语言模型不变量

杨天骥　华东师范大学

2023-01-08

目录

深度学习中的不变量：图神经网络

虽然我们存储向量 $x[1...n]$ 和图的邻接矩阵 $a[1...n][1...n]$ 时为节点编号，但是我们希望计算结果和节点编号无关，这就要用到置换不变性。
具体地，如果我们用如下函数为每个节点算特征：

$$f : \mathbb{R}^n \times \mathbb{R}^{n \times n} \to \mathbb{R}^n$$

节点 i 计算得到的特征

$$y = f(x[1...n], a[1...n][1...n])[i]$$

我们希望它满足

$$y = f(x[\sigma[1...n]], a[\sigma[1...n]][\sigma[1...n]])[\sigma[i]]$$

深度学习中的不变量：图神经网络

不甚严格地，对一个给定节点，上述等式如下图。

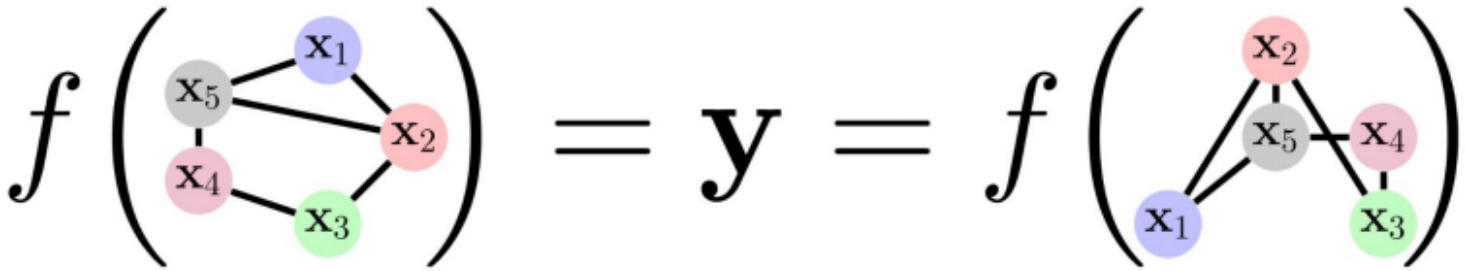
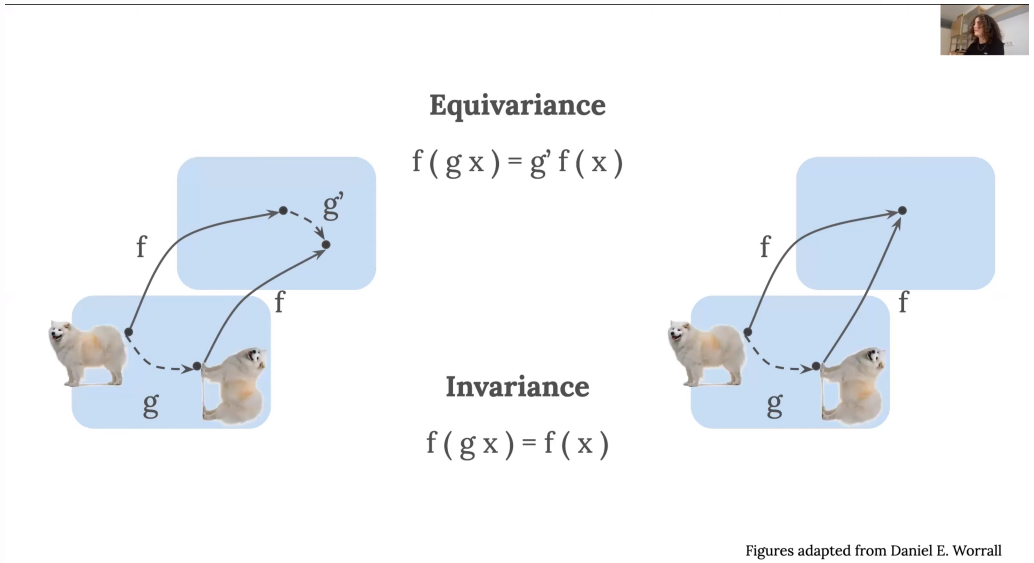


image source: https://dataroots.io/blog/a-gentle-introduction-to-geometric

深度学习中的不变量：计算机视觉

图像分割中的平移不变量：一个图像平移前后得到的特征相同，满足这个特点的运算是卷积。

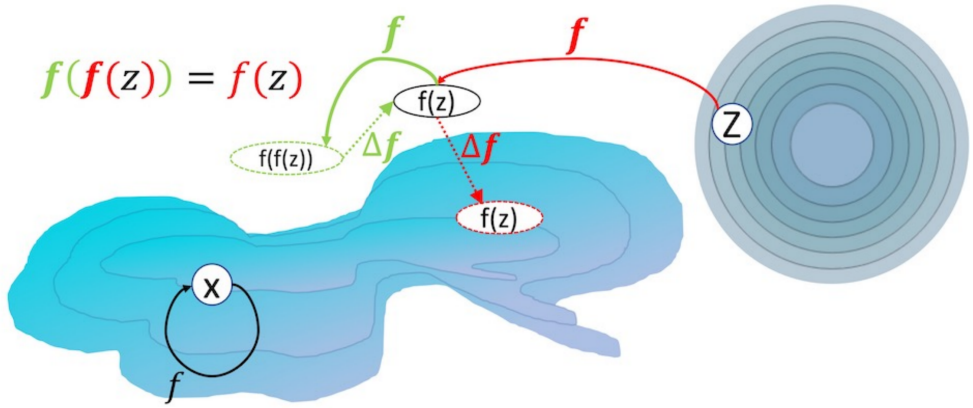


Figures adapted from Daniel E. Worrall

source: AMMI Seminar - Geometric Deep Learning and Reinforcement Learning (2021)

深度学习中的不变量：计算机视觉

类似还有图像恢复的幂等性，满足幂等性的函数 f 满足 $f(f(x)) = f(x)$，即模型会将恢复后的图像映射到自身.



source: https://openreview.net/forum?id=XIaS66XkNA

语言模型中的不变量

相比其他领域，语言模型当中的不变量被研究得不多。

这是因为先前这些内容中，不变量都是能够通过单个样本的变换或者应用于单个样本的数据增强直接学得的，而语言模型当中的不变量往往涉及到模型在样本分布的变换下的不变性，而分布本身是待学习的，这就导致表示这类不变性的等式中含有一些本身未知，也很难从模型的输出中计算的结果。

下面我先介绍一个先前唯一看到的关于不变量的研究，然后说说我的课题。

语言模型中的不变量：词元置换

出于理论上的兴趣，希望模型只表示词元之间的关系。(Lexinvariant Language Models)

形式上，若对一句句子中的每个词元应用词表上的双射 $\sigma : \Sigma \to \Sigma$，模型总是计算出相同的概率。

p(“a big banana”)
=
p(“e cop cekeke”)
=
p(“o lan lomomo”)

source: https://arxiv.org/pdf/2305.16349.pdf

语言模型中的不变量：平移变换

令语言 L 的词表为 Σ .
令语言模型为 $p_\theta : \Sigma^* \to [0,1]$ 。对序列 s , $p_\theta(s)$ 表示 s 在 L 中出现的概率。
假设样本是从一个Stationary Process中截取的，那么真实分布 $p : \Sigma^* \to [0,1]$ 满足

$$\sum_{x_{[1]} \in \Sigma} p(x_{[1\ldots n]}) = p(x_{[2\ldots n]}) \qquad (*)$$

而目前的工作中，研究者并没有让语言模型 p_θ 像前述的分布一样满足条件 $(*)$。

语言模型中的不变量：平移变换

实际上对最简单的Markovian模型，这个条件也不是平凡的。

$$p_\theta(x_{[1\ldots n]}) = q_\theta(x_{[1]}) \prod_{j=1}^{n-1} \pi_\theta(x_{[j+1]} \mid x_{[j]})$$

代入 $(*)$

$$\sum_{x_{[1]} \in \Sigma} q_\theta(x_{[1]}) \prod_{j=1}^{n-1} \pi_\theta(x_{[j+1]} \mid x_{[j]}) = q_\theta(x_{[2]}) \prod_{j=2}^{n-1} \pi_\theta(x_{[j+1]} \mid x_{[j]})$$

得到

$$\sum_{x_{[1]} \in \Sigma} q_\theta(x_{[1]}) \pi_\theta(x_{[2]} \mid x_{[1]}) = q_\theta(x_{[2]})$$

语言模型中的不变量：平移变换

采取这个不变量假设可能的好处：

- 减少位置偏置：对一种给定类型的文本段，它们可能总是出现在样本的特定位置。推理时如果在不常见的位置见到这类文本段，模型的推理效果可能变差。平移变换下的不变性使得无论这个文本段出现在什么位置，模型都能采取相同的表示。
- 增加参数效率：和上面一点基本上时相同的，对不满足平移不变性的模型，同一文本在不同位置出现，模型不需要反复学习新的表示，只需根据出现在前面的文本调整分布。这类似于卷积可以用更少的参数表示符合平移不变量条件的全连接层。

实现思路：

- 首先调研能否直接设计一些函数使得模型直接满足平移不变性。
- 如果前一种方案每成果，调研是否能用数据增强的方式遮蔽掉平移变换。