Wednesday, January 12, 2022

Note: “White Box” of Deep Learning

Facheng Yu

ReduNet: A White-box Deep Network from the Principle of Maximizing Rate Reduction

- 深度学习目标
  - 监督学习之cross entropy
    - 目的仅在于给出标签，尽管可能误标—经验表明，作为“黑盒”的深度学习产生的标签也能是随机的。
    - 这样端到端的拟合不能在一开始就保证网络能够在多大程度上捕捉数据的内蕴结构。
    - 数据表达会出现neural collapsing的现象,类内的结构会被抑制和忽略。（precise geometric and statistical properties of the learned features are also often obscured, which leads to the lack of the interpretability and subsequent performance guarantees）
  - 最小化—深度学习的低维表达
    - 人们长期认为the role of deep networks is to learn certain low-dimensional representations of the data.（这里低维是在非线性的意义下）
    - 因此有一个解释深度学习功能的角度是，将中间层的outputs看作选择数据x的特定低维潜在特征z，而z在多类中具有判别意义。通过学习z来facilitate接下来使用分类器的分类任务。
    - Information bottleneck(IB,信息瓶颈)进一步假设了网络是通过学习最小充分统计量z来预测y。形式上来说，其在于最大化互信息I(z,y)以及最小化互信息I(x,z)。
    - CE损失下网络会牺牲鲁棒性来防止“标签损失”或者转移。（因为CE通常用于“predicting only”任务，当任务改变时，就会发现网络的鲁棒性不足。）
      - 解决办法：use the label y as only side information来学习多样（而不是最小）的表达.
    - Reconciling contractive and contrastive learning（协调压缩学习和对比学习）

  - Auto-encoding，一种非监督学习框架: x -> z -> x^
    - 学习高维数据的低维表达，并从低维表达复原高维数据。
    - 通过施加特定的紧性（维度、能量、体积等）学习一种端到端的表达。
    - 例如最小化jacobian行列式来惩罚局部体积。
    - 在多模态结构中会出现mode collapsing现象（不能捕捉所有内在子类结构，或者不能为分类或者聚类做出精确判别），而论文提出的principled rate reduction measure 可以兼顾。
  - Contrastive learning： 当数据类别数K很大，任意数据对(x1, x2)大概率属于不同类，因此z1应该和z2高度不相关，而z1应当与transformed version(x1)的特征相关。由此得到contrastive loss
  - 收缩学习目的在于压缩整个类的特征，而对比学习则想要扩展任何样本对的特征。虽然目的相反，这两种机制效果都很好是否是因为其作用在数据的不同部分上?

- 深度网络结构
  - The ultimate goal of any good theory for deep learning is to facilitate a better understanding of deep networks and to design better network architectures and algorithms with performance guarantees.
  - Deep (convolution) neural networks的经验设计
    - parameters theta from random initialization via BP.
    - 从AlexNet开始，现代深度网络开始继续被经验修正和改进。
    - VGG, ResNet, DenseNet, Recurrent CNN or LSTM and MoE 持续产生性能突破。
    - 现在的提升网络性能的方法，都是基于经验性的。Some recent practices even take to the extreme by searching for effective network structures and training strategies through extensive random search techniques, such as Neural Architecture Search (Zoph and Le, 2016; Baker et al., 2017), AutoML (Hutter et al., 2019), and Learning to Learn (Andrychowicz et al., 2016).
  - Constructive approaches to deep (convolution) networks.

- However, it remains unclear about the role of the convolutions (dictionary) in each layer and exactly why such low-level sparse coding is needed for the high-level classification task. To a large extent, this work will provide a new perspective to elucidate the role of the sparsifying convolutions in a deep network: not only will we reveal why sparsity is needed for ensuring invariant classification but also the (multi-channel) convolution operators can be explicitly derived and constructed.
- 不使用BP对权重微调的方法：
  - Use wavelets to construct convolution networks(for signal invariance),ScatteringNets
    - 对数据和特征不敏感，需要的卷积核随深度指数增长
  - PCANets, using principal components of the input data as the convolutional kernels.
  - Both above networks is not directly related to a specific task.

- 区分表达的原则目标 via compression
  - 满足混合分布D = {D_j}的数据X是否能够高效地被分类取决于D_j的分散程度。
  - 一个主流的假设是，每类的分布都有着低维的内蕴结构。原因有：
    - 高维数据是高度冗余的。
    - 同类的数据应该彼此相似、相关。
    - 一般我们仅考虑在某些变形和增强下保持x不变的等效结构。
  - 分布D在每类上有支撑子流形M_j，我们则要学习将子流形M_j（R_D）映射到线性子空间S_j(R_n)的映射z = f(x, theta).
  - LDR(linear discriminative representation):
    - 类内可压缩：low-dimensional linear subspace
    - 类间可分别：highly uncorrelated
    - 多样性表达：Dimension of features for each class/cluster should be as large as possible
  - LDR前两个条件和经典分类方法(LDA)是一致的.

- we propose an information-theoretic measure that maximizes the coding rate difference between the whole dataset and the sum of each individual class, known as rate reduction. This new objective provides a more unifying view of above objectives such as cross-entropy, information bottleneck, contractive and contrastive learning.

- 深度网络的搭建方法 via optimization
  - “We contend that all key features and structures of modern deep (convolution) neural networks can be naturally derived from optimizing the rate reduction objective ”
  - 经典的迭代投影提督上升优化方法采取每层迭代一次的形式。
  - 在上述框架下可认为网络宽度表现为保证低维结构的统计资源，网络深度表现为线型判别表达的算力资源。
  - 导数指出在频域上建立这样的卷积网络计算上更高效。

The Principle of Maximal Coding Rate Reduction

- 线性表达紧致性测量
  - 来自信息论
    - Rate distortion: R(z, ε), the minimal number of binary bits needed to encode z s.t. the expected decoding error is less than ε
  - 子空间上有限样本的Rate distortion:
    - 计算困难：不知道z的分布，但是可以有限采样。
    - 平均意义上每个样本的coding length(m是样本数，n是特征维数，这里要求m足够大)

(2007) for proofs. Therefore, the compactness of learned features *as a whole* can be measured in terms of the average coding length per sample (as the sample size m is large), a.k.a. the *coding rate* subject to the distortion ϵ:

$$R(\boldsymbol{Z}, \epsilon) \doteq \frac{1}{2} \log \det \left(\boldsymbol{I} + \frac{n}{m\epsilon^2} \boldsymbol{Z}\boldsymbol{Z}^* \right). \tag{7}$$

See Figure 2 for an illustration.

- 1

- Rate distortion of samples on a mixture of subspaces
  - 一般来说，多维数据可能属于多种低维子空间，因此Z可以被分到多个子集Z1UZ2U..UZk，Zi属于低维子空间。

subsets: $\boldsymbol{Z} = \boldsymbol{Z}^1 \cup \boldsymbol{Z}^2 \cup \ldots \cup \boldsymbol{Z}^k$, with each $\boldsymbol{Z}^j$ containing samples in one low-dimensional subspace.[14] So the above coding rate (7) is accurate for each subset. For convenience, let $\boldsymbol{\Pi} = \{\boldsymbol{\Pi}^j \in \mathbb{R}^{m \times m}\}_{j=1}^k$ be a set of diagonal matrices whose diagonal entries encode the membership of the m samples in the k classes. More specifically, the diagonal entry $\boldsymbol{\Pi}^j(i,i)$ of $\boldsymbol{\Pi}^j$ indicates the probability of sample i belonging to subset j. Therefore $\boldsymbol{\Pi}$ lies in a simplex: $\Omega \doteq \{\boldsymbol{\Pi} \mid \boldsymbol{\Pi}^j \geq \mathbf{0},\ \boldsymbol{\Pi}^1 + \cdots + \boldsymbol{\Pi}^k = \boldsymbol{I}\}$. Then, according to Ma et al. (2007), with respect to this partition, the average number of bits per sample (the coding rate) is

$$R_c(\boldsymbol{Z}, \epsilon \mid \boldsymbol{\Pi}) \doteq \sum_{j=1}^{k} \frac{\operatorname{tr}(\boldsymbol{\Pi}^j)}{2m} \log\det\left(\boldsymbol{I} + \frac{n}{\operatorname{tr}(\boldsymbol{\Pi}^j)\epsilon^2}\boldsymbol{Z}\boldsymbol{\Pi}^j\boldsymbol{Z}^*\right). \tag{8}$$

  - Log det(.)有利于解决秩最小化问题

- 最大Coding Rate Reduction的原则
  - 考虑类间不相关性，不同类的特征共同张成的空间要足够大
  - 考虑类内相关性，每类张成的空间要足够小
  - the basic rule that similarity contracts and dissimilarity contrasts：

To be more precise, a good (linear) discriminative representation $\boldsymbol{Z}$ of $\boldsymbol{X}$ is one such that, given a partition $\boldsymbol{\Pi}$ of $\boldsymbol{Z}$, achieves a large difference between the coding rate for the whole and that for all the subsets:

$$\Delta R(\boldsymbol{Z}, \boldsymbol{\Pi}, \epsilon) \doteq R(\boldsymbol{Z}, \epsilon) - R_c(\boldsymbol{Z}, \epsilon \mid \boldsymbol{\Pi}). \tag{9}$$

- The role of nomalization
  - 注意ΔR 和Z同阶，为了让不同表达之间的reduction有可比性，我们需要normalize Z. 通过Frobenius norm来和特征个数匹配或单位化

) scale with the number of features in $\boldsymbol{Z}^j \in \mathbb{R}^{n \times m_j}$: $\|\boldsymbol{Z}^j\|_F^2 = m_j$ (

  - Besides normalizing the scale, normalization could also act as a precondition mechanism that helps accelerate gradient descent (Liu et al., 2021).

- 由此，我们的目标转变为学习使得ΔR最大化的Z(θ) = f(X,θ) 及其拆分 Π (if not given in advance, being nomalized) (the principle of maximal coding rate reduction (MCR^2)
  - 其思想正在于整体大于局部之和

- 和获得信息量(information gain, IG)的关系
  - The maximal coding rate reduction can be viewed as a generalization to IG
- Properties of Rate Reduction Function

Theorem 1 (Informal Statement) *Suppose $\boldsymbol{Z}_\star = \boldsymbol{Z}_\star^1 \cup \cdots \cup \boldsymbol{Z}_\star^k$ is the optimal solution that maximizes the rate reduction (11) with the rates R and R_c given by (7) and (8). Assume that the optimal solution satisfies $\mathsf{rank}(\boldsymbol{Z}_\star^j) \leq d_j$.[18] We have:*

- Between-class Discriminative*: As long as the ambient space is adequately large ($n \geq \sum_{j=1}^k d_j$), the subspaces are all orthogonal to each other,* i.e. *$(\boldsymbol{Z}_\star^i)^* \boldsymbol{Z}_\star^j = \boldsymbol{0}$ for $i \neq j$.*
- Maximally Diverse Representation*: As long as the coding precision is adequately high, i.e., $\epsilon^4 < \min_j \left\{ \frac{m_j}{m} \frac{n^2}{d_j^2} \right\}$, each subspace achieves its maximal dimension, i.e. $\mathsf{rank}(\boldsymbol{Z}_\star^j) = d_j$. In addition, the largest $d_j - 1$ singular values of $\boldsymbol{Z}_\star^j$ are equal.*

- Relation to neural collapse:
  - 相同类的特征是一致的，而不同类彼此最大分离。
  - R_c则是neural collapse的一种解
- 和OLE loss的比较

low-rank embedding (OLE) loss: $\max_{\boldsymbol{\theta}} \mathrm{OLE}(\boldsymbol{Z}(\boldsymbol{\theta}), \boldsymbol{\Pi}) \doteq \|\boldsymbol{Z}(\boldsymbol{\theta})\|_* - \sum_{j=1}^k \|\boldsymbol{Z}^j(\boldsymbol{\theta})\|_*$,

  - Unlike the rate reduction ΔR, OLE is always negative and achieves the maximal value 0 when the subspaces are orthogonal, regardless of their dimensions. 没有保证表达的多样性
- 和收缩学习及对比学习的关系
  - 样本对(xi,xj)，ΔR^{ij} = R($Z^i \cup Z^j$,ε)– $^1/_2$(R(Z^i,ε)+R(Z^j,ε))给出了一种类似距离的描述

- 深度网络 from Maximising Rate Reduction
  - To learn the feature mapping z(θ) = f(x,θ)
  -

- 梯度上升
  - 重新表达一下ΔR(coding rate reduction)目标函数：

$$\begin{aligned}\Delta R(\boldsymbol{Z}, \boldsymbol{\Pi}, \epsilon) &= R(\boldsymbol{Z}, \epsilon) - R_c(\boldsymbol{Z}, \epsilon \mid \boldsymbol{\Pi}) \\ &\doteq \underbrace{\frac{1}{2}\log\det\Big(\boldsymbol{I} + \alpha \boldsymbol{Z}\boldsymbol{Z}^*\Big)}_{R(\boldsymbol{Z},\epsilon)} - \underbrace{\sum_{j=1}^{k}\frac{\gamma_j}{2}\log\det\Big(\boldsymbol{I} + \alpha_j \boldsymbol{Z}\boldsymbol{\Pi}^j\boldsymbol{Z}^*\Big)}_{R_c(\boldsymbol{Z},\epsilon\mid\boldsymbol{\Pi})},\end{aligned} \tag{12}$$

where for simplicity we denote $\alpha = \frac{n}{m\epsilon^2}$, $\alpha_j = \frac{n}{\mathrm{tr}(\boldsymbol{\Pi}^j)\epsilon^2}$, $\gamma_j = \frac{\mathrm{tr}(\boldsymbol{\Pi}^j)}{m}$ for $j = 1, \ldots, k$.

- 优化方法
  - 最简单：PGA(projected gradient ascent)

$$\boldsymbol{Z}_{\ell+1} \propto \boldsymbol{Z}_\ell + \eta \cdot \frac{\partial \Delta R}{\partial \boldsymbol{Z}}\bigg|_{\boldsymbol{Z}_\ell} \quad \text{s.t.} \quad \boldsymbol{Z}_{\ell+1} \subset \mathbb{S}^{n-1},\ \ell = 1, 2, \ldots, \tag{13}$$

for some step size $\eta > 0$ and the iterate starts with the given data $\boldsymbol{Z}_1 = \boldsymbol{X}$[23]. This scheme

  - 计算梯度项实际上需要计算ΔR(Z) 中的两项：

Simple calculation shows that the gradient $\frac{\partial \Delta R}{\partial \boldsymbol{Z}}$ entails evaluating the following derivatives of the two terms in $\Delta R(\boldsymbol{Z})$:

$$\frac{1}{2}\frac{\partial \log\det(\boldsymbol{I}+\alpha\boldsymbol{Z}\boldsymbol{Z}^*)}{\partial \boldsymbol{Z}}\bigg|_{\boldsymbol{Z}_\ell} = \underbrace{\alpha(\boldsymbol{I}+\alpha\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell^*)^{-1}}_{\boldsymbol{E}_\ell \in \mathbb{R}^{n\times n}}\boldsymbol{Z}_\ell, \tag{14}$$

$$\frac{1}{2}\frac{\partial\left(\gamma_j \log\det(\boldsymbol{I} + \alpha_j \boldsymbol{Z}\boldsymbol{\Pi}^j\boldsymbol{Z}^*)\right)}{\partial \boldsymbol{Z}}\bigg|_{\boldsymbol{Z}_\ell} = \gamma_j \underbrace{\alpha_j(\boldsymbol{I} + \alpha_j \boldsymbol{Z}_\ell\boldsymbol{\Pi}^j\boldsymbol{Z}_\ell^*)^{-1}}_{\boldsymbol{C}_\ell^j \in \mathbb{R}^{n\times n}}\boldsymbol{Z}_\ell\boldsymbol{\Pi}^j. \tag{15}$$

  - 其中E_l仅取决于Z_l，其目的在于扩展特征以增加整体的coding rate
  - Cjl则取决于每一类特征，其目的在于压缩每一类的coding rate
  - 完整的梯度为以下形式：

$$\frac{\partial \Delta R}{\partial \boldsymbol{Z}}\bigg|_{\boldsymbol{Z}_\ell} = \underbrace{\boldsymbol{E}_\ell}_{\text{Expansion}}\boldsymbol{Z}_\ell - \sum_{j=1}^{k}\gamma_j \underbrace{\boldsymbol{C}_\ell^j}_{\text{Compression}}\boldsymbol{Z}_\ell\boldsymbol{\Pi}^j. \tag{16}$$

- 由上式知，梯度不能在一个成员未知的点上计算，因此我们会考虑在第l层特征上建造一个小的增量变化g(·,θl)来模拟上面的梯度方案。

$$\boldsymbol{z}_{\ell+1} \propto \boldsymbol{z}_\ell + \eta \cdot g(\boldsymbol{z}_\ell, \boldsymbol{\theta}_\ell) \quad \text{subject to} \quad \boldsymbol{z}_{\ell+1} \in \mathbb{S}^{n-1} \tag{19}$$

such that: $\big[g(\boldsymbol{z}_\ell^1, \boldsymbol{\theta}_\ell), \ldots, g(\boldsymbol{z}_\ell^m, \boldsymbol{\theta}_\ell)\big] \approx \frac{\partial \Delta R}{\partial \boldsymbol{Z}}\big|_{\boldsymbol{Z}_\ell}$. That is, we need to approximate the gradient

- 因此我们需要近似梯度流