ICLR2020 | 如何判断两个神经网络学到的知识是否一致

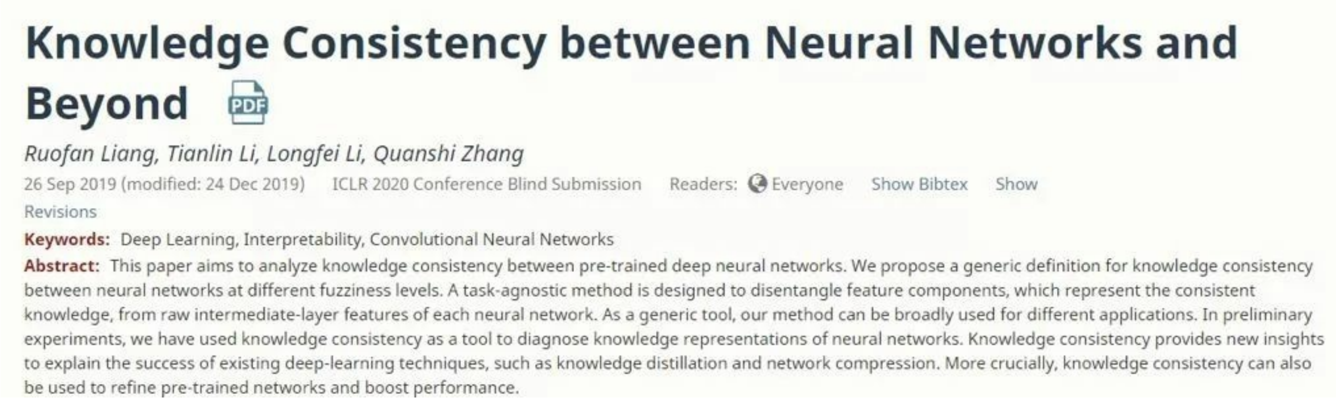
夕小瑶的卖萌屋 2月17日

一只小狐狸带你解锁**NLP/ML/DL**秘籍

正文来源：机器之心

前言

人工智能顶会 ICLR 2020 将于 4 月 26 日于埃塞俄比亚首都亚的斯亚贝巴举行。在最终提交的 2594 篇论文中，有 687 篇被接收，接收率为 26.5%。 本文介绍了上海交通大学张拳石团队的一篇接收论文——《Knowledge Consistency between Neural Networks and Beyond》。在本文中，研究者提出了一种对神经网络特征表达一致性、可靠性、知识盲点的评测与解释方法。



Knowledge Consistency between Neural Networks and Beyond

Ruofan Liang, Tianlin Li, Longfei Li, Quanshi Zhang

26 Sep 2019 (modified: 24 Dec 2019) ICLR 2020 Conference Blind Submission Readers: Everyone Show Bibtex Show Revisions

Keywords: Deep Learning, Interpretability, Convolutional Neural Networks

Abstract: This paper aims to analyze knowledge consistency between pre-trained deep neural networks. We propose a generic definition for knowledge consistency between neural networks at different fuzziness levels. A task-agnostic method is designed to disentangle feature components, which represent the consistent knowledge, from raw intermediate-layer features of each neural network. As a generic tool, our method can be broadly used for different applications. In preliminary experiments, we have used knowledge consistency as a tool to diagnose knowledge representations of neural networks. Knowledge consistency provides new insights to explain the success of existing deep-learning techniques, such as knowledge distillation and network compression. More crucially, knowledge consistency can also be used to refine pre-trained networks and boost performance.

论文链接：https://arxiv.org/pdf/1908.01581.pdf

概览

深度神经网络（DNN）已经在很多任务中表现出了强大的能力，但目前仍缺乏诊断其中层表征能力的数学工具，如发现表征中的缺陷或识别可靠/不可靠的特征。由于数据泄漏或数据集发生变化，基于测试准确率的传统 DNN 评测方法无法深入评估 DNN 表征的正确性。

因此，在本论文中，来自上海交大的研究者提出了一种从知识一致性的角度来诊断 DNN 中层网络表征能力的方法。即，给定两个为同一任务训练的 DNN（无论二者架构是否相同），目标是检验两个 DNN 的中间层是否编码相似的视觉概念。

该研究实现了：（1）定义并量化了神经网络之间知识表达的不同阶的一致性；（2）对强弱神经网络中层知识进行分析；（3）对中层特征的诊断，在不增加训练样本标注的前提下进一步促进神经网络分类准确率；（4）为解释神经网络压缩和知识蒸馏提供了一种新的思路。

算法简介

该论文定义了两个神经网络之间在知识表达层面的一致性，即分析两个独立训练的神经网络是否建模了相同或相似的知识。研究者关注的是两个神经网络所建模的知识的相似性，而非特征的相似性（比如，将一个神经网络的中层卷积核的顺序打乱，并相应的重新排列其对应的上层卷积核的顺序，经过上层卷积后特征与原始神经网络对应特征相同，这时，这两神经网络具有不同的中层特征，但事实上建模了相同的知识）。

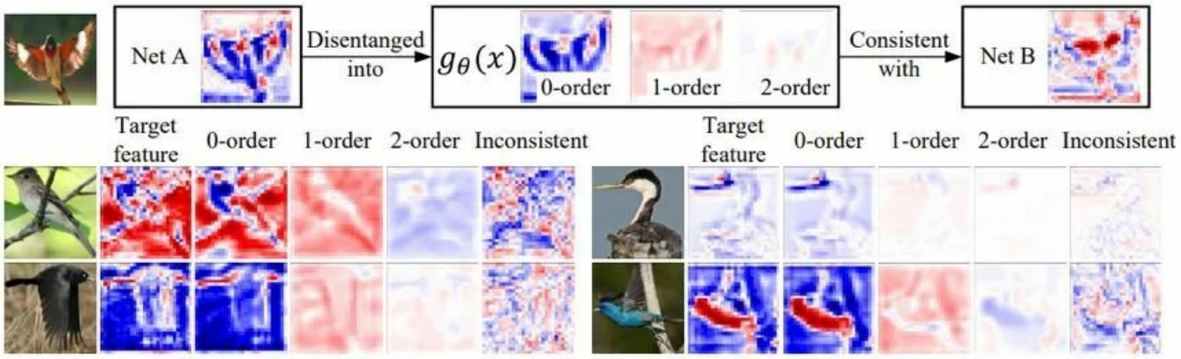


Figure 1: Knowledge consistency. We define, disentangle, and quantify consistent features between two DNNs. Consistent features disentangled from a filter are visualized at different orders (levels) of fuzziness.

另一方面，可以利用神经网络知识表达的一致性，直接对神经网络内部特征表达可靠性进行评测，而不需要额外标注新的监督信息，此评测标准也与具体任务设置无关。如果没有可靠的数学工具去评测神经网络的特征的可靠性，仅仅通过最终的分类正确率来评测神经网络，对深度学习的未来发展是远远不够的。

因而，针对同一任务训练多个不同的神经网络，此研究量化出各神经网络间彼此一致的知识表达，并拆分出其所对应的特征分量。具体来说，f_A 和 f_B 分别表示神经网络 A 与神经网络 B 的中层特征，当 f_A 可以通过线性变换得到 f_B 时，可认为 f_A 和 f_B 零阶一致；当 f_A 可以通过一次非线性变换得到 f_B 时，可认为 f_A 和 f_B 一阶一致；类似的，当 f_A 可以通过 n 次非线性变换得到 f_B 时，可认为 f_A 和 f_B 为 n 阶一致。

如下图所示，可以通过以下神经网络，将神经网络中层特征 f_A 拆分为 0-K 阶不同的一致性特征分量，以及不一致特征分量。

$$x^* = g_\theta(x) + x^\Delta, \qquad g_\theta(x) = x^{(0)} + x^{(1)} + \cdots + x^{(K)}$$

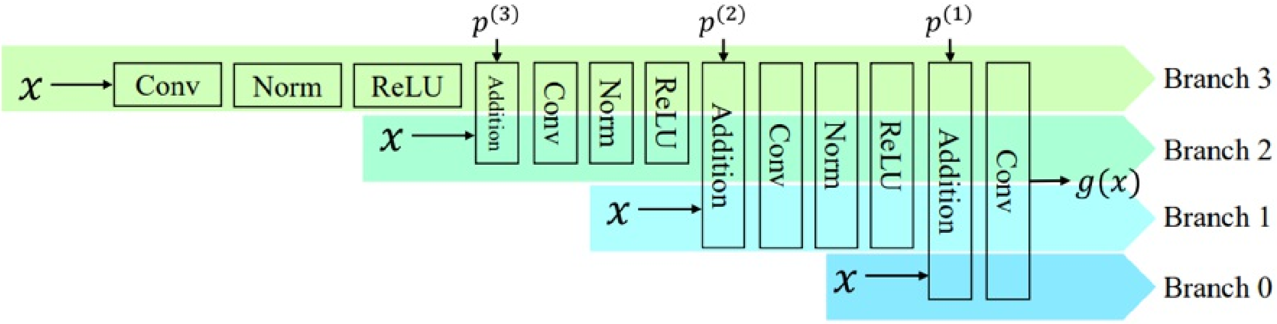


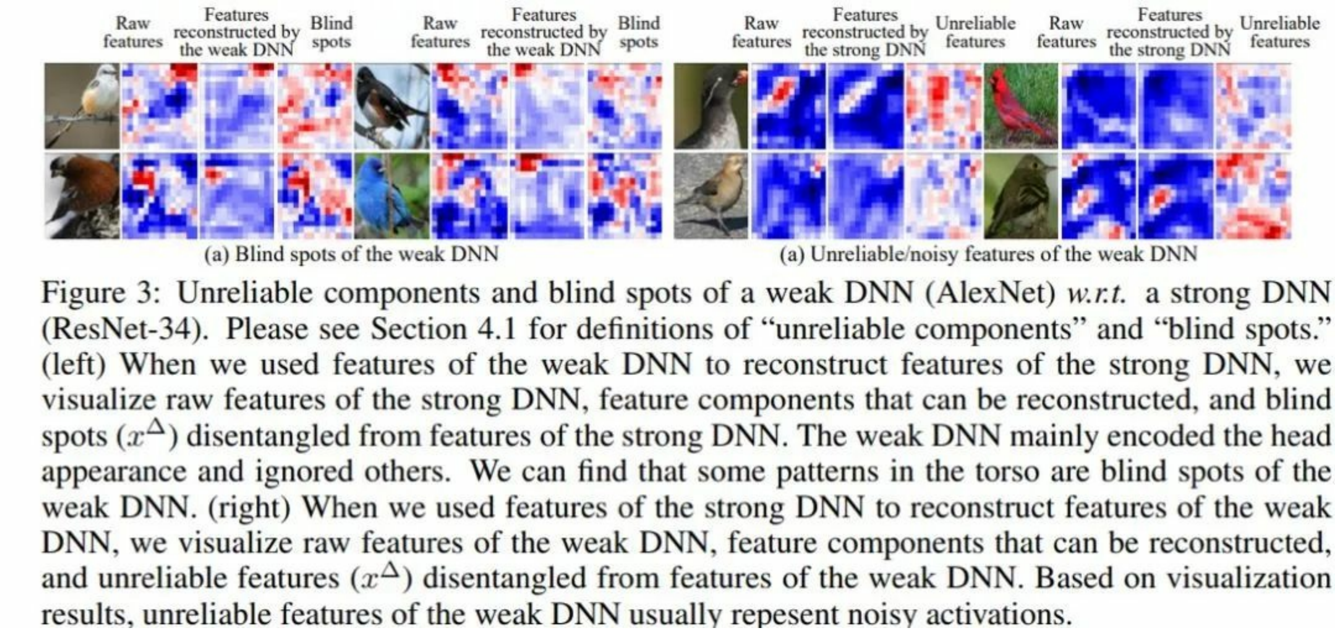
Figure 2: Neural network for disentanglement of consistent features ($K = 3$).

低阶一致性分量往往表示相对可靠的特征，而不一致分量则表示神经网络中的噪声信号。

在应用层面，知识一致性可以用来发现神经网络中的不可靠特征和知识盲点。将一个深层高性能网络作为标准的知识表达，去分析诊断一个相对浅层的神经网络的知识表达缺陷（浅层神经网络有自己特定的应用价值，比如用在移动端）。当利用浅层神经网络 (DNN A) 特征去重建深层神经网络 (DNN B) 特征时，深层神经网络中的不一致特征分量 (δ=f_B-g(f_A)) 往往代表着浅层神经网络的知识盲点；相对应地，当利用深层神经网络特征去重建浅层神经网络特征时，浅层神经网络中的不一致特征分量 (δ=f_A-g(f_B)) 往往代表着其中不可靠的特征分量。

实验结果

下图显示了算法所预测的浅层神经网络的知识盲点与不可靠特征。

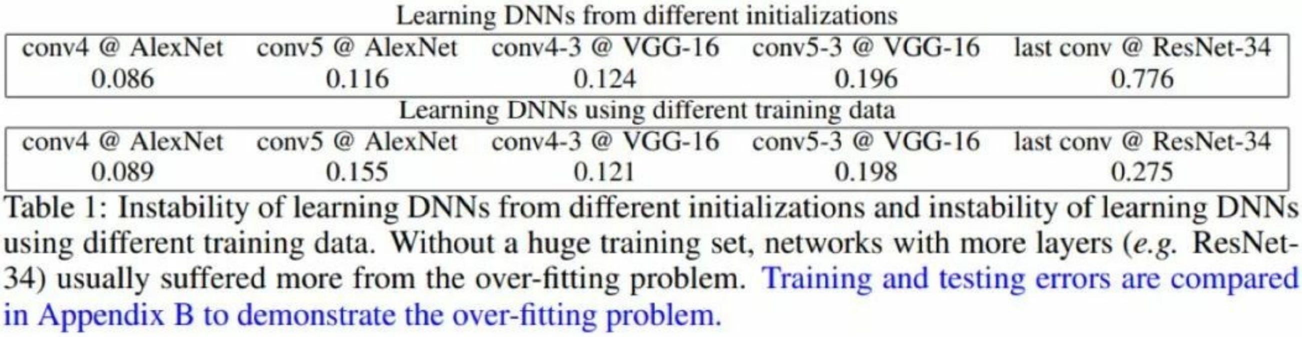

Raw features
Features reconstructed by the weak DNN
Blind spots
Raw features
Features reconstructed by the weak DNN
Blind spots
Raw features
Features reconstructed by the strong DNN
Unreliable features
Raw features
Features reconstructed by the strong DNN
Unreliable features

(a) Blind spots of the weak DNN

(a) Unreliable/noisy features of the weak DNN

Figure 3: Unreliable components and blind spots of a weak DNN (AlexNet) *w.r.t.* a strong DNN (ResNet-34). Please see Section 4.1 for definitions of "unreliable components" and "blind spots." (left) When we used features of the weak DNN to reconstruct features of the strong DNN, we visualize raw features of the strong DNN, feature components that can be reconstructed, and blind spots (x^{Δ}) disentangled from features of the strong DNN. The weak DNN mainly encoded the head appearance and ignored others. We can find that some patterns in the torso are blind spots of the weak DNN. (right) When we used features of the strong DNN to reconstruct features of the weak DNN, we visualize raw features of the weak DNN, feature components that can be reconstructed, and unreliable features (x^{Δ}) disentangled from features of the weak DNN. Based on visualization results, unreliable features of the weak DNN usually repesent noisy activations.

下表从知识一致性的角度，分析神经网络训练的稳定性。当训练样本相对较少时，浅层的神经网络的训练有更强的稳定性。



Learning DNNs from different initializations

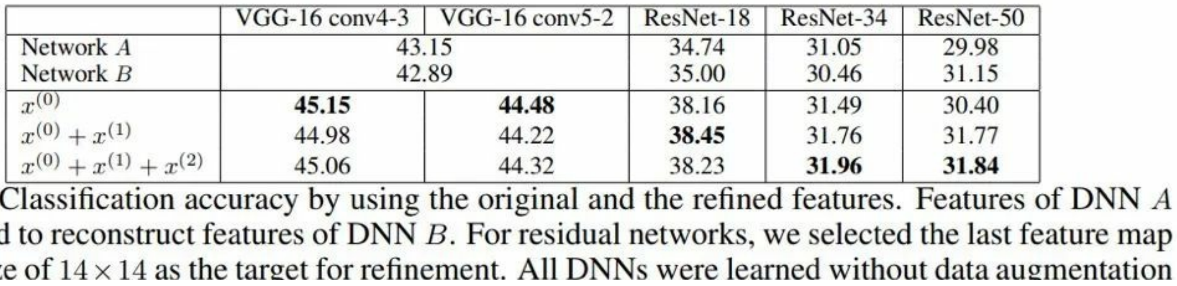
conv4 @ AlexNet	conv5 @ AlexNet	conv4-3 @ VGG-16	conv5-3 @ VGG-16	last conv @ ResNet-34
0.086	0.116	0.124	0.196	0.776

Learning DNNs using different training data

conv4 @ AlexNet	conv5 @ AlexNet	conv4-3 @ VGG-16	conv5-3 @ VGG-16	last conv @ ResNet-34
0.089	0.155	0.121	0.198	0.275

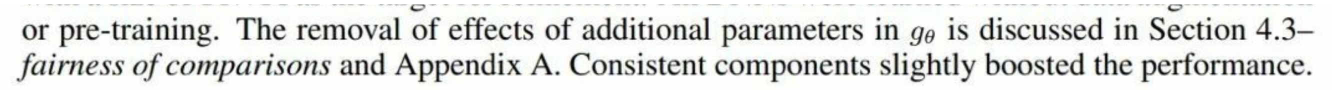
Table 1: Instability of learning DNNs from different initializations and instability of learning DNNs using different training data. Without a huge training set, networks with more layers (*e.g.* ResNet-34) usually suffered more from the over-fitting problem. Training and testing errors are compared in Appendix B to demonstrate the over-fitting problem.

如下图所示，一致的特征分量往往代表更可靠的信息，可以进一步提升神经网络的分类精度。即，在不增加训练样本标注的前提下，利用知识一致性进一步提升模型的分类正确率。



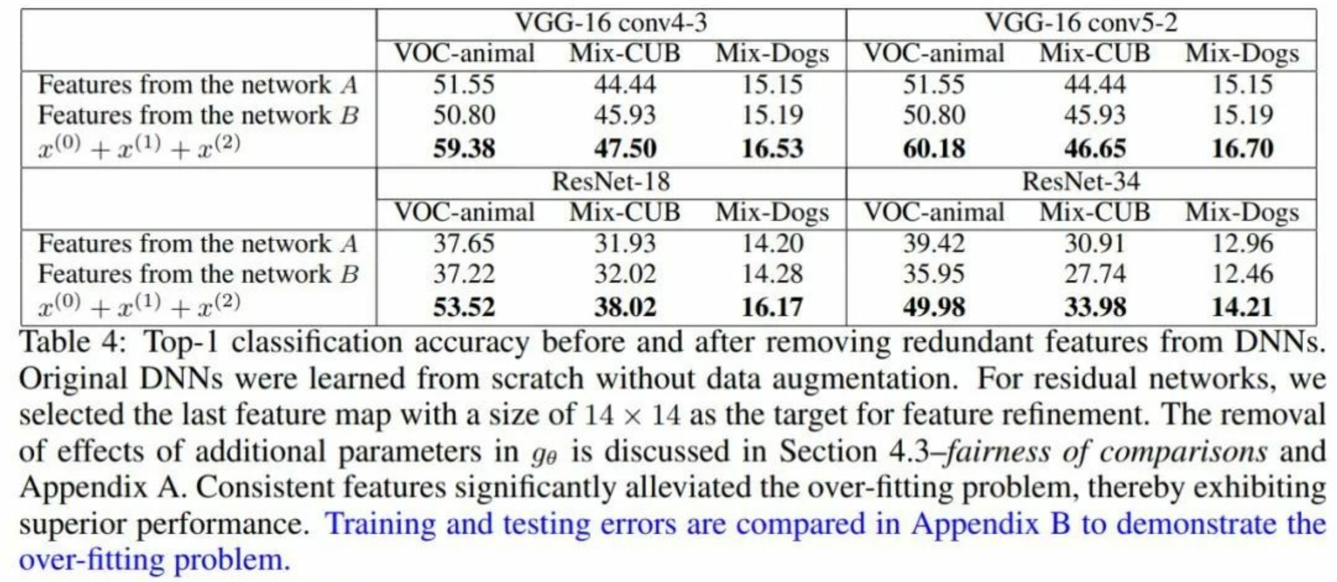
	VGG-16 conv4-3	VGG-16 conv5-2	ResNet-18	ResNet-34	ResNet-50
Network *A*	43.15		34.74	31.05	29.98
Network *B*	42.89		35.00	30.46	31.15
$x^{(0)}$	**45.15**	**44.48**	38.16	31.49	30.40
$x^{(0)}+x^{(1)}$	44.98	44.22	**38.45**	31.76	31.77
$x^{(0)}+x^{(1)}+x^{(2)}$	45.06	44.32	38.23	**31.96**	**31.84**

Table 3: Classification accuracy by using the original and the refined features. Features of DNN *A* were used to reconstruct features of DNN *B*. For residual networks, we selected the last feature map with a size of 14×14 as the target for refinement. All DNNs were learned without data augmentation



or pre-training. The removal of effects of additional parameters in g_θ is discussed in Section 4.3–*fairness of comparisons* and Appendix A. Consistent components slightly boosted the performance.

知识一致性算法可以消除神经网络中的冗余特征。预训练的神经网络(如利用 ImageNet 训练的神经网络)往往建模了海量类别的分类信息，当目标应用只针对少量类别时，预训练的特征中表达无关类别的特征分量则可视为冗余信息。如下图所示，知识一致性算法可以有效的去除与目标应用无关的冗余特征分量，进一步提升目标应用的性能。



	VGG-16 conv4-3			VGG-16 conv5-2		
	VOC-animal	Mix-CUB	Mix-Dogs	VOC-animal	Mix-CUB	Mix-Dogs
Features from the network A	51.55	44.44	15.15	51.55	44.44	15.15
Features from the network B	50.80	45.93	15.19	50.80	45.93	15.19
$x^{(0)}+x^{(1)}+x^{(2)}$	**59.38**	**47.50**	**16.53**	**60.18**	**46.65**	**16.70**
	ResNet-18			ResNet-34		
	VOC-animal	Mix-CUB	Mix-Dogs	VOC-animal	Mix-CUB	Mix-Dogs
Features from the network A	37.65	31.93	14.20	39.42	30.91	12.96
Features from the network B	37.22	32.02	14.28	35.95	27.74	12.46
$x^{(0)}+x^{(1)}+x^{(2)}$	**53.52**	**38.02**	**16.17**	**49.98**	**33.98**	**14.21**

Table 4: Top-1 classification accuracy before and after removing redundant features from DNNs. Original DNNs were learned from scratch without data augmentation. For residual networks, we selected the last feature map with a size of 14×14 as the target for feature refinement. The removal of effects of additional parameters in g_θ is discussed in Section 4.3–*fairness of comparisons* and Appendix A. Consistent features significantly alleviated the over-fitting problem, thereby exhibiting superior performance. Training and testing errors are compared in Appendix B to demonstrate the over-fitting problem.

此外，知识一致性算法可以分析不同任务训练得到模型中的一致/不一致特征。如下图所示，研究者训练网络 A 进行 320 类细分类(包括 CUB 中的 200 类鸟与 Stanford Dog 中的 120 类狗)，训练网络 B 进行简单的二分类(鸟或狗)，通过特征相互重构，可以看到网络 A 中建模了更多的知识，网络 A 的特征能够更好地重构网络 B 的特征。

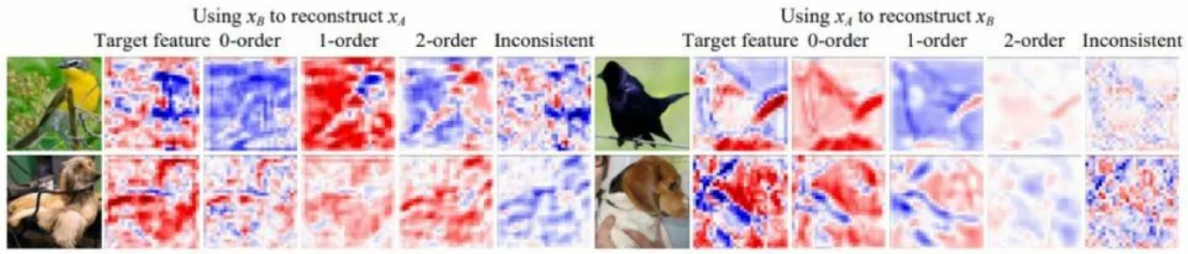
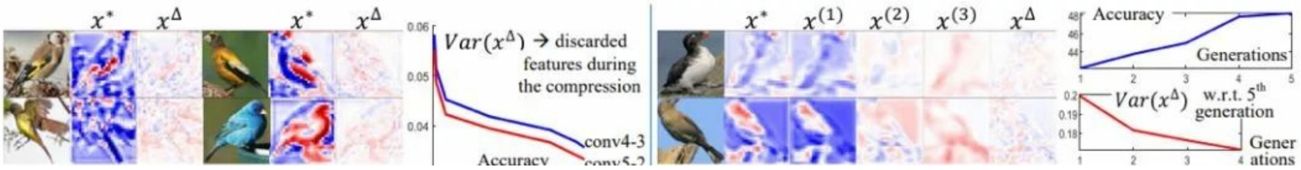


Figure 5: Consistent and inconsistent feature components disentangled between DNN B learned for binary classification and DNN A learned for fine-grained classification.

知识一致性算法可以用于分析网络压缩中的信息损失。研究者使用压缩后模型生成的特征来重建原始模型的特征，不一致的特征分量往往对应着压缩过程中被舍弃的知识。如下图（左）所示，通过量化这部分被舍弃的知识，他们发现在压缩过程中较小的知识损失会有更高的分类正确率。

此外，还可以通过知识一致性算法解释蒸馏。如下图（右），通过量化不同代的再生神经网络中不一致的特征分量，可以发现随着蒸馏代数的增加，不可靠的特征分量逐渐减少。



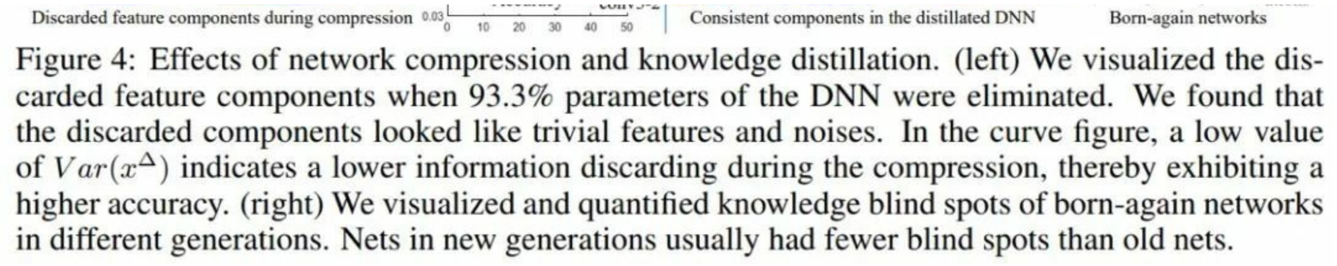


Figure 4: Effects of network compression and knowledge distillation. (left) We visualized the discarded feature components when 93.3% parameters of the DNN were eliminated. We found that the discarded components looked like trivial features and noises. In the curve figure, a low value of $Var(x^{\Delta})$ indicates a lower information discarding during the compression, thereby exhibiting a higher accuracy. (right) We visualized and quantified knowledge blind spots of born-again networks in different generations. Nets in new generations usually had fewer blind spots than old nets.

可 能 喜 欢

- 搜索引擎核心技术与算法 —— 倒排索引初体验
- 文本匹配相关方向打卡点总结
- 多任务学习时转角遇到Bandit老虎机
- 万万没想到，我的炼丹炉玩坏了
- 还在随缘炼丹？一文带你详尽了解机器学习模型可解释性的奥秘
- 后BERT时代：15个预训练模型对比分析与关键点探究

夕小瑶的卖萌屋

关注&星标小夕，带你解锁AI秘籍
内容过于专业，胆小者慎入