Explaining Knowledge Distillation by Quantifying the Knowledge

Xu Cheng
Shanghai Jiao Tong University
xcheng8@sjtu.edu.cn

Zhefan Rao
Huazhong University of Science & Technology
rzf19971013@gmail.com

Yilan Chen
Xi'an Jiaotong University
chenyilan@stu.xjtu.edu.cn

Quanshi Zhang
Shanghai Jiao Tong University
zqs1022@sjtu.edu.cn

Abstract

This paper presents a method to interpret the success of knowledge distillation by quantifying and analyzing task-relevant and task-irrelevant visual concepts that are encoded in intermediate layers of a deep neural network (DNN). More specifically, three hypotheses are proposed as follows. ***1.*** *Knowledge distillation makes the DNN learn more visual concepts than learning from raw data.* ***2.*** *Knowledge distillation ensures that the DNN is prone to learning various visual concepts simultaneously. Whereas, in the scenario of learning from raw data, the DNN learns visual concepts sequentially.* ***3.*** *Knowledge distillation yields more stable optimization directions than learning from raw data. Accordingly, we design three types of mathematical metrics to evaluate feature representations of the DNN. In experiments, we diagnosed various DNNs, and above hypotheses were verified.*

1. Introduction

The success of knowledge distillation [16] has been demonstrated in various studies [31, 45, 11]. It transfers knowledge from a well-learned deep neural network (DNN), namely the teacher network, to another DNN, namely the student network. However, explaining how and why knowledge distillation outperforms learning from raw data still remains a challenge.

In this work, we aim to analyze the success of knowledge distillation from a new perspective, *i.e.* quantifying the knowledge encoded in the intermediate layer of a DNN. We quantify and compare the amount of the knowledge encoded in the DNN learned via knowledge distillation and the DNN learned from raw data, respectively. Here, the DNN learned from raw data is termed the ***baseline network***. In this research, the amount of the knowledge of a specific layer is measured as the number of visual concepts (*e.g.* object parts like tails, heads and etc.), which is shown in Figure 1. These visual concepts activate the feature map of this specific layer and are used for prediction.

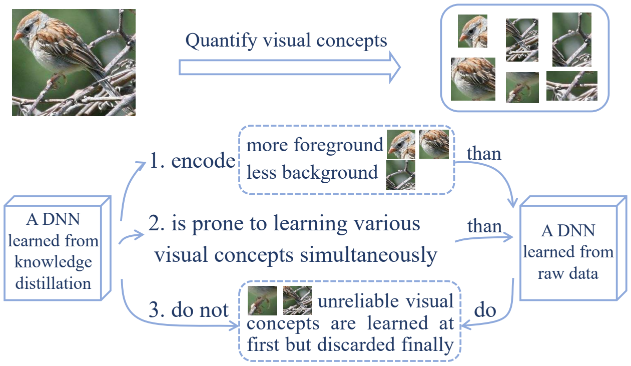
Quantify visual concepts
1. encode
more foreground
less background
than
A DNN learned from knowledge distillation
2. is prone to learning various visual concepts simultaneously
than
A DNN learned from raw data
3. do not
unreliable visual concepts are learned at first but discarded finally
do

Figure 1. Explanation of knowledge distillation by quantifying visual concepts. Three hypotheses are proposed and verified as follows. 1. Knowledge distillation makes the DNN learn more visual concepts than learning from raw data. 2. Knowledge distillation ensures that the DNN is prone to learning various visual concepts simultaneously. 3. Knowledge distillation yields more stable optimization directions than learning from raw data.

We design three types of mathematical metrics to analyze task-relevant and task-irrelevant visual concepts. Then, these metrics are used to quantitatively verify three hypotheses as follows.

Hypothesis 1: Knowledge distillation makes the DNN learn more visual concepts. In this paper, a visual concept is defined as an image region, whose information is significantly less discarded than the average information discarded in background regions, and is mainly used by the DNN. We distinguish visual concepts that are relevant to the task away from other concepts, *i.e.* task-irrelevant concepts. Let us take the classification task as an example. As is vividly shown in Figure 1, visual concepts on the foreground are usually regarded task-relevant, while those on the background are considered task-irrelevant.

According to the information-bottleneck theory [41, 36], DNNs tend to expose task-relevant visual concepts and discard task-irrelevant concepts to learn discriminative features. Compared to the baseline network (learned from raw data), a well-trained teacher network is usually considered to encode more task-relevant visual concepts and/or less task-irrelevant concepts. Because the student network mimics the logic of the teacher network, the student network is supposed to contain more task-relevant visual concepts and less task-irrelevant concepts.

Hypothesis 2: Knowledge distillation ensures the DNN is prone to learning various visual concepts simultaneously. In comparison, the baseline network tends to learn visual concepts sequentially, *i.e.* paying attention to different concepts in different epochs.

Hypothesis 3: Knowledge distillation usually yields more stable optimization directions than learning from raw data. When learning from raw data, the DNN usually tries to model various visual concepts in early epochs and then discard non-discriminative ones in later epochs [41, 36], which leads to unstable optimization directions. We name the phenomenon of inconsistent optimization directions through different epochs ***"detours"***[1] for short in this paper. In comparison, during the knowledge distillation, the teacher network directly guides the student network to target visual concepts without *significant detours*[1]. Let us take the classification of birds as an example. The baseline network tends to extract features from the head, belly, tail, and tree-branch in early epochs, and later discards features from the tree-branch. Whereas, the student network directly learns features from the head and belly with less detours[1].

Methods: We propose three types of mathematical metrics to quantify visual concepts hidden in intermediate layers of a DNN, and analyze how visual concepts are learned during the learning procedure. These metrics measure 1. the number of visual concepts, 2. the learning speed of different concepts, 3. the stability of optimization directions, respectively. We use these metrics to analyze the student network and the baseline network in comparative studies to prove above three hypotheses. More specifically, the student network is learned via knowledge distillation, and the baseline network learned from raw data is constructed to have the same architecture as the student network.

Note that visual concepts should be quantified without subjective manual annotations. There are mainly two reasons. 1) It is impossible for people to annotate all kinds of potential visual concepts in the world. 2) For a rigorous research, the subjective bias in human annotations should not affect the quantitative metric. To this end, [14, 26] leverages the entropy to quantify visual concepts encoded in an intermediate layer.

Contributions: Our contributions can be summarized as follows.

1. We propose a method to quantify visual concepts encoded in intermediate layers of a DNN.

2. Based on the quantification of visual concepts, we propose three types of metrics to interpret knowledge distillation from the view of knowledge representations.

3. Three hypotheses about knowledge distillation are proposed and verified, which shed light on the explanation of knowledge distillation.

2. Related Work

Although deep neural networks have exhibited superior performance in various tasks, they are still regarded as black boxes. Previous studies of interpreting DNNs can be roughly summarized into semantic explanations and mathematical explanations of the representation capacity.

Semantic explanations for DNNs: An intuitive way to interpret DNNs is to visualize visual concepts encoded in intermediate layers of DNNs. Feature visualization methods usually show concepts that may significantly activate a specific neuron of a certain layer. Gradient-based methods [47, 37, 46, 27] used gradients of outputs *w.r.t.* the input image to measure the importance of intermediate-layer activation units or input units. Inversion-based [5] methods inverted feature maps of convolutional layers into images. From visualization results, people roughly understand visual concepts encoded in intermediate layers of DNNs. For example, filters in low layers usually encode simple visual concepts such as edges and textures, and filters in high layers usually encode concepts like objects and patterns.

Other methods usually estimated the pixel-wise attribution/importance/saliency on an input image, which measured the influence of each input pixel to the final output [30, 25, 20, 9]. Some methods explored the saliency of the input image using intermediate-layer features, such as CAM [52], Grad-CAM [34], and Grad-CAM++ [2]. Zhou *et al*. [51] computed the actual image-resolution receptive field of neural activations in a feature map.

Bau *et al*. [1] disentangled feature representations into semantic concepts using human annotations. Fong and Vedaldi [8] demonstrated that a DNN used multiple filters to represent a specific semantic concept. Zhang *et al*. used an explanatory graph [48] and a decision tree [50] to represent hierarchical compositional part representations in CNNs. TCAV [19] measured the importance of user-defined concepts to classification.

Another direction of explainable AI is to learn a DNN with interpretable feature representations in an unsupervised or weakly-supervised manner. In the capsule network [33], activities of each capsule encoded various properties. The interpretable CNN [49] learned object part features without part annotations. InfoGAN [4] and β-

[1] "Detours" indicate the phenomenon that a DNN tries to model various visual concepts in early epochs and discard non-discriminative ones later.

VAE [15] learned interpretable factorised latent representations for generative networks.

In contrast, in this research, the quantification of intermediate-layer visual concepts requires us to design metrics with coherency and generality. *I.e.* unlike previous studies compute importance/saliency/attention [47, 37, 46, 27] based on heuristic assumptions or using massive human-annotated concepts [1] to explain network features, we quantify visual concepts using the conditional entropy of the input. The entropy is a generic tool with strong connections to various theories, *e.g.* the information-bottleneck theory [41, 36]. Moreover, the coherency allows the same metric to ensure fair comparisons between layers of a DNN, and between DNNs learned in different epochs.

Mathematical explanations for the representation capacity of DNNs: Evaluating the representation capacity of DNNs mathematically provides a new perspective for explanations. The information-bottleneck theory [41, 36] used the mutual information to evaluate the representation capacity of DNNs [13, 43]. The stiffness [10] was proposed to diagnose the generalization of a DNN. The CLEVER score [40] was used to estimate the robustness of neural networks. The Fourier analysis [44] was applied to explain the generalization of DNNs learned by stochastic gradient descent. Novak *et al.* [28] investigated the correlation between the sensitivity of trained neural networks and generalization. Canonical correlation analysis (CCA) [21] was used to measure the similarity between representations of neural networks. Chen *et al.* [3] proposed instance-wise feature selection via mutual information for model interpretation. Zhang *et al.* [23] explored knowledge consistency between DNNs.

Different from previous methods, our research aims to bridge the gap between mathematical explanations and semantic explanations. We use the entropy of the input to measure the number of visual concepts in a DNN. Furthermore, we quantify visual concepts on the background and the foreground *w.r.t.* the input image, explore whether a DNN learn various concepts simultaneously or sequentially, and analyze the stability of optimization directions.

Knowledge distillation: knowledge distillation is a popular and successful technique in knowledge transferring. Hinton *et al.* [16] considered "soft targets" led to the superior performance of knowledge distillation. Furlanello *et al.* [11] explained the dark knowledge transferred from the teacher to the student as importance weighting.

From a theoretical perspective, Lopez-Paz *et al.* [24] interpreted knowledge distillation as a form of learning with privileged information. Phuong *et al.* [29] explained the success of knowledge distillation from the view of data distribution, optimization bias, and the size of the training set.

However, to the best of our knowledge, the mathematical explanations for knowledge distillation are rare. In this paper, we interpret knowledge distillation from a new perspective, *i.e.* quantifying, analyzing, and comparing visual concepts encoded in intermediated layers between DNNs learned by knowledge distillation and DNNs learned purely from raw data mathematically.

3. Algorithm

In this section, we are given a pre-trained DNN (*i.e.* the teacher network) and then distill it into another DNN (*i.e.* the student network). In this way, we aim to compare and explain the difference between the student network and the DNN learned from raw data (*i.e.* the baseline network). To simplify the story, we limit our attention to the task of object classification. Let $x \in R^n$ denote the input image, and $f_T(x), f_S(x) \in R^L$ denote intermediate-layer features of the teacher network and its corresponding student network, respectively. Knowledge distillation is conducted to force $f_S(x)$ to approximate $f_T(x)$. Classification results of the teacher and the student are given as $y_T = g_T(f_T(x))$ and $y_S = g_S(f_S(x)) \in R^c$, respectively.

We compare visual concepts encoded in the baseline network and those in the student network to explain knowledge distillation. For a fair comparison, the baseline network has the same architecture as the student network.

3.1. Preliminaries: Quantification of Information Discarding

According to the information-bottleneck theory [41, 36], the information of the input image is gradually discarded through layers. [14, 26] proposed a method to quantify the input information that was encoded in a specific intermediate layer of a DNN, *i.e.* measuring how much input information was neglected when the DNN extracted the feature of this layer. The information discarding is formulated as the conditional entropy $H(X')$ of the input, given the intermediate-layer feature $f^* = f(x)$, as follows.

$$H(X') \quad s.t. \ \forall x' \in X', \quad \| f(x') - f^* \|^2 \leq \tau \tag{1}$$

X' denotes a set of images, which correspond to the concept of a specific object instance. The concept of the object is assumed to be represented by a small range of features $\| f(x') - f^* \|^2 \leq \tau$, where τ is a small positive scalar. It was assumed that x' follows an $i.i.d.$ Gaussian distribution, $x' \sim \mathcal{N}(x, \Sigma = diag(\sigma_1^2, \ldots, \sigma_n^2))$, where σ_i controls the magnitude of the perturbation at each i-th pixel. n denotes the number of pixels of the input image. In this way, the assumption of the Gaussian distribution ensures that the entropy $H(X')$ of the entire image can be decomposed into pixel-level entropies $\{H_i\}$ as follows.

$$H(X') = \sum_{i=1}^{n} H_i \tag{2}$$

where $H_i = \log \sigma_i + \frac{1}{2}\log(2\pi e)$ measures the discarding of pixel-wise information. Please see [14, 26] for details.

3.2. Quantification of visual concepts

> **Hypothesis 1**: Knowledge distillation makes the DNN learn more reliable visual concepts than learning from raw data.

In this section, we aim to compare the number of visual concepts that are encoded in the baseline network and those in the student network, so as to verify the above hypothesis.

Using the annotated concepts or not: For comparison, we try to define and quantify visual concepts encoded in the intermediate layer of a DNN (either the student network or the baseline network). Note that, we do **not** study visual concepts defined by human annotations. For example, Bau *et al.* [1] defined visual concepts of objects, parts, textures, scenes, materials, and colors by using manual annotations. However, this research requires us to use and quantify visual concepts without explicit names, which cannot be accurately labeled. These visual concepts are usually referred to as ***"Dark Matters"*** [42].

There are mainly two reasons to use dark-matter visual concepts, instead of traditional semantic visual concepts. 1. There exist no standard definitions for semantic visual concepts, and the taxonomy of semantic visual concepts may have significant subjective bias. 2. The cost of annotating all visual concepts is usually unaffordable.

Metric: In this paper, we quantify dark-matter visual concepts from the perspective of the information theory. Given a pre-trained DNN, a set of training images $\mathbf{I}$ and an input image $x \in \mathbf{I}$, let us focus on the pixel-wise information discarding H_i *w.r.t.* the intermediate-layer feature $f^* = f(x)$. High pixel-wise entropies $\{H_i\}$ (shown in Equation (2)), indicate that the DNN neglects more information of these pixels. Whereas, the DNN mainly utilizes pixels with low entropies $\{H_i\}$ to compute the feature f^*. In this way, image regions with low pixel-wise entropies $\{H_i\}$ can be considered to represent relatively valid visual concepts. For example, the head and wings of the bird in Figure 2 are mainly used by the DNN for fine-grained classification. Therefore, metrics are defined as follows.

$$\begin{aligned} N^{\text{bg}}_{\text{concept}}(x) &= \sum_{i\in\Lambda_{\text{bg}}\ w.r.t.\ x} \mathbb{1}(\overline{H} - H_i > b), \\ N^{\text{fg}}_{\text{concept}}(x) &= \sum_{i\in\Lambda_{\text{fg}}\ w.r.t.\ x} \mathbb{1}(\overline{H} - H_i > b), \qquad (3) \\ \lambda &= \mathbb{E}_{x\in\mathbf{I}}[N^{\text{fg}}_{\text{concept}}(x)/(N^{\text{fg}}_{\text{concept}}(x) + N^{\text{bg}}_{\text{concept}}(x))] \end{aligned}$$

where $N^{\text{bg}}_{\text{concept}}(x)$ and $N^{\text{fg}}_{\text{concept}}(x)$ denote the number of visual concepts encoded on the background and the foreground, respectively. Λ_{bg} and Λ_{fg} are sets of pixels on the background and the foreground *w.r.t.* the input image x, respectively. $\mathbb{1}(\cdot)$ is the indicator function. If the condition inside is valid, $\mathbb{1}(\cdot)$ returns 1, and otherwise 0. $\overline{H} = \mathbb{E}_{i\in\Lambda_{\text{bg}}}[H_i]$ denotes the average entropy value on the background, which measures the significance of information discarding *w.r.t.* background pixels. Those pixels on the background are considered to represent task-irrelevant visual concepts. Therefore, we can use $\overline{H}$ as a baseline entropy. Image regions with significantly lower entropy values than $\overline{H}$ can be considered as valid visual concepts, where b is a positive scalar. The metric λ is used to measure the discriminative power of features. As shown in Figure 2, in order to improve the stability and efficiency of the computation, $\{H_i\}$ is computed in 16×16 grids, *i.e.* all pixels in each local grid share the same σ_i. The dark color in Figure 2 indicates the low entropy value H_i.

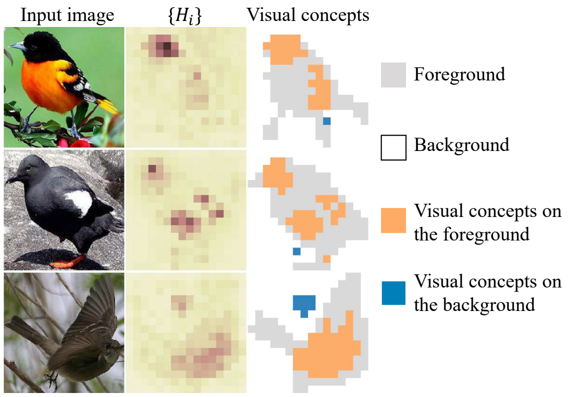
Input image
$\{H_i\}$
Visual concepts
Foreground
Background
Visual concepts on the foreground
Visual concepts on the background

Figure 2. Visualization of visual concepts. The second column shows $\{H_i\}$ of different images. Image regions with low pixel-wise entropies $\{H_i\}$ are considered as visual concepts, which are shown in the third column.

In statistics, visual concepts on the foreground are usually task-relevant, while those on the background are mainly task-irrelevant. In this way, a well-learned DNN is supposed to encode a large number of visual concepts on the foreground and very few ones on the background. Thus, a larger λ value indicates a more discriminative DNN.

Generality and coherency: The design of a metric should consider both the generality and the coherency. The generality is referred to as that the metric is supposed to have strong connections to existing mathematical theories. The coherency ensures comprehensive and fair comparisons in different cases. In this paper, we aim to quantify and compare the number of visual concepts between different network architectures and between different layers. As discussed in [14, 26], existing methods of explaining DNNs usually depend on specific network architecture or specific tasks, such as gradient-based methods [47, 37, 46, 27], perturbation-based methods [9, 20] and inversion-based methods [5]. Unlike previous methods, the conditional entropy of the input ensures fair comparisons between different network architectures and between different layers, which is summarized in Table 1.

	Generality	Coherency	
		Layers	Nets
Gradient-based [47, 37, 46, 27]	No	No	No
Perturbation-based [9, 20]	No	No	No
Inversion-based [5]	No	No	No
Entropy-based	Yes	Yes	Yes

Table 1. Comparisons of different methods in terms of generality and coherency. The entropy-based method provides coherent results across layers and networks.

3.3. Learning simultaneously or sequentially

Hypothesis 2: Knowledge distillation ensures that the DNN is prone to learning various concepts simultaneously. Whereas, the DNN learned from raw data learns concepts sequentially through different epochs.

In this section, we propose two metrics to prove Hypothesis 2. Given a set of training images **I**, $g_1, g_2, \ldots, g_M$ denote DNNs learned in different epochs. This DNN can be either the student network or the baseline network. g_M obtained after the last epoch M is regarded as the final DNN. For each specific image $I \in \mathbf{I}$, we quantify visual concepts on the foreground, which are learned by the DNN after different epochs $N_1^{\text{fg}}(I), N_2^{\text{fg}}(I), \ldots, N_M^{\text{fg}}(I)$.

In this way, whether or not a DNN learns visual concepts simultaneously can be analyzed in following two terms: 1. whether $N_j^{\text{fg}}(I)$ increases fast along with the epoch number; 2. whether $N_j^{\text{fg}}(I)$ of different images increases simultaneously. The first term indicates whether a DNN learns various visual concepts of a specific image quickly, while the second term evaluates whether a DNN learns visual concepts of different images simultaneously.

For a rigorous evaluation, as shown in Figure 3, we calculate the epoch number $\hat{m} = \arg\max_k N_k^{\text{fg}}(I)$, where a DNN obtains richest visual concepts on the foreground. Let w_0 and w_k denote initial parameters and parameters learned after the k-th epoch, respectively. We utilize $\sum_{k=1}^{m} \frac{\|w_k - w_{k-1}\|}{\|w_0\|}$, named ***"weight distance"*** , to measure the learning effect at m-th epoch [12, 7]. Compared to using the epoch number, the weight distance better quantifies the total path of updating the parameter w_k at each epoch k. Thus, we use the average value D_{mean} and standard deviation value D_{std} of weight distances to quantify whether a DNN learns visual concepts simultaneously.

$$D_{\text{mean}} = \mathop{\mathbb{E}}_{I \in \mathbf{I}} \left[\sum_{k=1}^{\hat{m}} \frac{\|w_k - w_{k-1}\|}{\|w_0\|} \right],$$
$$D_{\text{std}} = \mathop{Var}_{I \in \mathbf{I}} \left[\sum_{k=1}^{\hat{m}} \frac{\|w_k - w_{k-1}\|}{\|w_0\|} \right] \quad (4)$$

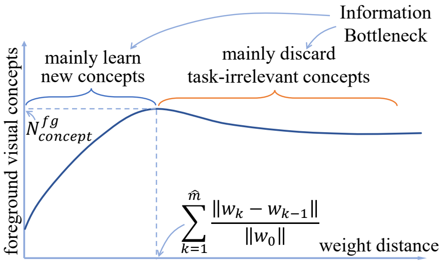

Figure 3. Procedure of learning foreground visual concepts *w.r.t.* weight distances. According to the information-bottleneck theory, a DNN tends to learn various visual concepts mainly in early stages and then mainly discard task-irrelevant concepts later. Strictly speaking, a DNN learns new concepts and discards old concepts during the entire stages. We can consider that the learning stage of $\hat{m}$ encodes richest concepts.

D_{mean} represents the average weight distance, where the DNN obtains the richest task-relevant visual concepts. The value of D_{mean} indicates whether a DNN learns visual concepts quickly. D_{std} describes the variation of the weight distance *w.r.t* different images, and its value denotes whether a DNN learns various visual concepts simultaneously. Hence, small values of D_{mean} and D_{std} indicate that the DNN learns various concepts quickly and simultaneously.

3.4. Learning with Less Detours

Hypothesis 3: Knowledge distillation yields more stable optimization directions than learning from raw data.

During the knowledge distillation, the teacher network directly guides the student network to learn target visual concepts without significant detours[1]. In comparison, according to the information-bottleneck theory [41, 36], when learning from raw data, the DNN usually tries to model various visual concepts and then discard non-discriminative ones, which leads to unstable optimization directions.

In order to quantify the stability of optimization directions of a DNN, a new metric is proposed. Let $S_1(I), S_2(I), \ldots, S_M(I)$ denote the set of visual concepts on the foreground of image I encoded by $g_1, g_2, \ldots, g_M$, respectively. Here, each visual concept $a \in S_j(I)$ is referred to as a specific pixel i on the foreground of image I, which satisfies $\overline{H} - H_i > b$. The stability of optimization directions can be measured as follows.

$$\rho = \frac{\|S_M(I)\|}{\|\bigcup_{j=1}^{M} S_j(I)\|} \quad (5)$$

The numerator reflects the number of visual concepts, which have been chosen ultimately for object classification (shown as the black box in Figure 4). The denominator represents visual concepts temporarily learned during the

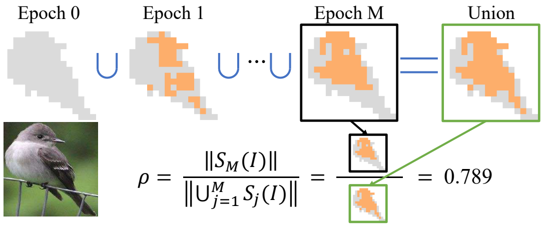
Epoch 0
Epoch 1
Epoch M
Union
$\rho = \frac{\|S_M(I)\|}{\|\bigcup_{j=1}^{M} S_j(I)\|} = \frac{\quad}{\quad} = 0.789$

Figure 4. Detours[1] of learning visual concepts. We visualize sets of foreground visual concepts learned after different epochs. The green box indicates the union of visual concepts learned during all epochs. The $(1-\rho)$ value denotes the ratio of visual concepts that are discarded during the learning procedure *w.r.t.* the union set of concepts. Thus, a larger ρ value indicates that the DNN learns with less detours[1].

learning procedure, which is shown as the green box in Figure 4. $(\bigcup_{j=1}^{M} S_j(I) \setminus S_M(I))$ denotes the set of visual concepts, which have been tried, but finally are discarded by the DNN. A high value of ρ indicates that the DNN is optimized with less detours[1] and more stably; vice versa.

4. Experiment

4.1. Implementation Details

Datasets & DNNs: We designed comparative experiments to verify three proposed hypotheses. For comprehensive comparisons, we conducted experiments based on AlexNet [22], VGG-11, VGG-16, VGG-19 [38], ResNet-50, ResNet-101, and ResNet-152 [18]. Given each DNN as the teacher network, we distilled knowledge from the teacher network to the student network, which had the same architecture as the teacher network for fair comparisons. Meanwhile, the baseline network was also required to have the same architecture as the teacher network.

We trained these DNNs based on the ILSVRC-2013 DET dataset [35], the CUB200-2011 dataset [39], and the Pascal VOC 2012 dataset [6]. All teacher networks in Sections 4.3, 4.4, 4.5 were pre-trained on the ImageNet dataset [32], and then fine-tuned using all three datasets, respectively. Meanwhile, all baseline networks were learned from scratch. For training on the ILSVRC-2013 DET dataset, we conducted the classification of terrestrial mammal categories for comparative experiments, considering the high computational burden. For the ILSVRC-2013 DET dataset and the Pascal VOC 2012 dataset, data augmentation [17] was applied to prevent overfitting. For the CUB200-2011 dataset, we used object images cropped by object bounding boxes for both training and testing. In particular, for the Pascal VOC 2012 dataset, images were cropped by using $1.2\,width \times 1.2\,height$ of the original object bounding box for stable results. For the ILSVRC-2013 DET dataset, we cropped each image by using $1.5\,width \times 1.5\,height$ of the original object bounding box. Because there existed no ground-truth annotations of object segmentation in the ILSVRC-2013 DET dataset, we used the object bounding box as the foreground region. Pixels within the object bounding box were regarded as the foreground Λ_{fg}, and pixels outside the object bounding box were referred to as the background Λ_{bg}.

Distillation: In the procedure of knowledge distillation, we selected a fully-connected (FC) layer l as the target layer. $\|f_T(x) - f_S(x)\|^2$ was used as the distillation loss to mimic the feature of the corresponding layer of the teacher network, where $f_T(x)$ and $f_S(x)$ denoted the l^{th}-layer features of the teacher network and its corresponding student network, respectively.

Parameters of the student network under the target FC layer l were learned exclusively using the distillation loss. Hence, the learning process was not affected by the information of additional human annotations except the knowledge encoded in the teacher network, which ensured fair comparisons. Then we froze parameters under the target layer l and learned parameters above the target layer l merely using the classification loss.

Selection of layers: For each pair of the student network and the baseline network, we aimed to quantify visual concepts encoded in FC layers and thus conducted comparative experiments. We found that these selected DNNs usually had three FC layers. For sake of brevity, we named three FC layers $\text{FC}_1, \text{FC}_2, \text{FC}_3$ for short, respectively. Note that, for the ILSVRC-2013 DET dataset and the Pascal VOC 2012 dataset, dimensions of intermediate-layer features encoded in the FC_3 layer were much smaller than feature dimensions of the FC_1 and FC_2 layers. Hence, the target layer was chosen from the FC_1 and FC_2 layer, when DNNs were learned on the ILSVRC-2013 DET dataset and the Pascal VOC 2012 dataset. For the CUB200-2011 dataset, all three FC layers were selected as target layers. Note that, ResNets usually only had one FC layer. In this way, we replaced the only FC layer into a block with two convolutional and three FC layers, each followed by a ReLU layer. Thus, we could measure visual concepts in the student network and the baseline network *w.r.t.* each FC layer. For the hyper-parameter b (shown in Equation (3)), it was set to 0.25 for AlexNet, and was set to 0.2 for other DNNs. It was because AlexNet had much less layers than other DNNs.

4.2. Quantification of Visual Concepts in the Teacher Network, the Student Network and the Baseline Network

According to our hypotheses, the teacher network was learned from a large number of training data. Hence, the teacher network had learned better representations, *i.e.* encoding more visual concepts on the foreground and less concepts on the background than the baseline network. Thus, the student network learned from the teacher was supposed to contain more visual concepts on the foreground

Dataset	Layer		$N^{\text{fg}}_{\text{concept}}\uparrow$	$\lambda\uparrow$
CUB	VGG-16 FC$_1$	T	**34.00**	**0.78**
		S	29.57	0.75
		B	22.50	0.68
	VGG-16 FC$_2$	T	**34.62**	**0.80**
		S	32.92	0.75
		B	23.31	0.67
	VGG-16 FC$_3$	T	**33.97**	**0.81**
		S	29.78	0.63
		B	23.26	0.71
ILSVRC	VGG-16 FC$_1$	T	**36.80**	**0.87**
		S	35.98	0.84
		B	36.47	0.81
	VGG-16 FC$_2$	T	38.76	**0.89**
		S	**42.74**	0.82
		B	36.35	0.82

Table 2. Comparisons of visual concepts encoded in the teacher network (T), the student network (S) and the baseline network (B). The teacher network encoded more visual concepts on the foreground $N^{\text{fg}}_{\text{concept}}$ and obtained a larger ratio λ than the student network. Meanwhile, the student network had larger values of $N^{\text{fg}}_{\text{concept}}$ and λ than the baseline network.

than the baseline network. In this section, we aimed to compare the number of visual concepts encoded in the teacher network, the student network, and the baseline network.

We learned a teacher network from scratch, on the ILSVRC-2013 DET dataset and the CUB200-2011 dataset. In order to boost the performance of the teacher network, data augmentation [17] was used. The student network was distilled in the same way as Section 4.1, which had the same architecture as the teacher network and the baseline network. Results based on VGG-16 were reported in Table 2. We found that the number of concepts on the foreground $N^{\text{fg}}_{\text{concept}}$ and the ratio λ of the teacher network were larger than those of the student network. Meanwhile, the student network obtained larger values of $N^{\text{fg}}_{\text{concept}}$ and λ than the baseline network. In this way, the assumed relationship between the teacher network, the student network, and the baseline network was roughly verified. We also noticed that there was an exception that the $N^{\text{fg}}_{\text{concept}}$ value of the teacher network was smaller than that of the student network. It was because the teacher network had a larger average background entropy value $\overline{H}$ than the student network.

4.3. Verification of Hypothesis 1

Hypothesis 1 assumed that knowledge distillation ensured the student network to learn more task-relevant visual concepts and less task-irrelevant visual concepts. Thus, we utilized $N^{\text{fg}}_{\text{concept}}$, $N^{\text{bg}}_{\text{concept}}$ and λ metrics in Equation (3) to verify this hypothesis.

Values of $N^{\text{fg}}_{\text{concept}}$, $N^{\text{bg}}_{\text{concept}}$ and λ, which evaluated at the FC$_1$ and FC$_2$ layers of each DNN learned using the CUB200-2011 dataset, the ILSVRC-2013 dataset and the Pascal VOC 2012 dataset, were shown in Table 3. Most results proved Hypothesis 1. *I.e.* the student network tended to encode more visual concepts on the foreground and less concepts on the background, thereby exhibiting a larger ratio λ than the baseline network. Figure 5 showed visual concepts encoded in the FC$_1$ layer of VGG-11, which also proved Hypothesis 1. Note that very few student networks encoded more background visual concepts $N^{\text{bg}}_{\text{concept}}$. It was because that DNNs used as the teacher network were pre-trained on the ImageNet dataset in Sections 4.3, 4.4, 4.5 to verify Hypotheses 1-3. Pre-trained teacher networks encoded visual concepts of 1000 categories, which were much more than necessary. This would make the student network exhibited a larger $N^{\text{bg}}_{\text{concept}}$ value than the baseline network.

4.4. Verification of Hypothesis 2

For Hypothesis 2, we aimed to verify that knowledge distillation enabled the student network to have a higher learning speed, *i.e.* learning different concepts simultaneously. We used D_{mean} and D_{std} to prove this hypothesis.

As shown in Table 3, the D_{mean} value and D_{std} value of the student network were both smaller than that of the baseline network, which verified Hypothesis 2. Note that there were still failure cases. For example, the D_{mean} and D_{std} were measured at the FC$_1$ layer of AlexNet or at the FC$_2$ layer of VGG-11. The reason was that AlexNet and VGG-11 both had relatively shallow network architectures. When learning from raw data, DNNs with shallow architectures would learn more concepts and avoid overfitting. Nevertheless, besides very few exceptional cases, knowledge distillation outperformed learning from raw data for most DNNs.

4.5. Verification of Hypothesis 3

Hypothesis 3 aimed to prove that compared to the baseline network, knowledge distillation made the student network optimized with less detours[1]. The metric ρ depicted the stability of optimization directions and was used to verify above hypothesis. Results reported in Table 3 demonstrated that in most cases, the ρ value of the student network was larger than that of the baseline network. When we measured ρ by AlexNet and VGG-11, failure cases emerged due to the shallow architectures of these two networks. Hence, the optimization directions of the student network tended to be unstable and took more detours[1].

5. Conclusion and Discussions

In this paper, we interpret the success of knowledge distillation from the perspective of quantifying the knowledge encoded in the intermediate layer of a DNN. Three types of metrics were proposed to verify three hypotheses in the scenario of classification. *I.e.* knowledge distillation ensures DNNs learn more task-relevant concepts and less task-irrelevant concepts, have a higher learning speed, and optimize with less detours[1] than learning from raw data.

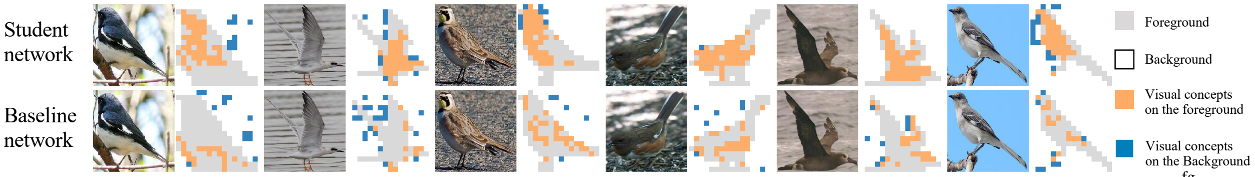
Student network
Baseline network
Foreground
Background
Visual concepts on the foreground
Visual concepts on the Background

Figure 5. Visualization of visual concepts encoded in the FC$_1$ layer of VGG-11. Generally, the student network had a larger $N^{\text{fg}}_{\text{concept}}$ value and a smaller $N^{\text{bg}}_{\text{concept}}$ value than the baseline network.

Network	Layer		CUB200-2011 dataset						ILSVRC-2013 DET dataset						Pascal VOC 2012 dataset					
			$N^{\text{fg}}_{\text{concept}} \uparrow$	$N^{\text{bg}}_{\text{concept}} \downarrow$	$\lambda \uparrow$	$D_{\text{mean}} \downarrow$	$D_{\text{std}} \downarrow$	$\rho \uparrow$	$N^{\text{fg}}_{\text{concept}} \uparrow$	$N^{\text{bg}}_{\text{concept}} \downarrow$	$\lambda \uparrow$	$D_{\text{mean}} \downarrow$	$D_{\text{std}} \downarrow$	$\rho \uparrow$	$N^{\text{fg}}_{\text{concept}} \uparrow$	$N^{\text{bg}}_{\text{concept}} \downarrow$	$\lambda \uparrow$	$D_{\text{mean}} \downarrow$	$D_{\text{std}} \downarrow$	$\rho \uparrow$
AlexNet	FC$_1$	S	**36.60**	**4.00**	**0.90**	8.35	25.09	**0.57**	**49.46**	**0.66**	**0.99**	**0.48**	**0.10**	**0.62**	**25.84**	**5.86**	**0.79**	**1.14**	**0.56**	0.43
		B	24.13	5.65	0.81	**4.81**	**14.54**	0.52	41.00	0.92	0.98	1.32	0.31	0.61	20.30	6.08	0.77	2.00	2.21	**0.44**
	FC$_2$	S	**38.13**	**3.50**	**0.92**	**3.77**	**3.97**	0.49	**57.86**	1.70	0.98	**0.28**	**0.01**	0.60	**31.81**	7.29	**0.81**	**0.62**	**0.07**	**0.47**
		B	23.33	5.48	0.80	5.36	20.79	**0.49**	42.24	**0.96**	**0.98**	1.15	0.15	**0.60**	21.85	**6.56**	0.77	2.04	1.46	0.44
	FC$_3$	S	**33.20**	**4.31**	**0.89**	**8.13**	**39.79**	**0.51**	–	–	–	–	–	–	–	–	–	–	–	–
		B	22.73	4.94	0.83	13.57	137.74	0.42	–	–	–	–	–	–	–	–	–	–	–	–
VGG-11	FC$_1$	S	**30.69**	**10.65**	**0.75**	**1.21**	**0.61**	**0.56**	**44.48**	**4.68**	**0.91**	**0.26**	**0.06**	0.50	**30.56**	8.36	**0.78**	**1.09**	**0.30**	0.38
		B	24.26	10.77	0.70	2.01	3.18	0.55	28.27	7.80	0.80	0.93	0.08	**0.53**	20.31	**7.28**	0.73	1.41	0.54	**0.44**
	FC$_2$	S	**36.51**	**10.66**	**0.78**	**5.22**	19.32	0.49	**54.20**	**6.98**	**0.89**	**0.18**	**0.02**	**0.48**	**38.08**	10.34	**0.79**	**0.70**	**0.29**	**0.45**
		B	26.86	10.71	0.72	6.62	**16.21**	**0.54**	29.68	8.64	0.79	1.19	0.52	0.47	20.03	**7.42**	0.72	1.65	1.80	0.36
	FC$_3$	S	**34.53**	14.21	**0.72**	**4.15**	**4.55**	**0.50**	–	–	–	–	–	–	–	–	–	–	–	–
		B	24.53	**10.95**	0.69	20.66	95.29	0.49	–	–	–	–	–	–	–	–	–	–	–	–
VGG-16	FC$_1$	S	**43.77**	**8.73**	**0.84**	**0.64**	**0.06**	**0.66**	**56.29**	**3.13**	**0.95**	**0.02**	**0.0001**	**0.47**	**42.26**	11.54	**0.80**	**0.33**	**0.09**	**0.52**
		B	22.50	11.27	0.68	2.38	4.98	0.50	36.06	7.71	0.83	0.40	0.13	0.44	26.87	**8.26**	0.76	1.65	0.61	0.48
	FC$_2$	S	**36.83**	**11.03**	**0.77**	**0.80**	**0.37**	**0.54**	37.79	**4.31**	**0.90**	**0.17**	**0.02**	0.32	**31.19**	8.70	0.78	**0.83**	**0.45**	0.35
		B	23.31	11.56	0.67	5.43	22.96	0.50	**38.41**	9.66	0.80	0.79	0.52	**0.43**	29.37	**8.04**	**0.78**	2.65	1.90	**0.46**
	FC$_3$	S	**32.32**	10.21	**0.77**	**6.17**	**32.63**	**0.47**	–	–	–	–	–	–	–	–	–	–	–	–
		B	23.26	**9.97**	0.71	17.53	216.05	0.46	–	–	–	–	–	–	–	–	–	–	–	–
VGG-19	FC$_1$	S	**40.74**	**10.42**	**0.80**	**0.66**	**0.15**	**0.60**	**46.50**	**2.52**	**0.95**	**0.16**	**0.0002**	**0.39**	**46.38**	14.05	0.77	**0.25**	**0.07**	**0.45**
		B	22.42	11.19	0.67	2.33	3.67	0.47	29.71	5.83	0.84	0.33	0.12	0.39	28.65	**7.93**	**0.78**	1.10	0.80	0.41
	FC$_2$	S	**40.20**	**9.03**	**0.82**	**1.16**	**0.63**	**0.56**	**50.90**	**5.96**	**0.91**	**0.06**	**0.0006**	**0.37**	**47.03**	13.66	**0.78**	**0.10**	**0.03**	**0.45**
		B	24.00	10.40	0.70	4.64	19.07	0.47	30.31	6.15	0.84	0.45	0.18	0.37	28.46	**8.20**	0.78	2.14	1.92	0.41
	FC$_3$	S	**28.60**	**6.37**	**0.82**	**4.89**	**11.57**	**0.48**	–	–	–	–	–	–	–	–	–	–	–	–
		B	21.29	7.77	0.74	20.61	143.61	0.46	–	–	–	–	–	–	–	–	–	–	–	–
ResNet-50	FC$_1$	S	**43.02**	**10.15**	**0.81**	24.43	166.76	0.48	**56.00**	6.50	**0.90**	3.45	**4.74**	**0.45**	**42.54**	10.76	0.80	3.43	19.60	**0.40**
		B	42.15	11.83	0.79	**20.78**	**122.79**	**0.53**	43.80	**5.75**	0.89	**2.73**	6.82	0.36	39.65	**9.81**	**0.81**	**1.64**	**15.20**	0.39
	FC$_2$	S	**48.58**	**9.75**	**0.83**	37.62	**206.22**	**0.55**	**52.57**	**6.54**	**0.90**	0.25	1.45	**0.40**	**41.03**	12.37	0.77	**1.85**	**13.03**	**0.41**
		B	42.06	11.88	0.79	**29.28**	248.03	0.52	43.63	6.93	0.87	**0.02**	**0.02**	0.35	38.00	**10.00**	**0.80**	2.68	30.91	0.38
	FC$_3$	S	41.38	11.73	0.77	926.61	142807.00	0.43	–	–	–	–	–	–	–	–	–	–	–	–
		B	**42.03**	**11.48**	**0.79**	**111.18**	**3299.20**	**0.53**	–	–	–	–	–	–	–	–	–	–	–	–
ResNet-101	FC$_1$	S	**45.93**	**11.14**	**0.81**	**23.32**	**236.76**	0.51	**48.59**	**5.06**	**0.91**	**1.99**	**2.20**	**0.39**	42.54	9.37	0.82	**1.39**	**32.87**	0.35
		B	44.18	12.55	0.78	40.41	828.72	**0.52**	42.94	8.16	0.84	5.41	10.39	0.35	**43.33**	**9.30**	**0.83**	15.28	48.71	**0.39**
	FC$_2$	S	**51.59**	**9.02**	**0.85**	67.60	**947.85**	**0.54**	**49.27**	**6.39**	**0.89**	**0.98**	**0.65**	**0.37**	**41.71**	9.16	0.82	3.30	100.97	0.38
		B	43.22	12.32	0.78	**43.40**	1155.22	0.50	41.79	7.30	0.85	6.58	17.16	0.34	41.35	**8.32**	**0.84**	**2.26**	**48.61**	**0.39**
	FC$_3$	S	**47.71**	**10.24**	**0.82**	**73.33**	**2797.15**	**0.53**	–	–	–	–	–	–	–	–	–	–	–	–
		B	42.40	10.53	0.80	162.68	16481.93	0.49	–	–	–	–	–	–	–	–	–	–	–	–
ResNet-152	FC$_1$	S	44.81	12.09	0.79	**26.35**	**289.59**	0.48	**44.90**	5.63	**0.89**	6.25	**5.86**	**0.36**	**41.09**	**10.09**	**0.81**	**0.33**	**3.59**	**0.39**
		B	**45.62**	**10.68**	**0.81**	36.92	767.58	**0.54**	39.93	**5.40**	0.89	**6.08**	6.74	0.33	40.15	10.82	0.79	0.59	11.39	0.37
	FC$_2$	S	43.79	**10.04**	**0.81**	**7.13**	**42.77**	0.52	**40.98**	6.90	0.86	**4.64**	**5.71**	0.32	**41.36**	**12.04**	**0.78**	**14.29**	**17.33**	**0.38**
		B	**45.08**	10.85	0.81	44.59	1200.97	**0.52**	40.29	**5.56**	**0.89**	7.86	12.24	**0.33**	38.57	12.07	0.77	18.03	67.52	0.36
	FC$_3$	S	44.21	11.89	0.79	**47.28**	**1463.55**	0.50	–	–	–	–	–	–	–	–	–	–	–	–
		B	**44.89**	**10.77**	**0.81**	167.41	16331.28	**0.52**	–	–	–	–	–	–	–	–	–	–	–	–

Table 3. Comparisons between the student network (S) and the baseline network (B). ↑/↓ indicates that larger/smaller values were better. In general, the student network had larger values of $N^{\text{fg}}_{\text{concept}}$, λ, ρ, and smaller values of $N^{\text{bg}}_{\text{concept}}$, D_{mean}, D_{std} than the baseline network, which proved Hypotheses 1-3.

There are several limitations of our work. We only focus on the classification task in this paper. However, applying our methods to other tasks (*e.g.* object segmentation), or other types of data (*e.g.* the video) is theoretically feasible. Meanwhile, for these tasks, side information may be required. In this paper, our proposed metrics are implemented by using the entropy-based analysis, which has strong connections to the information-bottleneck theory. Unlike the information-bottleneck theory, the proposed metrics can measure the pixel-wise discarding. However, the learning procedure of DNNs cannot be precisely divided into the learning phase and the discarding phase. In each epoch, the DNN may simultaneously learn new visual concepts and discard old task-irrelevant concepts. Thus, the target epoch $\hat{m}$ in Figure 3 is just a rough estimation of the division of two learning phases.

Acknowledgements

The corresponding author Quanshi Zhang is with the John Hopcroft Center and MoE Key Lab of Artificial Intelligence AI Institute, Shanghai Jiao Tong University. He thanks the support of National Natural Science Foundation of China (U19B2043 and 61906120) and Huawei Technologies. Zhefan Rao and Yilan Chen made equal contributions to this work as interns at Shanghai Jiao Tong University.

References

[1] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017. 2, 3, 4

[2] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018. 2

[3] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 882–891, 2018. 3

[4] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016. 2

[5] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4829–4837, 2016. 2, 4, 5

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 6

[7] Sebastian Flennerhag, Pablo G Moreno, Neil D Lawrence, and Andreas Damianou. Transferring knowledge across learning processes. *arXiv preprint arXiv:1812.01054*, 2018. 5

[8] Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8730–8738, 2018. 2

[9] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017. 2, 4, 5

[10] Stanislav Fort, Paweł Krzysztof Nowak, and Srini Narayanan. Stiffness: A new perspective on generalization in neural networks. *arXiv preprint arXiv:1901.09491*, 2019. 3

[11] Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. *arXiv preprint arXiv:1805.04770*, 2018. 1, 3

[12] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, pages 8789–8798, 2018. 5

[13] Ziv Goldfeld, Ewout Van Den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In *International Conference on Machine Learning*, pages 2299–2308, 2019. 3

[14] Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. Towards a deep and unified understanding of deep neural models in nlp. In *International Conference on Machine Learning*, pages 2454–2463, 2019. 2, 3, 4

[15] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5):6, 2017. 3

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 3

[17] Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-revnet: Deep invertible networks. *arXiv preprint arXiv:1802.07088*, 2018. 6, 7

[18] Shaoqing Ren Kaiming He, Xiangyu Zhang and Jian Sun. Deep residual learning for image recognition. *In CVPR*, 2016. 6

[19] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *arXiv preprint arXiv:1711.11279*, 2017. 2

[20] Pieter-Jan Kindermans, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: Patternnet and patternattribution. *arXiv preprint arXiv:1705.05598*, 2017. 2, 4, 5

[21] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. *arXiv preprint arXiv:1905.00414*, 2019. 3

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 6

[23] Ruofan Liang, Tianlin Li, Longfei Li, Jing Wang, and Quanshi Zhang. Knowledge consistency between neural networks and beyond. *In International Conference on Learning Representations*, 2020. 3

[24] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *In ICLR*, 2016. 3

[25] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017. 2

[26] Haotian Ma, Yinqing Zhang, Fan Zhou, and Quanshi Zhang. Quantifying layerwise information discarding of neural networks. *In arXiv:1906.04109*, 2019. 2, 3, 4

[27] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015. 2, 3, 4, 5

[28] Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018. 3

[29] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International Conference on Machine Learning*, pages 5142–5151, 2019. 3

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016. 2

[31] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 1

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6

[33] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017. 2

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 2

[35] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 6

[36] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017. 2, 3, 5

[37] K Simonyan, A Vedaldi, and A Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2017. 2, 3, 4, 5

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *In ICLR*, 2015. 6

[39] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[40] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018. 3

[41] Natalie Wolchover. New theory cracks open the black box of deep learning. *In Quanta Magazine*, 2017. 2, 3, 5

[42] Dan Xie, Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. Learning and inferring dark matter?and predicting human intents and trajectories in videos. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1639–1652, 2017. 4

[43] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017. 3

[44] Zhiqin John Xu. Understanding training and generalization in deep learning by fourier analysis. *arXiv preprint arXiv:1808.04295*, 2018. 3

[45] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017. 1

[46] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural net-

works through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015. 2, 3, 4, 5

[47] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 2, 3, 4, 5

[48] Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu. Interpreting cnn knowledge via an explanatory graph. *In the 32nd AAAI Conference on Artificial Intelligence*, 2018. 2

[49] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[50] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[51] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *In ICLR*, 2015. 2

[52] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2