Early-Fusion and End-to-End Training

- FLAVA (CVPR 2022)
- UniT (ICCV 2021)
- Align before Fuse (NeurIPS 2021)

Presenters: Mengyu Yang & Qingyu Xiao

Presenters

Mengyu Yang

- **ML PhD**
- **Advisor**: James Hays
- **Interests**: Multimodal learning with vision, audio, and language

Qingyu Xiao

- **Robotics PhD**
- **Advisor**: Matthew Gombolay
- **Interests**: Agile robotics, robot perceptions

FLAVA: A Foundational Language and Vision Alignment Model

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, Douwe Kiela

CVPR 2022

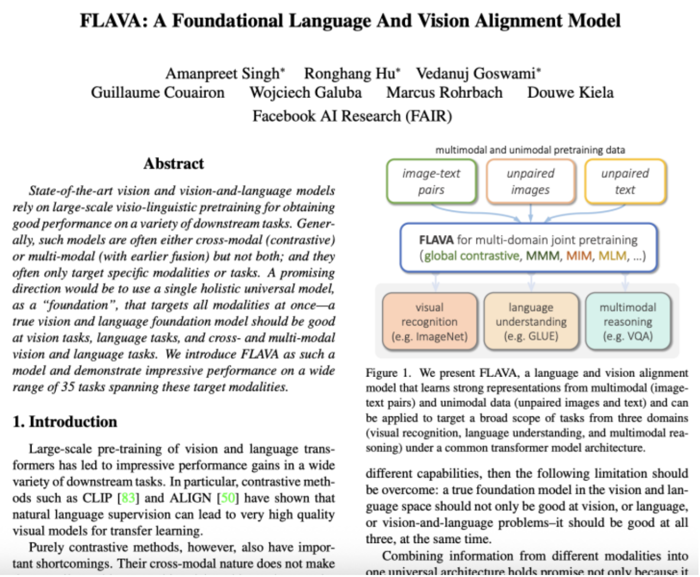
Timeline

- (2018 - 2020) Transformer + pretraining
    - **ImageBERT**: Applying BERT-style masked modeling to image-text
    - **ViLBERT**: Two-stream model for vision and language using cross-attention

- (2020 - 2021) Scaling up and contrastive learning
    - **CLIP**: Contrastive learning at scale (400M image-text pairs)
    - **ALIGN**: Similar to CLIP, but with even more (noisy) web data

Drawbacks of these models

Definition – Domain: combination of modalities used (i.e. vision, language, vision-language)



FLAVA: A Foundational Language And Vision Alignment Model

Amanpreet Singh* Ronghang Hu* Vedanuj Goswami*
Guillaume Couairon Wojciech Galuba Marcus Rohrbach Douwe Kiela
Facebook AI Research (FAIR)

Abstract

State-of-the-art vision and vision-and-language models rely on large-scale visio-linguistic pretraining for obtaining good performance on a variety of downstream tasks. Generally, such models are often either cross-modal (contrastive) or multi-modal (with earlier fusion) but not both; and they often only target specific modalities or tasks. A promising direction would be to use a single holistic universal model, as a "foundation", that targets all modalities at once—a true vision and language foundation model should be good at vision tasks, language tasks, and cross- and multi-modal vision and language tasks. We introduce FLAVA as such a model and demonstrate impressive performance on a wide range of 35 tasks spanning these target modalities.

1. Introduction

Large-scale pre-training of vision and language transformers has led to impressive performance gains in a wide variety of downstream tasks. In particular, contrastive methods such as CLIP [83] and ALIGN [50] have shown that natural language supervision can lead to very high quality visual models for transfer learning.

Purely contrastive methods, however, also have important shortcomings. Their cross-modal nature does not make

Figure 1. We present FLAVA, a language and vision alignment model that learns strong representations from multimodal (image-text pairs) and unimodal data (unpaired images and text) and can be applied to target a broad scope of tasks from three domains (visual recognition, language understanding, and multimodal reasoning) under a common transformer model architecture.

different capabilities, then the following limitation should be overcome: a true foundation model in the vision and language space should not only be good at vision, or language, or vision-and-language problems–it should be good at all three, at the same time.

Combining information from different modalities into one universal architecture holds promise not only because it

- While these vision-language (V&L) models are on the right track, certain shortcomings prevent it from being ***foundational***
- Lack of *domain* and *task* diversity
  - Single domain
  - Only 1 unimodal + V&L domain
  - All domains but small set of tasks
- Some models (e.g. CLIP) trained on proprietary data

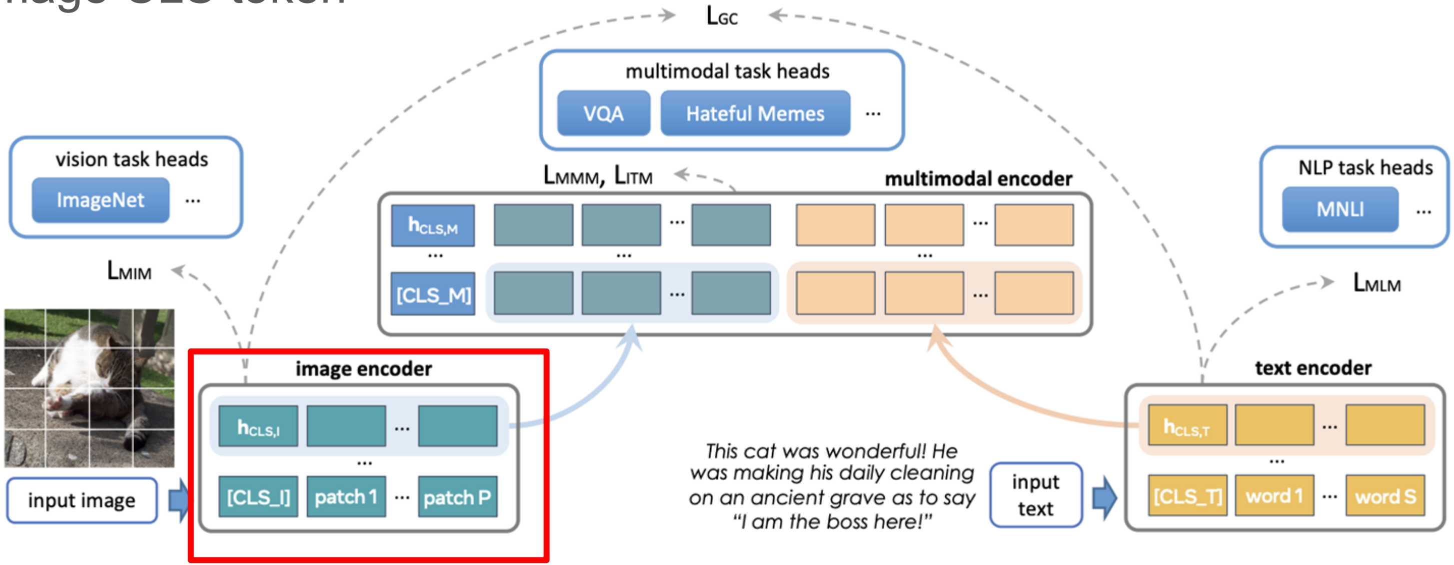
Comparison of V&L models

Method	Multimodal Pretraining data			Pretraining Objectives				Target Modalities			
	public	dataset(s)	size	Contr.	ITM	Masking	Unimodal	V	CV&L	MV&L	L
CLIP [83]	✗	WebImageText	400M	✓	–	–	–	✓	✓	–	–
ALIGN [50]	✗	JFT	1.8B	✓	–	–	–	✓	✓	–	–
SimVLM [109]	✗	JFT	1.8B	–	–	PrefixLM	CLM	*	✓	✓	✓
UniT [43]	–	None	–	–	–	–	–	*	–	✓	✓
VinVL [118]	✓	Combination	9M	✓	–	MLM	–	–	✓	✓	–
ViLT [54]	✓	Combination	10M	–	✓	MLM	–	–	✓	✓	–
ALBEF [62]	✓	Combination	5M	✓	✓	MLM	–	–	✓	✓	–
FLAVA (ours)	✓	PMD (Tbl. 2)	70M	✓	✓	MMM	MLM+MIM	✓	✓	✓	✓

- FLAVA covers unimodal, cross-modal, and multi-modal domains across 35 tasks
- 22 vision-only, 8 language-only, 5 V&L

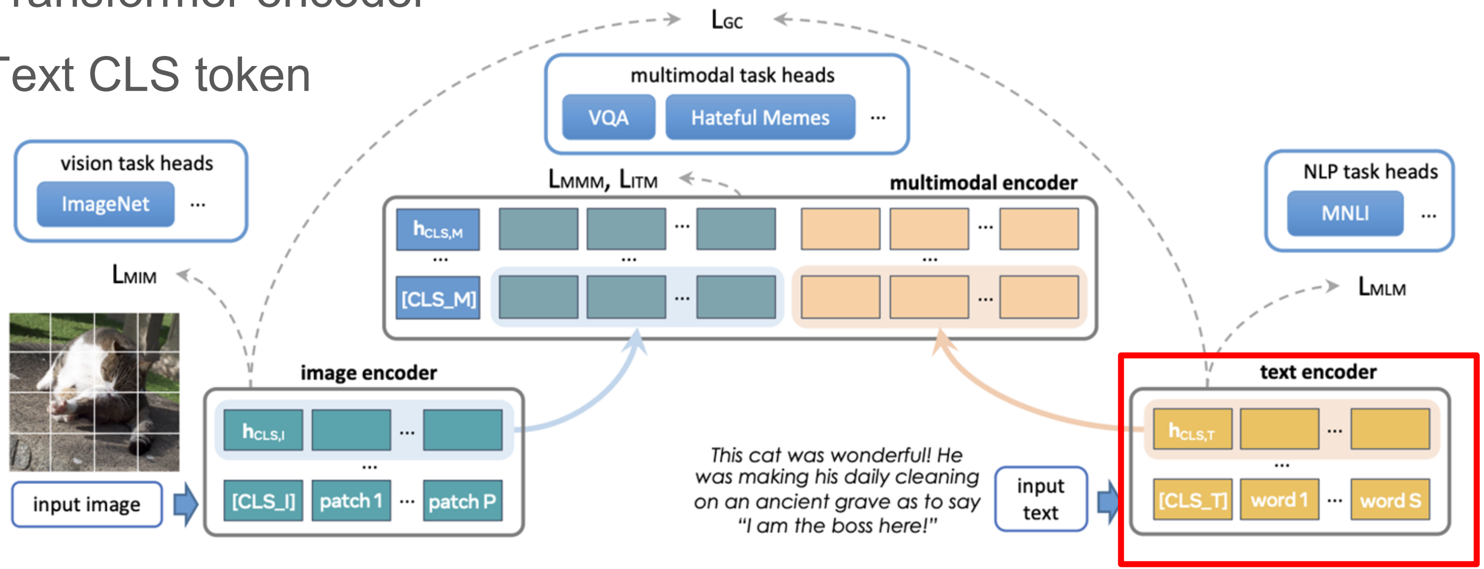
Architecture - Vision encoder

- ViT-B/16 encoder
- Image CLS token



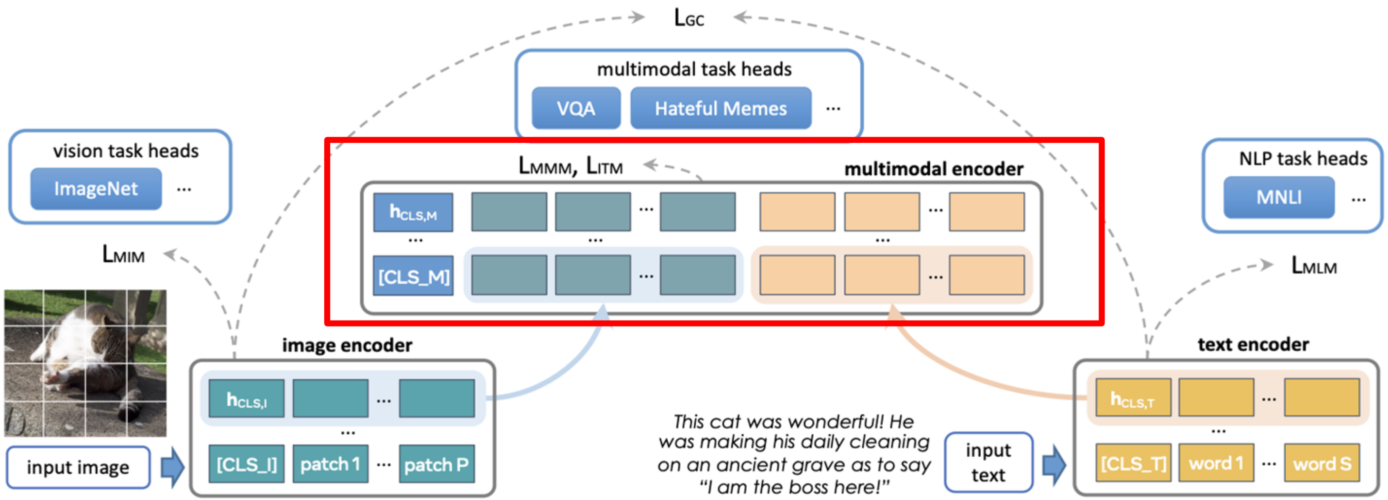
Architecture - Text encoder

- BERT tokenizer
- Transformer encoder
- Text CLS token



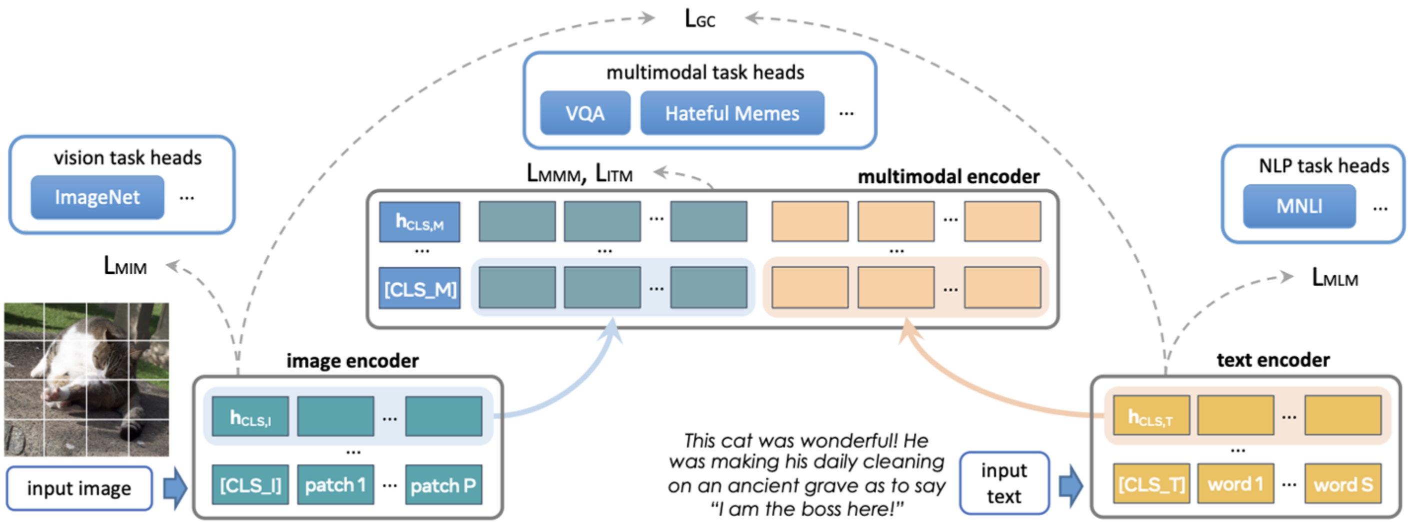
Architecture - Multimodal encoder

- Transformer model
- Input: Concat multimodal CLS token, image, text hidden states *[CLS_M | H_I | H_T]*
- Cross-attention between modalities



Architecture

Q: What do you think about this architecture?

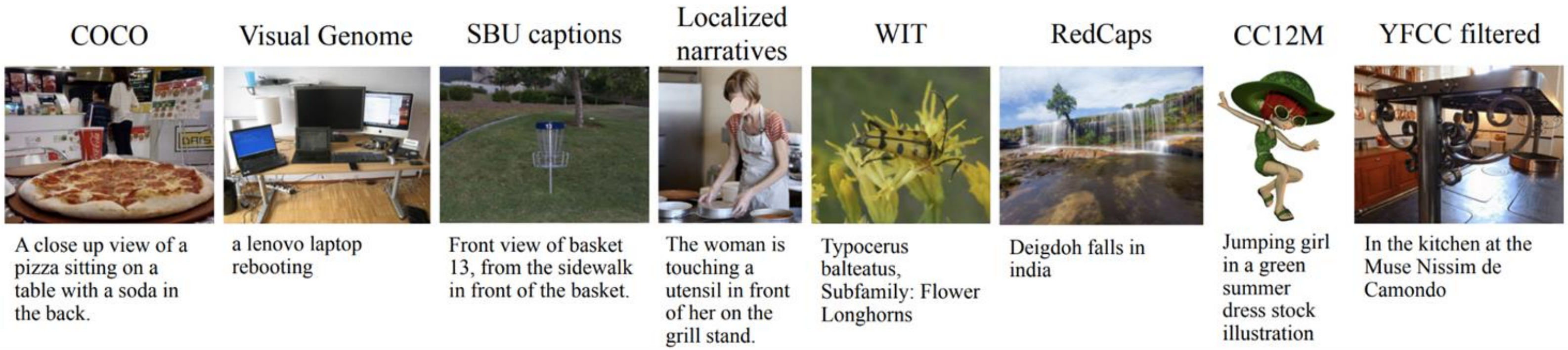


Key takeaway: 1 encoder per domain

Training

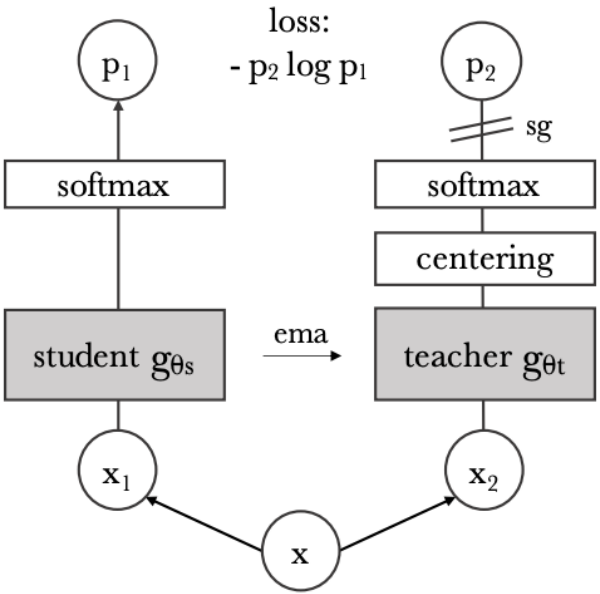
Q: Why pretrain the unimodal encoders first?

1. Unimodal pretraining of image and text encoders
   - DINO initialization + masked image modelling, masked language modelling
1. Joint training on all 3 domains
   - Global contrastive loss, masked multimodal modelling
   - Round-robin sampling between unimodal text, unimodal image, and multimodal objectives
1. Evaluation via finetuning, linear probing, or zero-shot inference



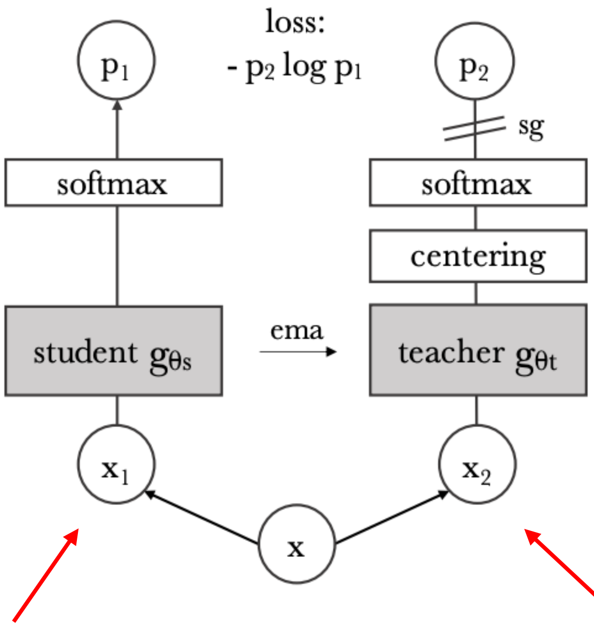
Qualitative examples from the multimodal training datasets

DINO



- Self-supervised learning w/ knowledge distillation:
  - SSL: Data itself provides the labels for training
  - Knowledge distill.: Train **student** network to match distribution of **teacher** network
- Apply different augmentations for student vs. teacher

DINO



Student gets all views

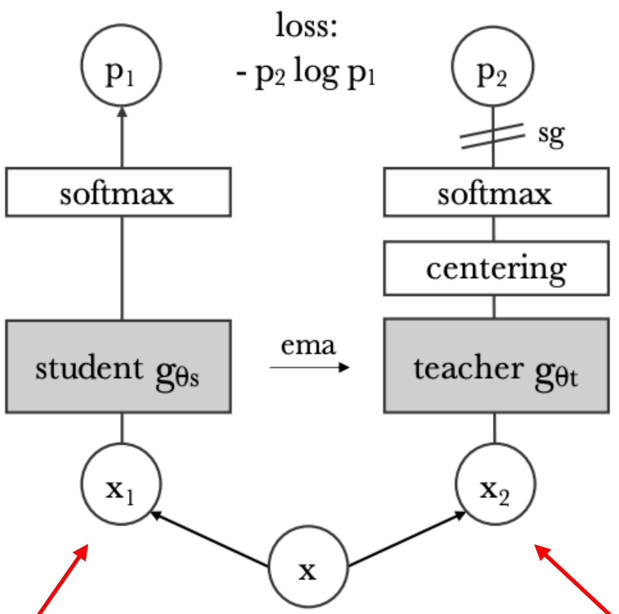
Teacher only gets global views

Multi-crop strategy:

Global views:

Local views:

DINO



Student gets all views

Teacher only gets global views

- "Local-to-global" correspondence
- Each encoder outputs a distribution over K dimensions

1. Take softmax:

$$P_s(x)^{(i)} = \frac{\exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^{K}\exp(g_{\theta_s}(x)^{(k)}/\tau_s)},$$

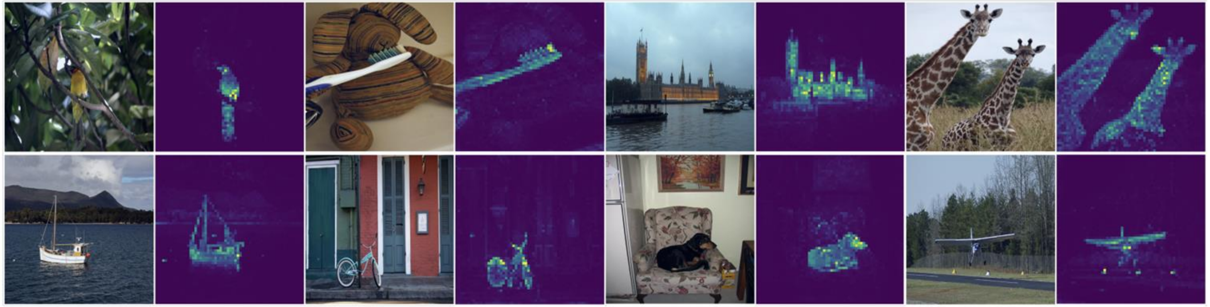
1. Minimize cross-entropy b/t teacher and student distributions:

$$\min_{\theta_s} H(P_t(x), P_s(x)), \quad \text{where } H(a,b) = -a\log b.$$

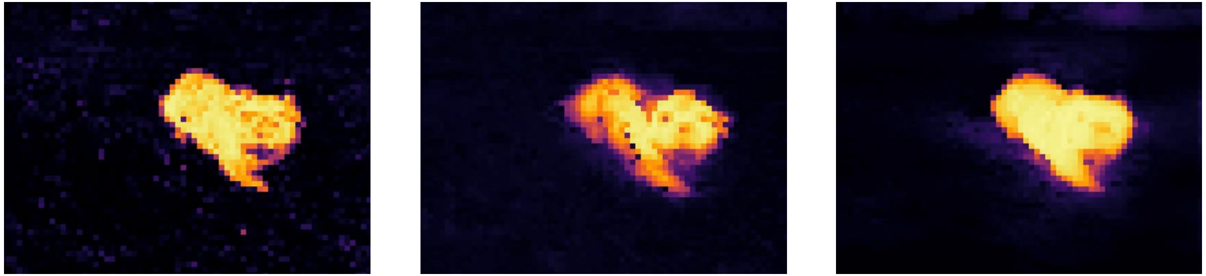
DINO

- Features have strong object boundary priors

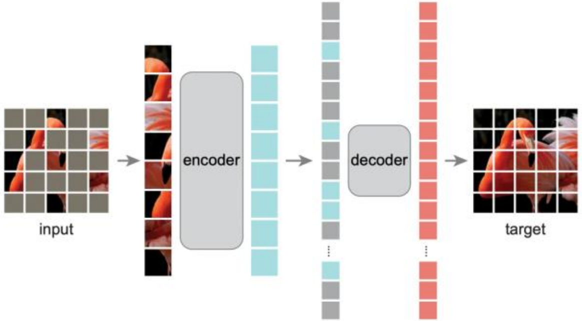
DINO



DINO
vs.
DINOv2
vs.
DINOv3



Pretraining objectives



Masking

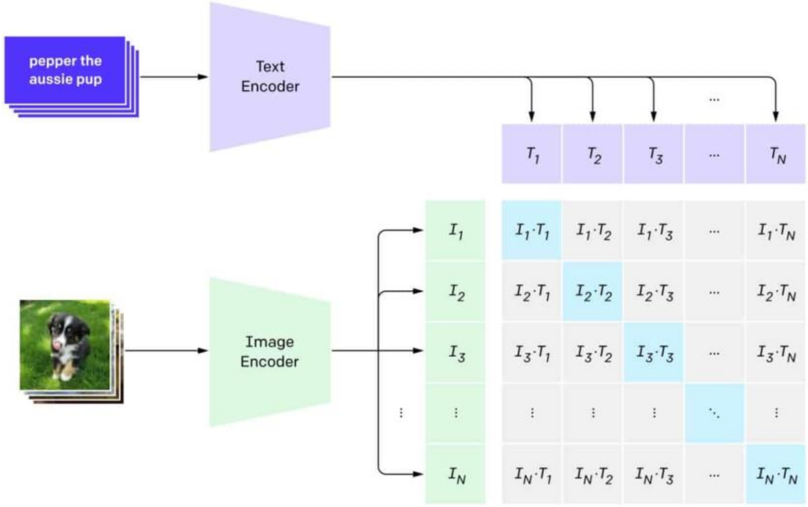
- Mask subset of the input token sequence and then predict using context of non-masked tokens
- **Image**: Encode image as discrete tokens (dVAE) and classify masked tokens
- **Text**: Tokenize text and classify tokens
- Applicable to both uni- and multi-modal settings

Discrete variation autoencoders (dVAE)

- Image patches are encoded and mapped to a discrete latent code
- Learnable latent codes form a **codebook**: e_1, e_1, …, e_K
- During masked pretraining with dVAE, simply need to classify the codebook index k ∈ [1, K] of the masked patches

Q: Why classification vs. reconstruction?

Pretraining objectives



Contrastive learning

- Method for aligning different modalities through positive and negative pairs
- Based on user-defined similarity metric (typically dot product)
- Sensitive to batch size
  - FLAVA uses ***global*** contrastive loss: examples gathered across all GPUs for loss calculation
  - Open-source CLIP: only examples in local GPU used in loss calculation

Pretraining objectives

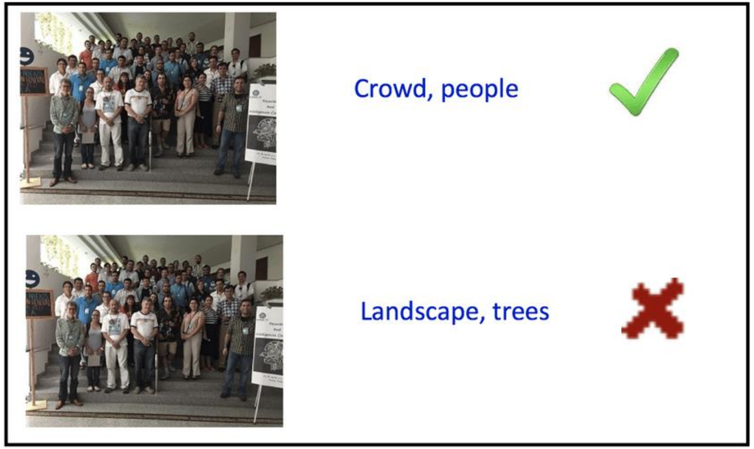


Image-text matching (ITM)

- Classification to determine whether image-text pairs match (**yes**/**no**)

Multimodal dataset

- Only trains on *publicly available* datasets
- Significantly smaller than data used by previous models (e.g. CLIP w/ 400M)
- **Caveat**: Does not take into account unimodal datasets used

	#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

		MIM 1	MLM 2	FLAVA$_{C}$ 3	FLAVA$_{MM}$ 4	FLAVA w/o init 5	FLAVA 6	CLIP 7	CLIP 8
Datasets	Eval method	PMD	PMD	PMD	PMD	(PMD+IN-1k+CCNews+BC)	PMD	PMD	400M [83]
MNLI [111]	fine-tuning	–	73.23	70.99	76.82	78.06	**80.33**	32.85	33.52
CoLA [110]	fine-tuning	–	39.55	17.58	38.97	44.22	**50.65**	11.02	25.37
MRPC [29]	fine-tuning	–	73.24	76.31	79.14	78.91	**84.16**	68.74	69.91
QQP [49]	fine-tuning	–	86.68	85.94	88.49	98.61	**88.74**	59.17	65.33
SST-2 [97]	fine-tuning	–	87.96	86.47	89.33	90.14	**90.94**	83.49	88.19
QNLI [88]	fine-tuning	–	82.32	71.85	84.77	86.40	**87.31**	49.46	50.54
RTE [7, 25, 36, 40]	fine-tuning	–	50.54	51.99	51.99	54.87	**57.76**	53.07	55.23
STS-B [1]	fine-tuning	–	78.89	57.28	84.29	83.21	**85.67**	13.70	15.98
NLP Avg.		–	71.55	64.80	74.22	75.55	**78.19**	46.44	50.50
ImageNet [90]	linear eval	41.79	–	74.09	74.34	73.49	**75.54**	72.95	80.20
Food101 [11]	linear eval	53.30	–	87.77	87.53	87.39	**88.51**	85.49	91.56
CIFAR10 [58]	linear eval	76.20	–	**93.44**	92.37	92.63	92.87	91.25	94.93
CIFAR100 [58]	linear eval	55.57	–	**78.37**	78.01	76.49	77.68	74.40	81.10
Cars [56]	linear eval	14.71	–	**72.12**	72.07	66.81	70.87	62.84	85.92
Aircraft [74]	linear eval	13.83	–	**49.74**	48.90	44.73	47.31	40.02	51.40
DTD [20]	linear eval	55.53	–	76.86	76.91	75.80	**77.29**	73.40	78.46
Pets [79]	linear eval	34.48	–	**84.98**	84.93	82.77	84.82	79.61	91.66
Caltech101 [32]	linear eval	67.36	–	94.91	95.32	94.95	**95.74**	93.76	95.51
Flowers102 [76]	linear eval	67.23	–	96.36	**96.39**	95.58	96.37	94.94	97.12
MNIST [60]	linear eval	96.40	–	98.39	98.58	**98.70**	98.42	97.38	99.01
STL10 [21]	linear eval	80.12	–	98.06	98.31	98.32	**98.89**	97.29	99.09
EuroSAT [41]	linear eval	95.48	–	97.00	96.98	97.04	**97.26**	95.70	95.38
GTSRB [100]	linear eval	63.14	–	78.92	77.93	77.71	**79.46**	76.34	88.61
KITTI [35]	linear eval	86.03	–	87.83	88.84	88.70	**89.04**	84.89	86.56
PCAM [106]	linear eval	85.10	–	85.02	85.51	**85.72**	85.31	83.99	83.72
UCF101 [98]	linear eval	46.34	–	82.69	82.90	81.42	**83.32**	77.85	85.17
CLEVR [52]	linear eval	61.51	–	79.35	**81.66**	80.62	79.66	73.64	75.89
FER 2013 [38]	linear eval	50.98	–	59.96	60.87	58.99	**61.12**	57.04	68.36
SUN397 [113]	linear eval	52.45	–	81.27	81.41	81.05	**82.17**	79.96	82.05
SST [83]	linear eval	57.77	–	56.67	**59.25**	56.40	57.11	56.84	74.68
Country211 [83]	linear eval	8.87	–	27.27	26.75	27.01	**28.92**	25.12	30.10
Vision Avg.		57.46	–	79.14	79.35	78.29	**79.44**	76.12	82.57
VQAv2 [39]	fine-tuning	–	–	67.13	71.69	71.29	**72.49**	59.81	54.83
SNLI-VE [114]	fine-tuning	–	–	73.27	78.36	78.14	**78.89**	73.53	74.27
Hateful Memes [53]	fine-tuning	–	–	55.58	70.72	**77.45**	76.09	56.59	63.93
Flickr30K [81] TR R@1	zero-shot	–	–	68.30	**69.30**	64.50	67.70	60.90	82.20
Flickr30K [81] TR R@5	zero-shot	–	–	93.50	92.90	90.30	**94.00**	88.90	96.60
Flickr30K [81] IR R@1	zero-shot	–	–	60.56	63.16	60.04	**65.22**	56.48	62.08
Flickr30K [81] IR R@5	zero-shot	–	–	86.68	87.70	86.46	**89.38**	83.60	85.68
COCO [66] TR R@1	zero-shot	–	–	43.08	**43.48**	39.88	42.74	37.12	52.48
COCO [66] TR R@5	zero-shot	–	–	75.82	**76.76**	72.84	**76.76**	69.48	76.68
COCO [66] IR R@1	zero-shot	–	–	37.59	**38.46**	34.95	38.38	33.29	33.07
COCO [66] IR R@5	zero-shot	–	–	67.28	**67.68**	64.63	67.47	62.47	58.37
Multimodal Avg.		–	–	66.25	69.11	67.32	**69.92**	62.02	67.29
Macro Avg.		19.15	23.85	70.06	74.23	73.72	**75.85**	61.52	66.78

Ablations show overall best model includes all pretraining objectives

- Contrastive learning
- Unimodal + multimodal pretraining
- Initialize unimodal encoders with pretrained models

Quantitative results

	public data		Multimodal Tasks			Language Tasks								ImageNet
			VQAv2	SNLI-VE	HM	CoLA	SST-2	RTE	MRPC	QQP	MNLI	QNLI	STS-B	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**

UniT: Multimodal Multitask Learning with a **Uni**fied **T**ransformer

Ronghang Hu Amanpreet Singh

ICCV 2021

UniT vs FLAVA

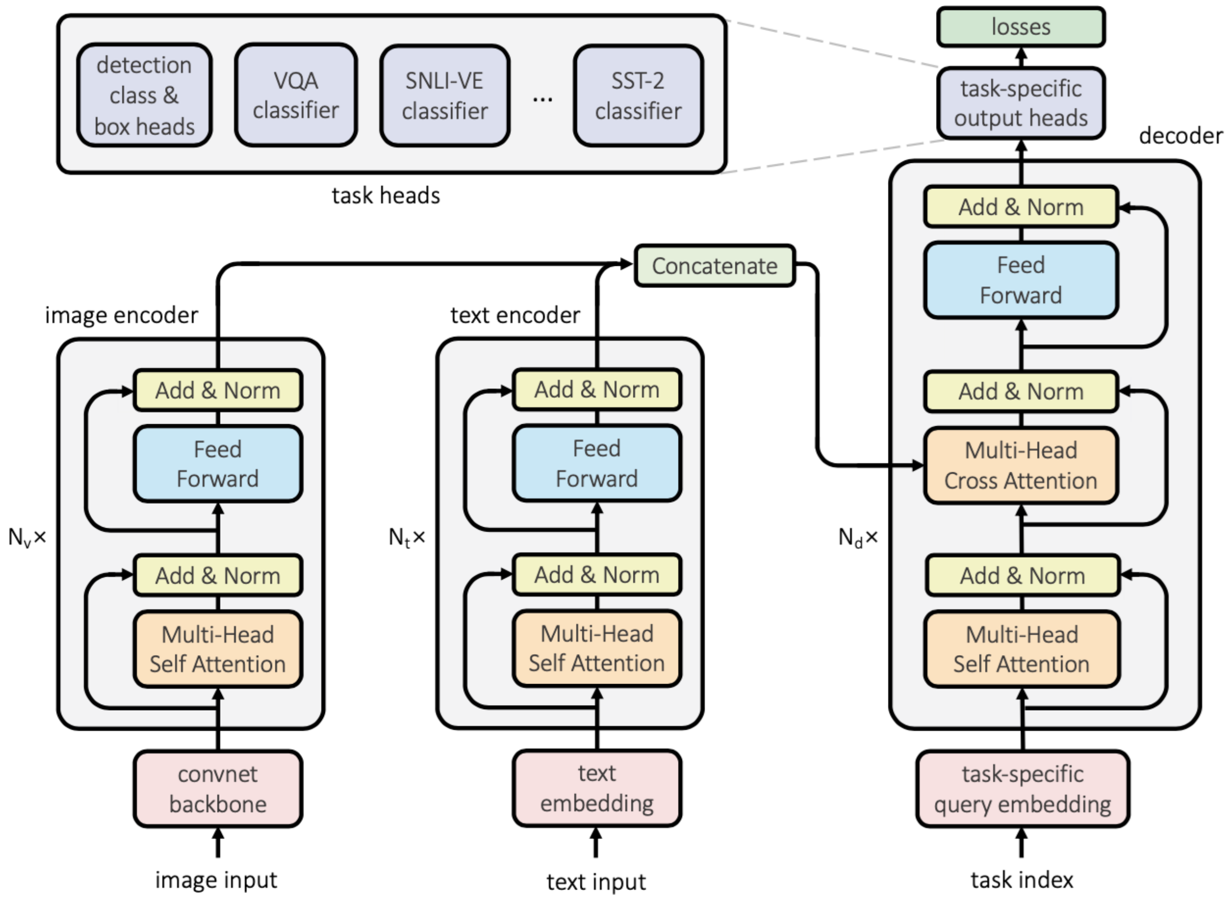
- Both UniT and FLAVA are multimodal and multitask.
- Let’s start with authors

UniT: Multimodal Multitask Learning with a Unified Transformer

Ronghang Hu   Amanpreet Singh
Facebook AI Research (FAIR)
{ronghanghu,asg}@fb.com

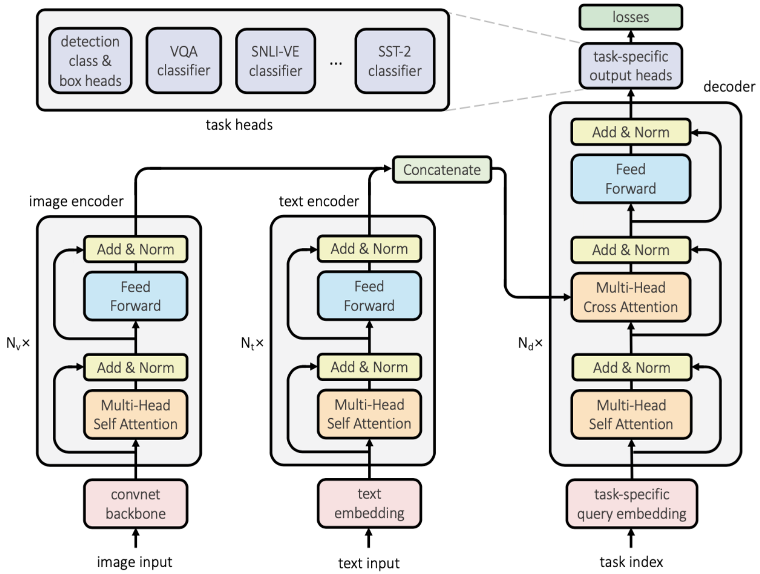
FLAVA: A Foundational Language And Vision Alignment Model

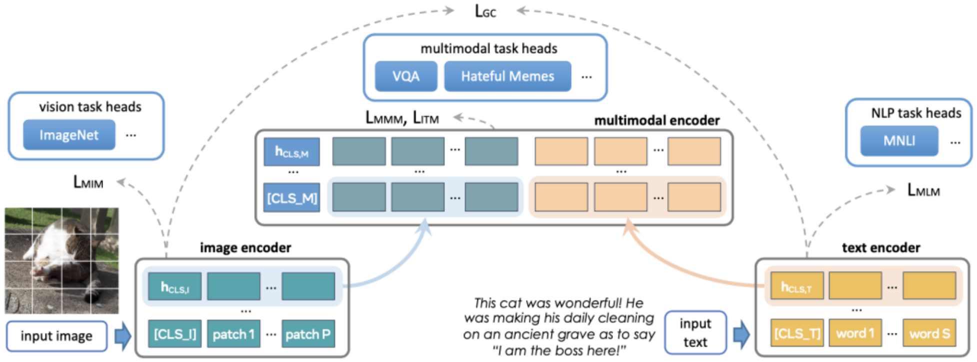
Amanpreet Singh* Ronghang Hu* Vedanuj Goswami*
Guillaume Couairon   Wojciech Galuba   Marcus Rohrbach   Douwe Kiela
Facebook AI Research (FAIR)



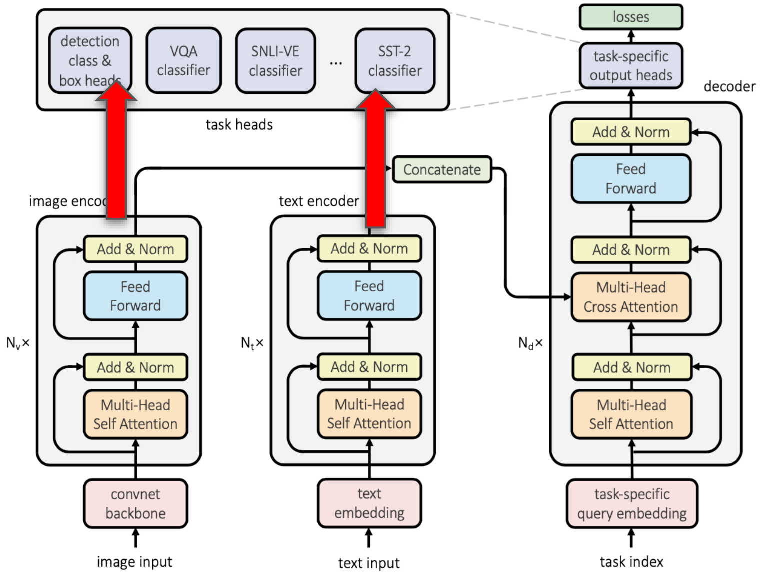
losses
task-specific output heads
decoder
detection class & box heads
VQA classifier
SNLI-VE classifier
...
SST-2 classifier
task heads
Concatenate
image encoder
text encoder
Add & Norm
Feed Forward
Add & Norm
Multi-Head Self Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Self Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Cross Attention
Add & Norm
Multi-Head Self Attention
$N_v\times$
$N_t\times$
$N_d\times$
convnet backbone
text embedding
task-specific query embedding
image input
text input
task index

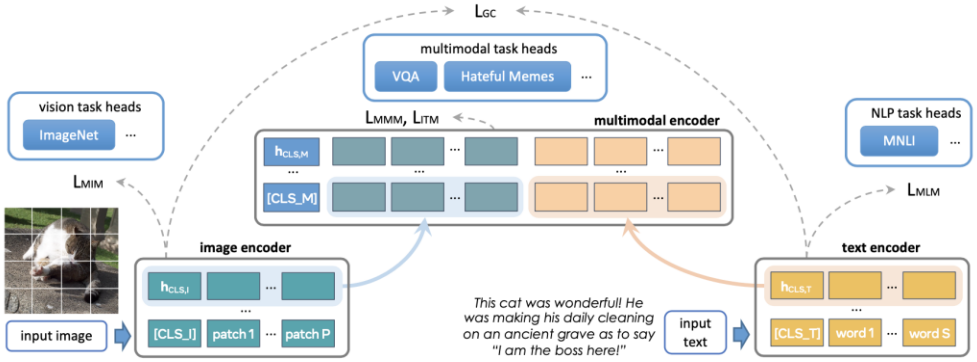
What's the difference between UniT and FLAVA?

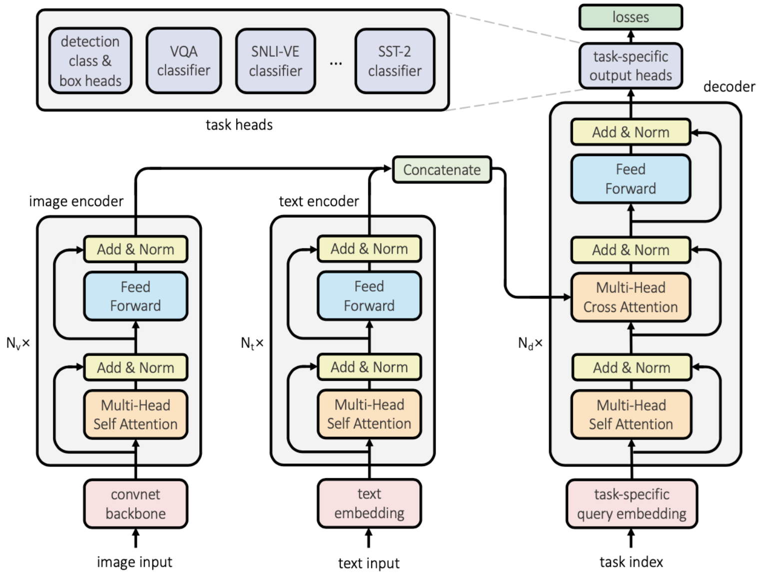




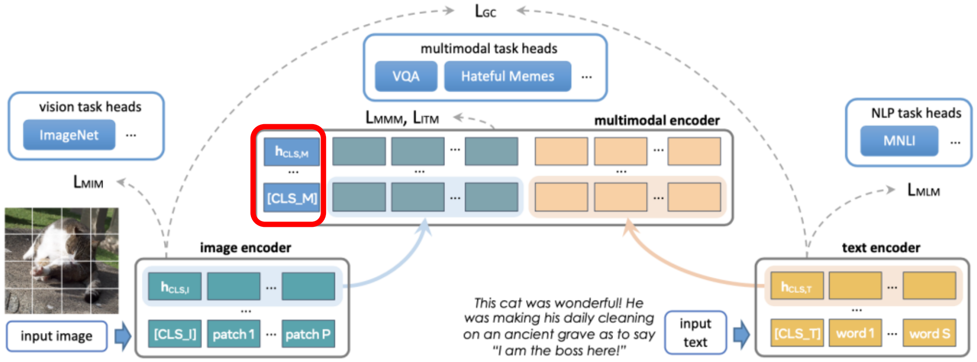
What's the difference between UniT and FLAVA?



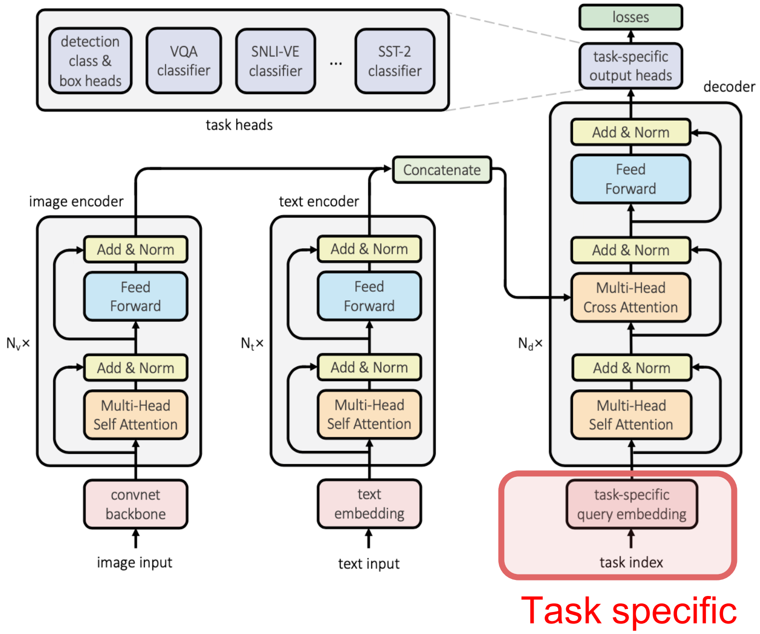


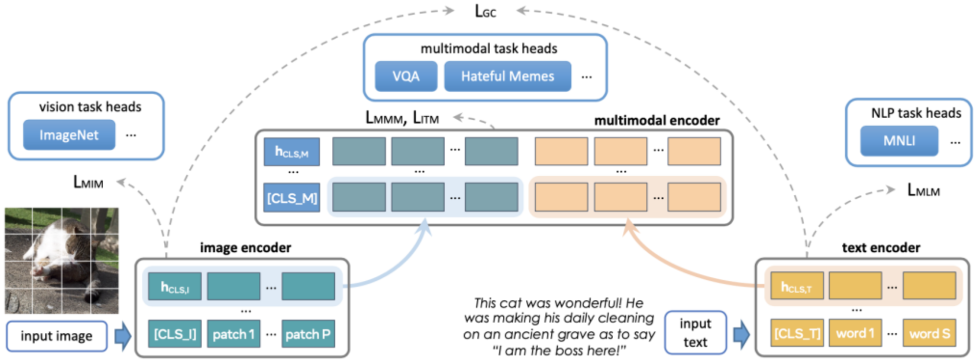


What's the difference between UniT and FLAVA?

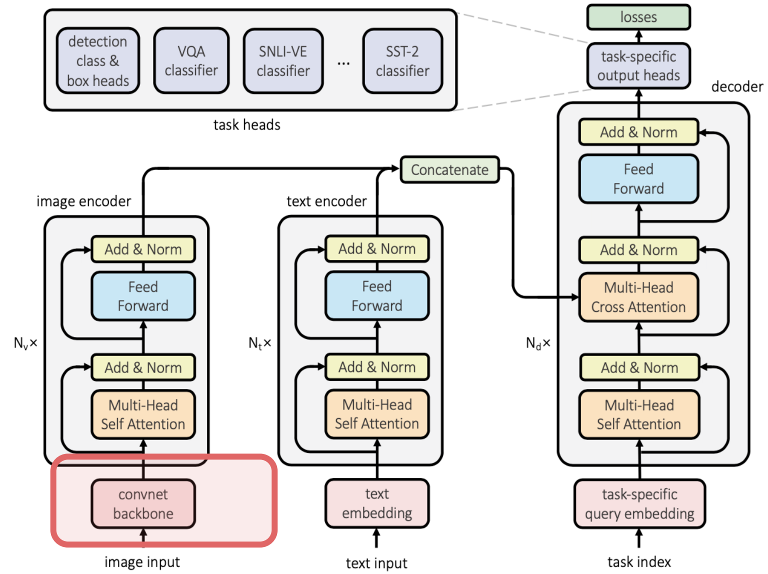


What's the difference between UniT and FLAVA?

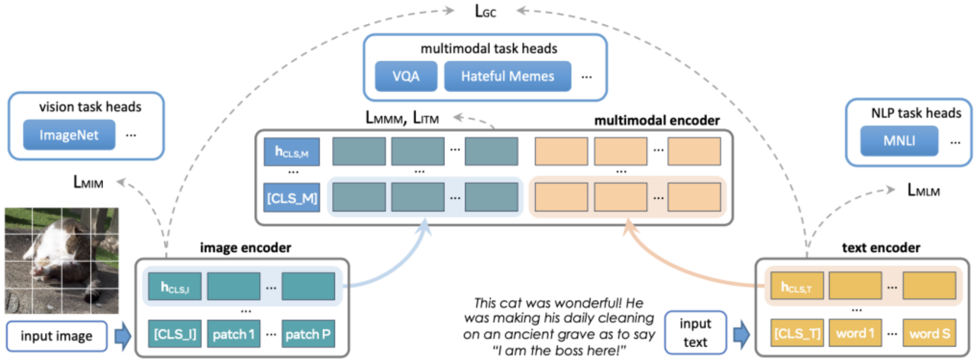




What's the difference between UniT and FLAVA?



CNN based backbone



	public data		Multimodal Tasks VQAv2	SNLI-VE	HM	CoLA	SST-2	RTE	Language Tasks 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**

What is the other biggest difference between UniT and FLAVA other than model architecture?

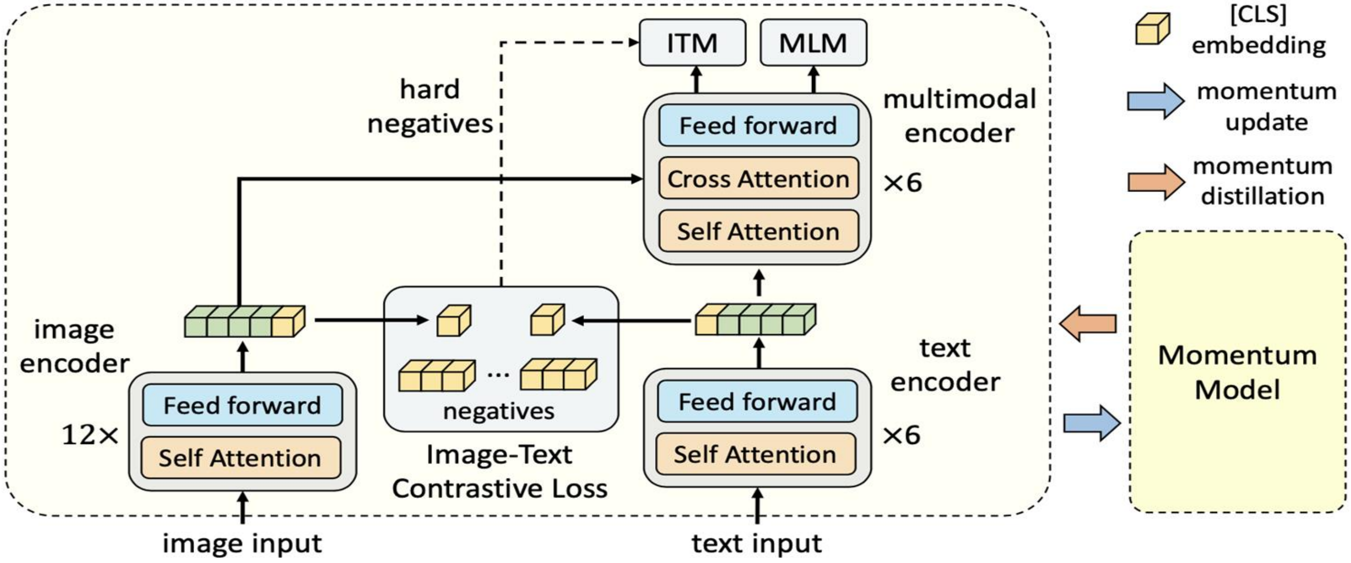
<u>Al</u>ign **<u>Be</u>**fore **<u>F</u>**use: Vision and Language Representation Learning with Momentum Distillation

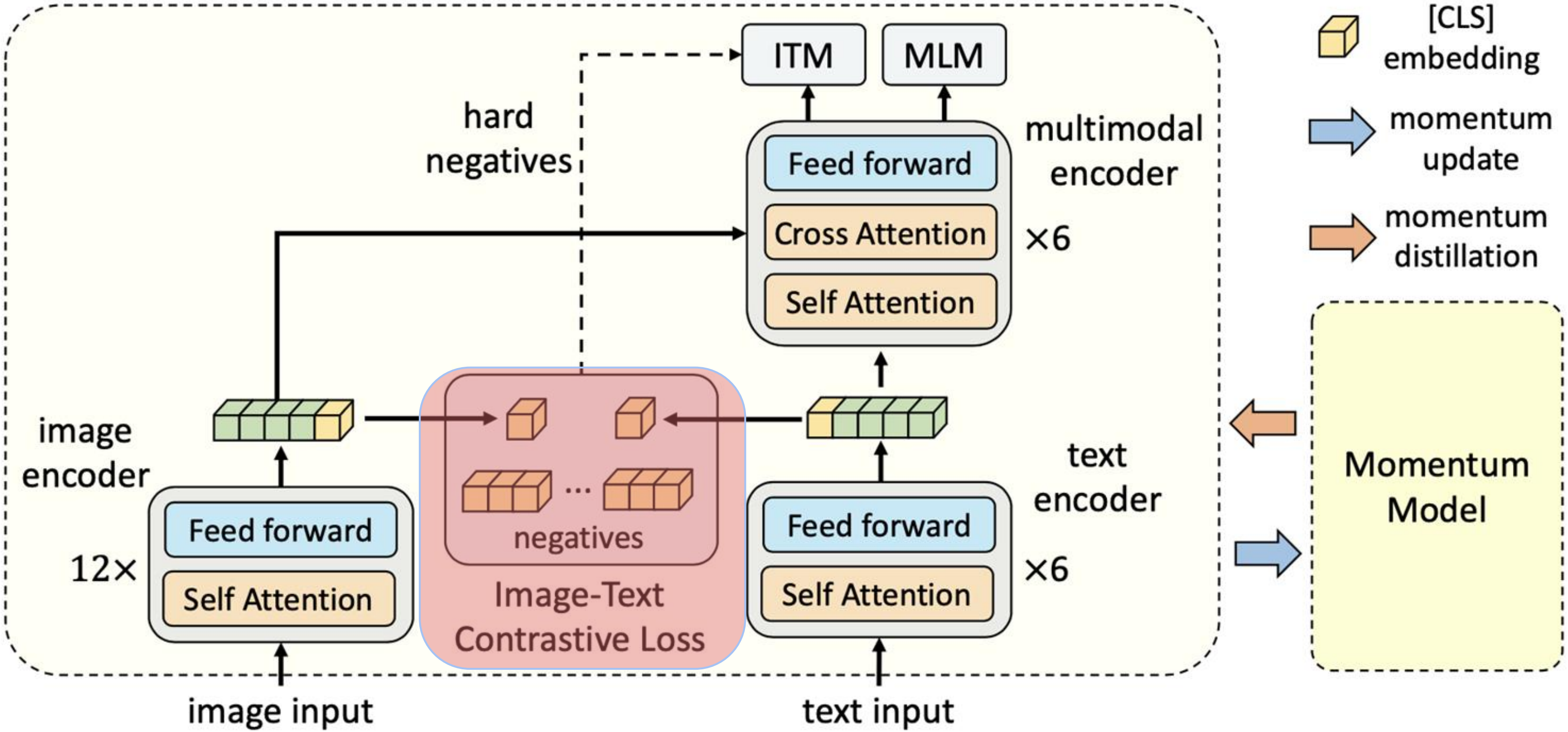
Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare
Shafiq Joty, Caiming Xiong, Steven C.H. Hoi

NeurIPS 2021

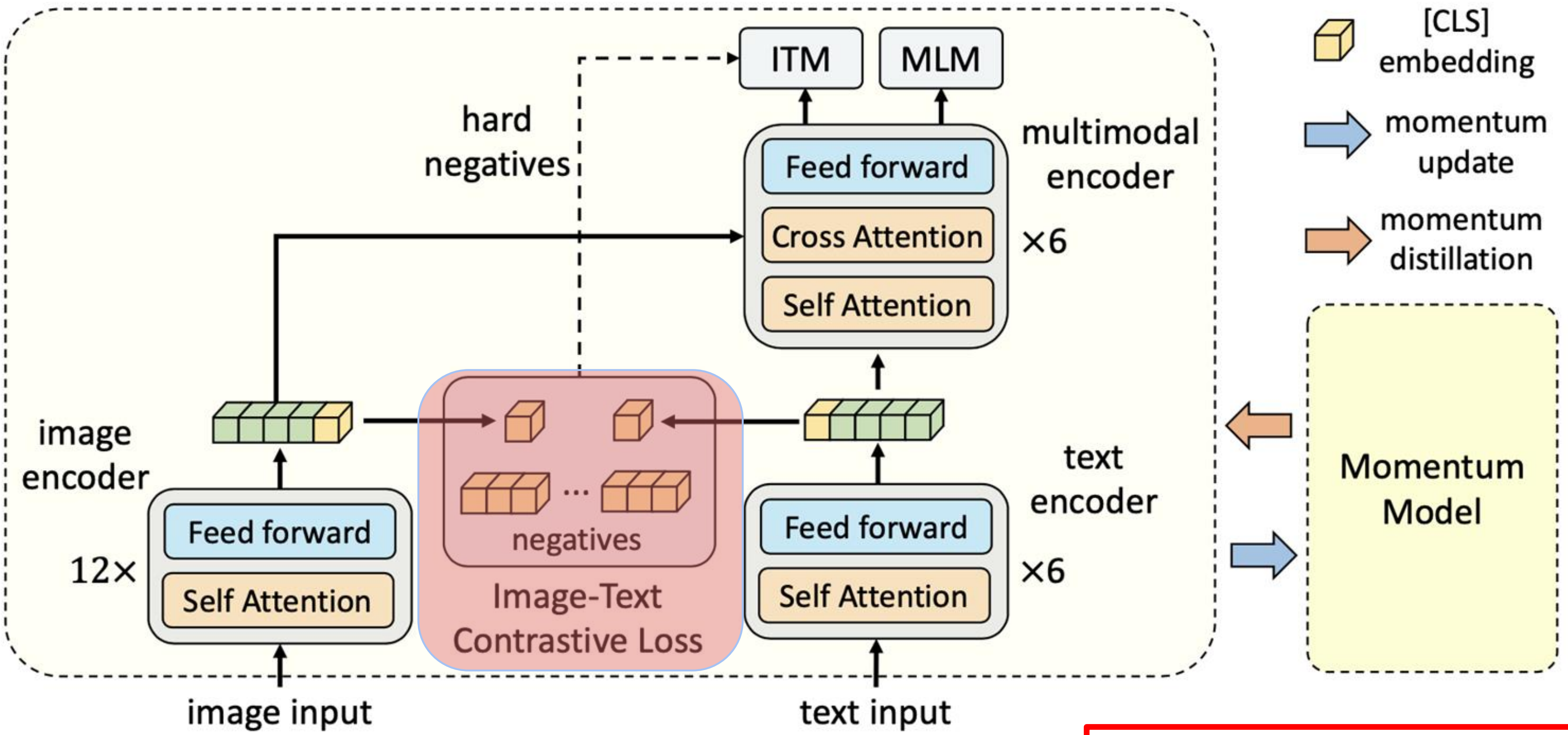
UniT is an end-to-end multimodal multitask learning framework

ALBEF pretrains the model and fine tune on downstream tasks.





What could be the objectives for pretraining?

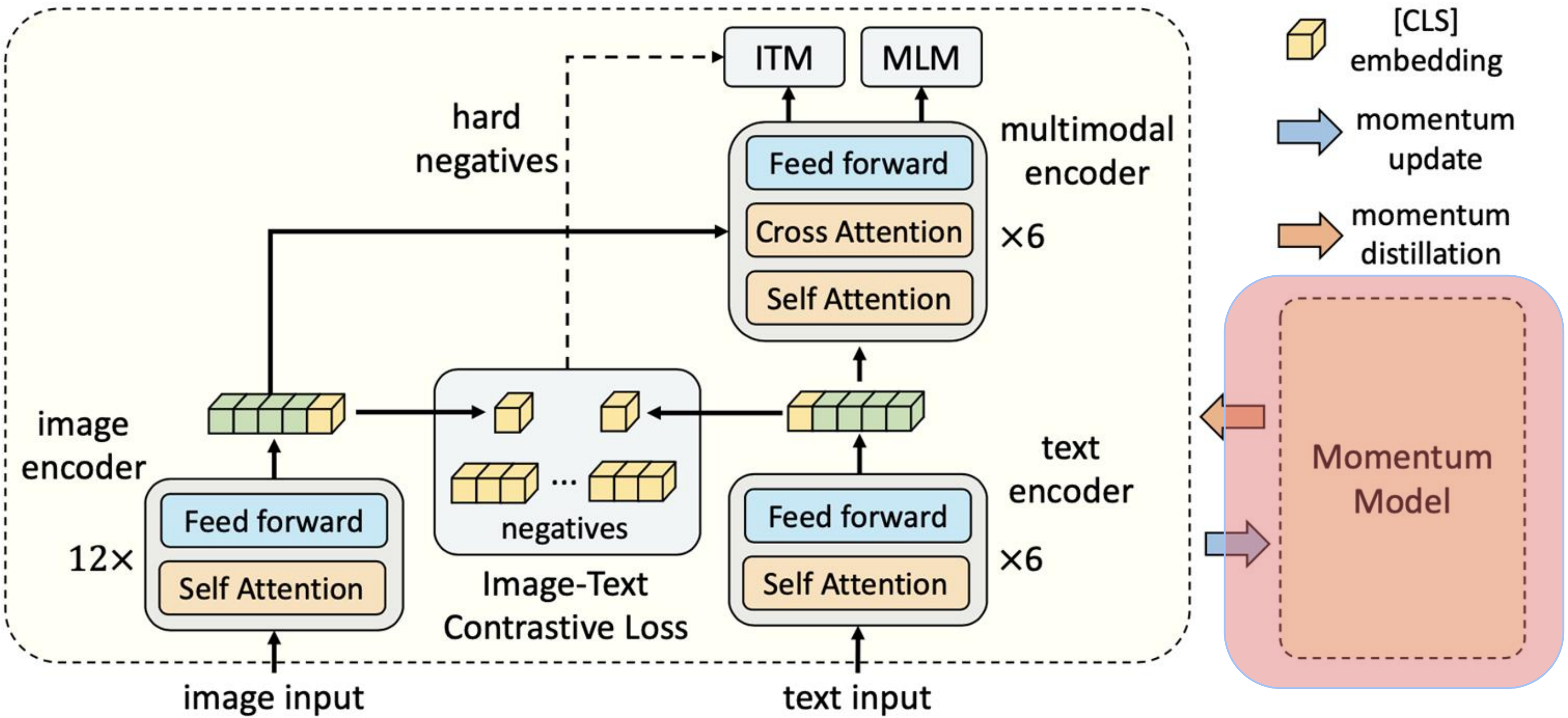


Contrastive!

AND Mask Language Modeling!

AND Image-Text Matching!

$$\boxed{\mathcal{L} = \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{itm}}}$$



Data are noisy

$$\mathcal{L}_{\text{itc}}^{\text{mod}} = (1-\alpha)\mathcal{L}_{\text{itc}} + \frac{\alpha}{2}\mathbb{E}_{(I,T)\sim D}\left[\text{KL}(\boldsymbol{q}^{\text{i2t}}(I) \parallel \boldsymbol{p}^{\text{i2t}}(I)) + \text{KL}(\boldsymbol{q}^{\text{t2i}}(T) \parallel \boldsymbol{p}^{\text{t2i}}(T))\right]$$

$$\mathcal{L}_{\text{mlm}}^{\text{mod}} = (1-\alpha)\mathcal{L}_{\text{mlm}} + \alpha\mathbb{E}_{(I,\hat{T})\sim D}\text{KL}(\boldsymbol{q}^{\text{msk}}(I,\hat{T}) \parallel \boldsymbol{p}^{\text{msk}}(I,\hat{T}))$$

Ablation Study

#Pre-train Images	Training tasks	TR (flickr test)	IR (flickr test)	SNLI-VE (test)	NLVR^2 (test-P)	VQA (test-dev)
4M	MLM + ITM	93.96	88.55	77.06	77.51	71.40
	ITC + MLM + ITM	96.55	91.69	79.15	79.88	73.29
	ITC + MLM + ITM_{hard}	97.01	92.16	79.77	80.35	73.81
	ITC_{MoD} + MLM + ITM_{hard}	97.33	92.43	79.99	80.34	74.06
	Full (ITC_{MoD} + MLM_{MoD} + ITM_{hard})	97.47	92.58	80.12	80.44	74.42
	ALBEF (Full + $\text{MoD}_{\text{Downstream}}$)	97.83	92.65	80.30	80.50	74.54
14M	ALBEF	98.70	94.07	80.91	83.14	75.84

Results

Method	# Pre-train Images	Flickr30K (1K test set)					
		TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
UNITER [2]	4M	83.6	95.7	97.7	68.7	89.2	93.9
CLIP [6]	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN [7]	1.2B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	4M	90.5	98.8	99.7	76.8	93.7	96.7
ALBEF	14M	**94.1**	**99.5**	**99.7**	**82.8**	**96.3**	**98.1**

Thank you