1

多模态表示学习关注于所学习到的多模态表示是否具有良好性质，并且能够很好地应用于下游任务。
对原始数据提取一个好的特征表示一直是机器学习关注的重要问题，好的特征表示主要有平滑性、时间和空间一致性、稀疏性和自然聚类等特性。特征表示代表了一个实体数据，一般用张量来表示。实体可以是一个图像，音频样本，单个词，或一个句子。多模态的特征表示是使用来自多个此类实体的信息，主要存在的问题有：
（1）如何组合来自不同模态的数据
（2）如何处理不同模态不同程度的噪音
（3）如何处理缺失数据。

PS：相较于多模态，基于单模态的表征学习已被广泛且深入地研究。在Transformer出现之前，不同模态所适用的最佳表征学习模型不同，例如，CNN广泛适用CV领域，LSTM占领NLP领域。较多的多模态工作仍旧局限在使用N个异质网络单独提取N个模态的特征，之后采用Joint或Coordinated结构进行训练。不过这种思路在很快改变，随着越来越多工作证实Transformer在CV和NLP以及Speech领域都可以获得极佳的性能，仅使用Transformer统一多个模态、甚至多个跨模态任务成为可能。基于Transformer的多模态预训练模型在2019年后喷涌而出，如LXMERT, Oscar, UNITER属于Joint结构，CLIP, BriVL属于Coordinated结构。
（现在感觉用大模型做全表征效果更好，比如GPT，DALL·E）

2

协同（Coordinated）多模态表示是指使用一个资源丰富的模态信息来辅助另一个资源相对贫瘠的模态进行学习，将每个模态投影到分离但相关的空间，这种方法适用于推理时仅有一种模态出现的情况。

PS：比如迁移学习（Transfer Learning）就是属于这个范畴，绝大多数迈入深度学习的初学者尝试做的一项工作就是将 ImageNet 数据集上学习到的权重，在自己的目标数据集上进行微调。
协同学习中还有一类工作叫做协同训练，它负责研究如何在多模态数据中将少量的标注进行扩充，得到更多的标注信息。

3

理解高维数据中关系：
多元统计分析是研究多个随机变量之间相互依赖关系及其内在统计规律的一门学科，它能够在多个对象和多个指标互相关联的情况下分析它们的统计规律。有如下应用：

- 变量之间的相依性分析
- 构造预测模型，进行预报控制。

- 进行数值分类，构造分类模型。
- 简化系统结构，探讨系统内核

在深度学习中，一般用来对变量（特征）进行简化，抽取主要特征。

4

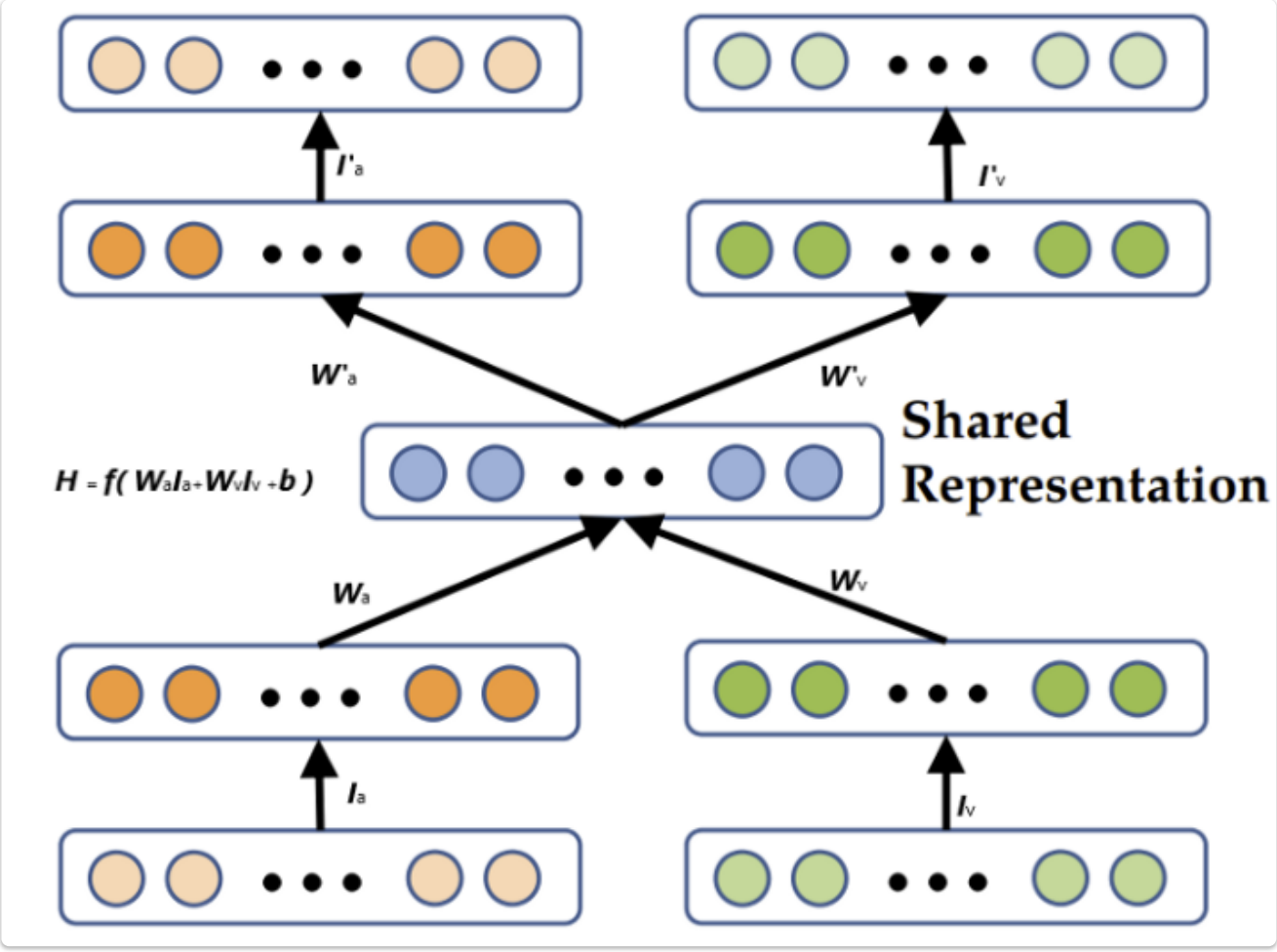
典型相关分析的核心如下：

1. Data Reduction：用少量的线性组合来解释两组变量之间的相关作用。
2. Data Interpretation：寻找特征值，这些特征值对于解释两个变量集合之间的相互作用十分关键。
   典型相关分析的基本思想和主成分分析的基本思想相似，它将一组变量与另一组变量之间单变量的多重线性相关性研究，转换为少数几对综合变量之间的简单线性相关性的研究，并且这少数几对变量所包含的线性相关性的信息几乎覆盖了原变量组所包含的全部相应信息。

在相关分析中，当考察的一组变量仅有两个时，可用简单相关系数来衡量它们;当考察的一组变量有多个时，可用复相关系数来衡量它们。大量的实际问题需要我们把指标之间的联系扩展到两组变量，即两组随机变量之间的相互依赖关系。典型相关分析就是用来解决此类问题的一种分析方法。它实际上是利用主成分的思想来讨论两组随机变量的相关性问题，把两组变量间的相关性研究化为少数几对变量之间的相关性研究，而且这少数几对变量之间又是不相关的，以此来达到化简复杂相关关系的目的。

5

让我们考虑一个双视图输入，$Z = [Ia，Iv]$其中Ia和Iv是两个不同的数据视图，例如音频和视频。在下图中，显示了具有此数据的深度相关的简单架构。



其中编码器和解码器都是单层的。 H 是编码表示。 $Ha = f(Wa.Ia + b)$ 是编码表示$Ia.$ f\$是非线性激活函数。\$$Hv = f(Wa.Ia + b)$ 。双峰数据Z的常见表示形式如下:
$\mathrm{H} = f(W\mathrm{a} \cdot \mathrm{Ia} + Wv \cdot Iv + b)$。
在解码器部分, 模型通过 $I'a = g(W'a \cdot H' + b')$ 和 $I'v = g(W'vH + b')$, 其中 g 是激活函数, I'a 和I'v是重建的输入。
在训练期间, 梯度是根据三个损失计算的:
i) 最小化自重构误差, 即最小化从 Ia 重建 Ia和从 Iv 重建 Iv 的误差。
ii）最小化交叉重建误差, 即最小化从Ia重建Ia和从Ia重建Ia的误差。
iii）最大化两个视图的隐藏表示之间的相关性, 即最大化 Ha 和 Hv 之间的相关性。
深度相关网络就是如此构成

6

从多个视图(模式)学习数据分区（算多模态融合的一种方法）
共识原则:最大化多个不同视图的一致性
互补原则:需要多种视图才能获得更全面、更准确的描述