Understanding Convolutional Neural Networks with Information Theory: An Initial Exploration

Shujian Yu, *Student Member, IEEE,* Kristoffer Wickstrøm,
Robert Jenssen, *Member, IEEE,* and José C. Príncipe, *Fellow, IEEE.*

Abstract

The matrix-based Rényi's α-entropy functional and its multivariate extension were recently developed in terms of the normalized eigenspectrum of a Hermitian matrix of the projected data in a reproducing kernel Hilbert space (RKHS). However, the utility and possible applications of these new estimators are rather new and mostly unknown to practitioners. In this paper, we first show that our estimators enable straightforward measurement of information flow in realistic convolutional neural networks (CNNs) without any approximation. Then, we introduce the partial information decomposition (PID) framework and develop three quantities to analyze the synergy and redundancy in convolutional layer representations. Our results validate two fundamental data processing inequalities and reveal some fundamental properties concerning the training of CNNs.

Index Terms

Convolutional Neural Networks, Data Processing Inequality, Multivariate Matrix-based Rényi's α-entropy, Partial Information Decomposition.

I. Introduction

There has been a growing interest in understanding deep neural networks (DNNs) mapping and training using information theory [1], [2], [3]. According to Schwartz-Ziv and Tishby [4], a DNN should be analyzed by measuring the information quantities that each layer's representation T preserves about the input signal X with respect to the desired signal Y (i.e., $\mathbf{I}(X;T)$ with respect to $\mathbf{I}(T;Y)$, where $\mathbf{I}$ denotes mutual information), which has been called the Information Plane (IP). Moreover, they also empirically show that the common stochastic gradient descent (SGD) optimization undergoes two separate phases in the IP: an early "fitting" phase, in which both $\mathbf{I}(X;T)$ and $\mathbf{I}(T;Y)$ increase rapidly along with the iterations, and a later "compression" phase, in which there is a reversal such that $\mathbf{I}(X;T)$ and $\mathbf{I}(T;Y)$ continually decrease. However, the observations so far have been constrained to a simple multilayer perceptron (MLP) on toy data, which were later questioned by some counter-examples in [5].

In our most recent work [6], we use a novel matrix-based Rényi's α-entropy [7] to analyze the information flow in stacked autoencoders (SAEs). We observed that the existence of "compression" phase associated with $\mathbf{I}(X;T)$ and $\mathbf{I}(T;Y)$ in IP is predicated to the proper dimension of the bottleneck layer size S of SAEs: if S is larger than the intrinsic dimensionality d [8] of training data, the mutual information values start to increase up to a point and then go back approaching the bisector of IP; if S is smaller than d, the mutual information values increase consistently up to a point, and never go back.

Despite the great potential of earlier works [4], [5], [6], there are several open questions when it comes to the applications of information theoretic concepts to convolutional neural networks (CNNs). These include but are not limited to:

1) The accurate and tractable estimation of information quantities in CNNs. Specifically, in the convolutional layer, the input signal X is represented by multiple feature maps, as opposed to a single vector in the fully connected layers. Therefore, the quantity we really need to measure is the *multivariate*

Shujian Yu and José C. Príncipe are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA. (email: yusjlcy9011@ufl.edu; principe@cnel.ufl.edu)

Kristoffer Wickstrøm and Robert Jenssen are with the Machine Learning Group at UiT - The Arctic University of Norway, Tromsø 9037, Norway. (email: {kwi030,robert.jenssen}@uit.no)

mutual information (MMI) between a single variable (e.g., X) and a group of variables (e.g., different feature maps)[1]. Unfortunately, the reliable estimation of MMI is widely acknowledged as an intractable or infeasible task in machine learning and information theory communities [9], especially when each variable is in a high-dimensional space.

2) A systematic framework to analyze CNN layer representations. By interpreting a feedforward DNN as a Markov chain, the existence of data processing inequality (DPI) is a general consensus [4], [6]. However, it is necessary to identify more inner properties on CNN layer representations using principled approach or framework, beyond DPI.

In this paper, we answer these questions and make the following contributions:

1) By suggesting the multivariate extension of the matrix-based Rényi's α-entropy functional [10], we show that the information flow, especially the MMI, in CNNs can be easily measured without approximation or accurate probability density function (PDF) estimation.

2) By introducing the partial information decomposition (PID) framework [11], we develop three quantities that enable identifying the synergy and redundancy tradeoff amongst different feature maps in convolutional layers. Our result has direct impact on the design of CNNs.

II. Information Quantity Estimation in CNNs

In this section we give a brief introduction to the recently proposed matrix-based Rényi's α-entropy functional estimator [7] and its multivariate extension [10]. Benefiting from the novel definition, we present a simple method to measure MMI between any pairwise layer representations in CNNs. The theoretical foundations for our estimators are proved in [7], [10].

A. Matrix-based Rényi's α-entropy functional and its multivariate extension

In information theory, a natural extension of the well-known Shannon's entropy is Rényi's α-order entropy [12]. For a random variable X with probability density function (PDF) $f(x)$ in a finite set $\mathcal{X}$, the α-entropy $\mathbf{H}_\alpha(X)$ is defined as:

$$\mathbf{H}_\alpha(f) = \frac{1}{1-\alpha}\log\int_{\mathcal{X}} f^\alpha(x)dx. \tag{1}$$

Rényi's entropy functional evidences a long track record of usefulness in machine learning and its applications [13]. Unfortunately, the accurate PDF estimation impedes its more widespread adoption in data driven science. To solve this problem, [7], [10] suggest similar quantities that resembles quantum Rényi's entropy [14] in terms of the normalized eigenspectrum of the Hermitian matrix of the projected data in RKHS, thus estimating the entropy, joint entropy among two or more variables directly from data without PDF estimation. For brevity, we directly give the definition.

Definition 1: Let $\kappa : \mathcal{X}\times\mathcal{X} \mapsto \mathbb{R}$ be a real valued positive definite kernel that is also infinitely divisible [15]. Given $X = \{x^1, x^2, ..., x^n\}$ and the Gram matrix K obtained from evaluating a positive definite kernel κ on all pairs of exemplars, that is $(K)_{ij} = \kappa(x^i, x^j)$, a matrix-based analogue to Rényi's α-entropy for a normalized positive definite (NPD) matrix A of size $n\times n$, such that $\mathrm{tr}(A) = 1$, can be given by the following functional:

$$\mathbf{S}_\alpha(A) = \frac{1}{1-\alpha}\log_2\left(\mathrm{tr}(A^\alpha)\right) = \frac{1}{1-\alpha}\log_2\Big[\sum_{i=1}^{n}\lambda_i(A)^\alpha\Big], \tag{2}$$

where $A_{ij} = \frac{1}{n}\frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}$ and $\lambda_i(A)$ denotes the i-th eigenvalue of A.

Definition 2: Given a collection of n samples $\{s_i = (x_1^i, x_2^i, \cdots, x_C^i)\}_{i=1}^n$, where the superscript i denotes the sample index, each sample contains C ($C \geq 2$) measurements $x_1 \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$, $\cdots$,

[1]By variable, we mean a random element, which can be vector valued random variable for instance.

$x_C \in \mathcal{X}_C$ obtained from the same realization, and the positive definite kernels $\kappa_1 : \mathcal{X}_1 \times \mathcal{X}_1 \mapsto \mathbb{R}$, $\kappa_2 : \mathcal{X}_2 \times \mathcal{X}_2 \mapsto \mathbb{R}$, $\cdots$, $\kappa_C : \mathcal{X}_C \times \mathcal{X}_C \mapsto \mathbb{R}$, a matrix-based analogue to Rényi's α-order joint-entropy among C variables can be defined as:

$$\mathbf{S}_\alpha(A_1, A_2, \cdots, A_C) = \mathbf{S}_\alpha \left(\frac{A_1 \circ A_2 \circ \cdots \circ A_C}{\mathrm{tr}(A_1 \circ A_2 \circ \cdots \circ A_C)} \right), \tag{3}$$

where $(A_1)_{ij} = \kappa_1(x_1^i, x_1^j)$, $(A_2)_{ij} = \kappa_2(x_2^i, x_2^j)$, $\cdots$, $(A_C)_{ij} = \kappa_k(x_C^i, x_C^j)$, and $\circ$ denotes the Hadamard product.

B. Multivariate mutual information estimation in CNNs

Suppose there are C filters in the convolutional layer, given an input image, it is represented by C different feature maps, each characterizing a specific property of the input. This suggests that the amount of information that the convolutional layer gained from input X is preserved in C information sources T^1, T^2, $\cdots$, T^C. Therefore, the amount of information that input gained from C feature maps is:

$$\mathbf{I}(X; \{T^1, T^2, \cdots, T^C\}) = \mathbf{H}(X) + \mathbf{H}(T^1, T^2, \cdots, T^C) - \mathbf{H}(X, T^1, T^2, \cdots, T^C), \tag{4}$$

where $\mathbf{H}$ denotes entropy for a single variable or joint entropy for a group of variables.

Given Eq. (2) and Eq. (3), $\mathbf{I}(X; \{T^1, T^2, \cdots, T^C\})$ in a mini-batch of size n can be estimated with:

$$\mathbf{I}_\alpha(B; \{A_1, A_2, \cdots, A_C\}) = \mathbf{S}_\alpha(B) + \mathbf{S}_\alpha \left(\frac{A_1 \circ A_2 \circ \cdots \circ A_C}{\mathrm{tr}(A_1 \circ A_2 \circ \cdots \circ A_C)} \right) - \mathbf{S}_\alpha \left(\frac{A_1 \circ A_2 \circ \cdots \circ A_C \circ B}{\mathrm{tr}(A_1 \circ A_2 \circ \cdots \circ A_C \circ B)} \right). \tag{5}$$

Here, B, $A_1, \cdots, A_C$ denote Gram matrices evaluated on input tensor and C feature maps tensors, respectively. Specifically, x_p^i (in *Definition 2*) refers to the feature map generated from the i-th input sample using the p-th ($1 \leq p \leq C$) filter, and A_p is evaluated exactly on $\{x_p^i\}_{i=1}^n$. Obviously, instead of estimating the joint PDF on $\{X, T^1, T^2, \cdots, T^C\}$ which is typically unattainable, one just needs to compute $(C+1)$ Gram matrices using a real valued positive definite kernel that is also infinitely divisible [15].

III. Main Results

This section presents two sets of experiments to validate the existence of two DPIs in CNNs, and the novel nonparametric information theoretic estimators put forth in this work. Specifically, Section III-A validates the existence of two DPIs in CNNs, whereas Section III-B illustrate, via the application of PID framework, some interesting observations associated with different CNN topologies in the training phase. Following this, we present two implications to the design and training of CNNs motivated by these results. We finally point out, in Section III-C, an advanced interpretation to the information plane (IP) that deserve more (theoretical) investigations. Two benchmark datasets, namely MNIST [16] and Fashion-MNIST [17], are selected for evaluation. To avoid influencing the flow of presentation, the results on Fashion-MNIST are demonstrated in Appendix A.

The baseline CNN architecture to be considered in this work is a LeNet-5 [16] like network with two convolutional layers, two pooling layers, and two fully connected layers (thus constituting 6 hidden layers). Each convolutional layer consists of 5×5 filters. We train the CNN using the basic SGD with momentum 0.95 and mini-batch size 128. In both datasets, we select learning rate 0.1 and 10 training epochs. Both "sigmoid" and "ReLU" activation functions are tested. For the estimation of MMI, we fix $\alpha = 1.01$ to approximate Shannon's definition, and use the radial basis function (RBF) kernel $\kappa(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$ to obtain the Gram matrices. The kernel size σ is determined based on the Silverman's rule of thumb [18] $\sigma = h \times n^{-1/(4+d)}$, where n is the number of samples in the mini-batch (128 in this work), d is the sample dimensionality and h is an empirical value selected experimentally by taking into consideration the data's average marginal variance. In this paper, we select $h = 5$ for the input signal forward propagation chain and $h = 0.1$ for the error backpropagation chain.

A. Two DPIs and their validation

We expect the existence of two DPIs in any feedforward CNNs with K hidden layers, i.e., $\mathbf{I}(X, T_1) \geq \mathbf{I}(X, T_2) \geq \cdots \geq \mathbf{I}(X, T_K)$ and $\mathbf{I}(\delta_K, \delta_{K-1}) \geq \mathbf{I}(\delta_K, \delta_{K-2}) \geq \cdots \geq \mathbf{I}(\delta_K, \delta_1)$, where T_1, T_2, $\cdots$, T_K are successive hidden layer representations from the first hidden layer to the output layer and δ_K, δ_{K-1}, $\cdots$, δ_1 are errors from the output layer to the first hidden layer. This is because both $X \to T_1 \to \cdots \to T_K$ and $\delta_K \to \delta_{K-1} \to \cdots \to \delta_1$ form a Markov chain [4], [6].

Fig. 1 shows the DPIs at the initial training stage, after 3 epochs' training and at the final training stage, respectively. As can be seen, DPIs hold in most of the cases. Note that, there are a few disruptions in the error backpropagation chain. One possible reason is that when training converges, the error becomes tiny such that Sliverman's rule of thumb is no longer a reliable choice to select scale parameter σ in our estimator.

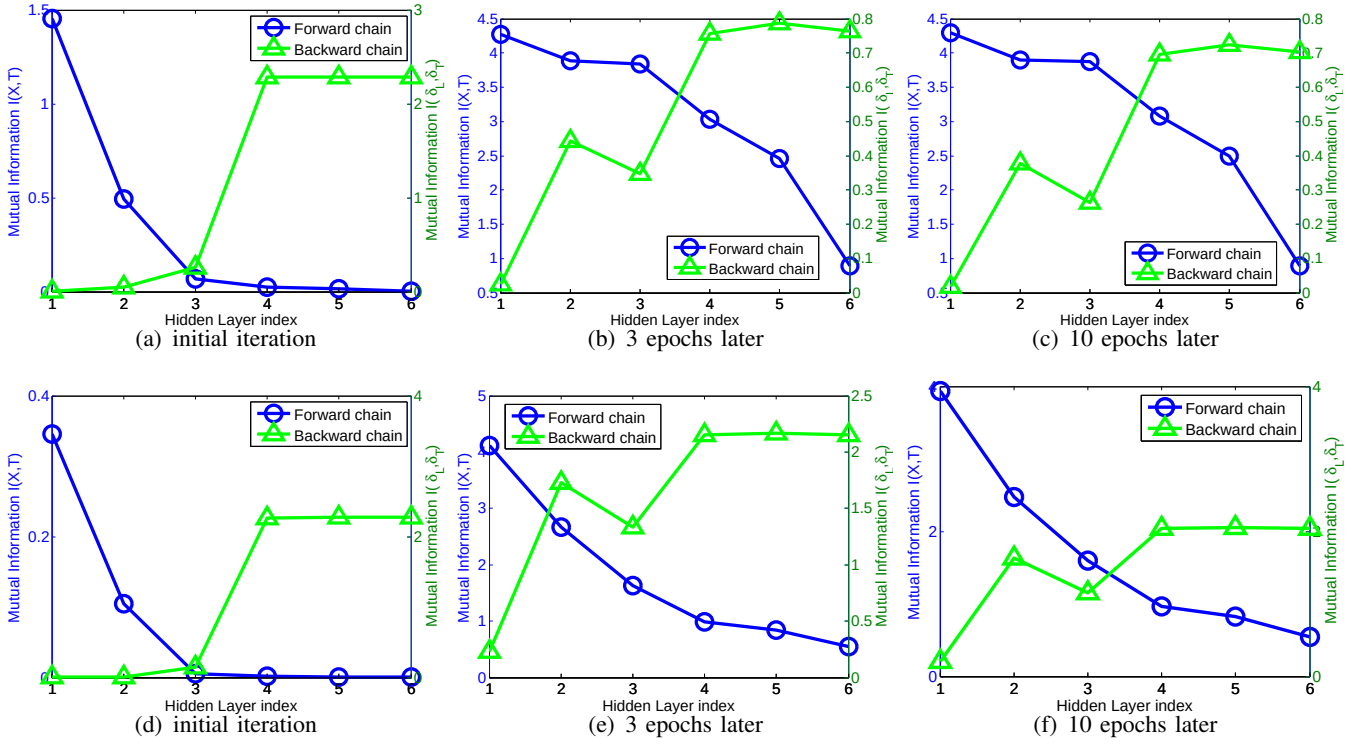


Fig. 1. Two DPIs in CNNs. (a)-(c) show the validation results, using a CNN with 2 filters in the first convolutional layer and 2 filters in the second convolutional layer; (d)-(f) show the validation results, using a CNN with 6 filters in the first convolutional layer and 12 filters in the second convolutional layer. In each subfigure, the blue curves show the MMI values between input and different layer representations, whereas the green curves show the MMI values between errors in the output layer and different hidden layers.

B. Redundancy and Synergy in Layer Representations

In this section, we explore some hidden properties, with the help of the PID framework, associated with different information theoretic quantities of convolutional layer representations in the training phase of CNNs. Particularly, we are interested in determining the redundancy and synergy amongst different feature maps and how their tradeoffs evolve with training in different CNN topologies. Moreover, we are also interested in identifying some upper or lower limits (if they exist) for these quantities.

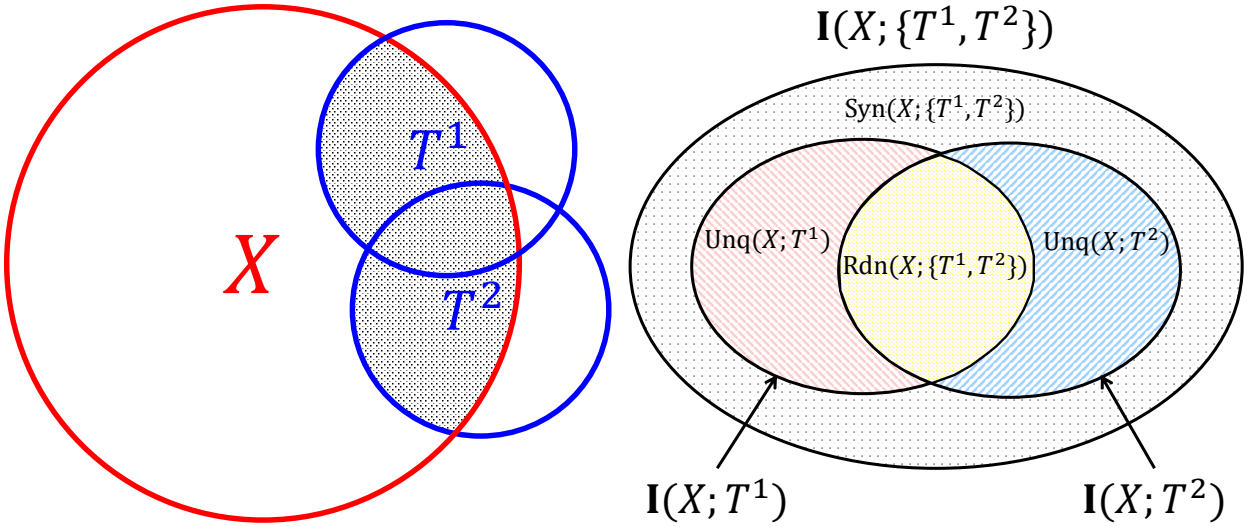
Given input signal X and two feature maps T^1 and T^2, the PID framework indicates that the MMI $\mathbf{I}(X; \{T^1, T^2\})$ can be decomposed into four non-negative components: the synergy $\mathbf{Syn}(X; \{T^1, T^2\})$ that measures the information about X provided by the coalition or combination of T^1 and T^2 (i.e., the information that cannot be captured by either T^1 or T^2 alone); the redundancy $\mathbf{Rdn}(X; \{T^1, T^2\})$ that measures the shared information about X that can be provided by either T^1 or T^2; the unique information

$\mathbf{Unq}(X;T^1)$ (or $\mathbf{Unq}(X;T^2)$) that measures the information about X that can only be provided by T^1 (or T^2). Moreover, the unique information, the synergy and the redundancy satisfy (see Fig. 2 for better understanding):

$$\mathbf{I}(X;\{T^1,T^2\}) = \mathbf{Syn}(X;\{T^1,T^2\}) + \mathbf{Rdn}(X;\{T^1,T^2\}) + \mathbf{Unq}(X;T^1) + \mathbf{Unq}(X;T^2); \quad (6)$$

$$\mathbf{I}(X;T^1) = \mathbf{Rdn}(X;\{T^1,T^2\}) + \mathbf{Unq}(X;T^1); \quad (7)$$

$$\mathbf{I}(X;T^2) = \mathbf{Rdn}(X;\{T^1,T^2\}) + \mathbf{Unq}(X;T^2). \quad (8)$$



(a) $\mathbf{I}(X;\{T^1,T^2\})$ (b) PID to $\mathbf{I}(X;\{T^1,T^2\})$

Fig. 2. Synergy and redundancy amongst different feature maps. (a) shows the interactions between input signal and two feature maps. The shadow area indicates the MMI $\mathbf{I}(X;\{T^1,T^2\})$. (b) shows the PID to $\mathbf{I}(X;\{T^1,T^2\})$.

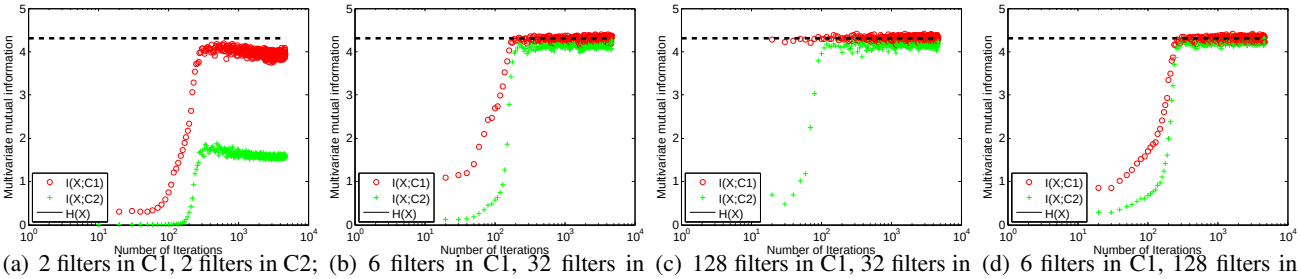
The intuitive framework for $\mathbf{I}(X;\{T^1,T^2\})$ can be straightforwardly extended for more than three variables, thus decomposing $\mathbf{I}(X;\{T^1,T^2,\cdots,T^C\})$ into much more components. For example, if $C=4$, there will be 166 individual non-negative items. Admittedly, the PID framework coupled with its Lattice decomposition offer us an intuitive manner to understand the interactions between input and different feature maps, the reliable estimation of each PID term still remains a big challenge. In fact, there is no universal agreement on the definition of synergy and redundancy among one-dimensional 3-way interactions, let alone the estimation of each synergy or redundancy item among numerous variables in high-dimensional space [19], [20]. To this end, we develop three quantities, that avoid the direct computation of synergy and redundancy, to characterize intrinsic properties of CNN layer representations. They are:

1) $\mathbf{I}(X;\{T^1,T^2,\cdots,T^C\})$, which is exactly the MMI. This quantity measures the amount of information about X that is captured by all feature maps (in one convolutional layer).

2) $\frac{2}{C(C-1)}\sum_{i=1}^{C}\sum_{j=i+1}^{C}\mathbf{I}(X;T^i)+\mathbf{I}(X;T^j)-\mathbf{I}(X;\{T^i,T^j\})$, which is referred to redundancy-synergy tradeoff. This quantity measures the (average) redundancy-synergy tradeoff in different feature maps. This is because, by Eqs. (6)-(8),

$$\mathbf{I}(X;T^i)+\mathbf{I}(X;T^j)-\mathbf{I}(X;\{T^i,T^j\}) = \mathbf{Rdn}(X;\{T^i,T^j\}) - \mathbf{Syn}(X;\{T^i,T^j\}). \quad (9)$$

Obviously, a positive value of this tradeoff implies redundancy, whereas a negative value signifies synergy [21]. Here, instead of measuring all PID terms that increase polynomially with C, we sample pairs of feature maps, calculate the information quantities for each pair, and finally compute averages

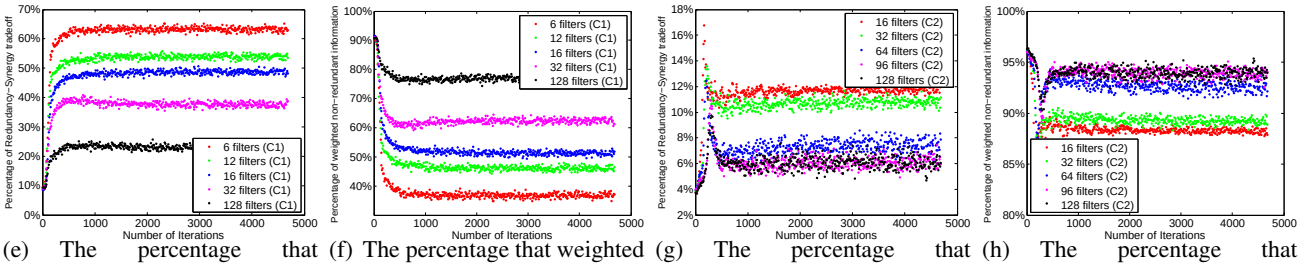


(a) 2 filters in C1, 2 filters in C2; ACA: 94.19%

(b) 6 filters in C1, 32 filters in C2; ACA: 98.18%

(c) 128 filters in C1, 32 filters in C2; ACA: 97.89%

(d) 6 filters in C1, 128 filters in C2; ACA: 98.13%



(e) The percentage that Redundancy-Synergy tradeoff accounts for the multivariate mutual information in each pair of feature maps. The network differ in the number of filters in C1.

(f) The percentage that weighted non-redundant information accounts for the multivariate mutual information in each pair of feature maps. The network differ in the number of filters in C1.

(g) The percentage that Redundancy-Synergy tradeoff accounts for the multivariate mutual information in each pair of feature maps. The network differ in the number of filters in C2.

(h) The percentage that weighted non-redundant information accounts for the multivariate mutual information in each pair of feature maps. The network differ in the number of filters in C2.

Fig. 3. The multivariate mutual information (MMI), the redundancy-synergy tradeoff, and the weighted non-redundant information in CNNs trained on MNIST dataset. (a)-(d) show the MMI in the first and the second convolutional layer representations C_1 and C_2, respectively. The dashed black line indicates the upper bound of MMI, i.e., the average mini-batch input entropy. We also report the average classification accuracy (ACA) on testing set (over 10 Monte-Carlo simulations) in each subtitle. (e) and (f) demonstrate the percentages of redundancy-synergy tradeoff and the weighted non-redundant information, with respect to MMI in each pair of feature maps, for CNNs with different number of filters in C_1, but 12 filters in C_2. (g) and (h) demonstrate the percentages of redundancy-synergy tradeoff and the weighted non-redundant information, with respect to MMI in each pair of feature maps, for CNNs with 6 filters in C_1, but different number of filters in C_2.

over all pairs to determine if synergy dominates in the training phase. Note that, the pairwise sampling procedure has been widely used in neuroscience [22].

3) $\frac{2}{C(C-1)}\sum_{i=1}^{C}\sum_{j=i+1}^{C} 2\times\mathbf{I}(X;\{T^i,T^j\})-\mathbf{I}(X;T^i)-\mathbf{I}(X;T^j)$, which is referred to weighted non-redundant information. This quantity measures the (average) amount of non-redundant information about X that is captured by pairs of feature maps. Again, by Eqs. (6)-(8),

$$2\times\mathbf{I}(X;\{T^i,T^j\})-\mathbf{I}(X;T^i)-\mathbf{I}(X;T^j)=\mathbf{Unq}(X;T^i)+\mathbf{Unq}(X;T^j)+2\times\mathbf{Syn}(X;\{T^i,T^j\}). \quad (10)$$

We call this quantity "weighted" because we overemphasized the role of synergy. Note that, the actual amount of non-redundant information is $\mathbf{Unq}(X;T^i)+\mathbf{Unq}(X;T^j)+\mathbf{Syn}(X;\{T^i,T^j\})$, rather than $\mathbf{Unq}(X;T^i)+\mathbf{Unq}(X;T^j)+2\times\mathbf{Syn}(X;\{T^i,T^j\})$.

We compute these three quantities in the training phase, and compare their values with respect to different CNN topologies. Fig. 3(a)-3(d) demonstrate the MMI values in two convolutional layers. By DPI, the maximum amount of information that each layer representation can capture is exactly the entropy of input. As can be seen, with the increase of the number of filters, the total amount of information that each convolutional layer captured also increases correspondingly. However, it is interesting to find that MMI values approach their theoretical maximum (i.e., the ensemble average entropy of mini-batch input) with only 6 filters in the first convolutional layer and 32 filters in the second convolutional layer. More filters (in a reasonable range) can improve the classification performance. However, if we blindly increase the number of filters, the classification accuracy cannot increase anymore or even becomes worse.

We argue that this phenomenon can be explained by the percentage that the redundancy-synergy tradeoff or the weighted non-redundant information accounts for the MMI in each pair of feature maps,

i.e., $\frac{2}{C(C-1)}\sum_{i=1}^{C}\sum_{j=i+1}^{C}\frac{\mathbf{I}(X;T^i)+\mathbf{I}(X;T^j)-\mathbf{I}(X;\{T^i,T^j\})}{\mathbf{I}(X;\{T^i,T^j\})}$ or $\frac{2}{C(C-1)}\sum_{i=1}^{C}\sum_{j=i+1}^{C}\frac{2\times\mathbf{I}(X;\{T^i,T^j\})-\mathbf{I}(X;T^i)-\mathbf{I}(X;T^j)}{\mathbf{I}(X;\{T^i,T^j\})}$. In fact, by referring to Fig. 3(e)-3(h), it is obvious that more filters can push the network towards an improved redundancy-synergy tradeoff, i.e., the synergy gradually dominates in each pair of feature maps with the increase of filter numbers. That is perhaps one of the main reasons why the increased number of filters can lead to better classification performance, even though the total multivariate mutual information stays the same. However, if we look deeper, it seems that the redundancy is always larger than the synergy such that their tradeoff can never cross the x-axis. This may suggest a (virtual) lower bound on the redundancy-synergy tradeoff. On the other hand, one should note that the amount of non-redundant information is always less than (or upper bounded by) the MMI no matter the number of filters, therefore it is impossible to improve the classification performance by blindly increasing the number of filters.

Having illustrated the DPIs and the redundancy-synergy tradeoffs, it is easy to summarize some implications concerning the design and training of CNNs. First, as a possible application of DPI in the error backpropagation chain, one has to realize that the DPI provides an indicator on where to perform the "bypass" in the recently proposed Relay backpropagation [23]. Second, the DPIs and the redundancy-synergy tradeoff may give some guidelines on the depth and width of CNNs. Indeed, we need multiple layers to denoise the input and to extract representations from different abstract levels. However, more layers will lead to severe information loss. The same interpretation goes for the number of filters in convolutional layers, we need sufficient number of filters to ensure the layer representations can extract and transfer input information as much as possible and to learn a good redundancy-synergy tradeoff. However, too many filters do not always lead to the increased amount of the non-redundant information, as the minimum probability of classification error is upper bounded by the mutual information expressed in different forms (e.g., [24], [25]).

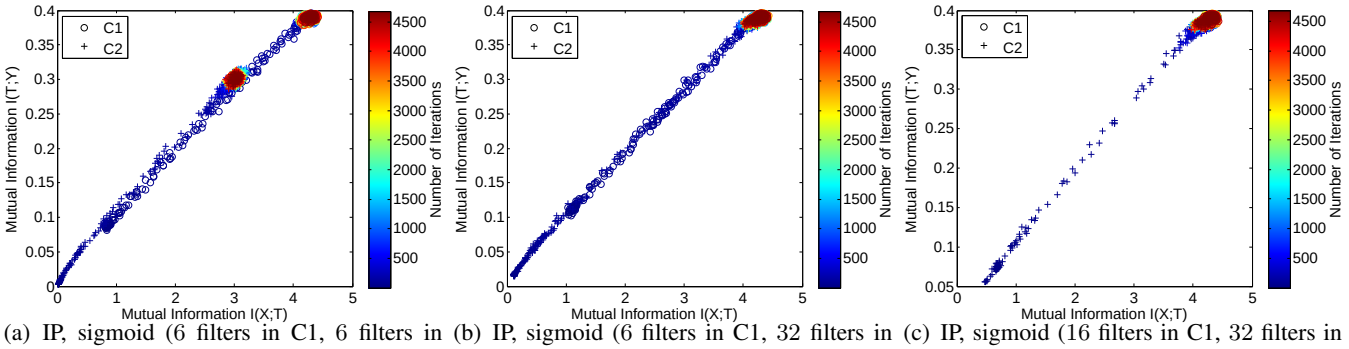
C. Revisiting the Information Plane (IP)

The behaviors of curves in the IP is currently a controversial issue. Recall the discrepancy reported by Saxe *et al.* [5], the existence of compression phase observed by Shwartz-Ziv and Tishby [4] depends on the adopted nonlinearity functions: double-sided saturating nonlinearities like "tanh" or "sigmoid" yield a compression phase, but linear activation functions and single-sided saturating nonlinearities like the "ReLU" do not. Interestingly, Noshad *et al.* [26] employed dependence graphs to estimate mutual information values and observed the compression phase even using "ReLU" activation functions. On the other hand, Goldfeld *et al.* [27] argued that compression is due to layer representations clustering, but it is hard to observe the compression in large network. We disagree with this attribution of different behavior to the nonlinear activation functions. Instead, we often forget that, rarely, estimators share all the properties of the statistically defined quantities [28]. Hence, variability in the displayed behavior is mostly likely attributed to different estimators[2], although this argument is rarely invoked in the literature. This is the reason we suggest that a first step before analyzing the information plane curves, is to show that the employed estimators meet the expectation of the DPI (or similar known properties of the statistical quantities). We show above that our Rényi's entropy estimator passes this test.

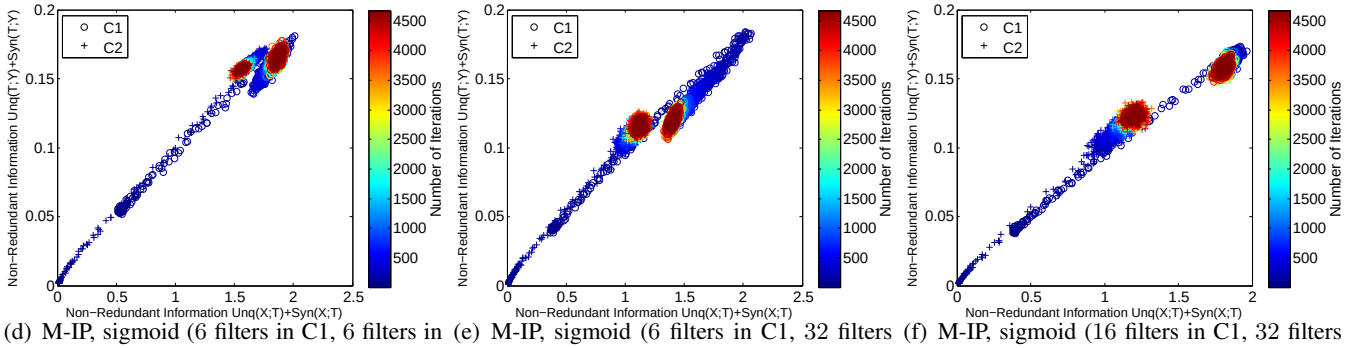
The IPs for different CNN topologies on MNIST dataset are shown in Fig. 4. From the first row, both $\mathbf{I}(X;T)$ and $\mathbf{I}(T;Y)$ increase rapidly up to a certain point with the SGD iterations, independently of the adopted activation functions or the number of filters in the convolutional layers. This result conforms to the description in [27], suggesting that the behaviour of CNNs in the IP not being the same as that of the MLPs in [4], [5], [26] and our intrinsic dimensionality hypothesis in [6] is specific to SAEs. However, if we remove the redundancy in $\mathbf{I}(X;T)$ and $\mathbf{I}(T;Y)$, and only preserve the unique information and the synergy (i.e., substituting $\mathbf{I}(X;T)$ and $\mathbf{I}(T;Y)$ with their corresponding (average) weighted non-redundant information defined in Section III-B), it is easy to observe the compression phase in the modified

[2]Shwartz-Ziv and Tishby [4] use the basic Shannon's definition and estimate mutual information by dividing neuron activation values into 30 equal-interval bins, whereas the base estimator used by Saxe *et al.* [5] provides Kernel Density Estimator (KDE) based lower and upper bounds on the true mutual information [29], [26].

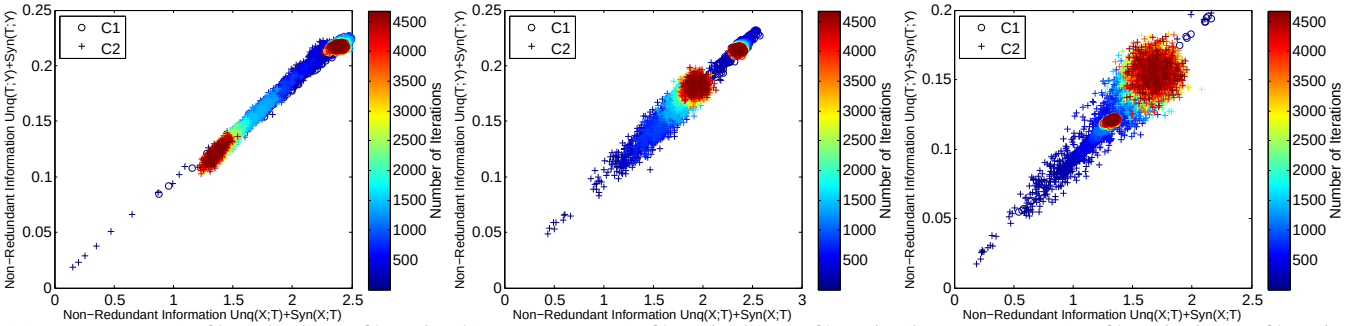
IP. Moreover, it seems that "sigmoid" is more likely to incur the compression, compared with "ReLU", where this intensity can be attributed to the nonlinearity. Our result shed light on the discrepancy in [4] and [5], and refined the argument in [26].



(a) IP, sigmoid (6 filters in C1, 6 filters in C2) (b) IP, sigmoid (6 filters in C1, 32 filters in C2) (c) IP, sigmoid (16 filters in C1, 32 filters in C2)



(d) M-IP, sigmoid (6 filters in C1, 6 filters in C2) (e) M-IP, sigmoid (6 filters in C1, 32 filters in C2) (f) M-IP, sigmoid (16 filters in C1, 32 filters in C2)



(g) M-IP, ReLU (6 filters in C1, 6 filters in C2) (h) M-IP, ReLU (6 filters in C1, 32 filters in C2) (i) M-IP, ReLU (16 filters in C1, 32 filters in C2)

Fig. 4. The Information Plane (IP) and modified Information Plane (M-IP) of different CNN topologies trained on MNIST dataset. The # of filters in $C1$, the # of filters in $C2$, and the adopted activation function are indicated in the subtitle of each plot. The curves in IP increase rapidly up to a point without compression (see (a)-(c)). By contrast, it is easy to observe the compression in M-IP (see (d), (e), (g) and (h)). Moreover, compared with ReLU, sigmoid is more likely to incur the compression (e.g., comparing (e) with (h), or (f) with (i)).

IV. Conclusions and Future Work

This paper presents a systematic method to analyze convolutional neural networks (CNNs) mapping and training from an information theoretic perspective. Using the multivariate extension of the matrix-based Rényi's α-entropy functional, we validated two data processing inequalities in CNNs. The introduction of partial information decomposition (PID) framework enables us to pin down the redundancy-synergy tradeoff in layer representations. We also analyzed the behaviors of curves in the information plane, aiming at clarify the debate on the existence of compression in DNNs. Future works are twofold:

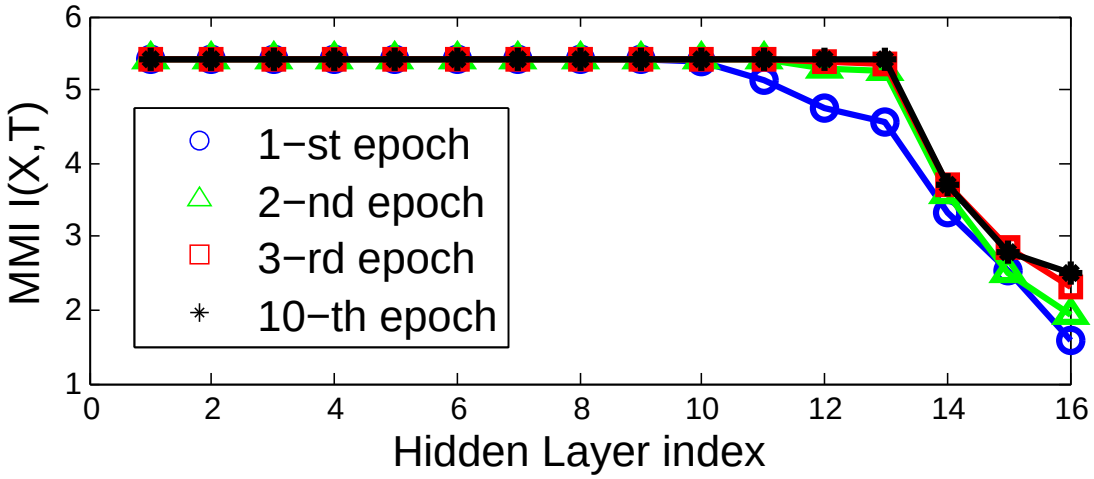


Fig. 5. DPI in VGG-16 on CIFAR-10: $\mathbf{I}(X,T_1) \geq \mathbf{I}(X,T_2) \geq \cdots \geq \mathbf{I}(X,T_K)$. Layer 1 to Layer 13 are convolutional layers, whereas Layer 14 to Layer 16 are fully-connected layers.

1) All the information quantities mentioned in this paper are estimated based on a vector rastering of samples, i.e., each layer input (e.g., an input image, a feature map) is first converted to a single vector before entropy or mutual information estimation. Albeit its simplicity, we distort spatial relationships amongst neighboring pixels. Therefore, a question remains on the reliable information theoretic estimation that is feasible to a tensor structure.

2) We look forward to evaluating our estimators on more complex CNN architectures, such as VGGNet [30] and ResNet [31]. According to our observation, it is easy to validate the DPI and the rapid increase of mutual information (in top layers) in VGG-16 on CIFAR-10 dataset [32] (see Fig. 5). However, it seems that the MMI values in bottom layers are likely to be "saturated". The problem arises when we try to take the Hadamard product of the kernel matrices of each feature map in Eq. (5). The elements in these (normalized) kernel matrices have values between 0 and 1, and taking the entrywise product of, e.g., 512 such matrices like in the convolutional layer of VGG-16, will tend towards a matrix with diagonal entries $1/n$ and nearly zero everywhere else. The eigenvalues of the resulting matrix will quickly have almost the same value across training epochs. We aim to solve this limitation in future works.

REFERENCES

[1] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *IEEE ITW*, 2015, pp. 1–5.

[2] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," *JMLR*, vol. 19, no. 1, pp. 1947–1980, 2018.

[3] T. Tax, P. A. Mediano, and M. Shanahan, "The partial information decomposition of generative neural network models," *Entropy*, vol. 19, no. 9, p. 474, 2017.

[4] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[5] A. M. Saxe *et al.*, "On the information bottleneck theory of deep learning," in *ICLR*, 2018.

[6] S. Yu and J. C. Principe, "Understanding autoencoders with information theoretic concepts," *arXiv preprint arXiv:1804.00057*, 2018.

[7] L. G. Sanchez Giraldo, M. Rao, and J. C. Principe, "Measures of entropy from data using infinitely divisible kernels," *IEEE Transactions on Information Theory*, vol. 61, no. 1, pp. 535–548, 2015.

[8] F. Camastra and A. Staiano, "Intrinsic dimension estimation: Advances and open problems," *Information Sciences*, vol. 328, pp. 26–41, 2016.

[9] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *JMLR*, vol. 13, no. Jan, pp. 27–66, 2012.

[10] S. Yu, L. G. Sanchez Giraldo, R. Jenssen, and J. C. Principe, "Multivariate extension of matrix-based renyi's α-order entropy functional," *arXiv preprint arXiv:1808.07912*, 2018.

[11] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," *arXiv preprint arXiv:1004.2515*, 2010.

[12] A. Rényi, "On measures of entropy and information," in *Proc. of the 4th Berkeley Sympos. on Math. Statist. and Prob.*, vol. 1, 1961, pp. 547–561.

[13] J. C. Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

[14] M. Müller-Lennert, F. Dupuis, O. Szehr, S. Fehr, and M. Tomamichel, "On quantum rényi entropies: A new generalization and some properties," *J. Math. Phys.*, vol. 54, no. 12, p. 122203, 2013.
[15] R. Bhatia, "Infinitely divisible matrices," *The American Mathematical Monthly*, vol. 113, no. 3, pp. 221–235, 2006.
[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[17] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.
[18] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26.
[19] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, "Quantifying unique information," *Entropy*, vol. 16, no. 4, pp. 2161–2183, 2014.
[20] V. Griffith and C. Koch, "Quantifying synergistic mutual information," in *Guided Self-Organization: Inception*. Springer, 2014, pp. 159–190.
[21] A. J. Bell, "The co-information lattice," in *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA*, vol. 2003, 2003.
[22] N. Timme, W. Alford, B. Flecker, and J. M. Beggs, "Synergy, redundancy, and multivariate information measures: an experimentalists perspective," *J. Comput. Neurosci.*, vol. 36, no. 2, pp. 119–140, 2014.
[23] L. Shen and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *ECCV*, 2016, pp. 467–482.
[24] M. Hellman and J. Raviv, "Probability of error, equivocation, and the chernoff bound," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 368–372, 1970.
[25] I. Sason and S. Verdú, "Arimoto–rényi conditional entropy and bayesian m-ary hypothesis testing," *IEEE Transactions on Information Theory*, vol. 64, no. 1, pp. 4–25, 2018.
[26] M. Noshad and A. O. Hero III, "Scalable mutual information estimation using dependence graphs," *arXiv preprint arXiv:1801.09125*, 2018.
[27] Z. Goldfeld *et al.*, "Estimating information flow in neural networks," *arXiv preprint arXiv:1810.05728*, 2018.
[28] L. Paninski, "Estimation of entropy and mutual information," *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003.
[29] A. Kolchinsky and B. Tracey, "Estimating mixture entropy with pairwise distances," *Entropy*, vol. 19, no. 7, p. 361, 2017.
[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[32] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

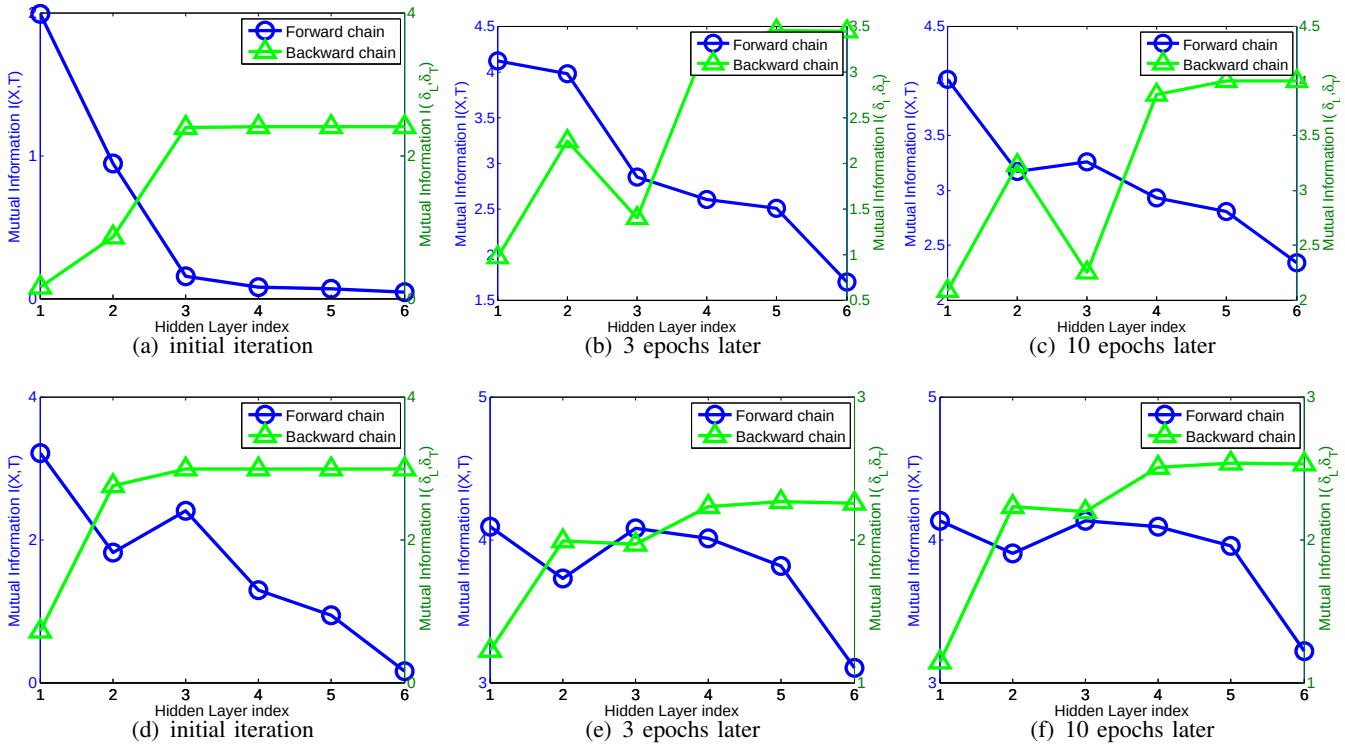
APPENDIX

A. Results on Fashion-MNIST

We represent results on Fashion-MNIST dataset. Specifically, Fig. 6 validates the existence of two data processing inequalities (DPIs) at the initial training stage, after 3 epochs' training and at the final training stage, respectively. Obviously, two DPIs hold in most of the cases. Again, there are a few disruptions in the error backpropagation chain due to the improper scale parameter σ in our estimator selected using the Silverman's rule of thumb.

Fig. 7 demonstrates the values of our developed three quantities, i.e., the multivariate mutual information (MMI), the redundancy-synergy tradeoff, and the weighted non-redundant information, in the training phase of CNNs. As can be seen in Fig. 7(a)-7(d), with the increase of the number of filters, the total amount of information that each convolutional layer captured also increases correspondingly. However, it is interesting to find that MMI values approach their theoretical maximum (i.e., the ensemble average entropy of mini-batch input) with only 6 filters in the first convolutional layer and 32 filters in the second convolutional layer. More filters (in a reasonable range) can improve the classification performance. However, if we blindly increase the number of filters in two convolutional layers, the classification accuracy cannot increase anymore or becomes even worse. The tendency of redundancy-synergy tradeoff and weighted non-redundant information are shown in Fig. 7(e)-7(h), it is obvious that more filters can push the network towards an improved redundancy-synergy tradeoff, i.e., the synergy gradually dominates in each pair of feature maps with the increase of number of filters.

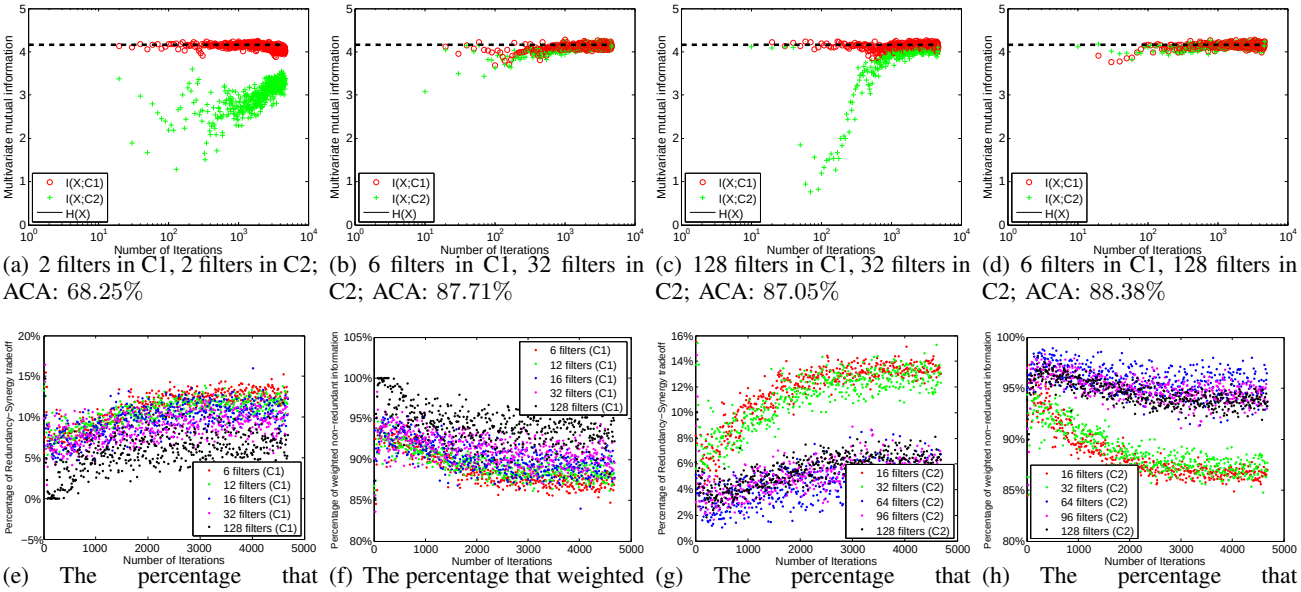
We finally show the information planes (IPs) for different CNN topologies in Fig. 8. Again, we did not observe any compression in the original IP using different activation functions. However, in the modified IP that only only monitors the non-redundant information, it is easy to observe that sigmoid is more likely to incur the compression, compared with ReLU.



(a) initial iteration (b) 3 epochs later (c) 10 epochs later

(d) initial iteration (e) 3 epochs later (f) 10 epochs later

Fig. 6. Two DPIs in CNNs. (a)-(c) show the validation results, using a CNN with 2 filters in the first convolutional layer and 2 filters in the second convolutional layer; (d)-(f) show the validation results, using a CNN with 6 filters in the first convolutional layer and 12 filters in the second convolutional layer. In each subfigure, the blue curves show the MMI values between input and different layer representations, whereas the green curves show the MMI values between errors in the output layer and different hidden layers.



(a) 2 filters in C1, 2 filters in C2; ACA: 68.25%

(b) 6 filters in C1, 32 filters in C2; ACA: 87.71%

(c) 128 filters in C1, 32 filters in C2; ACA: 87.05%

(d) 6 filters in C1, 128 filters in C2; ACA: 88.38%

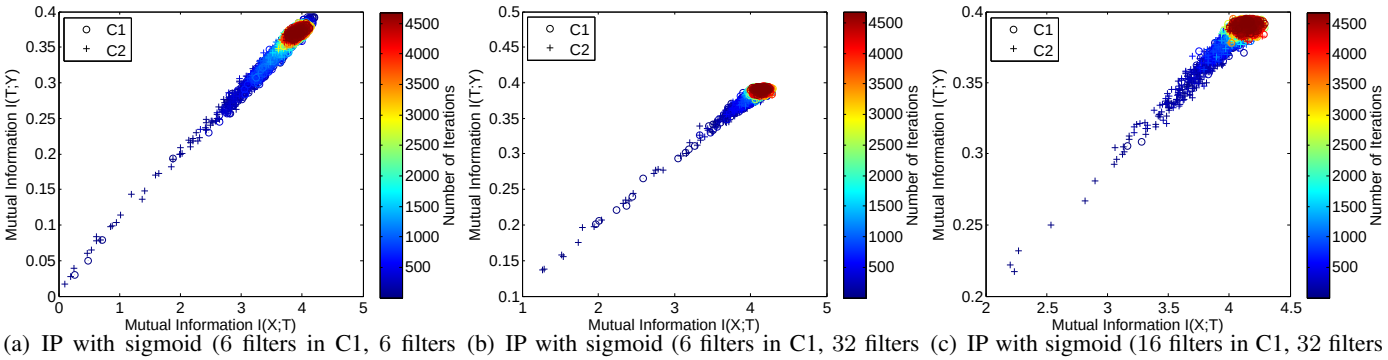
(e) The percentage that Redundancy-Synergy tradeoff accounts for the multivariate mutual information in each pair of feature maps. The network differ in the number of filters in C1.

(f) The percentage that weighted non-redundant information accounts for the multivariate mutual information in each pair of feature maps. The network differ in the number of filters in C1.

(g) The percentage that Redundancy-Synergy tradeoff accounts for the multivariate mutual information in each pair of feature maps. The network differ in the number of filters in C2.

(h) The percentage that weighted non-redundant information accounts for the multivariate mutual information in each pair of feature maps. The network differ in the number of filters in C2.

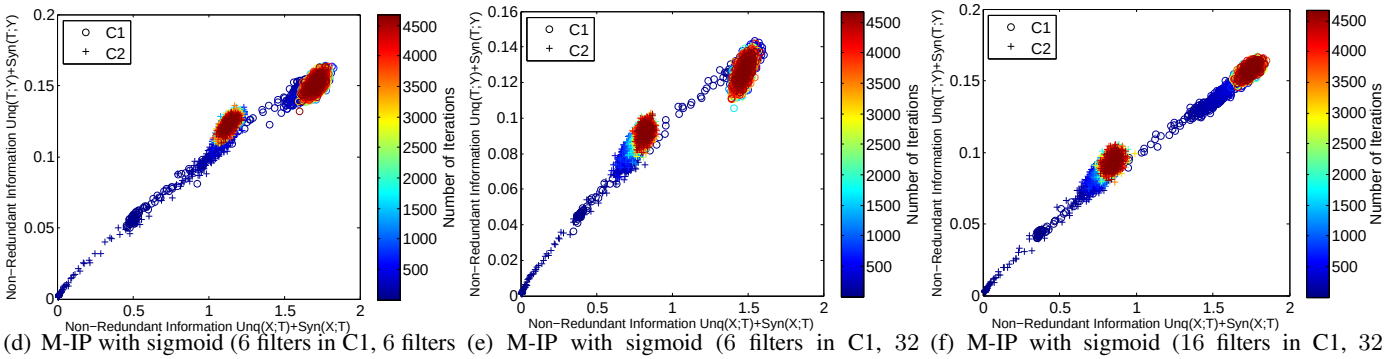
Fig. 7. The multivariate mutual information (MMI), the redundancy-synergy tradeoff, and the weighted non-redundant information in CNNs trained on MNIST dataset. (a)-(d) show the MMI in the first and the second convolutional layer representations C_1 and C_2, respectively. The dashed black line indicates the upper bound of MMI, i.e., the average mini-batch input entropy. We also report the average classification accuracy (ACA) on testing set (over 10 Monte-Carlo simulations) in each subtitle. (e) and (f) demonstrate the percentages of redundancy-synergy tradeoff and the weighted non-redundant information, with respect to MMI in each pair of feature maps, for CNNs with different number of filters in C_1, but 12 filters in C_2. (g) and (h) demonstrate the percentages of redundancy-synergy tradeoff and the weighted non-redundant information, with respect to MMI in each pair of feature maps, for CNNs with 6 filters in C_1, but different number of filters in C_2.



(a) IP with sigmoid (6 filters in C1, 6 filters in C2)

(b) IP with sigmoid (6 filters in C1, 32 filters in C2)

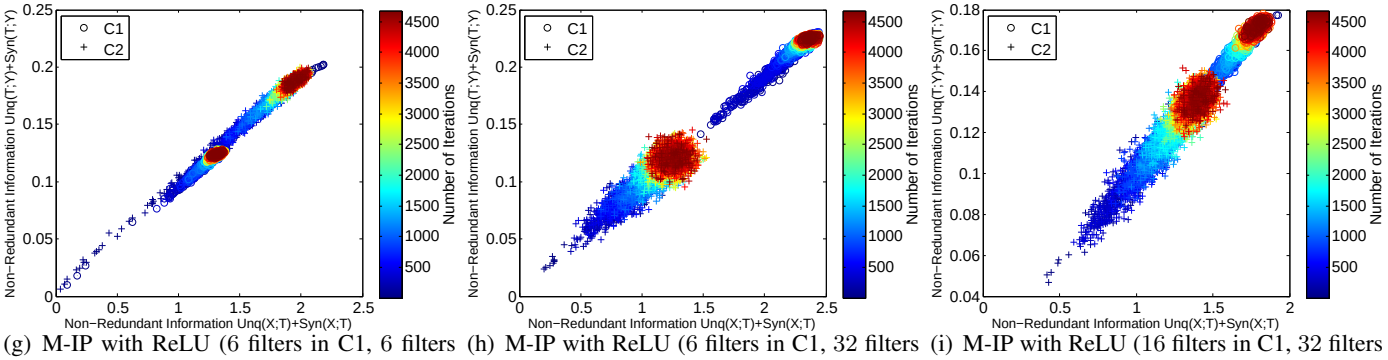
(c) IP with sigmoid (16 filters in C1, 32 filters in C2)



(d) M-IP with sigmoid (6 filters in C1, 6 filters in C2)

(e) M-IP with sigmoid (6 filters in C1, 32 filters in C2)

(f) M-IP with sigmoid (16 filters in C1, 32 filters in C2)



(g) M-IP with ReLU (6 filters in C1, 6 filters in C2)

(h) M-IP with ReLU (6 filters in C1, 32 filters in C2)

(i) M-IP with ReLU (16 filters in C1, 32 filters in C2)

Fig. 8. The Information Plane (IP) and modified Information Plane (M-IP) of different CNN topologies trained on Fashion-MNIST dataset. The # of filters in $C1$, the # of filters in $C2$, and the adopted activation function are indicated in the subtitle of each plot. The curves in IP increase rapidly up to a point without compression (see (a)-(c)). By contrast, it is easy to observe the compression in M-IP (see (d), (e) and (g)). Moreover, compared with ReLU, sigmoid is more likely to incur the compression (e.g., comparing (e) with (h)).