周报（2023. 8.18-2023.8.23）姓名：孙瑞阳（请假）

科研详情

文献阅读

文献 1

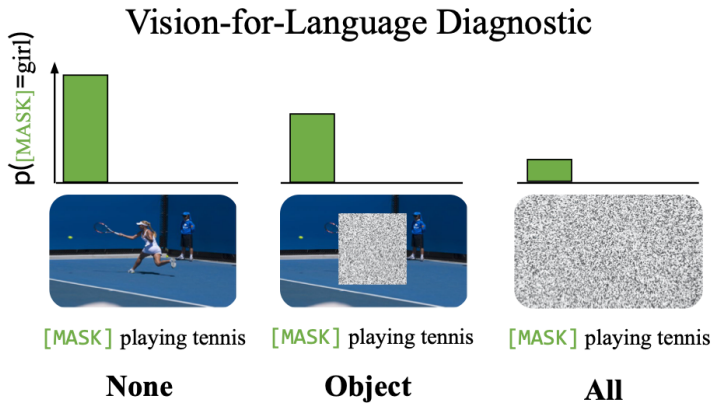
题目：Vision-and-Language or Vision-for-Language? On Cross-Modal Inflfluence in Multimodal Transformers

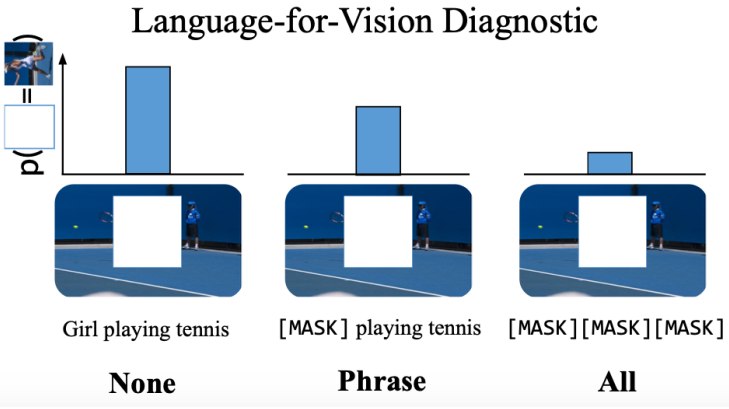
作者：Stella Frank, Emanuele Bugliarello, Desmond Elliott

出处：arxiv2021

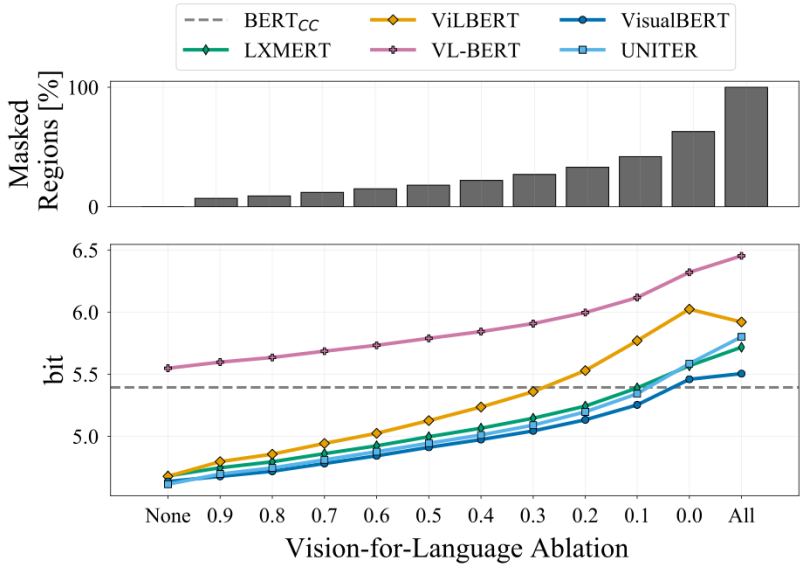
方法：

提出了 Multimodal Bottleneck Transformer(MBT)，利用 self-attention 在中间层对多模态数据进行信息交换。为了减少计算量，将最需要和另一个模态分享的信息编码在一个 4 维隐向量中，使用它分别与两个模态的向量做 self-attention 以实现信息交换。





作者引入了一种**跨模态输入消融（cross-modal input ablation）** 方法来量化预训练的模型学习使用跨模态信息的程度。本文的方法**不需要额外的训练，通过消融一个模态的信息来预测另一个模态的输出，比较不同消融程度对结果的变化，来探究模态之间的相互作用。**



文本片段的表示受到视觉输入的强烈影响，而视觉区域的表示受到伴随的文本输入的较小影响 。这表明跨模态信息交换的水平不是对称的：**模型已经学会了使用 vision-for-language 而不是 language-for-vision** 。

本文介绍的跨模态输入消融诊断证明了预训练的视觉和语言模型中的不对称性：**mask 文本的预测受到消融视觉输入的强烈影响；而在预测 mask 图像区域时，消融文本输入几乎没有影响** 。这些结果提供了对实际模型行为的一个有用的 check，并与平衡的跨模态激活的假设背道而驰。

启发：

1. 不同模态的不同消融程度对结果的变化，来探究模态之间的相互作用，对这部分实验分析比较感兴趣，在看代码

文献2

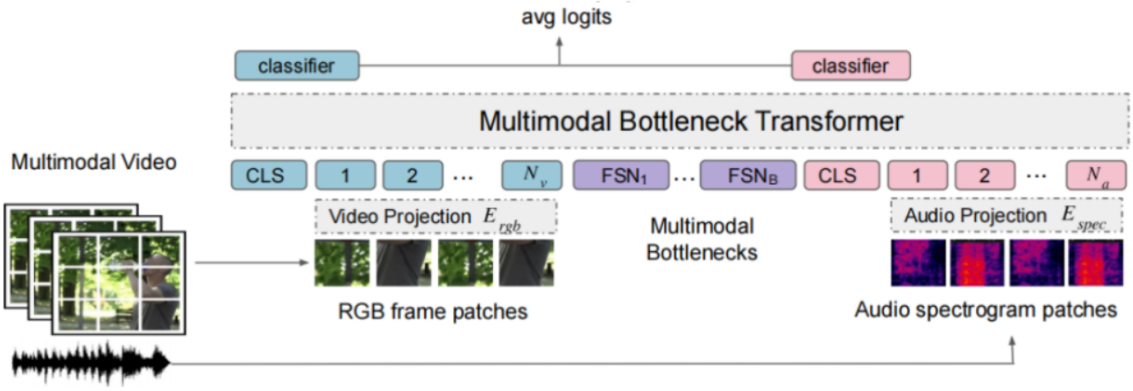
题目：Attention Bottlenecks for Multimodal Fusion

作者：Arsha Nagrani Shan Yang Anurag Arnab Aren Jansen　Cordelia Schmid　Chen Sun

出处：NIPS 2021

方法：

提出了 Multimodal Bottleneck Transformer(MBT)，利用 self-attention 在中间层对多模态数据进行信息交换。为了减少计算量，将最需要和另一个模态分享的信息编码在一个 4 维隐向量中，使用它分别与两个模态的向量做 self-attention 以实现信息交换。

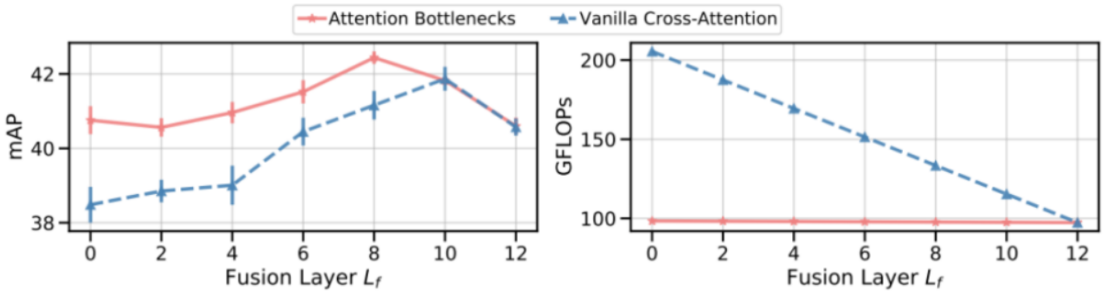


引入了一种基于 transformer 的新型架构，该架构将 fusion bottlenecks 安置在多个 layer 中以实现模态融合。与传统的 pairwise self-attention 相比，该模型迫使不同模态之间的信息

通过少量的 bottleneck latents 形成交流，要求模型在每个模态中整理和压缩相关信息，并共享必要的信息。
在多模态学习中，一般都是限制网络的早期层专注于单模态处理，并且只在后期层引入交叉模态连接。
相信早期的层（lower layer）参与处理低水平特征,而 higher layers 的重点是学习语义概念。低级视觉特征,如边缘和角落图像可能没有一个特定的声音签名,因此不可能受益于与音频在早期融合。
如何实现：
在前几层在每个模态的 token 上单独使用 transformer 层，之后将两者拼接起来再输入提出的 Multimodal Bottleneck Transformer

实验结果：

随着单独操作层数变深，效果先升后降，说明middle fusion最有效：



启发：

1. 看代码，学习多模态中的 token fusion 策略，接下来可以在这个方向做一些工作。

Authors	Title	内容
Zhicheng Shi，深圳大学	A Density-Peak-Based Clustering Method for Multiple Densities Dataset	密度峰聚类，跟着之前那篇 science 做的研究
李唯嘉，中山大学	Fine-grained building attribute mapping based on deep learning and a satellite to-street view matching method	做了一个遥感和街景配好的数据集（刚刚公开），还使用 OSM，用 RentinaNet、FasterRCNN、Mask RCNN 等做目标检测和分割，街景考虑多角度
涂伟	Spatial Synergistic Simulalon of Land Use - Populalon -Economy in the GBA	人口、OSM 等多源协同的 GBA 模型
Yu chen，深圳大学	The Study on the knowledge Graph and synergistic development of industrial clusters in the GBA	构建知识图谱，产业集群协同发展研究
Zuopeng Xiao and Jingying Liao	Modeling individual travel behavior in the real-time context: An space-time prism approach with isochronous circle.	时空轨迹压缩，对 margin time 进行计算

章语之、杨军，清华	A user-friendly assessment of six commonly used urban growth models	CA-Markov、CLUE-S、FLUS、LCM、LUSD、SLEUTH 六种衡量城市增长模型的数据要求、灵活性和参数比较。
Tianyou Chu, Yumin Chen, Jianshen Ma, 武大资环	An Entity Recognition and Semantic Clustering of City Complaint Hotline Data for Uncovering Urban Hot Problem	Entity Recognition and Semantic Clustering (ERSC)方法：对比学习+聚类；无监督+有监督，之后会做 GPT 上的一些工作
Shiqi Wang, Yuze Li, Anthony Chen and Chengxiang Zhuge，港理工	A Data-Driven Approach to Deploying Wireless Charging Lanes on a Large-Scale Electrified Bus Network.	基于实际 EFVs 轨迹数据的 DWC 车道联合布置与作业模型，建模一个多目标混合整数规划模型，目标是最大化节省的充电时间、最小化充电成本和最小化对交通的负面影响。
Carlo RATTI, MIT	Senseable Cites	做城市建模，包括介绍了地下城市的建模技术
Tianhong Zhao, Zhengdong Huang and Wei Tu	A Multiview Spatiotemporal Model for Bus Travel Demand Prediction using Graph Neural Networks.	多视角时空图神经网络(MSTGNN)模型对公交短期出行需求进行预测。首先，构建由公交、地铁和出租车视图组成的多视图，每个视图包含一个局部图和全局图;其次，开发了基于多视角注意力的时序图卷积模块，以捕获不同传输模式之间的时空和跨视角交互依赖；**引入跨视角空间特征一致性损失作为辅助损失。（和之前曹瑞老师 ISPRS 的 loss 很相似，这个网络也能用）**

工作进展

1：阅读文献；

2: 期刊论文草稿差不多写完了，还差公式和网络图重写重画。

3: 审了一篇稿子：A Novel Approach to Incomplete Multimodal Learning for Remote Sensing Data Fusion

下周计划

1. 排版完期刊论文给您
2. 看论文，跑开源代码