悟道·文澜
大规模通用中文多模态预训练模型及其可视化解释

卢志武 教授
中国人民大学高瓴人工智能学院
（代表文澜团队）

文澜的出发点

- 学术界多模态数据集与真实世界的数据集不同

「水果蛋糕上有一些蜡烛在燃烧。」

「生日快乐！许个愿吧！」
「我的减肥大计又泡汤啦！」

文澜的出发点

- 面对千万~亿级的真实世界里的图文对
  - 怎样的模型才能比较好的刻画这种关系？
  - 是否能以已有单模态预训练大模型的成果为基础？
  - 能否节约资源，可以应用落地到大中小型企业？

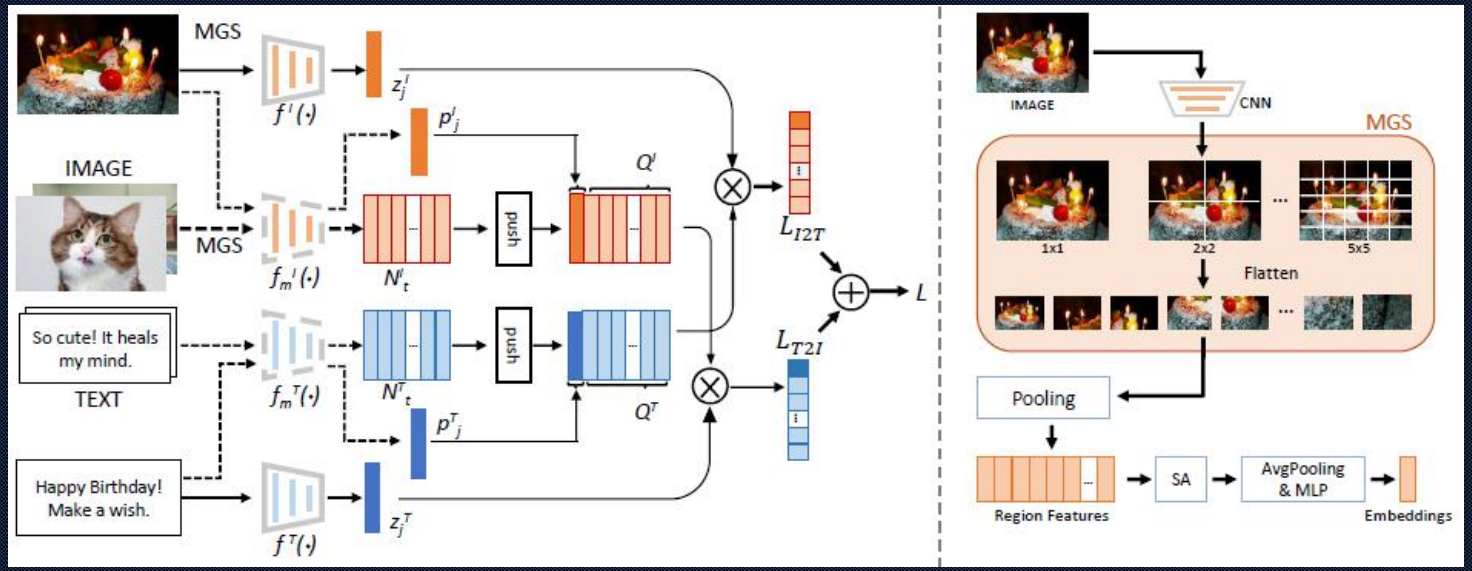
文澜 2.0 - 模型设计

文澜2.0的主要特点

- 首个中英文多模态双塔模型：采用图文弱相关假设，双塔模型的最大参数量为53亿，支持中英文双语，提供不同大小的多个版本
- 最大中文多模态预训练数据：共有6.5亿图文对，来自互联网和移动互联网，不经过特殊的数据清洗，契合图文弱相关假设
- 资源节约型的大规模预训练算法：提出基于DeepSpeed的分布式预训练算法，最大化利用GPU和CPU，并最优地支持大规模跨模态对比学习，在112卡上支持6.5亿数据的预训练

网络结构优化

- 替换常用的目标检测器（计算量非常大），采用Multi-Grid Split（MGS）池化技术，显著地提高计算效率



分布式预训练优化

- 文澜2.0面临的分布式预训练难题：
  - 预训练模型大：最大参数量为53亿，并有中英文多个版本
  - 预训练数据集大：6.5亿图文对，目前最大的中文多模态数据集

- 利用DeepSpeed的数据并行、混合精度训练及零冗余优化器（ZeRO）：
  - 减少预训练模型所占的GPU显存
  - 最大化GPU和CPU的利用率
  - 最优地支持大规模的跨模态对比学习

- 提出基于DeepSpeed的跨模态对比学习预训练算法
  - 112卡A100的预训练时间：6.5亿图文对，约7天/epoch

文澜 2.0 - 评测结果

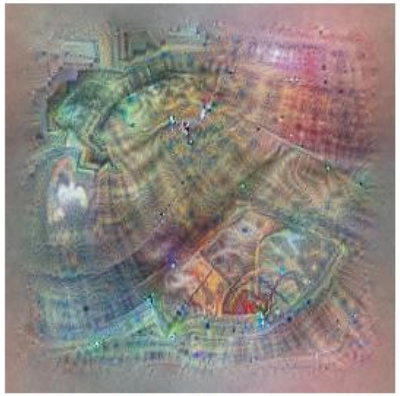
下游任务 – 遥感数据的零样本分类

- 利用文澜2.0的图文encoder，直接在遥感数据集AID上进行零样本分类，需要将30个遥感类名翻译成中文。测试时30个遥感类别被分成不同unseen/seen的分划。OpenAI CLIP直接在英文数据集上进行测试。

Method	Unseen/Seen Class Ratios				
	30 / 0	8 / 22	12 / 18	16 / 14	20 / 10
CLIP w/ ResNet-50	46.01	65.99 (0.08)	59.15 (0.05)	54.44 (0.05)	51.72 (0.04)
CLIP w/ ResNet-101	48.05	68.71 (0.07)	64.39 (0.06)	57.75 (0.06)	54.54 (0.05)
CLIP w/ ResNet-50x4	50.96	69.32 (0.08)	64.30 (0.05)	59.53 (0.06)	56.35 (0.04)
文澜2.0	**58.12**	**76.73 (0.09)**	**71.25 (0.07)**	**67.52 (0.06)**	**64.19 (0.04)**

下游任务 – 遥感数据的零样本分类

Remote sensing class “baseball field” from AID

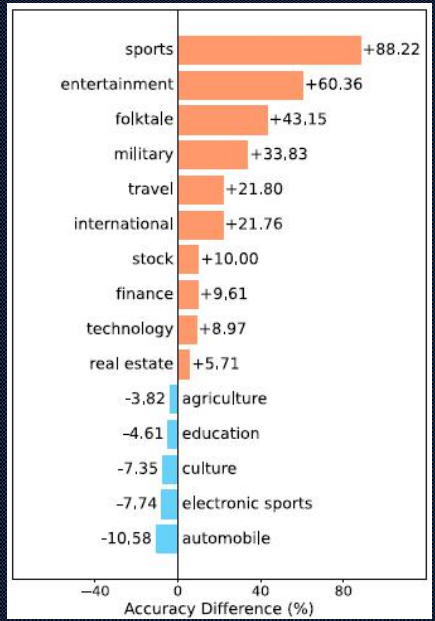


Visualization of “baseball field viewed from above” with BriVL

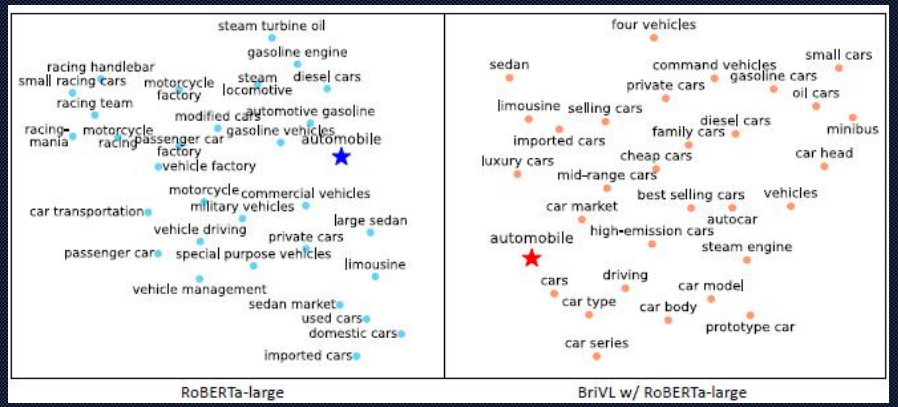
下游任务 – 中文新闻的零样本分类

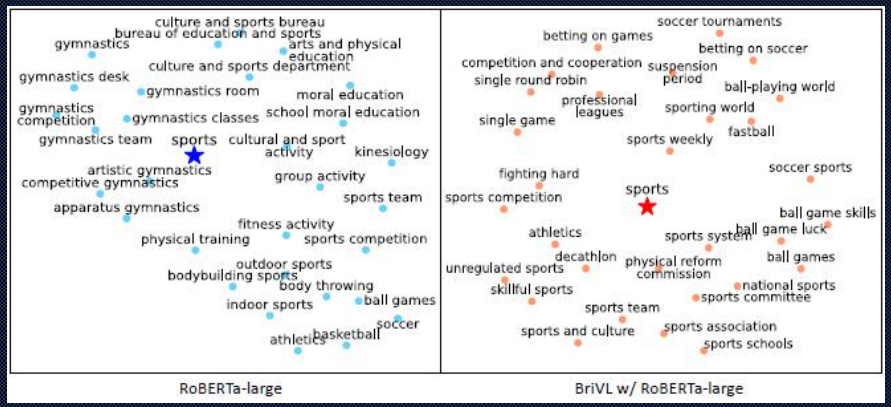
- 利用文澜2.0的文本encoder，在中文新闻数据Toutiao上进行零样本分类。对比实验考虑了单模态预训练的RoBERTa-base和RoBERTa-large。右图展示了文澜2.0相对于单模态RoBERTa-large在每类上的涨跌。

Method	Toutiao News
单模态RoBERTa-base	33.57
单模态RoBERTa-base (在文澜1.0的文本数据上微调)	32.13
文澜1.0（RoBERTa-base）	50.80
单模态RoBERTa-large	38.90
文澜2.0（RoBERTa-large）	**61.92**



下游任务 – 中文新闻的零样本分类





下游任务 – 视觉问答VQA

- 文澜2.0在VQA数据集Visual7w-telling上进行微调， 下游任务为视觉问答VQA，数据集需要提前翻译成中文，训练集：测试集=7:3。
- 基于跨模态对比学习的VQA建模：当前图像为anchor，正样本为问题+真实答案，负样本为问题+其他选项。

Method	Fix BN	# Unfixed Blocks	Question Type						
			What	Where	When	Who	Why	How	Overall
文澜2.0 (direct training)	no	4	70.51	71.99	81.88	77.05	78.36	68.62	72.16
文澜2.0 (pretrain & finetune)	no	2	**79.89**	**81.71**	87.78	**84.48**	82.66	**76.31**	**80.67**
文澜2.0 (pretrain & finetune)	no	4	79.41	81.66	87.31	84.46	**83.11**	74.44	80.16
文澜2.0 (pretrain & finetune)	yes	4	77.79	80.50	**87.99**	84.44	82.46	72.87	78.96

下游任务 – 视觉问答VQA

Why is the train blurry?
A. Moving fast.
B. Bad weather.
C. It's raining.
D. It's nighttime.
BriVL (direct training): C
BriVL (pre-train & finetune): A

How many people are playing?
A. Two.
B. Three.
C. Four.
D. Five.
BriVL (direct training): A
BriVL (pre-train & finetune): B

What are the boats doing?
A. Floating.
B. Sailing.
C. Getting cleaned.
D. Not moving.
BriVL (direct training): A
BriVL (pre-train & finetune): D

Why is the traffic stopped?
A. Car accident.
B. Traffic.
C. Red light.
D. Parade.
BriVL (direct training): B
BriVL (pre-train & finetune): C

文澜 2.0 - 可解释性

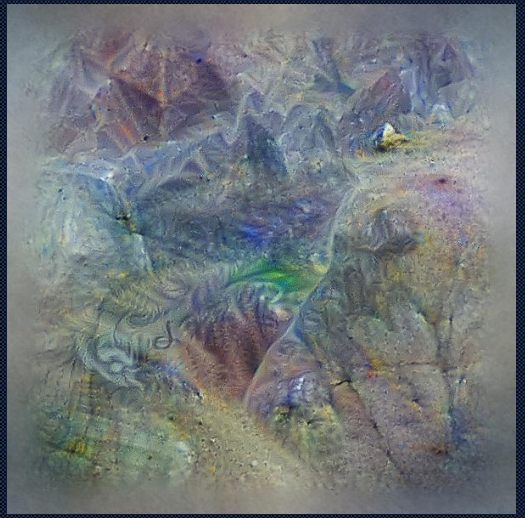
多模态神经元生成

1. 给定一个中文概念，输入一张随机噪声图像。
2. 通过文澜的文本Encoder 得到中文概念的Embedding。
3. 多模态神经元生成的目标函数为：最大化图像Encoder 最后一层某个神经元的输出，并同时让当前输入图像的视觉Embedding 逼近中文概念的文本Embedding 。
4. 固定文澜的图像Encoder，通过BP算法去更新输入的图像。
5. 算法收敛后，得到的输入图像即可看作图像Encoder最后一层某个神经元的可视化。

多模态神经元示例 – 具象概念

飞机场

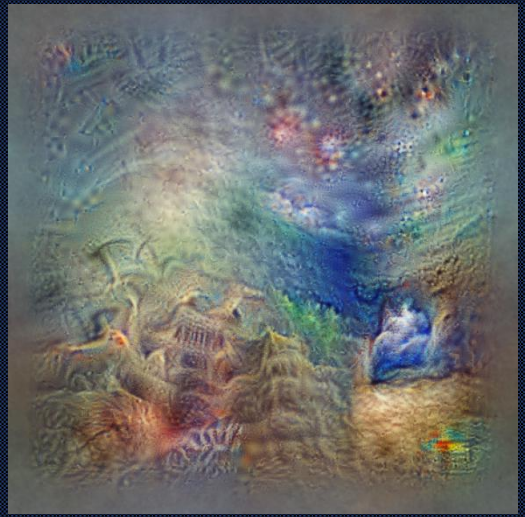
生日蛋糕



山脉

多模态预训练后的神经网络能“看到”具象的人类概念!

多模态神经元示例 – 抽象概念



梦境

科学

自然

多模态预训练后的神经网络也能“看到”抽象的人类概念!

多模态神经元示例 - 古诗句意境生成

江南可采莲，莲叶何田田

帘卷西风，人比黄花瘦

竹外桃花三两枝

多模态预训练后的神经网络甚至能“理解”古诗句的意境!

利用文澜实现VQGAN Inversion

1. 给定一个中文概念，输入一张随机噪声图像。
2. 通过文澜的文本Encoder 得到中文概念的Embedding。
3. VQGAN Inversion的目标函数为：当前输入图像经过VQGAN后输出图像，其视觉Embedding（通过文澜图像Encoder得到）必须逼近中文概念的文本Embedding 。
4. 固定VQGAN和文澜的图像Encoder，通过BP算法去更新当前输入图像。
5. 算法收敛后，最终得到的输入图像即可看作关于给定中文概念的VQGAN Inversion结果。

VQGAN Inversion示例 – 具象概念生成

喜马拉雅山

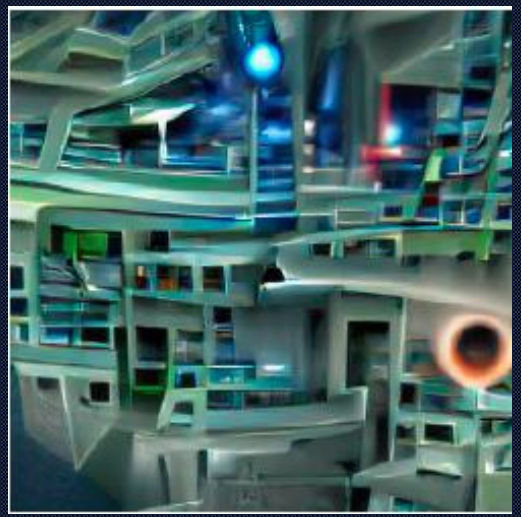
冬日的校园

乌云背后有阳光

VQGAN Inversion示例 – 超现实主义生成

赛博朋克城市

云中城堡



高维空间

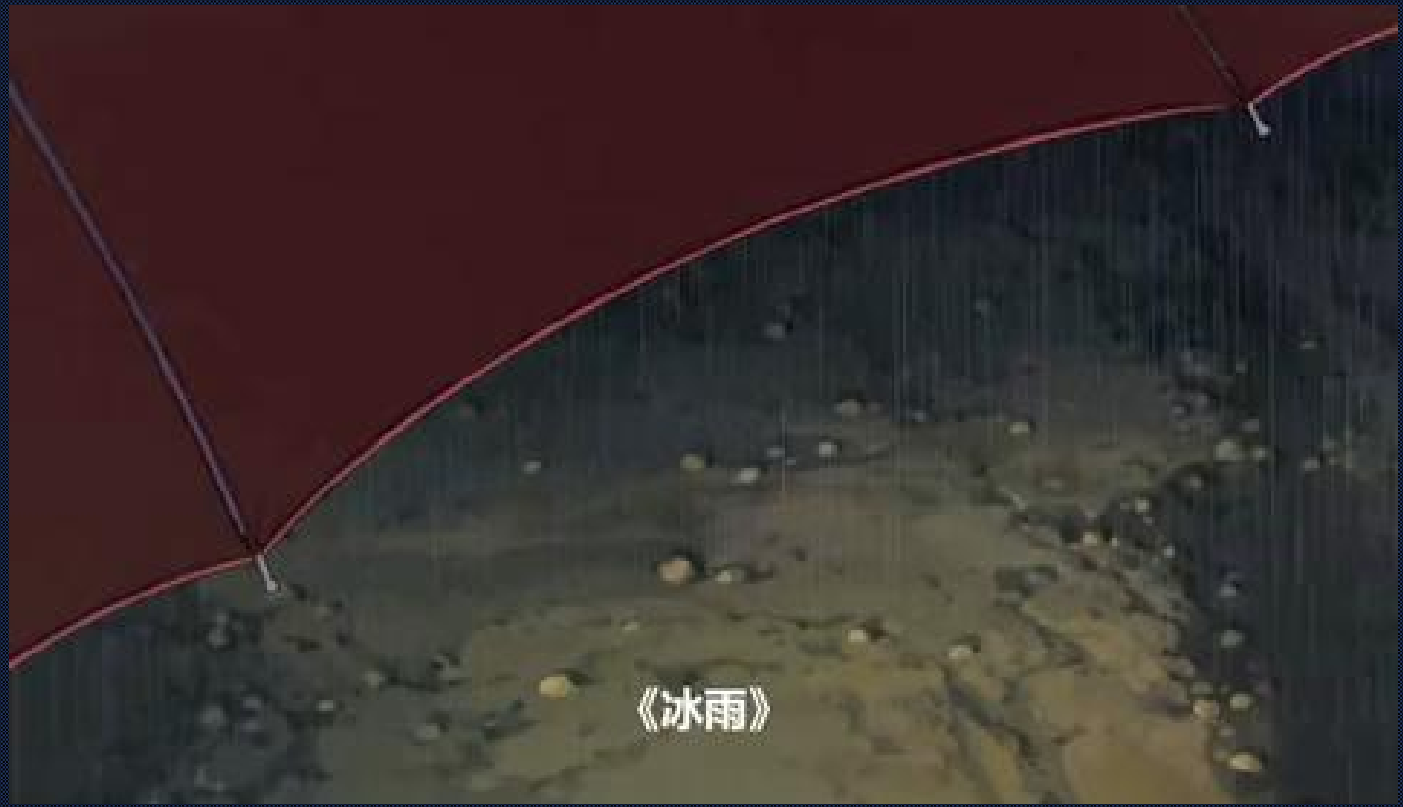
VQGAN Inversion示例 – 连环画生成

毛毛虫吃树叶长大

毛毛虫吐丝成茧

毛毛虫破茧而出，化成蝴蝶

文澜 2.0 – 应用落地



从文澜产品发布以来，用户累计调用公开API的次数逾千万次！

悟道·文澜
大规模通用中文多模态预训练模型
及其可视化解释

感谢整个文澜团队!