Facebook 推出多模态通用模型 FLAVA，吊打 CLIP 平均十个点！

原创 子龙 夕小瑶的卖萌屋 2022-01-05 12:05

文 | 子龙
编 | 智商掉了一地

> **厉害了！作者将单一模型运用于三个不同领域的不同任务，结构简单且训练直观，还能有看出色的表现。**

自Transformer横空出世，从NLP到CV，再到今天的多模态，无数基于Transformer的模型被应用于各类任务，似乎真的印证了当年文章的标题"Transformer is **ALL** you need"。然而，纯粹的NLP任务有BERT，RoBERTa；CV任务有ViT，多模态任务又有VLBERT，OSCAR，虽然都是基于Transformer的结构，但显然都是针对不同任务设计不同模型。那么"万能"的Transformer能否构建出一个统治各类任务的模型，实现真的一个模型解决所有问题呢？

今天文章的作者就关注到了当前各个模型的局限，提出了一个适用于**NLP+CV+多模态**的模型FLAVA，可运用于三种领域共计35个任务，且都有着出色的表现。

论文题目：

FLAVA: A Foundational Language And Vision Alignment Model

论文链接：

https://arxiv.org/abs/2112.04482

介绍

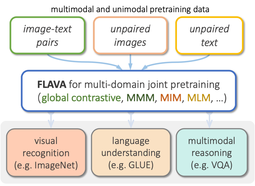
文章标题中，作者称模型为"Foundational"，他们不希望借助各种奇技淫巧的Tricks。而是通过尽可能简单的结构，配合直观的的训练手段，达到涵盖NLP、CV、多模态的目的。

FLAVA基于三种不同的输入：

- 匹配的图片-文本
- 单独文本
- 单独图片

解决三个领域的问题：

- NLP：语言理解（到GLUE）
- CV：视觉识别（到ImageNet）
- 多模态：多模态解释（到VQA）



图片编码器(Image Encoder)

FLAVA**直接使用现有模型ViT的结构，同时也照搬ViT的处理方法，分割图片进行编码**，在ViT输出的隐状态上，FLAVA利用单一模态数据集中的图片进行Masked Image Modeling。首先，利用dVAE将图片转化为类似词向量的token：再参照BEiT，对masked隐状态进行分类，即利用周围图片分块，预测masked的图片属于dVAE划分的哪一类，这样在图片上也可以做BERT那样很mask modeling。

文本编码器(Text Encoder)

FLAVA在文本部分多处理就相对简单，作者采取常见的Masked Language Modeling。对一部分masked token进行预测，和其他当对区别在于，FLAVA**没有采用BERT之流将文本语言模型的结构**，而是和图片编码器一样，使用了ViT的结构，不过因为是不同的模态，自然采用了不同的模型参数。

多模态编码器(Multimodal Encoder)

在图片编码器和文本编码器之上，FLAVA**添加了一层多模态编码器做模态融合**，多模态编码器将前两者输出的隐藏状态作为输入，同样利用ViT的模型结构进行融合。

多模态预训练

在文本编码器和图片编码器中，FLAVA在单一模态上进行了预训练，在多模态预训练方面，FLAVA使用了三种多模态预训练任务：

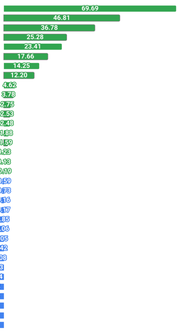
- 对比学习：FLAVA利用图片编码器和文本编码器的隐藏状态，增大相匹配的图片-文本对之间的余弦相似度，减小非匹配的图片-文本对之间的余弦相似度。
- Masked Multimodal Modeling：与图片编码器上的MIM类似，只不过改为利用多模态编码器的隐状态进行预测。
- 图片-文本匹配：与许多现有模型一样，FLAVA利用多模态编码器的[CLS]的隐状态，识别当前图片与文本是否匹配。

效果

从上述模型细节可以看出，无论是模型结构，还是预训练任务，文本与图片之间高度对称，同时也设计也十分直观。接下来看看在35个任务上的表现。

		MIM	MLM	FLAVA	FLAVA/Base	FLAVA w/o text	FLAVA	CLIP	CLIP
		1	2	3	4	5	6	7	8
Datasets	Eval method	PMD	PMD	PMD	PMD	(PMD+IN-1k+CC[Neo+RC)	PMD	PMD	400M [R21]
MNLI [11]	fine-tuning	-	75.25	76.98	76.82	78.06	**80.33**	52.85	10.92
QQ [12]	fine-tuning	-	39.55	17.58	38.97	44.22	**78.65**	11.02	25.37
QNLI [13]	fine-tuning	-	54.70	50.55	54.67	57.56	**84.26**	50.43	49.67
SQUF [14]	fine-tuning	-	86.08	85.56	86.49	88.02	**88.74**	59.17	65.93
SST [15]	fine-tuning	-	87.96	86.47	89.33	90.14	**90.94**	83.49	88.19
QNLI [16]	fine-tuning	-	82.52	75.85	84.77	86.40	**87.31**	49.48	50.34
RTE [2,3,7,8,9]	fine-tuning	-	56.94	53.98	51.99	54.87	**87.36**	53.07	55.28
MRPC [10]	fine-tuning	-	84.80	79.41	84.96	86.27	**91.18**	76.47	78.92
NLP Avg.	-	-	71.55	64.81	74.22	75.55	**78.10**	46.44	50.50
ImageNet [5+6]	linear eval	41.79	-	74.09	74.34	73.49	**75.54**	72.95	80.20
Food101 [1]	linear eval	35.30	-	87.77	87.33	87.59	**88.51**	85.69	91.26
CIFAR100 [4]	linear eval	76.20	-	**93.44**	92.37	92.65	92.87	91.25	94.93
CIFAR10 [4]	linear eval	87.08	-	**97.79**	97.53	97.75	97.77	96.93	98.29
Cars [18]	linear eval	16.71	-	75.53	72.07	68.81	70.81	62.84	85.82
Aircraft [19]	linear eval	13.83	-	**49.74**	44.90	44.73	47.51	40.02	52.40
DTD [20]	linear eval	70.20	-	**79.06**	76.81	75.80	77.27	74.03	78.46
Pets [21]	linear eval	84.32	-	**93.348**	91.51	91.07	91.85	89.62	92.65
Caltech101 [22]	linear eval	87.36	-	94.91	95.32	94.95	**95.74**	93.76	95.51
Flowers102 [23]	linear eval	67.25	-	96.36	**96.29**	95.58	96.17	94.94	97.12
MNIST [24]	linear eval	96.40	-	98.39	98.58	**98.70**	98.41	97.98	99.01
EuroSAT [25]	linear eval	94.07	-	97.67	97.42	97.09	**97.78**	97.99	98.67
EuroSAT [11:1]	linear eval	95.40	-	97.60	96.94	97.04	**97.26**	95.70	97.16
GTSRB [26]	linear eval	63.14	-	78.82	77.93	77.71	**79.44**	76.34	88.81
TSRC [27]	linear eval	64.02	-	85.74	84.07	83.86	**86.06**	83.41	92.67
PCAM [14:1]	linear eval	85.10	-	85.02	85.71	**85.72**	85.31	83.99	85.73
UCF101 [28]	linear eval	46.34	-	82.69	81.90	81.42	**83.52**	77.85	85.27
HAT [29]	linear eval	79.65	-	84.69	84.07	83.62	**85.29**	82.05	87.62
FSK 2K [13:1]	linear eval	50.98	-	59.36	60.67	58.59	63.32	57.04	68.26
SUN397 [30]	linear eval	52.45	-	81.27	81.41	81.05	**82.17**	79.96	82.05
SST [31]	linear eval	57.77	-	94.67	**95.35**	94.40	95.13	58.84	70.68
Country211 [17]	linear eval	8.47	-	35.27	36.75	37.01	**38.81**	25.12	36.10
Vision Avg.	-	57.40	-	78.14	79.35	78.29	**79.44**	76.12	82.57
VQAv2 [32]	fine-tuning	-	-	67.13	72.69	71.29	**72.49**	59.81	54.83
VCR [33]	fine-tuning	-	-	54.19	65.44	64.65	**65.659**	52.73	52.04
Intelnl Memes [34]	fine-tuning	-	-	55.94	70.72	**77.45**	76.09	56.59	63.93
Flickr30K [35] TR R@1	zero-shot	-	-	68.30	**89.30**	84.50	67.70	60.90	82.20
Flickr30K [35] TR R@5	zero-shot	-	-	90.70	**98.04**	96.70	94.00	85.40	96.00
Flickr30K [35] IR R@1	zero-shot	-	-	60.76	83.19	80.04	**85.22**	56.48	62.09
Flickr30K [35] IR R@5	zero-shot	-	-	86.68	87.70	86.46	**95.35**	83.60	85.68
COCO [36] TR R@1	zero-shot	-	-	44.36	76.76	74.77	74.09	58.20	79.44
COCO [36] TR R@5	zero-shot	-	-	73.83	**93.74**	72.84	**96.76**	69.68	70.68
COCO [36] IR R@1	zero-shot	-	-	27.59	**38.06**	34.95	38.38	33.29	55.07
COCO [36] IR R@5	zero-shot	-	-	67.28	**67.08**	64.63	67.47	62.47	38.37
Multimodal Avg.	-	-	-	66.25	69.11	67.32	**69.92**	62.02	67.29
Shares Avg.	19.15	23.87	70.06	74.23	73.72	**75.05**	74.52	61.52	66.78

图中**三到线**表示最优结果，**加粗**表示在分开数据集上训练的最优结果。



从各个任务平均上看，FLAVA能够取得整体上的最优结果。多模态任务平均比CLIP高出2个百分点左右，整体平均比CLIP高出10个百分点左右，从具体任务上看，在不少任务上都取得了十分显著的提高，如STS-B数据集提高了69.69，MNLI数据集提高了46.81。

小结

不同于现有模型，FLAVA最大的特点，也可以说是创新点，在于作者实现了**将单一模型运用于三个不同领域的不同任务，而且都有着不错的效果**。虽然FLAVA并没有在涵盖在所有任务上都达到SOTA，但是整体性能上并不弱于现有模型，同时有着更广阔的运用场景。模型设计也没有各种奇技淫巧，这对未来研究通用模型有着很大的启发。

萌屋作者：子龙(Ryan)
本科毕业于北大计算机系，曾混迹于商汤和MSRA，现在是可在US(Goose Geese)的在读PhD，主要关注多模态中的NLP和data mining，也在探索更多有意思的topic，期本只是想公众号的在线群众，被自种有意思的推送吸引就会沉睡物，希望借此达动小屋的灵气，paper++，早日成为有编的程序员！
作品推荐：
1. 别再挑剑文了！多模态推理解更被时代需要！
2. Transformer要要强？Google被挑衅优良！
3. 预训练语言真的是世界模型？