悟道·文澜：超大规模多模态预训练模型带来了什么？

WuDao-WenLan: What Do Very-Large Multimodal Pre-Training Models Bring?

卢志武/LU Zhiwu[1]，金琴/JIN Qin[2]，宋睿华/SONG Ruihua[1]，文继荣/WEN Jirong[1,2]
（1. 中国人民大学高瓴人工智能学院，中国 北京 100872；
2. 中国人民大学信息学院，中国 北京 100872）
(1. Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China;
2. School of Information Renmin University of China, Beijing 100872, China)

DOI：10.12142/ZTETJ.202202005
网络出版地址：https://kns.cnki.net/kcms/detail/34.1228.tn.20220411.1207.002.html
网络出版日期：2022-04-12
收稿日期：2022-02-20

摘要：提出了悟道·文澜的BriVL双塔模型。该模型利用6.5亿对互联网图文数据，通过自监督的任务来训练，是目前最大的中文通用图文预训练模型。同时，还提出了悟道·文澜的多语言多模态预训练单塔模型—MLMM。实验结果证明，这两个模型在多个国际公开数据集上均取得了最佳性能。设计了实验并讨论超大规模多模态预训练模型对文本编码、图像生成和图文互检带来的影响，以及文澜模型的落地应用与学科交叉成果。

关键词：多模态预训练；多语言预训练；双塔模型；单塔模型

Abstract: A multimodal pre-training two-tower model called WuDao-WenLan BriVL is proposed, which is trained through self-supervised learning over 650 million image-text pairs crawled from the Web. This is the largest open-sourced Chinese image-text pre-training model. Moreover, a multi-lingual pre-training single-tower model called WuDao-WenLan MLMM is also proposed. Extensive experiments show that these two models achieve the new state-of-the-art performance on multiple public benchmark datasets. In addition, experiments are conducted to discuss what very-large multimodal pre-training models bring to text encoding, text-to-image generation, and image-text retrieval, as well as in what applications WenLan can be applied in multiple fields.

Keywords: multimodal pre-training; multi-lingual pre-training; two-tower model; single-tower model

人脑是一个复杂的系统，能够处理多种感官模态例如视觉、听觉、嗅觉等的信息。这使得人们能够准确、有效地完成感知、理解和决策任务。为了模仿人类的这些核心认知能力，人工智能模型利用大规模多模态数据来进行训练。如何利用从互联网上爬取的大规模多模态数据进行模型训练，成为近期业界的研究热点。如何能有效地利用这些爬取数据是一个巨大的挑战，因为我们无法对其进行详细的人工标注。另外，这些数据不可避免地存在一定量的数据噪声。如图1所示，学术界数据集多为由人工编写的强相关文本，如“水果蛋糕上有一些蜡烛在燃烧”，规模多为几万到百万图文对。与此不同的是，从互联网上搜集到的图像的周边文本通常与内容弱相关。

多模态预训练的目标是对齐不同模式的大规模数据，从而可以将所学知识迁移到各种下游任务中，并最终接近通用人工智能。目前，多模态预训练模型已经在广泛的多模态任务中取得了巨大成功。然而，学术界往往只重视在有限规模的标注数据集上取得更好的效果，因此多采用单塔模型，并在英文数据集上进行训练。这使得其应用场景被规模、性能

[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！]

（a）强相关文本　　（b）弱相关文本

▲图1 两种不同的图文数据

和语言所局限。在北京智源研究院悟道项目的支持下，文继荣教授带领中国人民大学卢志武教授、宋睿华长聘副教授、金琴教授等师生团队搜集了6.5亿对中文图文数据，率先提出图文弱相关是更为现实的假设，并利用跨模态对比学习来自监督地训练超大规模图像-文本多模态预训练模型文澜BriVL。另外，我们认为：不同模态和不同语言都有可能表示相同的语义信息。如图2所示，中文单词“狗”、英文单词“dog”或是一张狗的视觉图像，都能表示狗这一动物。因此，我们研究了如何通过预训练来捕捉视觉与语言在语义上的共通点，提供更好的视觉和语言特征，以支持不同的多语言多模态下游任务；同时提出文澜多语言多模态预训练模型MLMM。实验证明，两个模型均能在多项下游任务中获得国际最佳性能。

此外，我们还着重讨论了超大规模多模态预训练带来的影响，包括对文本编码、图像生成和图文互检的影响。总之，多模态预训练带来的改变才刚刚开始，它在人工智能方面有着巨大的潜力。

1 文澜BriVL超大规模图文预训练模型

1.1 相关工作

自2018年以来，单模态预训练模型（如BERT[1]、GPT[2]、ViT[3]等）的出现，极大地促进了相关领域的发展。人们也在持续探索具有更强通用性的多模态预训练模型，具有代表性的工作有UNITER[4]、OSCAR[5]等。然而，由于视觉数据集的标注需要的成本高昂，多模态数据集往往维持在百万的数据量级，因此，难以在此基础上训练出具备良好通用性与泛化性的多模态模型。多模态预训练模型根据其框架可分为两类：单塔和双塔。

最近的UNITER[4]、Oscar[5]、M6[6]、VisualBERT[7]、Unicoder-VL[8]、VL-BERT[9]等模型都采用单塔网络，它们利用一个特征融合模块（例如Transformer）来得到图像-文本对的嵌入。其中，一些单塔模型还使用对象检测器来检测图像区域，并将这些区域与相应的单词进行匹配。UNITER作为单塔模型的代表，对560万图文对进行遮挡语言建模（MLM）、遮挡区域建模（MRM）和图像文本匹配（ITM）的联合训练，从而学到通用的图像文本表示。Oscar将语义相同的对象（名词）作为图像和文本对齐的基础，从而简化图像和文本语义对齐的学习任务，即使用快速目标检测器（Fast R-CNN）就可以将检测到的对象标签与文本中的单词建立关联。现有单塔结构通常依赖于强相关的图文对数据，而这一强相关假设对大规模网络数据集来说通常是无效的。

中文：一只躺在地板上的白色大狗。
英语：A white dog has its head on the ground.
德语：Ein weißer Hund hat seinen Kopf auf dem Boden.
日语：床に横たわっている大きな白い犬。
韩语：바닥에 누워 있는 커다란 흰색 개

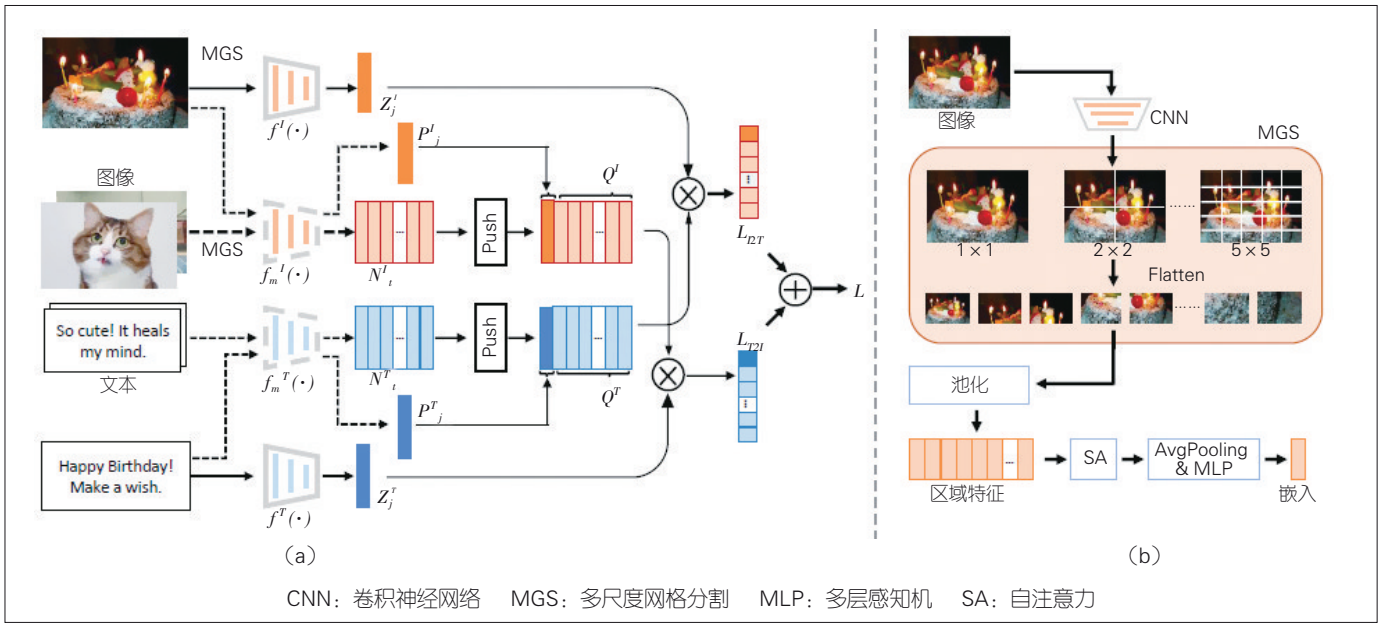
▲图2 不同语言和模态能够表达相同的语义

此外，单塔模型在推理阶段需要较高的计算成本。例如，需要将查询内容（图像或文本）输入到单塔模型中，计算它和所有候选对象的匹配分数。

相比之下，采用双塔结构的多模态预训练模型使用单独的图像和文本编码器，分别对图像和文本进行编码，然后进行图文对匹配来完成检索任务。这种模式的检索效率更高，但由于缺乏更深层次的图像-文本交互（即图像区域与单词的交互），通常只能达到次优性能。最近的双塔工作，如LigntningDot[10]，通过重新设计目标检测过程来应对这一挑战；CLIP[11]、ALIGN[12]、WenLan 1.0[13]和WenLan 2.0[14]则放弃了昂贵的对象检测器，利用跨模态对比学习任务来进行模型训练。

1.2 模型介绍

文澜BriVL模型在预训练数据的选择上，不再遵循强相关语义假设，而是转向弱相关假设；在网络架构上，选择双塔结构而不是单塔结构；使用了更加节约计算资源的跨模态对比学习算法来进行预训练。具体来说：（1）在弱相关语义假设下，图文数据不再需要任何人工标注，互联网上的海量多模态数据成为文澜BriVL模型的预训练数据来源。相比于人工标注的几百上千万强语义相关图文数据，文澜BriVL模型使用的预训练数据全部爬取自互联网，规模达到了6.5亿对。更重要的是，弱语义相关数据包含了复杂、抽象的人类情感和想法，能够帮助我们把文澜BriVL模型训练成一个更具认知能力的模型。（2）文澜BriVL模型不再需要耗时的目标检测器，使用的双塔网络架构在应用时也有明显的效率优势。双塔包含两个独立的编码器：一个用于图片，另一个用于文本。因此，在跨模态检索时，候选的图片或者文本可以提前计算出嵌入表示并做好索引，以满足现实应用的效率需求。（3）受到单模态对比学习算法MoCo的启发，文澜BriVL模型在使用跨模态对比学习的同时也引入Momentum机制以及动态维护负样本队列（如图3所示）。这样就解构了batch大小与负样本数量，从而在相对较小的batch下（即较少的图形处理器资源）就可以得到性能较好的预训练模型。



▲图3 文澜BriVL的网络架构图与图像编码器

1.3 实验分析

我们在图像零样本分类、文本零样本分类两个下游任务上进行实验，以验证文澜BriVL模型的迁移能力。

（1）下游任务1：ImageNet的零样本分类

我们利用文澜BriVL的图文编码器，可以直接在ImageNet数据集的200类图像子集上进行零样本分类。这需要提前将这200个类名翻译成中文。ImageNet 200类挑选的原则为：英文类名在翻译成中文时无明显错误。OpenAI CLIP则直接在英文数据集上进行测试。从表1可以发现，文澜BriVL 2.0的零样本图片分类准确率要高于CLIP。这说明我们的模型具有更好的泛化能力。

（2）下游任务2：中文学科的零样本分类

我们利用文澜BriVL1.0以及2.0的文本编码器，在中文学科分类数据集（CSLDCP）上进行小样本分类。我们采用被广泛使用的prompt-tuning方法来为1-shot分类。针对文澜BriVL模型，我们同时利用了视觉和文本两个模态的信息来进行prompt-tuning。对比实验考虑了单模态预训练的RoBERTa-base和RoBERTa-large。从表2可以发现，相比于单模态预训练模型RoBERTa，文澜BriVL模型具有更好的中文小样本分类能力。这说明多模态预训练在纯粹的NLP下游任务中也发挥了重要的作用。

1.4 模型可视化

文澜BriVL模型的可视化流程为：

（1）给定一个文本，输入一张随机噪声图像；

▼表1 ImageNet 200类的零样本分类结果

模型	ImageNet 200类	
	Top-1准确率/%	Top-5准确率/%
OpenAI CLIP	82.5	95.2
文澜BriVL 1.0（RoBERTa-base）	69.6	88.9
文澜BriVL 2.0（RoBERTa-large）	85.0	96.0

▼表2 中文学科的1-shot小样本分类结果

模型	CSLDCP
单模态RoBERTa-base	33.63
单模态RoBERTa-base (在文澜1.0的文本数据上微调)	33.40
文澜BriVL 1.0（RoBERTa-base）	35.59
单模态RoBERTa-large	38.24
文澜BriVL 2.0（RoBERTa-large）	46.47

CSLDCP：中文学科分类数据集

（2）通过模型的文本编码器得到文本的特征表示；

（3）多模态神经元可视化的目标函数为：让当前输入图像的视觉特征表示逼近文本特征；

（4）固定文澜的所有参数，通过反向传播来更新输入的噪声图像。

总之，算法收敛后，得到的图像是文澜BriVL认为的对输入文本最为接近的可视化处理结果。如图4所示，大规模多模态预训练后的神经网络已经能够理解古诗句的意境，展示了强大的中文理解能力。

2 文澜MLMM多语言多模态预训练模型

2.1 相关工作

目前，在多语言多模态的语义学习方面，已有一些工作陆续开展。M3P[15]首次采用了预训练来学习多语言多模态知识，以多任务学习的方式轮流将英文的图像描述数据和单模态的多语言语料输入到模型中，以进行预训练；UC2[16]使用机器翻译对现有的图像描述数据集进行多语言扩充，同时遮蔽两种语言相同意义的词来迫使模型根据图像内容进行还原。文献[17]采用英文图像描述数据和平行语料进行预训练，将Unicoder[18]扩展到多语言多模态上。

这些工作虽然取得了一定的成果，但其预训练规模仍局限于Conceptual Caption 3M数据集。较小规模的预训练使得模型的零样本跨语言迁移能力较弱。因此，我们致力于利用更大规模、更加开放领域的数据进行预训练，以获得更加通用、更加强大的多语言多模态预训练模型。

2.2 模型介绍

我们设计的MLMM模型的整体结构如图5所示。我们首先使用在Visual Genome数据集上预训练的Faster R-CNN目标检测器来提取图像中的区域特征，并将这些特征与相应的多语言文本Token一同输入到Transformer Encoder中。

为了捕获不同层次的视觉与语言特征，MLMM采用4个任务进行预训练：

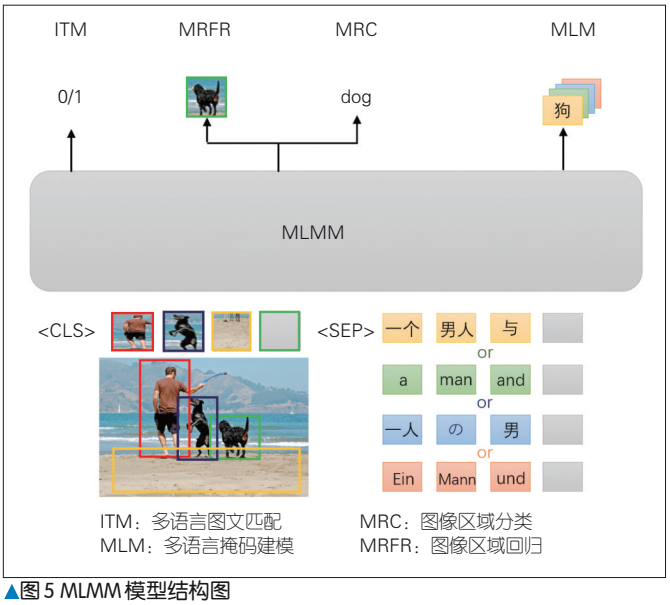
（1）ITM。为了建模图像与多语言文本的全局语义信息，我们使用ITM任务对MLMM模型进行预训练。该任务的目标是，判断输入的图像和多语言文本是否是语义匹配的。在ITM任务中，模型需要理解输入图像和多语言文本的全局语义信息，进而做出判断。

（2）MLM。我们采用MLM任务来建模多语言文本的细粒度语义信息。MLM的目标是根据图像区域信息和文本上下文，让模型来预测被遮蔽的多语言文本单词。

（3）图像区域回归（MRFR）。为了增强模型对图像的细致建模能力，MRFR任务要求模型根据文本和其他图像区域还原被遮蔽的图像区域特征。

（4）图像区域分类（MRC）。为了让模型能够细粒度地识别图像语义，我们实施了MRC任务，因此让模型来预测被遮蔽图像区域所属类别。虽然数据集中没有区域语义的标注信息，但是目标检测器检测得到的类别可以作为该任务的伪标注。目标检测器预测的类别并不是完美的，我们将目标检测器在目标类别上的分布作为软标签，通过计算MLMM预测分布与目标检测器软标签的KL divergence，来优化整个模型。

▲图4 文澜BriVL对诗句的神经元可视化



▲图5 MLMM模型结构图

我们使用的多语言多模态预训练数据集涵盖汉语、英语、德语、法语、捷克语、日语、韩语7种语言和与语义相匹配的图像，包含2.1亿对多语言图文数据。该数据集在以下两个数据集的基础上通过机器翻译进行构建：

（1）英文图文数据集Conceptual Caption 3M+12M。该数据集是目前图文预训练的通用数据集，约有1 500万图文对。数据集中的文本具体描述了图像中所包含的内容。针对该数据集，我们采用4种预训练任务进行训练。

（2）中文图文数据集RUC-CAS-WenLan。该数据集是我们构建的，涵盖新闻、百科、微博、微信等领域，文本内容与对应的图像呈弱相关关系。我们选取其中的1 500万图文对进行预训练。针对该数据集的特点，我们仅训练ITM任务。

2.3 实验分析

我们在多语言图文检索、多语言视觉问答两个下游任务中进行了实验，以验证MLMM的多语言多模态能力。

（1）下游任务1：多语言图文检索

多语言图文检索任务为：给定一段多语言文本，模型可以从数据库中找到与之语义最相关的一张图像，或通过一张图片找到与之最相关的多语言文本。对于多语言图文检索，我们在两个常用的多语言图文数据集 Multi30K 和 MSCOCO 上进行评测。Multi30K 是英文图文数据集 Flickr30K 的扩展，支持英语、德语、法语和捷克语 4 种语言；文献[19–20]分别将最初的英文 MSCOCO 数据集扩展到中文和日文。通常，多语言图文检索评测包含以下几个设定：

• Finetune on en。只使用英文下游数据对模型进行微调，然后测试模型在其他语言上的表现，以衡量模型在多语言上的扩展性。

• Finetune on each。使用多种语言的下游数据，分别对预训练模型进行微调，以衡量模型的单语言能力

• Finetune on all。同时使用多种语言的下游数据对一个预训练模型进行微调，以衡量模型的多语言容量。

与 M3P 和 UC2 相同，我们采用平均召回率，即图像检索文本、文本检索图像两个检索方向上的 Recall@1、5、10 的平均值，来衡量模型的检索效果。3 种微调设定下的实验结果如表 3 所示。

从表 3 中可以看出，在 3 种设定上，MLMM 都超过了现有最好的多语言预训练模型 M3P 和 UC2，达到当前最佳性能。尤其在英文上进行微调时，英文与其他语言之间的性能差距明显小于现有的工作中两者间的性能差距。这说明得益于更大规模的预训练，MLMM 能够表现出很强的跨语言迁移能力。

（2）下游任务 2：多语言视觉问答

给定一张图像和一个与图像内容相关的特定语言上的提问，多语言视觉问答任务要求模型能够给出正确的答案。我们采用 VQA 2.0 和 VQA VG Japanese 两个数据集进行多语言视觉问答的实验。其中，VQA 2.0 是英文视觉问答数据集，而 VQA VG JA 则是日文视觉问答数据集。与 UC2 相同，MLMM 将视觉问答任务视为多标签分类任务，即模型从一个固定的候选池中选择问题的答案。对于 VQA 2.0 数据集，我们选择最常见的 3 129 个回答作为答案候选池；对于 VQA VG Japanese，我们选择最常见的 3 000 个回答作为答案候选池。表 4 展示了 MLMM 在多语言视觉回答上的实验结果。

从表 4 中可以看出，MLMM 在多语言图文检索上超越了目前的预训练模型，在两个多语言视觉问答数据集上同样表现出色。这验证了通过大规模的预训练，MLMM 能够轻松适配各种多语言多模态的下游任务。

▼表 3 多语言图文检索平均召回率

数据集		Multi30K				MSCOCO		
微调策略	评测语言 模型	En	De	Fr	Cs	En	Zh	Ja
En	M3P	87.4	58.8	46.0	36.8	88.6	53.8	56.0
	UC2	87.2	74.9	74.0	67.9	88.1	82.0	71.7
	MLMM	91.9	86.7	86.9	85.6	90.6	90.3	86.6
Each	M3P	87.4	82.1	67.3	65.0	88.6	75.8	80.1
	UC2	87.2	83.8	77.6	74.2	88.1	84.9	87.3
	MLMM	91.9	88.1	85.3	83.8	90.6	89.0	90.9
All	M3P	87.7	82.7	73.9	72.2	88.7	87.9	86.2
	UC2	88.2	84.5	83.9	81.2	88.1	89.8	87.5
	MLMM	92.0	88.7	88.2	87.4	90.8	92.4	91.2

2.4 可视化分析

我们对 MLMM 学习到的跨语言跨模态的通用知识进行了可视化。我们将语义相匹配的多语言文本和图像输入到 MLMM 中，将最后一层 Transformer Encoder 的文本对图像区域的注意力权重进行可视化，如图 6 所示。对于中文和英文相同语义的单词，其注意力权重在图像区域上的分布基本一致。这说明通过大规模的预训练，MLMM 学习到了多语言单词之间以及和图像区域之间的语义对应关系。

3 超大规模多模态预训练模型带来的影响

3.1 多模态信息对文本编码的影响

当图像信息通过文澜预训练模型影响文本编码时，到底发生了怎样的改变？给定一个词，我们将该词在两个空间中的 K 近邻的词集合分别表示为 N_w^1 和 N_w^2，然后用笛卡尔相似度来计算该词在两个空间表示的相似性：

▼表 4 多语言视觉问答准确率

数据集 模型	VQA 2.0 test-dev	VQA VG Japanese
UNITER	71.22	22.70
UC2	71.48	34.20
MLMM	73.21	35.40

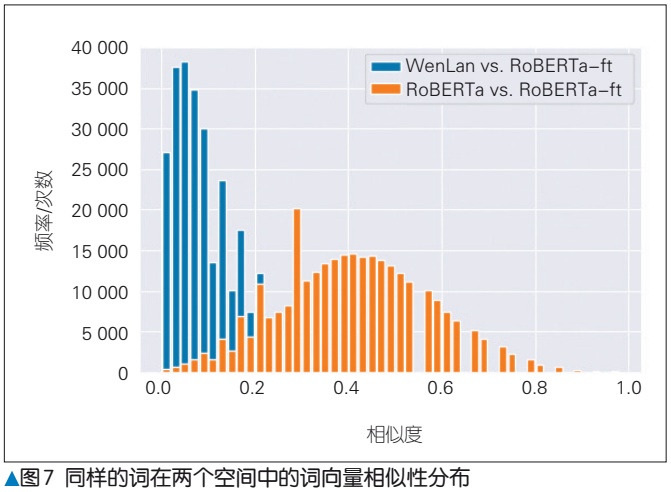
▲图 6 MLMM 模型在多语言图文检索中的注意力权重可视化

$$\text{Jaccard Similarity} = \frac{\left|N_w^1 \cap N_w^2\right|}{\left|N_w^1 \cup N_w^2\right|}。$$

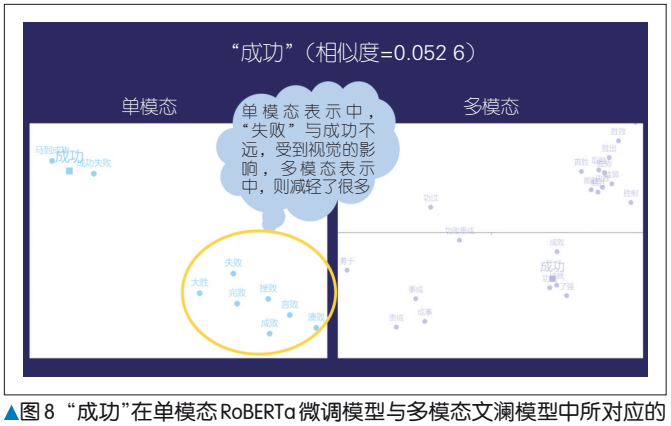
公平起见，哈尔滨工业大学的车万翔老师团队使用文澜的图文训练集中的所有文字，对RoBERTa进行了微调。在17万的词表上进行统计的结果如图7所示。和微调后的RoBERTa相比，RoBERTa看上去是一个相似度均值在0.4附近的正态分布；但和微调后的RoBERTa相比，WenLan的相似度明显变低，大部分样本集中在0.1以下。这说明图像对文本词向量有着显著的影响。

我们在查看了相似度较低的词语后发现了一些共同点：

（1）如图8所示，在单模态语言模型中，由于上下文类似，反义词的词嵌入向量经常会非常相似。例如，在图8的左部分中，当RoBERTa微调后，离“成功”不远的地方有一组与“失败”相关的词语；经过文澜多模态预训练，“成功”周围则以“成功”为主了（如图8右部分所示）。这可能是因为与“成功”和“失败”相关联的图像在色调和内容上相差较大。



▲图7 同样的词在两个空间中的词向量相似性分布



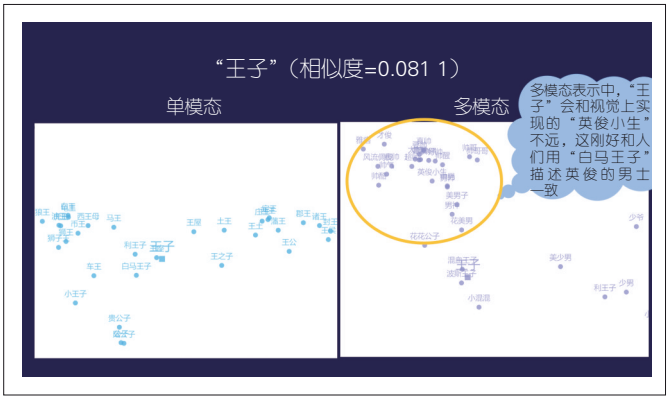
▲图8 “成功”在单模态RoBERTa微调模型与多模态文澜模型中所对应的空间上的邻近词语

（2）视觉上相似的词语会被拉近距离。以图9为例，RoBERTa微调模型会把“王子”与“王公”“狮子王”“贵公子”等语义上比较相近的词语拉近。多模态预训练模型会将“王子”和“美男子”“帅哥”“英俊小生”等词语拉近。这些概念在人们的印象中确实有很强的视觉语义相关性。

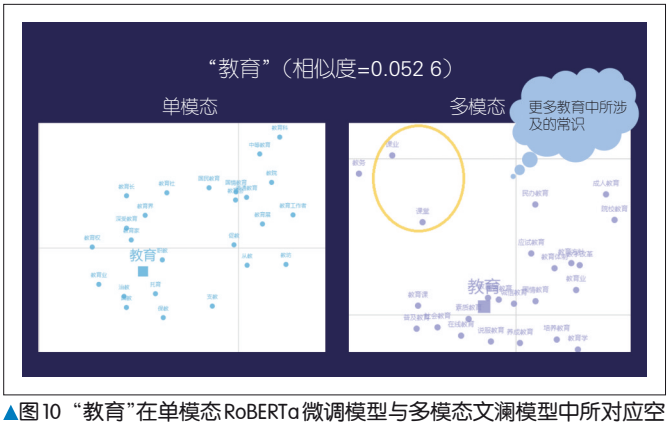
（3）同一情境的词语被拉近。如图10所示，RoBERTa微调模型通常会找到和“教育”同层次的近义词语，如“保育”“国民教育”“教育界”等；文澜模型则会找到一些“课业”“课堂”等词语，这些词语可能出现在类似的图片周围，并通过跨模态之间的对比学习拉近距离。

3.2 多模态预训练对图像生成的影响

基于单模态预训练生成模型的主要问题是，输入句子嵌入是由在单一模态中预先训练的文本编码器提取的，这在语义上与图像模态不一致。因此，单模态预训练生成模型需要学习、处理视觉和自然语言的不同统计特性，以便生成与给定文本对齐的真实图像。为此，现有方法采用了对比学习，并仔细设计了基于注意的单词和区域自我调节，以便更好地进行训练，这种方式是相当耗时的。在跨模态生成中（如文



▲图9 “王子”在单模态RoBERTa微调模型与多模态文澜模型中所对应的空间上的邻近词语



▲图10 “教育”在单模态RoBERTa微调模型与多模态文澜模型中所对应空间上的邻近词语

本生成图像），高效地弥补这两种模态之间的差距非常具有挑战性。

与以往方法不同，我们可以利用多模态预训练模型对图像和文本进行编码。例如，借助VQGAN inversion，可以实现基于文澜BriVL的文生成图。具体地，给定一个文本，输入一张随机噪声图像，通过文澜BriVL的文本编码器就可以得到文本的特征表示。VQGAN inversion的目标函数为：当前输入图像经过VQGAN后输出的图像，其视觉特征（通过文澜图像编码器得到）必须逼近输入文本的特征。固定VQGAN和文澜模型的所有参数，通过反向传播可以更新输入的噪声图像。算法收敛后，最终得到的图像即可看作关于给定文本的文生成图结果。如图11所示，借助VQGAN，文澜BriVL模型能够生成更贴近自然的图像。

这里的关键之处在于，由多模态预训练模型提取的文本嵌入可以自然地与图像模态对齐，这避免了之前方法中的额外复杂架构。总之，多模态预训练模型给文生成图任务带来了新的研究思路。

3.3 多模态预训练对文本-图像检索的影响

当文澜模型将图像和文本映射到同一空间时，文本与图像的互检就变得非常容易。当文本检索图像时，不再需要图像周围的文字作为桥梁，因此文澜模型可以匹配图像周围文字并没有描述的意境。图像检索文本也成为可能，不仅能识别出物体、场景或情感等类别标签，还可以和任意的句子、段落进行多模态共享语义空间上的匹配。这首次跨越了图文的语义鸿沟，实现了真正的跨模态检索。

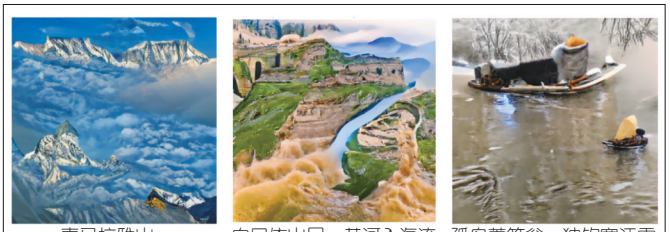
基于文澜BriVL模型，文澜团队实现了多个在线演示系统，具体见图12。

4 结束语

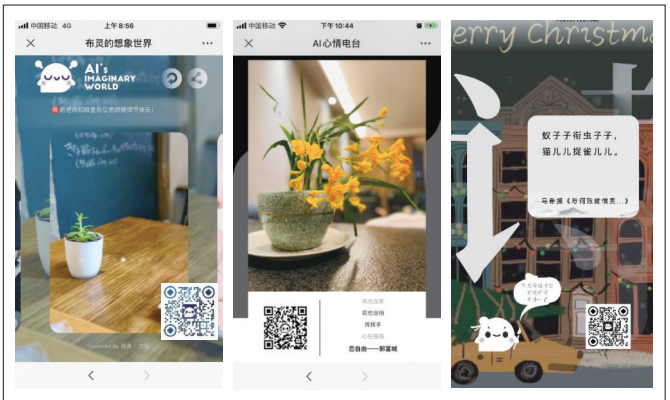
我们尝试了利用亿级的、来自互联网的图文对数据来训练多模态双塔模型BriVL和多语言多模态单塔模型MLMM。这两个预训练模型均在多个下游任务中获得了国际最佳性能。通过实验，我们发现多模态预训练模型将更多视觉相似或在同一场景中的词语拉近；能为文生成图提供统一的语义基础，提升图像生成的泛化能力和效果；能让文字和图像可以在映射到同一空间后实现真正的跨模态检索。目前，文澜BriVL 1.0已开源，可以通过以下网址访问或者申请下载：

• 文澜BriVL 1.0源码下载：https://github.com/BAAI-WuDao/BriVL

• 文澜BriVL 1.0模型申请：https://wudaoai.cn/model/detail/BriVL



▲图11 借助VQGAN inversion得到的文澜文生成图结果



（a）《布灵的想象世界》：可以用任意关键字、句子、段落和文章检索语义相关的图像

（b）《AI心情电台》：可以用任意图片检索意境相关的歌词

（c）《布灵的诗歌小馆》：可以用任意图片检索意境相关的古诗

▲图12 基于文澜模型开发的3款跨模态检索小应用

• 文澜BriVL 1.0在线API：https://github.com/chuhaojin/WenLan-api-document

自2021年3月发布以来，文澜受到了腾讯、酷我音乐、爱奇艺、网易等多家企业的关注。与长城汽车合作，文澜完成了由图像检索金句的“欧拉喵语”小应用，并在上海和成都车展以及ChinaJoy上与参观者进行现场的品牌互动；与OPPO合作，文澜模型实现了为视障人士读取收集图片的功能，践行科技向善的理念。

文澜模型的强大能力也产生了一些跨学科研究成果。由中国人民大学新闻学院和高瓴人工智能学院合作的《空间漫游与想象生产——线上影像策展中的网红城市建构：基于视觉·语言多模态预训练模型的计算传播研究》，获得了2021年计算传播学年会学生论文三等奖。中国人民大学艺术学院师生与上海大学教师组成的“云端艺术”团队，将文澜融合到他们的微信程序“红色夏天智能航宇”作品中，获得2021年上海图书馆开放数据竞赛优秀设计奖。

最后，如何平衡单双塔的有效性和效率是未来的重要问题，目前主要方法有两种：（1）对于单塔模型，可以在跨模式融合模块之前放置双塔体系结构，以减少巨大的检索延迟，同时尽可能保持高性能优势；（2）对于双塔模式，可以

考虑建立更精细/更紧密的模式相关性的学习目标，以提高其性能，同时保持高效率的优势。

参考文献

[1] NI M, HUANG H, SU L, et al. Learning universal representations via multitask multilingual multimodal pre-training [EB/OL]. [2022-01-12]. https://paperswithcode.com/paper/m3p-learning-universal-representations-via

[2] ZHOU M, ZHOU L, WANG S, et al. UC2: universal cross-lingual cross-modal vision-and-language pre-training [EB/OL]. [2022-01-10]. https://paperswithcode.com/paper/uc2-universal-cross-lingual-cross-modal

[3] FEI H, YU T, LI P. Cross-lingual cross-modal pretraining for multimodal retrieval [EB/OL]. [2022-01-10]. https://paperswithcode. com/paper/cross-lingual-cross-modal-pretraining-for

[4] HUANG H, LIANG Y, DUAN N, et al. Unicoder: a universal language encoder by pre-training with multiple cross-lingual tasks [EB/OL]. [2022-01-10]. https://paperswithcode.com/paper/unicoder-a-universal-language-encoder-by-pre

[5] LI X, XU C, WANG X, et al. COCO-CN for cross-lingual image tagging, captioning, and retrieval [EB/OL]. [2022-01-12]. https://paperswithcode.com/paper/coco-cn-for-cross-lingual-image-tagging

[6] YOSHIKAWA Y, SHIGETO Y, TAKEUCHI A. Stair captions: Constructing a large-scale Japanese image caption dataset [EB/OL]. [2022-01-10]. https://paperswithcode.com/paper/stair-captions-constructing-a-large-scale

[7] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2022-01-10]. https://paperswithcode.com/paper/bert-pre-training-of-deep-bidirectional

[8] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. [2022-01-10]. https://paperswithcode. com/paper/language-models-are-unsupervised-multitask

[9] DOSOVITSKIY A , BEYER L, KOLESNIKOV W, et al. An image is worth 16x16 words: transformers for image recognition at scale [EB/OL]. [2022-01-10]. https://paperswithcode. com/paper/an-image-is-worth-16x16-words-transformers-1

[10] CHEN Y, LI L, YU L, et al. Uniter: universal image-text representation learning [EB/OL]. [2022-01-10]. https://paperswithcode.com/paper/uniter-learning-universal-image-text-1

[11] LI X, YIN X, LI C, et al. Oscar: object-semantics aligned pre-training for vision-language tasks [EB/OL]. [2022-01-12]. https://paperswithcode.com/paper/oscar-object-semantics-aligned-pre-training

[12] LIN J, MEN R, YANG A, LIN J, et al. M6: A Chinese multimodal pretrainer [EB/OL]. [2022-01-12]. https://paperswithcode. com/paper/oscar-object-semantics-aligned-pre-training

[13] LI L, YATSKAR M, YIN D, et al. Visualbert: a simple and performant baseline for vision and language [EB/OL]. [2022-01-10]. https://paperswithcode.com/paper/oscar-object-semantics-aligned-pre-training

[14] LI G, DUAN N, FANG Y, et al. Unicoder-vl: a universal encoder for vision and language by cross-modal pre-training [EB/OL]. [2022-01-10]. https://paperswithcode.com/paper/unicoder-vl-a-universal-encoder-for-vision

[15] SU W, ZHU X, GAO Y, et al. Vl-bert: pre-training of generic visual-linguistic representations [EB/OL]. [2022-01-10]. https://paperswithcode.com/paper/unicoder-vl-a-universal-encoder-for-vision

[16] SUN S, CHEN Y, LI L, et al. Lightningdot: pre-training visual-semantic embeddings for real-time image-text retrieval [EB/OL]. [2022-01-10]. https://paperswithcode.com/paper/lightningdot-pre-training-visual-semantic

[17] RADFOR A, KIM J, HALLACY C, et al. Learning transferable visual models from natural language supervision [EB/OL]. [2022-01-10]. https://paperswithcode.com/paper/lightningdot-pre-training-visual-semantic

[18] JIA C, YANG Y, XIA Y. Scaling up visual and vision-language representation learning with noisy text supervision [EB/OL]. [2022-01-10]. https://paperswithcode. com/paper/lightningdot-pre-training-visual-semantic

[19] HUO Y, ZHANG M, LIU G, et al. Wenlan: bridging vision and language by large-scale multi-modal pre-training [EB/OL]. [2022-01-10]. https://paperswithcode.com/paper/wenlan-bridging-vision-and-language-by-large

[20] FEI F , LU Z, GAO Y, et al. Wenlan 2.0: make ai imagine via a multimodal foundation model [EB/OL]. [2022-01-10]. https://paperswithcode. com/paper/wenlan-bridging-vision-and-language-by-large

作 者 简 介

卢志武，中国人民大学高瓴人工智能学院教授、博士生导师；主要研究方向为机器学习、计算机视觉等；设计了首个公开的中文通用图文预训练模型文澜BriVL；发表论文90余篇。

金琴，中国人民大学信息学院教授、博士生导师；主要研究方向为多媒体智能计算、人机交互；在多媒体情感计算、视觉描述生成、跨模态交互等研究与应用中取得了突出成果，蝉联多项国际权威竞赛冠军，包括：2017—2021年TRECVID视频描述（VTT）评测冠军、2018—2020年CVPR ActivityNet Dense Video Captioning竞赛冠军、2017—2019年ACM Multimedia Audio-Visual Emotion Challenge（AVEC）竞赛冠军，获得2019年之江杯全球人工智能大赛视频内容描述生成冠军；发表论文100余篇。

宋睿华，中国人民大学高瓴人工智能学院长聘副教授，曾任微软亚洲研究院主管研究员和微软小冰首席科学家，担任SIGIR 2021短文的主席、EMNLP 2021的资深区域主席、《Information Retrieval Journal》的主编；提出的算法完成了人类史上第一本人工智能创作的诗集《阳光失了玻璃窗》，参与完成了首个公开的中文通用图文预训练模型文澜BriVL；发表论文90余篇，申请国际专利25项。

文继荣，中国人民大学长聘教授，现任高瓴人工智能学院执行院长和信息学院院长、大数据管理与分析方法研究北京市重点实验室主任，并担任北京智源人工智能研究院首席科学家、北京市第十三届政协委员、中央统战部党外知识分子建言献策专家组专家，入选首批“北京高校卓越青年科学家计划项目”，还担任AIRS 2016大会名誉主席、CCIR 2017大会主席、SIGIR 2018领域主席、SIGIR 2020程序委员会主席、WWW 2021领域主席、《ACM Transactions on Information Systems》《IEEE Transactions on Knowledge and Data Engineering》的编委；长期从事大数据和人工智能领域的研究；发表论文200余篇。