FLAVA: A Foundational Language And Vision Alignment Model.

Ariel Reyes Pardo[1]

[1]Pontificia Universidad Católica de Chile. Facultad de Matemáticas-Ingeniería.

6 de septiembre 2023

Contenido

FLAVA: A Foundational Language And Vision Alignment Model

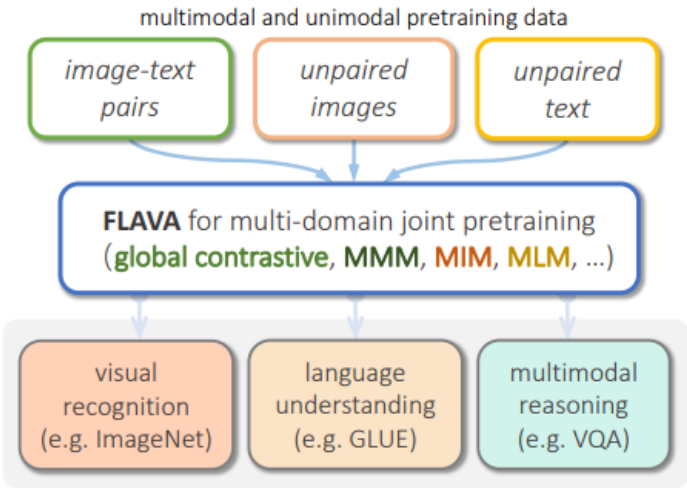
- Presentado en 2022 en CVPR (Conference on Computer Vision and Pattern Recognition)

Flava: A foundational language and vision alignment model
A Singh, R Hu, V Goswami... - Proceedings of the ..., 2022 - openaccess.thecvf.com
... We introduce **FLAVA** as such a model and demonstrate ... **FLAVA** combines dual and fusion encoder approaches into one holistic model that can be pretrained with our novel **FLAVA** ...
☆ Guardar 99 Citar Citado por 244 Artículos relacionados Las 7 versiones »

Introducción

- La supervisión en lenguaje natural puede conducir hacia alta calidad en los modelos visuales. (CLIP)

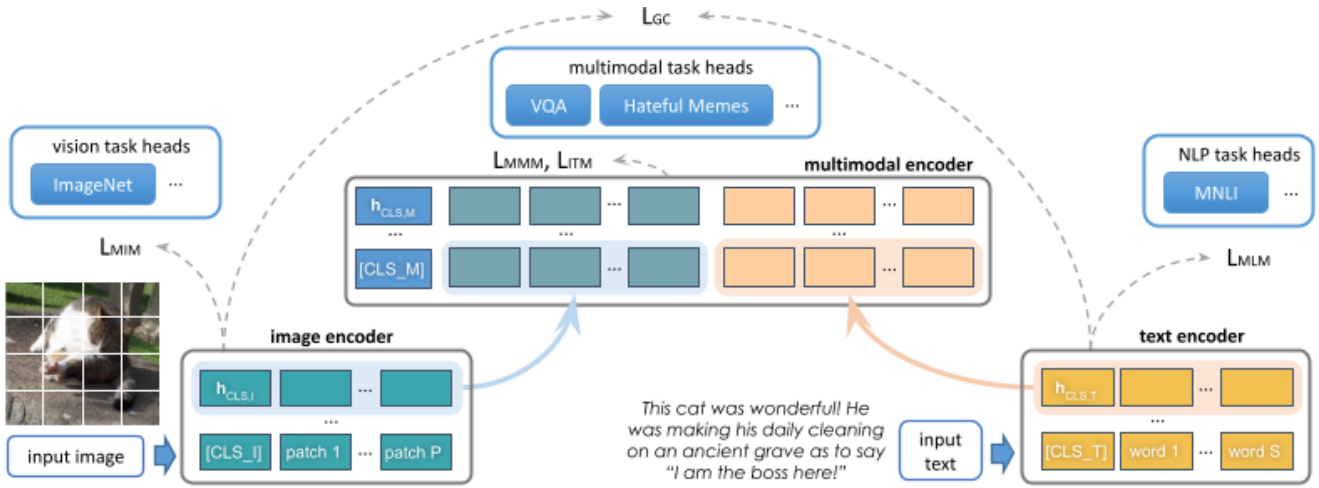
- Unimodalidad
  - Visión
  - Lenguaje
- Multimodalidad
  - Visual-Language Pretraining (VLP)



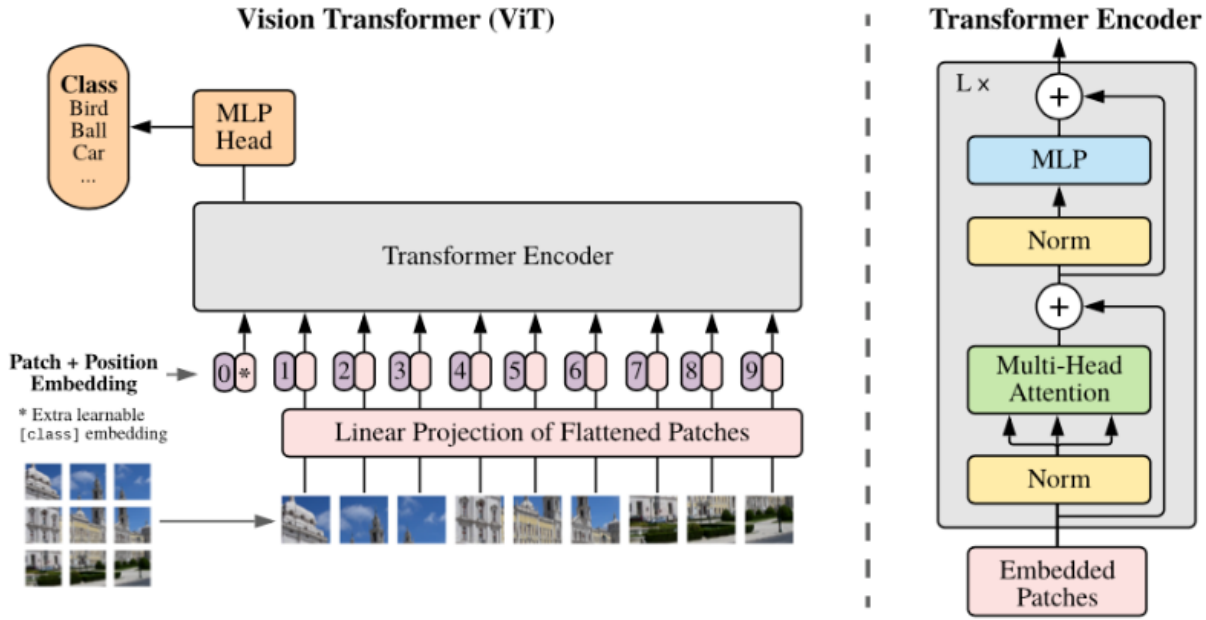
Contenido

Arquitectura

- Unimodalidad
  - Encoder imagen $h_{CLS,I}$
  - Encoder texto $h_{CLS,T}$
- Multimodalidad
  - Encoder fusión $h_{CLS,M}$



ViT encoder


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
Transformer Encoder
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

Contenido

Pretraining Tasks

- FLAVA está diseñado para ser capaz de tomar ventaja de datos unimodales en conjunto con datos pareados. Resulta un modelo que puede manejar tareas unimodales y como tareas de visión y lenguaje multimodales.
- Multimodal pretraining
  - Global contrastive (GC) loss (**CLIP**)
  - Masked multimodal modeling (MMM) (**BEiT**)
- Unimodal pretraining
  - Masked image modeling (MIM) (**BEiT**)
  - Masked language modeling (MLM) (**BERT**)

Contenido

Dataset

- Unimodal
  - ImageNet
  - CCNews
- Multimodal
  - Public Multimodal Datasets (PMD)

	#Image-Text Pairs	Avg. text length
COCO [66]	0.9M	12.4
SBU Captions [77]	1.0M	12.1
Localized Narratives [82]	1.9M	13.8
Conceptual Captions [92]	3.1M	10.3
Visual Genome [57]	5.4M	5.1
Wikipedia Image Text [99]	4.8M	12.8
Conceptual Captions 12M [14]	11.0M	17.3
Red Caps [27]	11.6M	9.5
YFCC100M [103], filtered	30.3M	12.7
Total	70M	12.1

Table 2. Public Multimodal Datasets (PMD) corpus used in FLAVA multimodal pretraining, which consists of publicly available datasets with a total size of 70M image and text pairs.

Contenido

Comparando con SOTA

	public data		Multimodal Tasks VQAv2	SNLI-VE	HM	Language Tasks CoLA	SST-2	RTE	MRPC	QQP	MNLI	QNLI	STS-B	ImageNet linear eval
1	✓	$\text{BERT}_{\text{base}}$ [28]	–	–	–	54.6	92.5	62.5	81.9/87.6	90.6/87.4	84.4	91.0	88.1	–
2	✗	CLIP-ViT-B/16 [83]	55.3	74.0	63.4	25.4	88.2	55.2	74.9/65.0	76.8/53.9	33.5	50.5	16.0	80.2
3	✗	$\text{SimVLM}_{\text{base}}$ [109]	77.9	84.2	–	46.7	90.9	63.9	75.2/84.4	90.4/87.2	83.4	88.6	–	80.6
4	✓	VisualBERT [63]	70.8	77.3†	74.1‡	38.6	89.4	56.6	71.9/82.1	89.4/86.0	**81.6**	87.0	81.8	–
5	✓	$\text{UNITER}_{\text{base}}$ [16]	72.7	78.3	–	37.4	89.7	55.6	69.3/80.3	89.2/85.7	80.9	86.0	75.3	–
6	✓	$\text{VL-BERT}_{\text{base}}$ [101]	71.2	–	–	38.7	89.8	55.7	70.6/81.8	89.0/85.4	81.2	86.3	82.9	–
7	✓	ViLBERT [70]	70.6	75.7†	74.1‡	36.1	90.4	53.7	69.0/79.4	88.6/85.0	79.9	83.8	77.9	–
8	✓	LXMERT [102]	72.4	–	–	39.0	90.2	57.2	69.7/80.4	75.3/75.3	80.4	84.2	75.3	–
9	✓	UniT [43]	67.0	73.1	–	–	89.3	–	–	90.6/ –	81.5	**88.0**	–	–
10	✓	CLIP-ViT-B/16 (PMD)	59.8	73.5	56.6	11.0	83.5	53.1	63.5/68.7	75.4/43.0	32.9	49.5	13.7	73.0
11	✓	FLAVA (ours)	**72.8**	**79.0**	76.7	50.7	90.9	**57.8**	81.4/86.9	90.4/87.2	80.3	87.3	85.7	**75.5**

- El mejor resultado general entre los enfoques multimodales está subrayado, mientras que la negrita significa el mejor modelo basado en datos públicos.
- FLAVA fue entrenado con 70M de datos mientras que CLIP fue entrenado con casi 6 veces más.
- SimVLM fue entrenado con 1.8B de datos.

Contenido

Conclusiones

- El rendimiento de FLAVA podría mejorar si es que se entrena con mayor cantidad de datos.
- FLAVA fue entrenado en un conjunto de datos de varios órdenes de magnitud más pequeños.
- Exploración de sesgos peligrosos en el dataset de entrenamiento.
- Se señala un camino a seguir hacia modelos generalizados pero abiertos que funcionan bien en un amplia variedad de tareas multimodales.

Fin

Muchas gracias :)