate" of $\boldsymbol{z}_i$ as

$$R_c(\boldsymbol{Z}, \boldsymbol{\Pi}) = \sum_{j=1}^{K} \frac{\mathrm{tr}(\boldsymbol{\Pi}_j)}{2N}\log\det\left(\boldsymbol{I} + \frac{D_{\rm in}}{\epsilon^2 \mathrm{tr}(\boldsymbol{\Pi}_j)}\boldsymbol{Z}\boldsymbol{\Pi}_j\boldsymbol{Z}^T\right)$$

bits per sample. $\boldsymbol{\Pi}_j$ is a diagonal matrix with entries "1" for samples that belong to the j-th class and "0" otherwise.

Our goal is to train the proposed model according to two loss functions: the MCR^2 loss $\mathcal{L}_{\rm MCR2}$ to regularize the latent encoding rate and the CE loss $\mathcal{L}_{\rm CE}$ to minimize the classification discrepancy between the Decoder output $\hat{c}$ and the true label c. A traditional training approach *would be* to superimpose the two losses to generate a sum loss function $\mathcal{L}_{CE} + \lambda\mathcal{L}_{\rm MCR2}$ using a regularization variable λ. Such a naïve joint loss function, however, requires the encoder and decoder to be trained jointly in an E2E manner, which is less practical when encoders are already **pre-deployed but the learning tasks are reconfigured**. Furthermore, traditional preset regularization requires careful tuning of the hyper-parameter λ to avoid local minimum that neither minimizes classification error, nor reduces the encoding rate. Instead, we propose a novel Dual-Phase Hierarchical Learning (DuPHiL) strategy to integrate the two loss functions beyond regularization. In each epoch of DuPHiL method:

Phase 1: To acquire diverse and discriminant features from inputs, the Encoder modules and the side branch are jointly updated to minimize $\mathcal{L}_{MCR2}$;

Phase 2: For accurate classification, the Decoder modules are jointly updated to minimize $\mathcal{L}_{CE}$, while Encoder modules are frozen after phase 1.

Because of the feature-preserving characteristics of MCR^2, Phase 1 shall retain the key features necessary for accurate

classification. This DuPHiL strategy leverages the strength of MCR^2 and induces models to be more robust to potential label corruptions and more flexible for integration with different classifier modules.

3.2. Discriminative Power Analysis

To better interpret the classification models, we adopt the concept of "discriminative power" [17] of neurons/filters in each layer. According to Fisher's linear discriminative analysis (LDA) [18], the within-class scatter matrix is defined as:

$$\boldsymbol{S}_w = \sum_{j=1}^{K} \sum_{\boldsymbol{x} \in \boldsymbol{\Pi}_j} (\boldsymbol{x} - \boldsymbol{m}_j)(\boldsymbol{x} - \boldsymbol{m}_j)^T. \tag{3}$$

The between-class scatter matrix is:

$$\boldsymbol{S}_b = \sum_{j=1}^{K} n_j (\boldsymbol{m} - \boldsymbol{m}_j)(\boldsymbol{m} - \boldsymbol{m}_j)^T. \tag{4}$$

where $\boldsymbol{m}_j$ is the class-wise sample mean of $\boldsymbol{\Pi}_j$, n_j is the number of samples in $\boldsymbol{\Pi}_j$ and $\boldsymbol{m}$ is the global sample mean of the dataset. We then define the discriminative power of each neuron/filter as [17] as $D \doteq \text{trace}(\boldsymbol{S}_b)/\text{trace}(\boldsymbol{S}_w)$. The neuron/filter with the highest score in a layer is the "best" neuron/filter of that layer.

In order to vary the encoding rate in response to different link rate constraints, we further prune neurons from the bottleneck layer in latent encoder, which is equivalent to nullifying entries in the latent representation $\hat{\boldsymbol{y}}$. We adopt the same neuron screening strategy in [17]: using training dataset, we evaluate the discriminative power D of each neuron in the bottleneck layer and prune those with lowest scores.

4. EXPERIMENTAL RESULTS

We train our proposed model on the popular CIFAR-10 and CIFAR-100 datasets [19]. We utilize a modified ResNet-18 backbone for CIFAR-10 and a modified ResNet-34 backbone for CIFAR-100. With side channel disabled, we first pre-train the AE from E2E using CE loss for 200 epochs to obtain the benchmark "CE-trained" model. Next, we optimize the pre-trained benchmark model using the proposed DuPHiL strategy for 100 extra epochs.

We adopt a simple arithmetic encoder to convert the quantized latent vector $\hat{\boldsymbol{y}}$ into bitstreams for bit rate measurement. We obtain the approximated cumulative distribution functions based on the histogram of $\hat{\boldsymbol{y}}$ of the training set.

4.1. Rate-Accuracy Performance

It is important to note that our MCR^2-guided architecture and DuPHiL strategy allow practical deployment of pre-trained encoders in distributed learning environment. However, it is important to examine whether this flexible re-training and reconfigurability of decoders enabled by our proposed learning strategy may lead to some performance loss.

The rate-accuracy trade-off results are illustrated in Fig. 2, where the suffixes "-10" and "-100" refer to CIFAR-10 and CIFAR-100 results, respectively. It is clear that, for classification, the input image can be compressed to a bit rate of around 0.3 bits-per-pixel (bpp) for CIFAR-10 and approximately 0.35 bpp for CIFAR-100 by using either DuPHiL or E2E CE training without accuracy loss. In comparison with the CE-trained baseline, our proposed DuPHiL strategy can achieve similar rate-accuracy performance on CIFAR-10 and consistently provide up to 1% higher accuracy at the same encoding rate on CIFAR-100 without any additional hardware cost. The ability to deploy pre-trained encoders and the reconfigurability of learning tasks using retrained decoders did not cause noticeable classification performance loss.

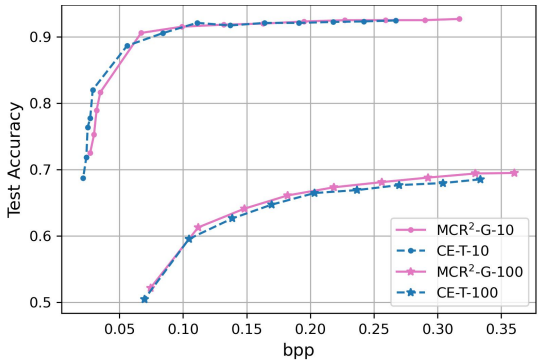


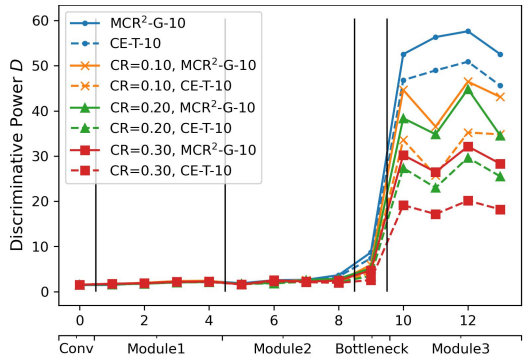
Fig. 2: Rate-accuracy performance comparison on CIFAR-10 and CIFAR-100 of MCR^2- and CE-guided AEs.

We observe that during Phase 1 training, both the overall rate R and group-wise rate R_c would grow, implying that the MCR^2-guided encoder tends to encode latent representations into more bits, while ensuring the in-class compactness and between-class discrimination of latent vectors.

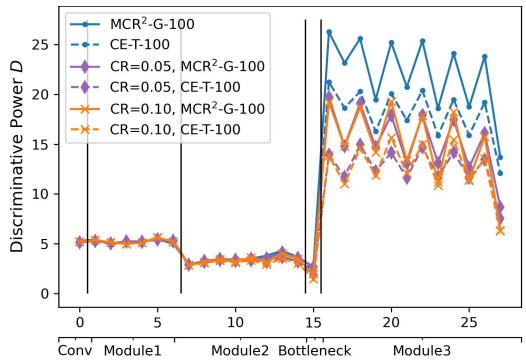
4.2. Robustness against Corrupted Labels

The authors of [11] have demonstrated that deep models under MCR^2 can learn well despite some fractions of corrupted labels during training. To also test our proposed model's robustness against label corruption, we pre-train the AE with corrupted labels firstly with CE loss in an E2E manner, which are used as the baseline models. We then enhance the pre-trained model with the proposed DuPHiL method. For evaluation, we use the correct ground truth labels. Our experiments include label corruption ratio (CR) of 10%, 20% and 30% on CIFAR-10 and CR of 5% and 10% on CIFAR-100. We present the layer-wise discriminant power scores in Fig. 3. From both plots, it is evident that MCR^2-guided models exhibit higher discriminative power than CE-trained models subject to the same level of label corruption. In addition, we can observe a fluctuation of discriminative power in a period of 2 layers, as the shortcut connections in ResNet modules directly add outputs of previous layers to latter layers, which lowers the discriminative power of latter layers.

From Fig. 4, we can observe that on CIFAR-10 dataset, with 10% training labels corrupted, the MCR^2-guided model



(a) CIFAR-10



(b) CIFAR-100

Fig. 3: Layer-wise discriminative power of proposed distributed AE, computed on test set.

can deliver up to 1% higher test accuracy than the CE-trained model at the same encoding data rate. Meanwhile, the MCR^2-guided model achieves robust learning even under 20% label corruption and clearly delivers higher accuracy at the same data rate than the corresponding benchmark CE model. Even with 30% random label corruption, our proposed learning model still yields a comparable rate-accuracy performance to the benchmark CE model with 20% label corruption. Similarly, on CIFAR-100 dataset, with 5% training labels corrupted, our MCR^2-guided model can achieve up to 2% higher test accuracy at the same rate. However, under 10% label corruption, the improvement becomes negligible, indicating the useful information remaining in the corrupted dataset is not sufficient for our proposed method to demonstrate benefit.

These results demonstrate the robustness of the proposed DuPHiL method against noisy training data.

4.3. Compatibility with Decoder Re-Training

To illustrate the general compatibility of our Encoders, we freeze the Encoders after CE or MCR^2-guided training, but train two new Decoders/classifiers from scratch, including: (1) a Decoder of the same architecture as in Fig. 1 but optimized with Kullback-Leibler Divergence (KL-D) [20] loss and (2) a linear support vector machine (SVM) [21]. We present the obtained results in Table 1. The results show that the encoder modules from DuPHiL continue to deliver

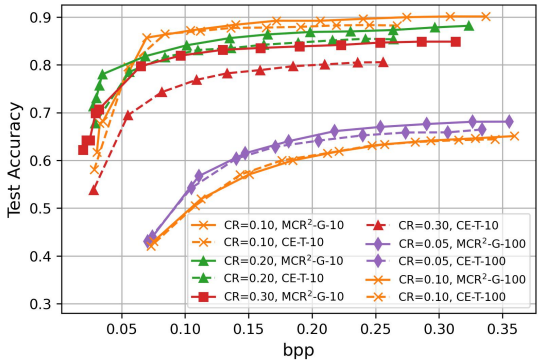


Fig. 4: Accuracy vs. bpp performance of proposed AEs with label corruption on CIFAR-10 & CIFAR-100.

Table 1: Accuracy performance of various classifiers based on fixed pre-trained Encoders.

Dataset	Classifier Model	Training Strategy	
		E2E (Baseline)	DuPHiL (Proposed)
CIFAR-10	As in Fig. 1 (CE loss)	92.64%	**92.77%**
	As in Fig. 1 (KL-D loss)	92.63%	**92.75%**
	Linear SVM	92.1%	**92.37%**
CIFAR-100	As in Fig. 1 (CE loss)	68.54%	**69.83%**
	As in Fig. 1 (KL-D loss)	68.45%	**69.79%**
	Linear SVM	64.69%	**65.37%**

robust performance. Using a linear SVM and the proposed decoder architecture trained by KL-D loss, we in fact observe up to 0.27% and 1.34% classification accuracy improvement on CIFAR-10 and CIFAR-100 datasets, respectively, over encoders from E2E training. Our results demonstrate the proposed MCR^2-guided Encoders are more flexible with various subsequent classifiers in comparison with E2E training.

5. CONCLUSIONS

We propose a hierarchical learning (DuPHiL) strategy to tackle the dual objectives of efficient discriminant feature extraction and accurate classification. Applying the information theoretic principle of MCR^2 in distributed DL configuration, our DuPHil training strategy consists of two phases. Phase one optimizes the encoding ResNet modules for efficient feature extraction. Phase two optimizes the decoding modules for learning tasks such as classification. Instead of naïvely summing loss functions of the two objectives, our DuPHiL strategy leverages the MCR^2 loss to guide encoder modules to acquire in-class-compact and between-class-separable features before minimizing the CE loss to optimize decoder modules. Results show that the proposed hierarchical learning not only achieves as good accuracy but also provides robustness to errors in training labels and flexible learning reconfigurability and applies directly to any existing AE-based approach.

6. REFERENCES

[1] Mark A Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[2] Mikolaj Jankowski, Deniz Gündüz, and Krystian Mikolajczyk, "Joint device-edge inference over wireless links with pruning," in *IEEE 21st Intl. Workshop on Signal Processing Advances in Wireless Comm.*, 2020, pp. 1–5.

[3] S. Yao, J. Li, D. Liu, T. Wang, S. Liu, H. Shao, and T. Abdelzaher, "Deep compressive offloading: Speeding up neural network inference by trading edge computation for network latency," in *Proc. 18th Conf. on Embedded Networked Sensor Systems*, 2020, pp. 476–488.

[4] J. Shao and J. Zhang, "Bottlenet++: An end-to-end approach for feature compression in device-edge co-inference systems," in *IEEE Intl. Conf. on Communications Workshops*. IEEE, 2020, pp. 1–6.

[5] Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He, "Supervised representation learning with double encoding-layer autoencoder for transfer learning," *ACM Trans. Intelligent Systems and Technology*, vol. 9, no. 2, pp. 1–17, 2017.

[6] Karren D Yang and Caroline Uhler, "Multi-domain translation by learning uncoupled autoencoders," *arXiv preprint arXiv:1902.03515*, 2019.

[7] Hui-huang Zhao and Han Liu, "Multiple classifiers fusion and cnn feature extraction for handwritten digits recognition," *Granular Computing*, vol. 5, no. 3, pp. 411–418, 2020.

[8] Jacob Steinhardt, Pang Wei Koh, and Percy Liang, "Certified defenses for data poisoning attacks," in *Proc. 31st International Conf. on Neural Information Processing Systems*, 2017, pp. 3520–3532.

[9] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[10] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel, "Using trusted data to train deep networks on labels corrupted by severe noise," *arXiv preprint arXiv:1802.05300*, 2018.

[11] Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma, "Learning diverse and discriminative representations via the principle of maximal coding rate reduction," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[12] Athanassios Skodras, Charilaos Christopoulos, and Touradj Ebrahimi, "The jpeg 2000 still image compression standard," *IEEE Signal processing magazine*, vol. 18, no. 5, pp. 36–58, 2001.

[13] Akshara Preethy Byju, Gencer Sumbul, Begüm Demir, and Lorenzo Bruzzone, "Remote-sensing image scene classification with deep neural networks in JPEG 2000 compressed domain," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 3458–3472, 2020.

[14] L. D. Chamain and Z. Ding, "Improving deep learning classification of JPEG2000 images over bandlimited networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2020, pp. 4062–4066.

[15] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli, "End-to-end optimized image compression," 2017, 5th International Conference on Learning Representations, ICLR 2017.

[16] Yi Ma, Harm Derksen, Wei Hong, and John Wright, "Segmentation of multivariate mixed data via lossy data coding and compression," *IEEE Trans. on Pattern analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1546–1562, 2007.

[17] Junhua Zou, Ting Rui, You Zhou, Chengsong Yang, and Sai Zhang, "Convolutional neural network simplification via feature map pruning," *Computers & Electrical Engineering*, vol. 70, pp. 950–958, 2018.

[18] Ronald A Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[19] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton, "CIFAR-10 (Canadian Institute for Advanced Research)," 2009.

[20] Solomon Kullback and Richard A Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[21] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.