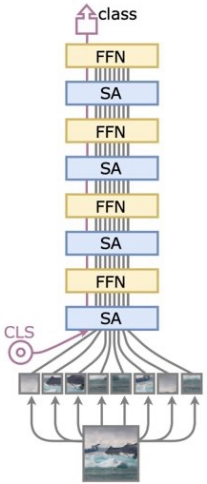
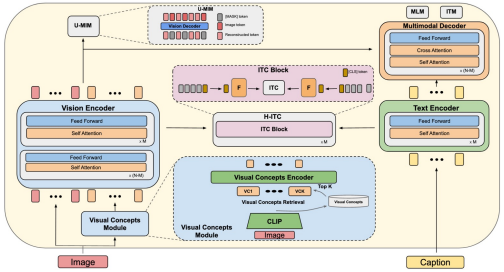
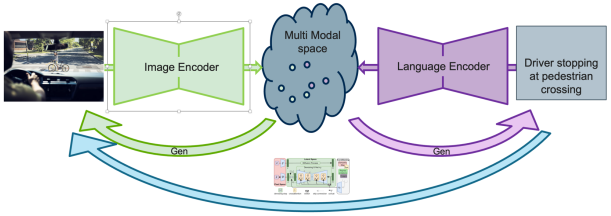
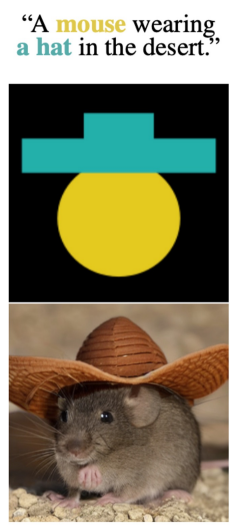








CAp - RFIAP

Vision&Language with transformers

Matthieu Cord
Prof. at Sorbonne University, ISIR lab.
Scientific Director of Valeo.ai

scai
SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

valeo.ai

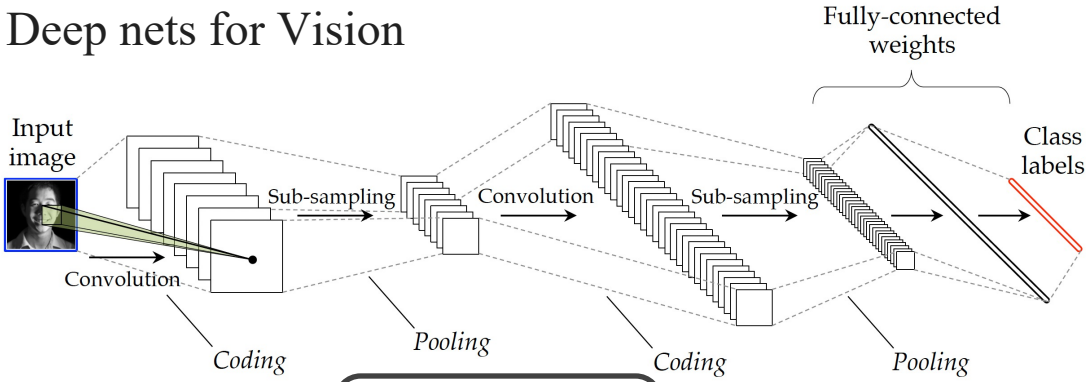
SCIENCES SORBONNE UNIVERSITÉ

Outline

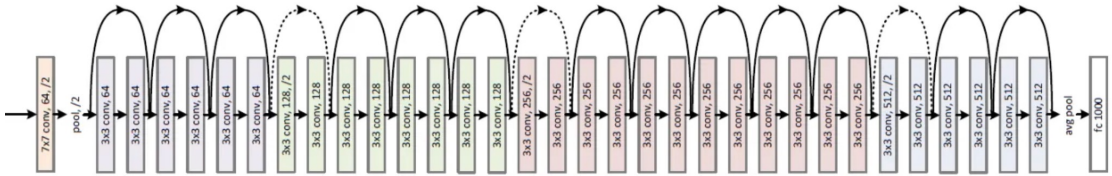
1. **Vision transformers: the basics**
2. From classification to detection, segmentation, …
3. Vision-Language Models in the era of LLMs

(Visual) Transformers

Deep nets for Vision

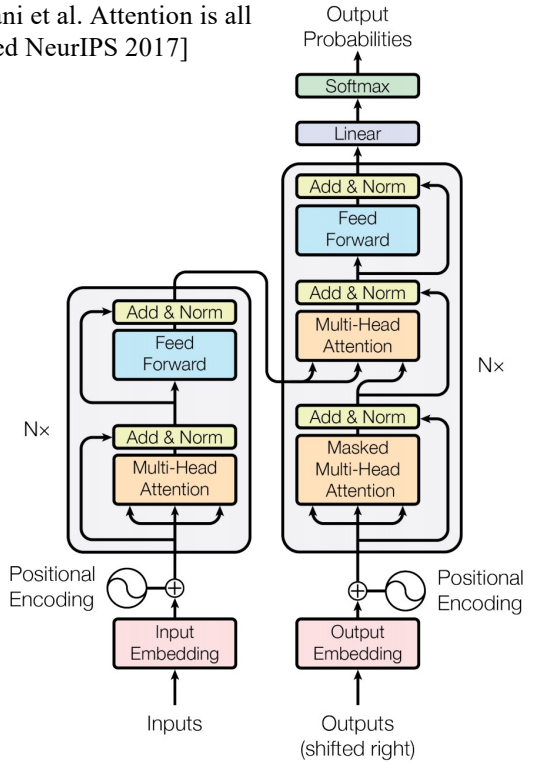


ConvNet
ResNet



Transformers

[Vaswani et al. Attention is all you need NeurIPS 2017]



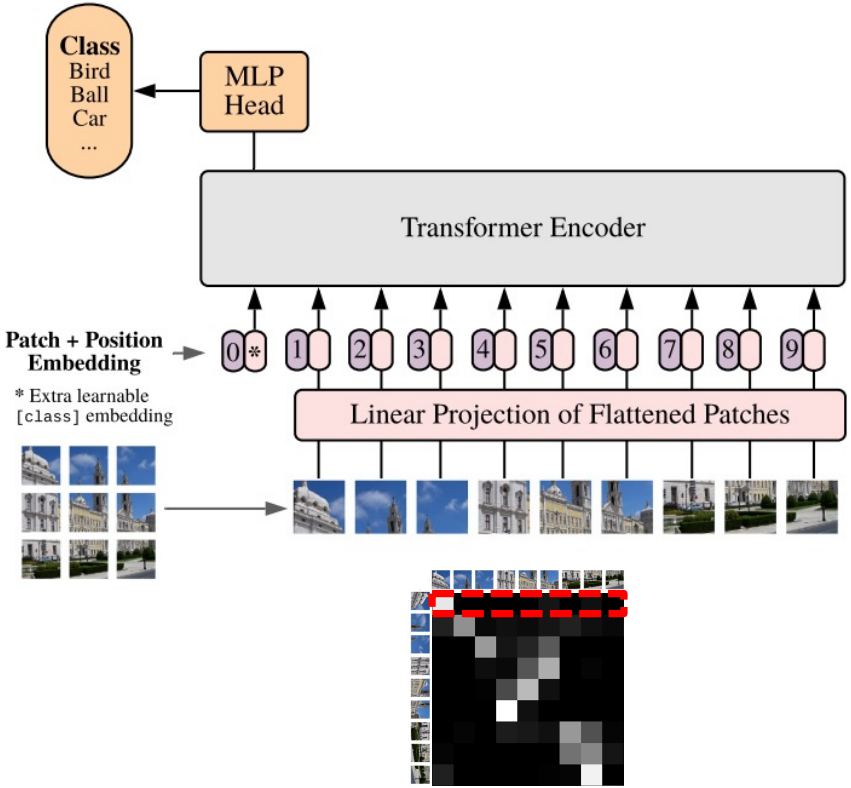
(Visual) Transformers

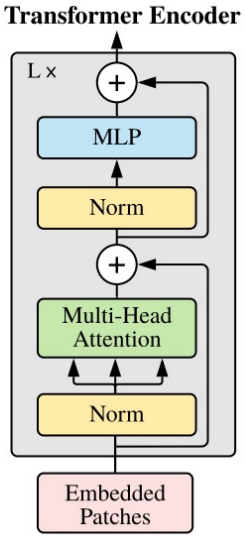
ViT (Vision Transformers) architecture
=> **Self attention encoder** modules for classification

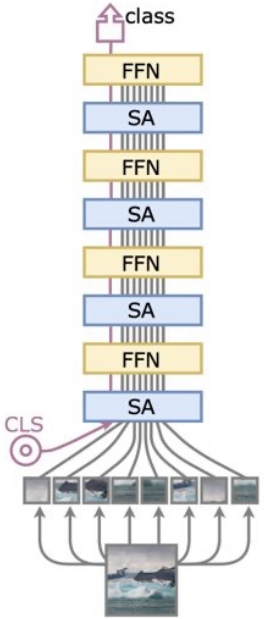
Published as a conference paper at ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com







(Visual) Transformers

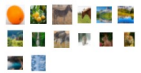
ViT OK but Requiring **hundreds of millions** of annotated images to reach convnets performance

Published as a conference paper at ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*,
Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

JFT-300M
300M images
18k labels
private



ImageNet-21k
14M images
21k labels

ImageNet
1.2M images
1k labels

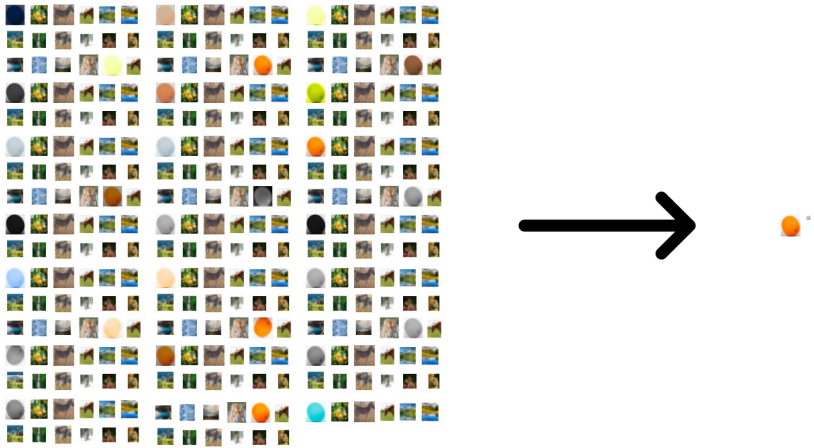
(Visual) Transformers

ViT OK but Requiring **hundreds of millions** of annotated images to reach convnets performance

Published as a conference paper at ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*,
Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com



Can we train an effective vision transformer model without a huge dataset like JFT?

(Visual) Transformers

Yes! **DeiT** paper

- Regularization (Stochastic d, Repeated A)
- Data augmentation (Mixup, CutMix)
- Distillation (From ConvNet)

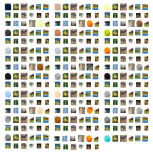
Published as a conference paper at ICML 2021

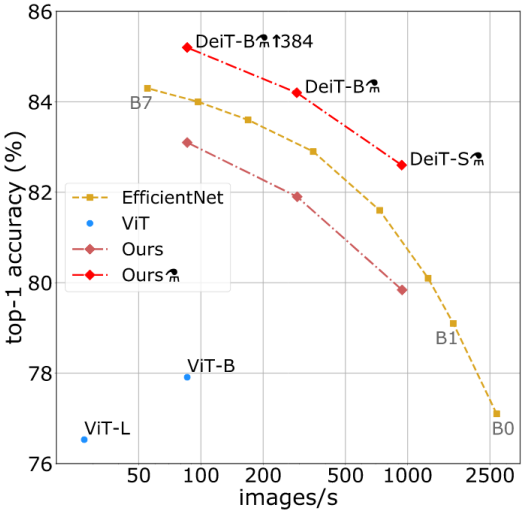
DeiT

Training data-efficient image transformers & distillation through attention

Hugo Touvron [1 2] Matthieu Cord [1 2] Matthijs Douze [1]
Francisco Massa [1] Alexandre Sablayrolles [1] Hervé Jégou [1]

State-of-the-art performance on ImageNet1k classification

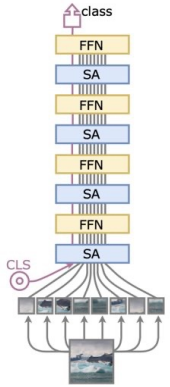
without huge  extra dataset!



(Visual) Transformers

Does ViT work with deeper models? No

Depth	Top-1
12	79.9
18	80.1
24	78.9†
36	78.9†
48	78.4†



Adaptation of DeiT: **CaiT**

Published as a conference paper at ICCV 2021

Going deeper with Image Transformers

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, Hervé Jégou

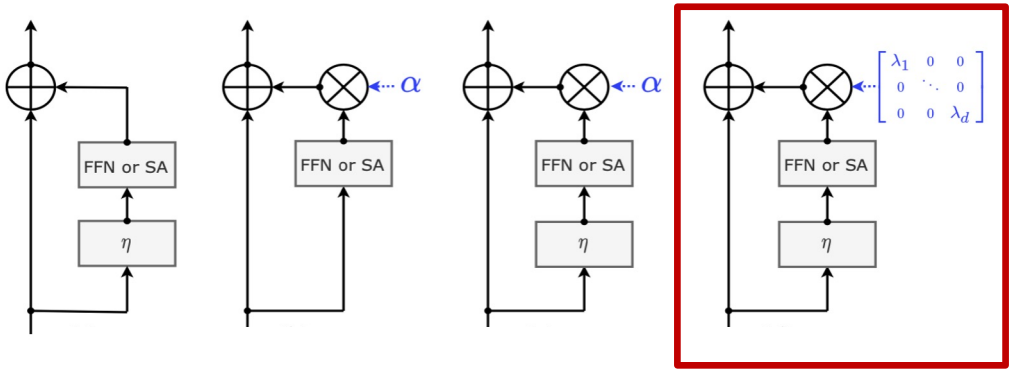
Main improvements:

-- Rescaling residual branch

-- Combined with stochastic depth:

$$H_\ell = \texttt{ReLU}(b_\ell f_\ell(H_{\ell-1}) + \mathrm{id}(H_{\ell-1}))$$

-- Class Activation architecture



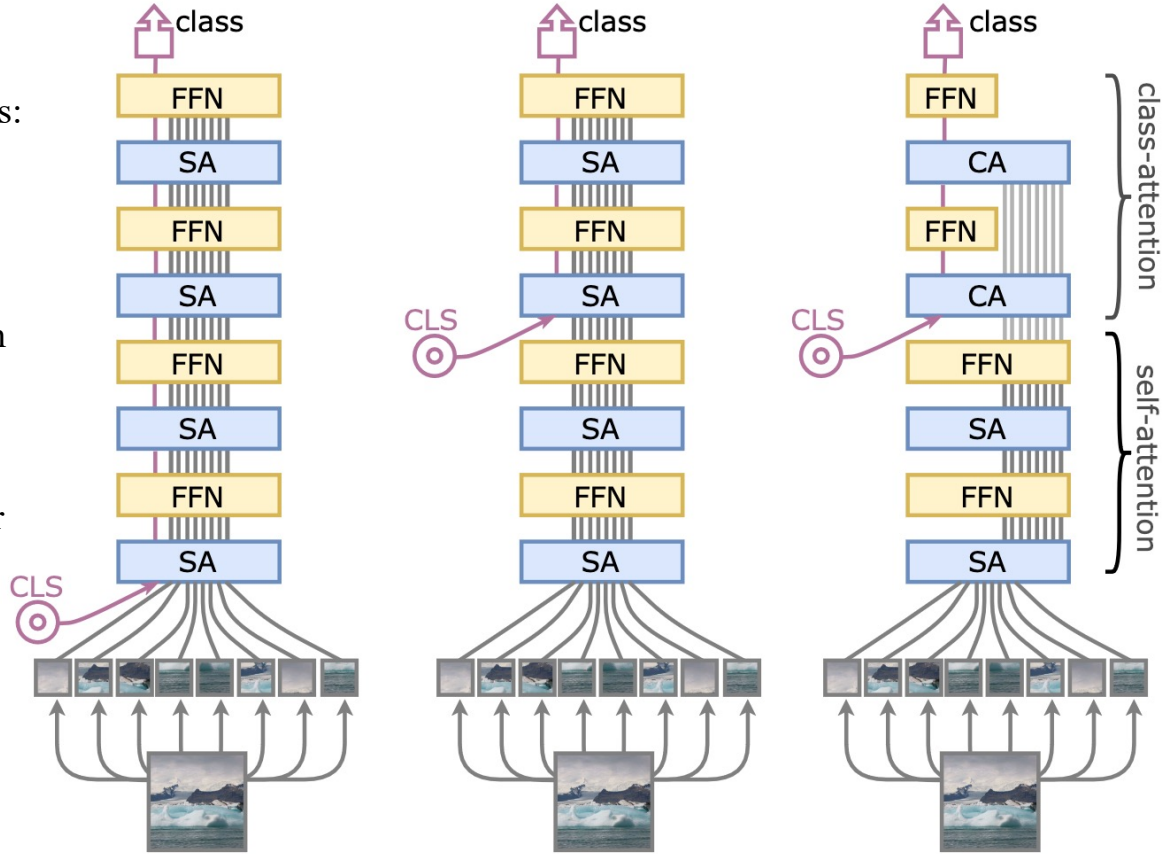
Class Activation architecture

In ViT class embedding CLS token inserted along with the patch embeddings:

- helping the attention process
- preparing the vector to be fed to the classifier

CaiT freezes the patch embeddings when inserting CLS:

- last part of the network (2 layers) dedicated to summarizing the information to be fed to the classifier
- save compute

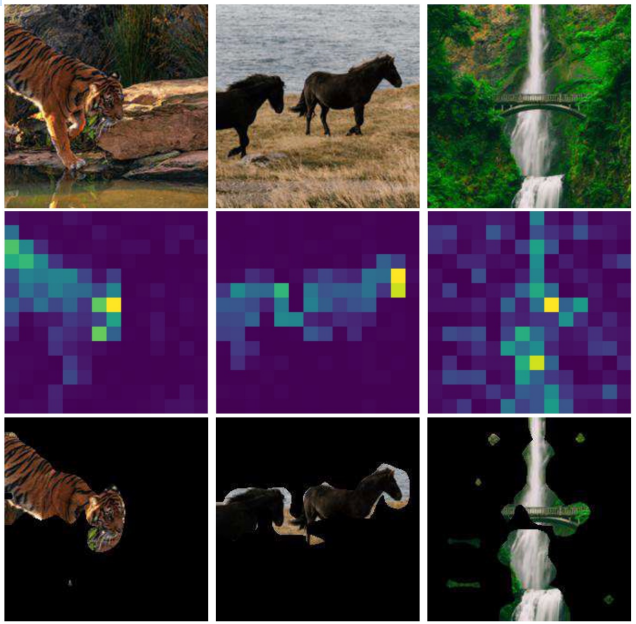


(Visual) Transformers

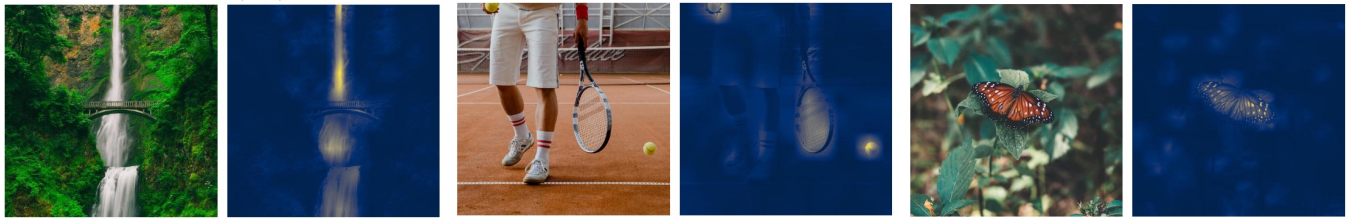
CaiT results:

Depth	Top-1	LayerScale
12	79.9	80.5
18	80.1	81.7
24	78.9†	82.4
36	78.9†	82.9

Network	nb of param.	nb of FLOPs	image size train	image size test	ImNet top-1	Real top-1	V2 top-1
CaiT-S36	68M	14B	224	224	83.3	88.0	72.5
CaiT-S36↑384	68M	48B	224	384	85.0	89.2	75.0
CaiT-S48↑384	89M	64B	224	384	85.1	89.5	75.5
CaiT-S36ϒ	68M	14B	224	224	84.0	88.9	74.1
CaiT-S36↑384ϒ	68M	48B	224	384	85.4	89.8	76.2
CaiT-M36↑384ϒ	271M	173B	224	384	86.1	90.0	76.3
CaiT-M36↑448ϒ	271M	248B	224	448	86.3	**90.2**	76.7
CaiT-M48↑448ϒ	356M	330B	224	448	**86.5**	**90.2**	**76.9**



Visualization of the class-attention:
Attention between the CLS and patches



(Visual) Transformers

Where we are:

Training deep vision transformers on ImageNet dataset (only) OK

Transformers not so interesting for small datasets (/convnets)

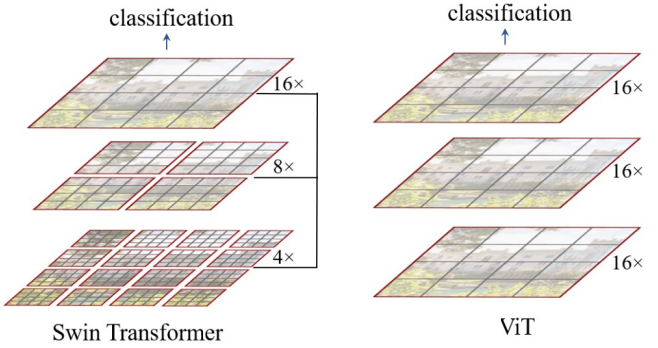
Q about modeling: Is an image 16x16 patches/words?

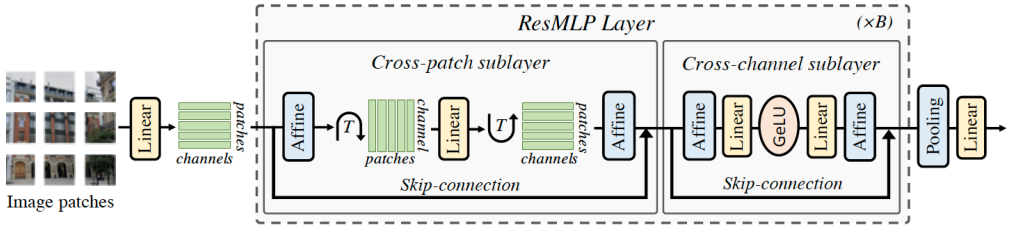
About Attention block: Quadratic complexity with the nb of patches

-- Limit the Attention complexity

-- Modify the architectures

- Swin Transformers
- MLP-like architectures
- Hybrid archi with convnets/transformers





This question also applies for NLP transformers:

Mamba paper at NeurIPS 2023

Flash Attention, …

But the entry ticket is very high!

From Image classification to segmentation, detection, …

Outline

1. Vision transformers: the basics
2. **From classification to detection, segmentation, …**
3. Vision-Language Models in the era of LLMs

(Visual) Transformers for detection, segmentation, …

Design output for classification, detection, …

- CLS token for classification
- CaiT strategy: CLS to decode the embeddings
- Extension to incremental classification task learning:

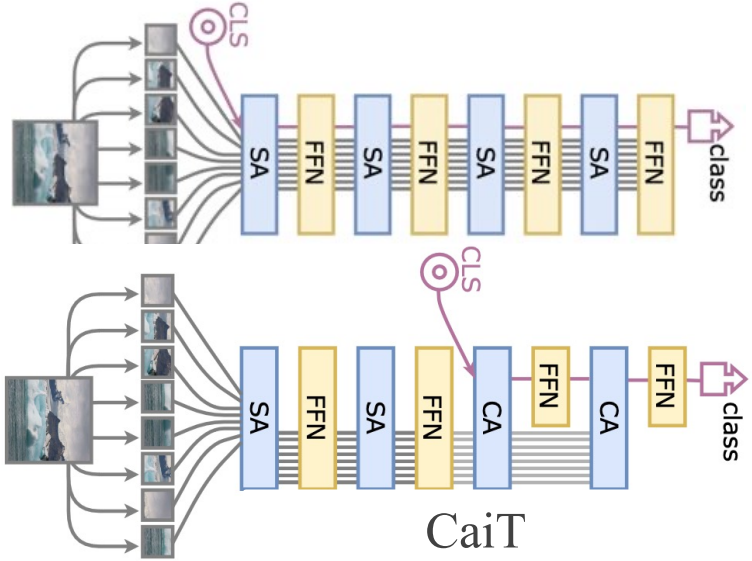
Published as a conference paper at CVPR 2022

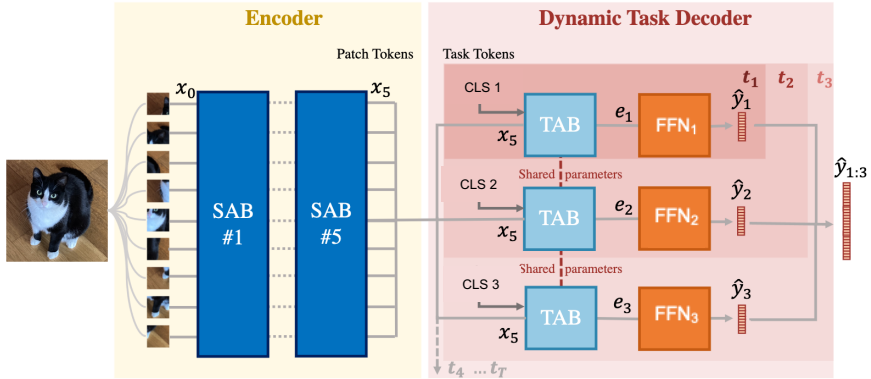
DyTox: Transformers for continual learning with dynamic token expansion

Arthur Douillard, Alexandre Ramé, Guillaume Couairon, Matthieu Cord

Multi inputs Possible with Several encoders+1 transformer for fusion&processing

And for other type of output as detection?





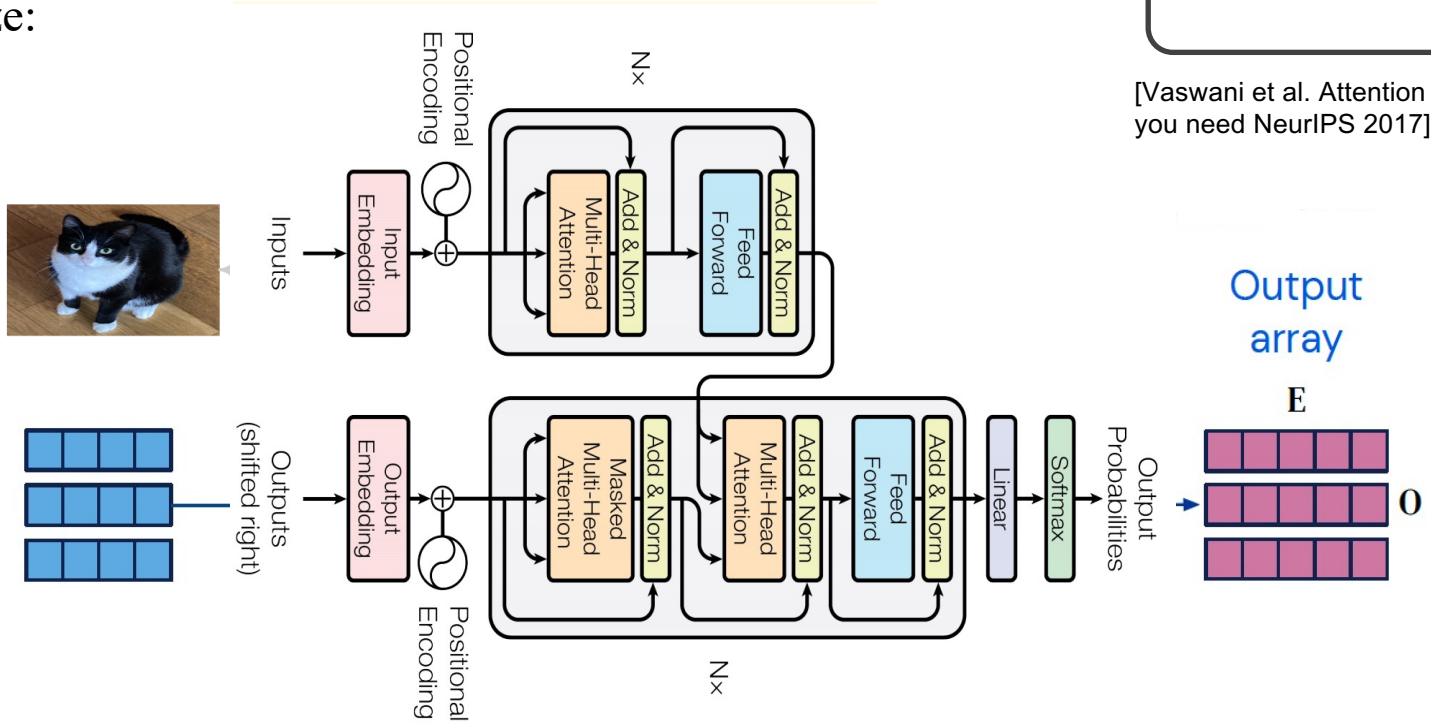
TAB: Task Attention Block

(Visual) Transformers for detection, segmentation, …

To summarize:

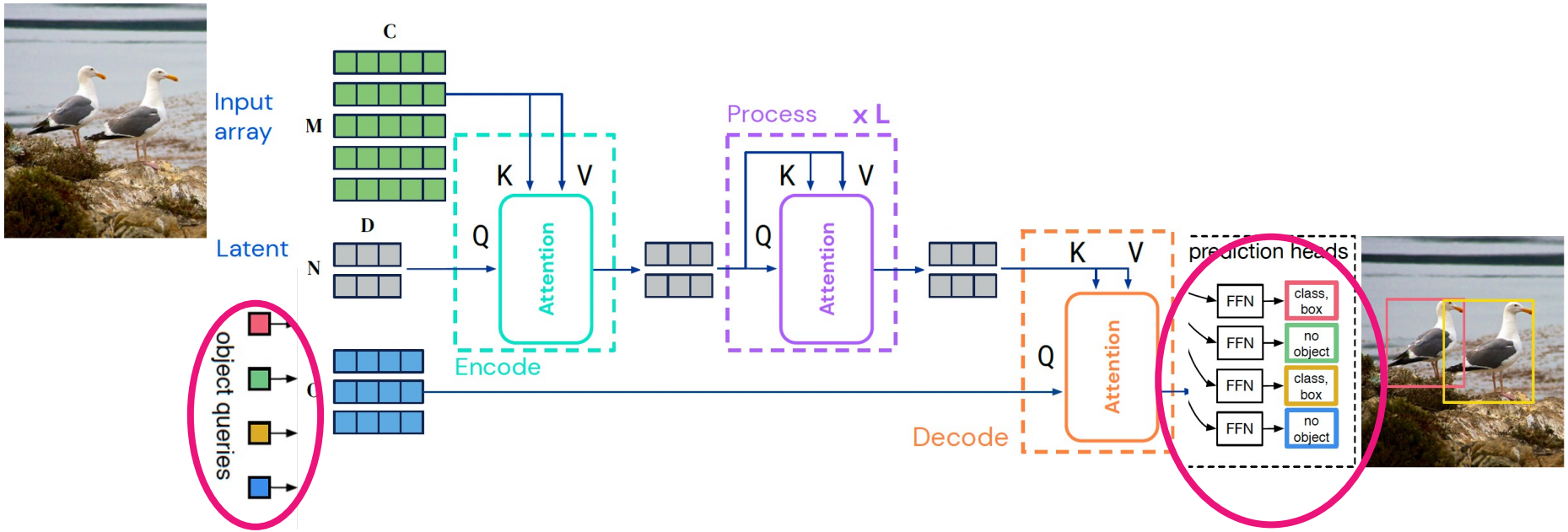
Transformers

[Vaswani et al. Attention is all you need NeurIPS 2017]



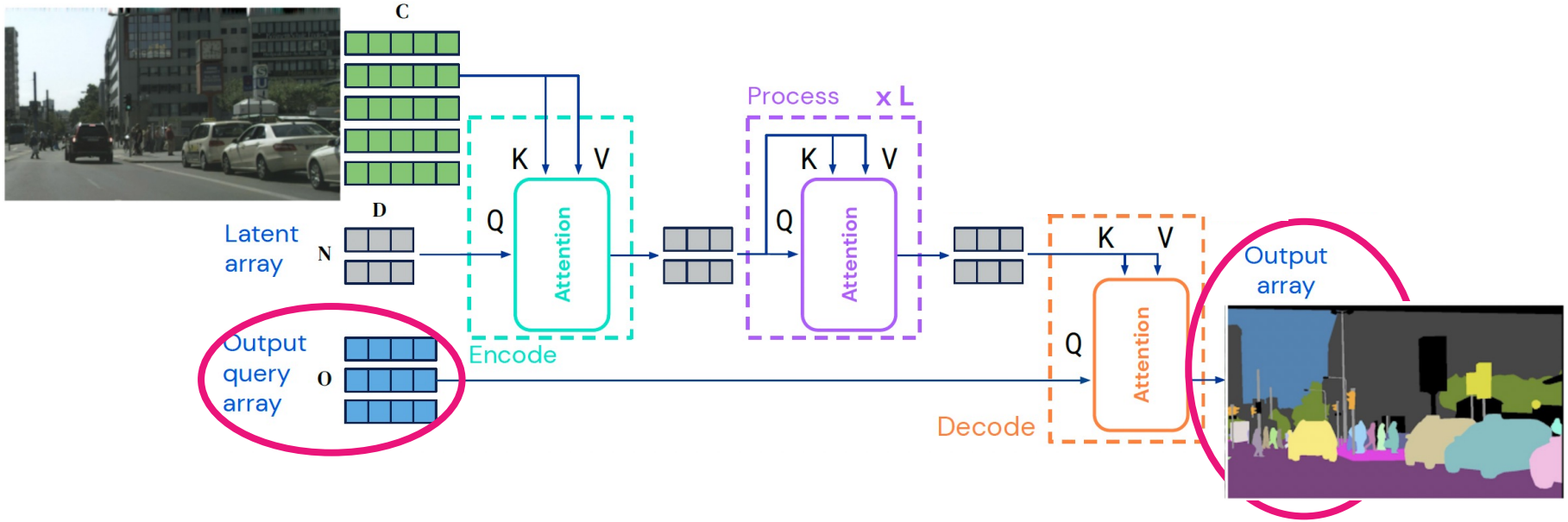
(Visual) Transformers for detection, segmentation, …

Output query array / Output array defines the downstream task: detection



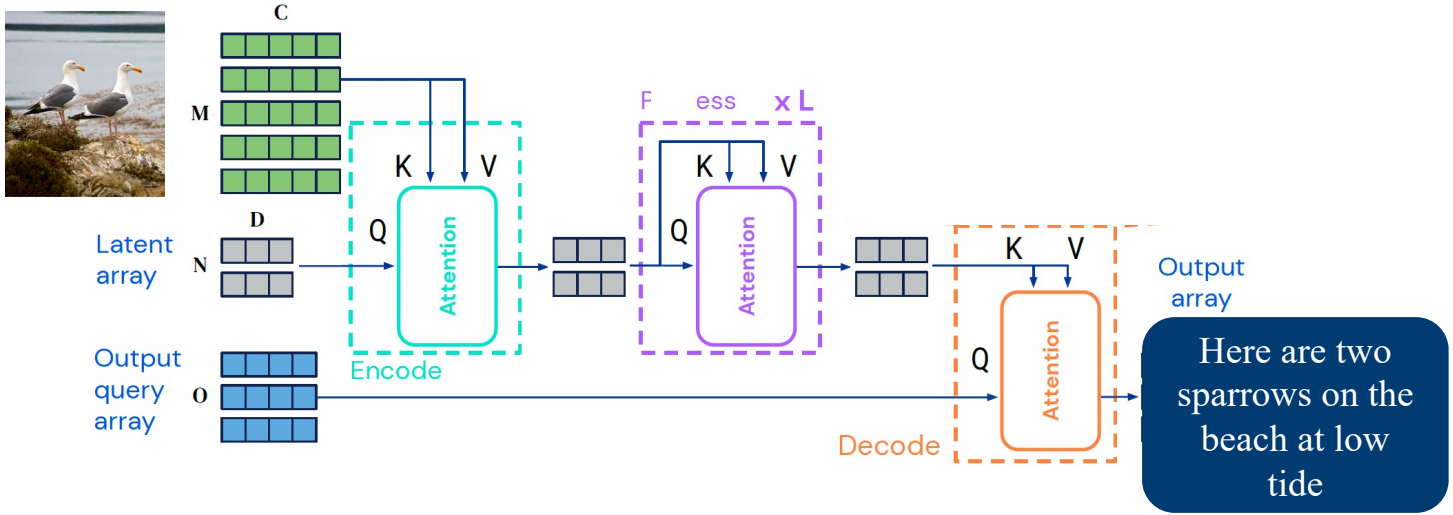
(Visual) Transformers for detection, segmentation, …

Output query array / Output array defines the downstream task: segmentation …



(Visual) Transformers for detection, segmentation, …

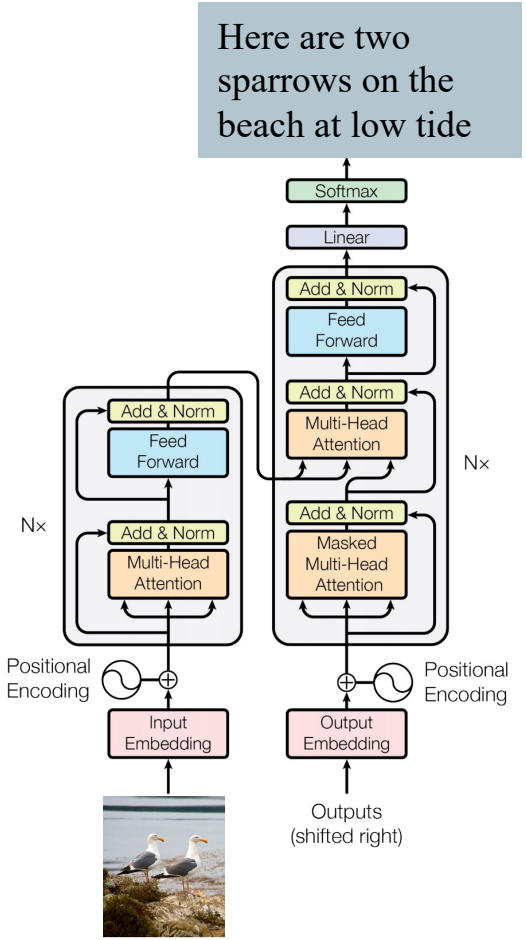
From Image to sentences!



Outline

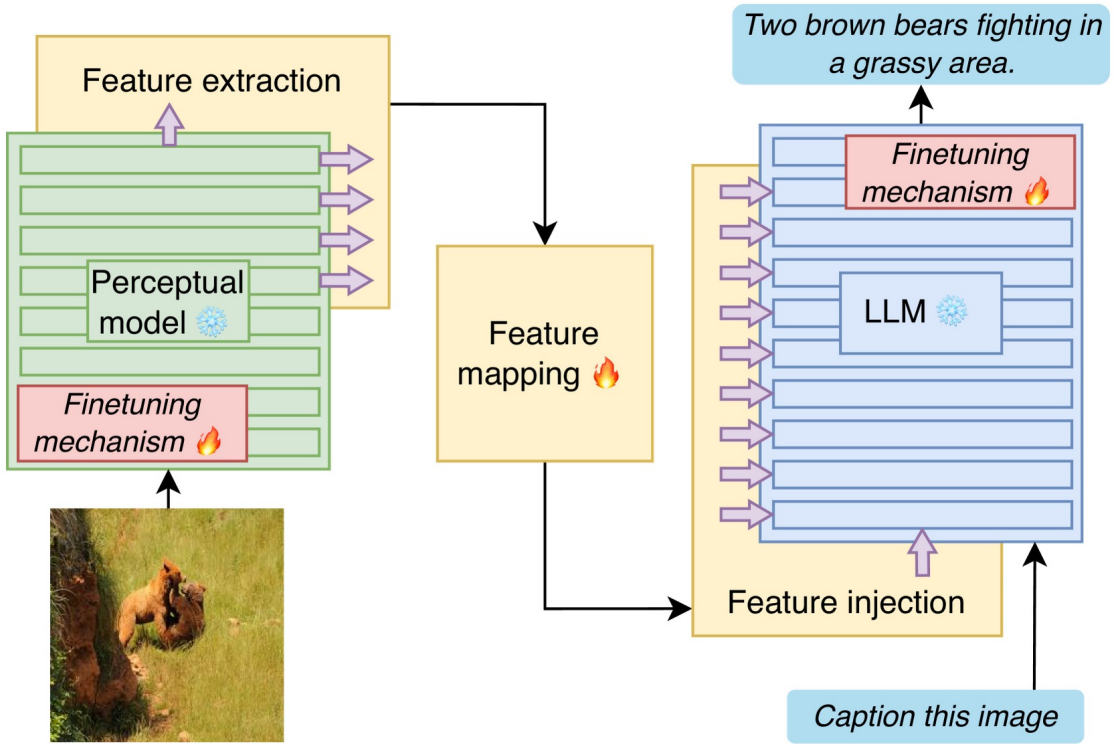
1. Vision transformers: the basics
2. From classification to detection, segmentation, …
3. **Vision-Language Models in the era of LLMs**
   - Unimodal models with connection
   - *One model for all*

Let's restrict the problem here to Visual (+Language) as input and Language as output => Image captioning, VQA



Vision Encoder + LLM Decoder

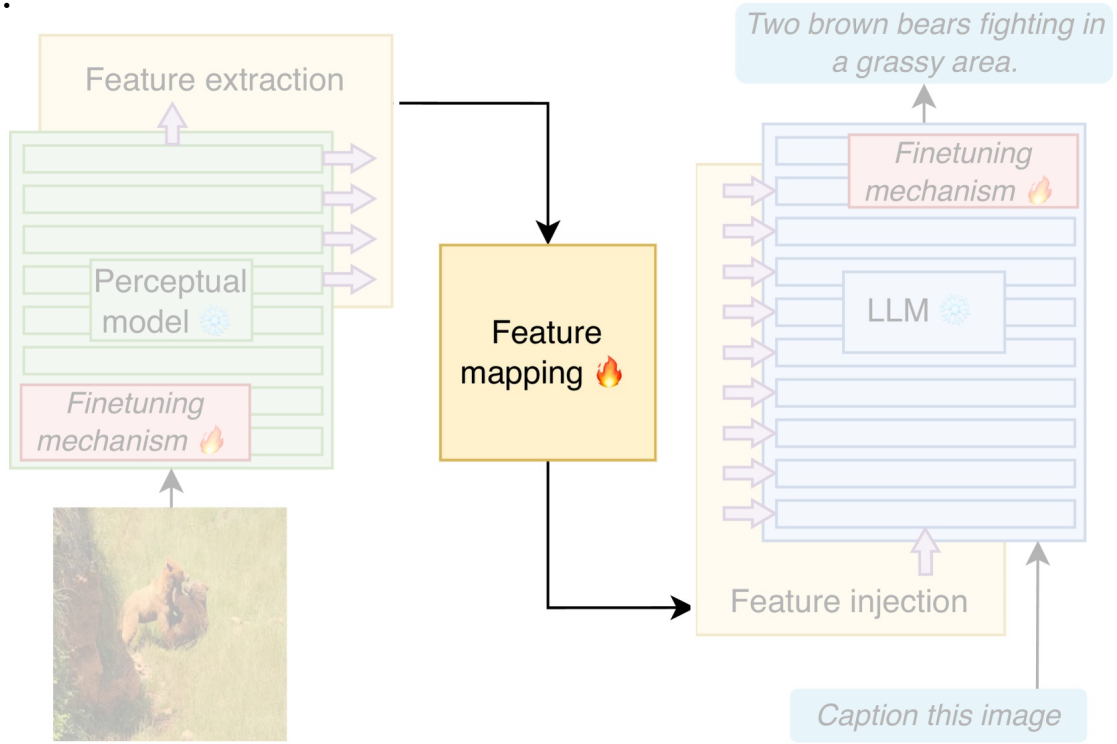
Image as input, textual caption as output



Why this modeling? Because the best LLM ever designed (and the plug&play update if a new LLM is released)

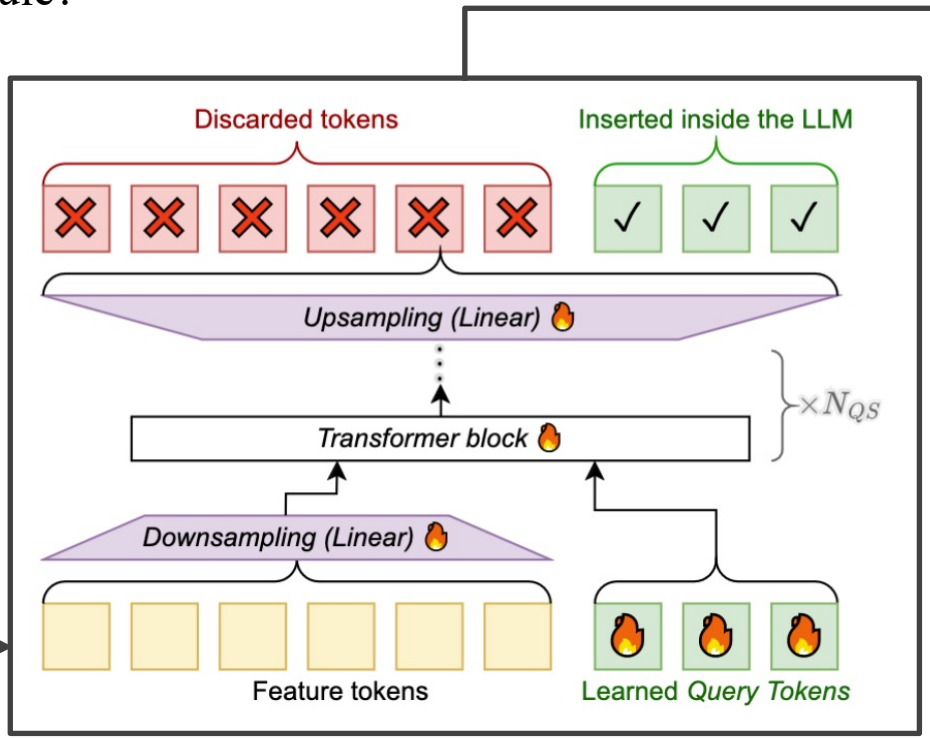
Vision Encoder + LLM Decoder

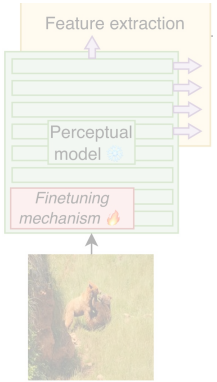
Feature mapping module?
A classic MLP, or:

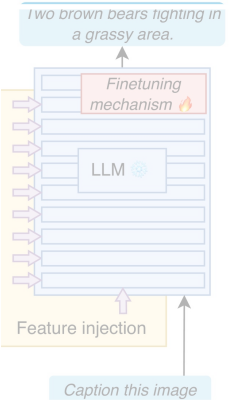


Vision Encoder + LLM Decoder

Feature mapping module?
A classic MLP,
Or CaiT-like:



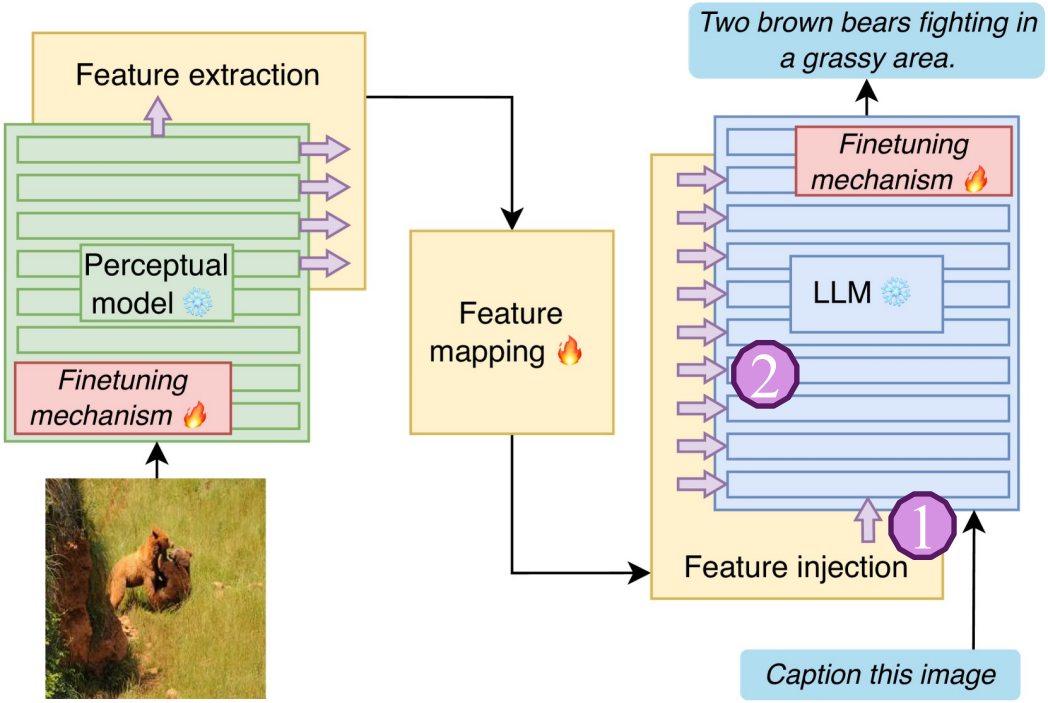


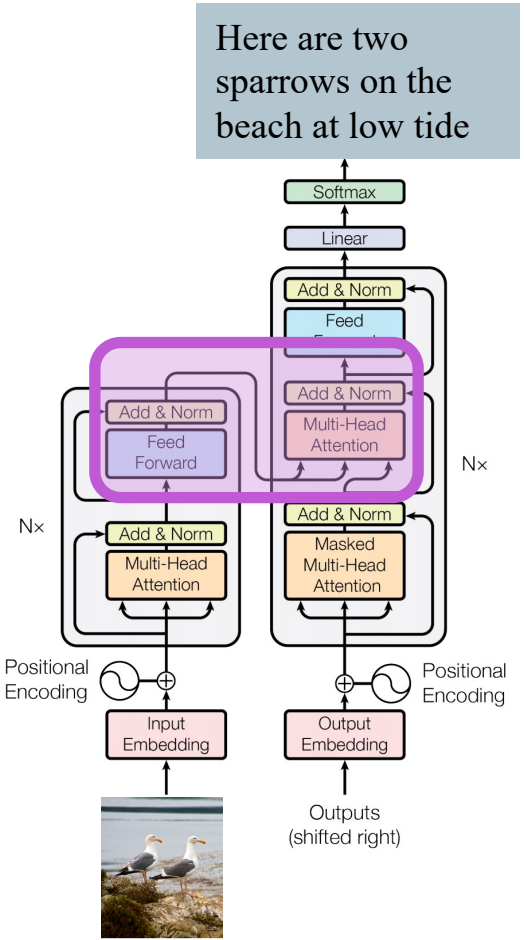


Improved Baselines for **D**ata-**e**fficient **P**erceptual **A**ugmentation of L**LM**s (DePALM)
Théophane Vallaeys, Mustafa Shukor, Matthieu Cord, Jakob Verbeek arXiv 2403.13499

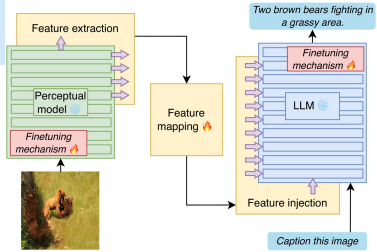
Vision Encoder + LLM Decoder

After feature mapping, feature **injection**!





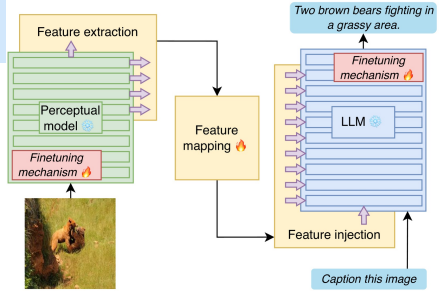
Vision Encoder + LLM Decoder



Improved Baselines for **D**ata-**e**fficient **P**erceptual **A**ugmentation of **LLM**s (DePALM)
Théophane Vallaeys, Mustafa Shukor, Matthieu Cord, Jakob Verbeek arXiv 2403.13499

Method	Backbones		Adaptation mechanism				# Tr.
	LLMs	Perceptual Enc.	Feature extraction	Feature mapping	Feature injection	Fine-tuning mechanisms	params.
Flamingo [1]	Chinchilla [33]	NFNet [5]	Tokens from last layer	Perceiver Resampler (Transformer)	GATED XATTN-DENSE (Cross-attention)	–	10B
BLIP-2 [43]	OPT [92], FlanT5 [13]	CLIP [65]	Tokens from last layer	Q-Former	1st layer token injection	–	1.2B
MAGMA [22]	GPT-J 6B [86]	CLIP [65] / NFNet [5]	Tokens from last layer	MLP	1st layer token injection	fine-tuning of perceptual model	243M
MAPL [58]	GPT-J 6B [86]	CLIP-L [65]	Tokens from last layer	QPMapper (d_{embed}=256, 4 layers)	1st layer token injection	–	3.4M
PromptFuse [46]	BART [42]	ViT [19]	Tokens from last layer	*nothing*	–	prompt tuning	15K
LiMBeR [60]	GTP-J 6B [86]	CLIP [65]	Tokens from last layer	Linear projection	1st layer token injection	–	12.5M
eP-ALM [72]	OPT-2.7B/6.7B [92]	ViT [77], AST [27], TimeSformer [4]	CLS tokens from n last layers	(Shared) linear projection	Token injection in intermediate layers	prompt tuning	4.2M
LLaMA-Adapter [25, 91]	LLaMA[82]	CLIP [65]	Tokens from last layer	Linear projection	Token injection in intermediate layers	inner-layer prompt tuning, bias tuning, norm tuning	14M
Frozen [84]	GPT-like [66]	NFNet [5]	Pooled output tokens	*nothing*	1st layer token injection	Fine-tune the NFNet	40.3M
ClipCap [61]	GPT-2[66]	CLIP [65]	Tokens from last layer	Transformer	1st layer token injection	–	43M
VL-Adapter [79]	BART [42], T5 [67]	CLIP [65]	Tokens from last layer	Linear projection	1st layer token injection	Adapters	5.8M
AnyMAL [62]	Llama 2-70B-chat [83],	CLIP [65], CLAP [23]	Tokens from last layer	Perceiver Resampler, or linear projection	1st layer token injection	LoRA [34]	–
DePALM$^{\text{QP,inner}}$	OPT-6.7B [92], LLaMA [82]	CLIP-L [65], DINOv2 [63], MAViL [36] TimeSformer [4]	Tokens from n last layers	QPMapper	Token injection in intermediate layers	prompt tuning	18.1M
DePALM							17.9M
DePALM$^{\text{R-rand,L0}}$, DePALM$^{\text{R-linear,L0}}$, DePALM$^{\text{R-QPMapper,L0}}$, DePALM$^{\text{R-avgpool,L0}}$			Tokens from last layer	Linear projection + Resampler	1st layer token injection		21M, 88M 18M, 21M
DePALM$^{\text{c-attn}}$			Tokens from n last layers	Projection + Small Transformer	Gated cross-attention		17.9M

Vision Encoder + LLM Decoder



Improved Baselines for **D**ata-**e**fficient **P**erceptual **A**ugmentation of **LLM**s (DePALM)
Théophane Vallaeys, Mustafa Shukor, Matthieu Cord, Jakob Verbeek arXiv 2403.13499

Take-home messages:

Parameter efficient approaches:
Leave the LLM and backbone frozen,
Train the mapping on (very) limited training sets to obtain very good results

Published as a conference paper at ICCV 2023

eP-ALM: Efficient Perceptual Augmentation of LMs

Mustafa Shukor, Corentin Dancette, Matthieu Cord

Simple design choices works best!
ie. passing all perceptual tokens at the input to the LLM

For efficiency DePALM mechanism:
compress perceptual to a few “summary tokens”
4 times faster to train and on par results

Vision Language Models

How to get the best VLM?
- Relax efficiency constraint
- Train on huge multimodal dataset

Joint work with my PhD student Hugo Laurençon (collab Hugging Face)

Published as a conference paper at NeurIPS 2023

OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents

Hugo Laurençon, ..., Matthieu Cord, Victor Sanh

Image-Text Pairs

Tottenham vs Chelsea Live Streaming

Tottenham Spurs vs Chelsea Live Streaming

Multimodal Document

The match between Tottenham Spurs vs Chelsea will kick off from 16:30 at Tottenham Hotspur Stadium, London.

The derby had been played 54 times and the Blues have dominated the Spurs. Out of 54 matches played, Chelsea has won 28 times and Spurs had only won 7 times. The remaining 19 matches had ended in draw.

However, in recent 5 meetings, Spurs had won 3 times where Chelsea had won the other two times. …

Dataset	Images	% unique images	Docs	Tokens	Open
KOSMOS-1	-	-	71M	-	✗
M3W	185M	-	43M	-	✗
mmc4-ff	385M	60.6%	79M	34B	✓
mmc4	**585M**	-	103M	43B	✓
OBELICS	353M	**84.3%**	**141M**	**115B**	✓

Table 1: General statistics of OBELICS and the current largest alternatives.

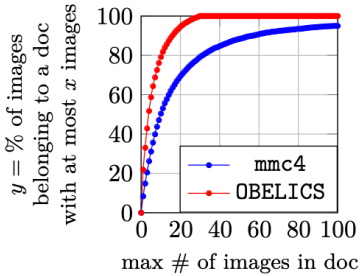


Figure 3: Distribution of images.

Vision Language Models

What is the best VLM and What matters when building vision-language models?
[**IDEFICS** arXiv 2024]

What matters when building vision-language models?

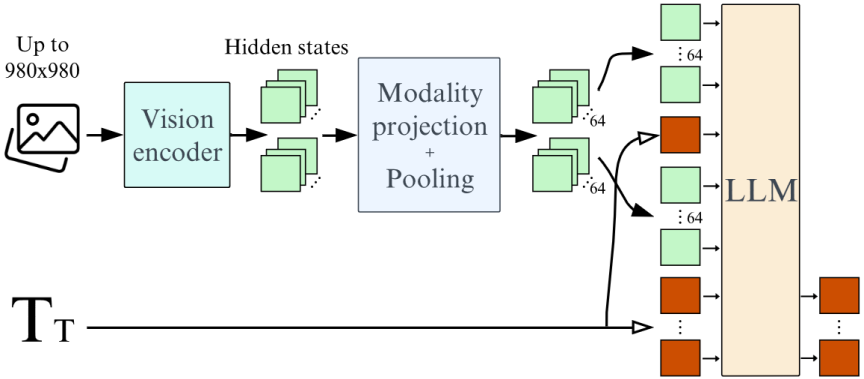
Hugo Laurençon[*,1,2] **Léo Tronchon**[*,1] **Matthieu Cord**[2] **Victor Sanh**[1]

*The order was chosen randomly.

[1]Hugging Face [2]Sorbonne Université

Model	Size	# tokens per image	MMMU	MathVista	TextVQA	MMBench
LLaVA-NeXT	13B	2880	36.2/-	35.3	67.1	70.0
DeepSeek-VL	7B	576	36.6/-	36.1	64.4	73.2
MM1-Chat	7B	720	37.0/35.6	35.9	72.8	72.3
Idefics2	8B	**64**	**43.5/37.9**	**51.6**	70.4	**76.8**
Idefics2	8B	320	43.0/37.7	51.4	**73.0**	76.7

Table 9: Performance of Idefics2 against state-of-the-art VLMs up to a size of 14B parameters. The evaluations are done in zero shot. Idefics2 with 64 or 320 tokens per image is the same model (same weights), only the inference differs. The full table is present in Appendix A.3.2.
(Benchmark, Split, Metric): (MMMU, val/test, MMMU score), (MathVista, testmini, MMMU score), (TextVQA, val, VQA acc.), (MMBench, test, accuracy).



Idefics2 fully-autoregressive (FA) architecture:

Vision encoder

Mapping to the LLM input space

Visual tokens (64 in our standard configuration, green) interleaved with the input sequence of text embeddings

LLM to predict the text tokens output

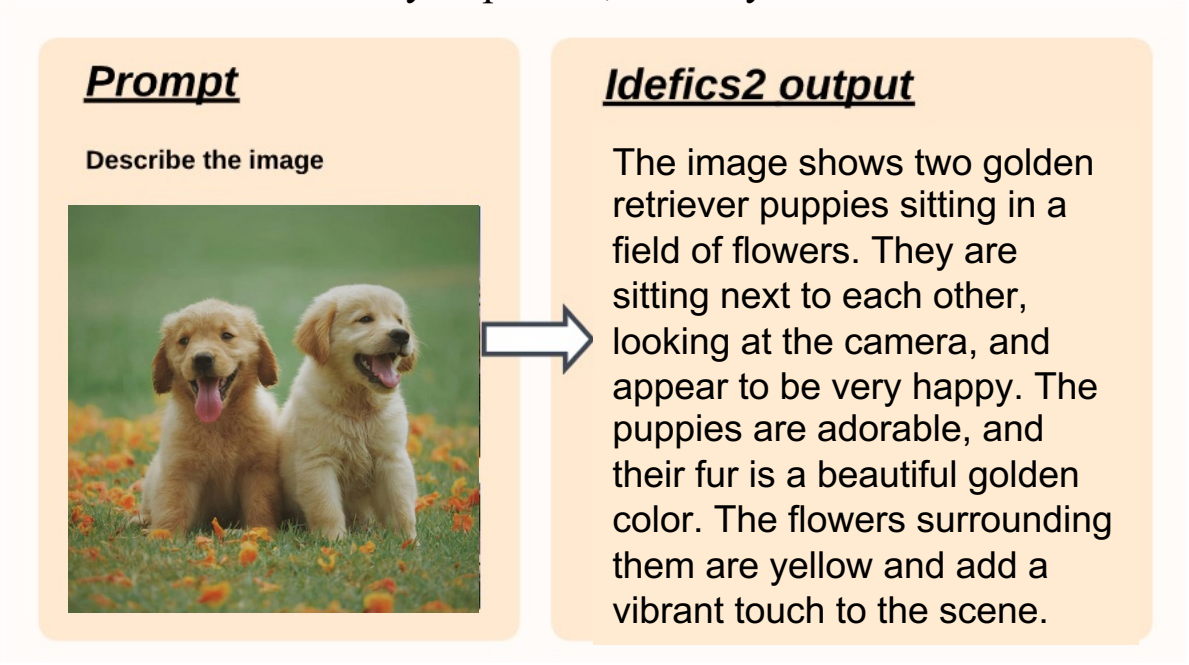
Vision Language Models

[**IDEFICS** arXiv 2024]
Quantitative results:

Model	Size	Archi.	# tokens per image	VQAv2	TextVQA	OKVQA	COCO
OpenFlamingo	9B	CA	-	54.8	29.1	41.1	96.3
Idefics1	9B	CA	-	56.4	27.5	47.7	97.0
Flamingo	9B	CA	-	58.0	33.6	50.0	99.0
MM1	7B	FA	144	63.6	46.3	51.4	**116.3**
Idefics2-base	8B	FA	**64**	**70.3**	**57.9**	**54.6**	116.0

Evaluation very important, not easy for Generative models

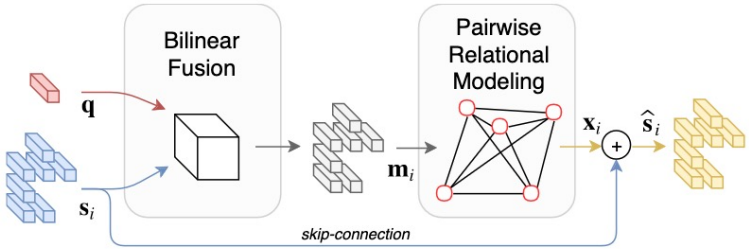
Qualitative results:

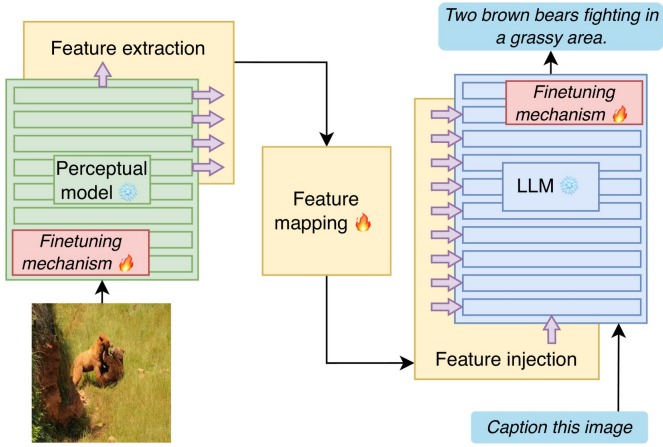


Conclusion / Perspectives

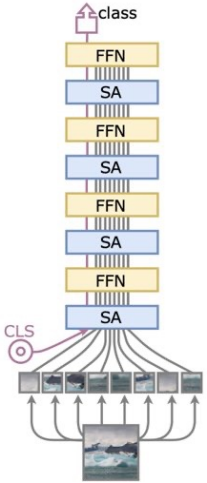
Multimodal (LMM) is the new “thing” (from NeurIPS 2023) but:

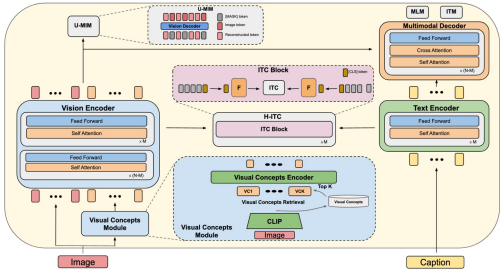
- Architectures/models
  - Transformers: the end?
  - Vision-Language interaction/representation

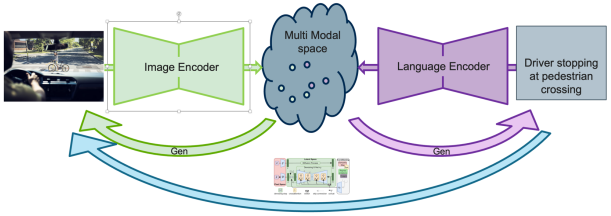




- Learning LMM: data, loss, optim., evaluation, generalization …
  - *Alexandre Ramé PhD*
- Compression/embedded
  - *Edouard Yvinec PhD*







"A mouse wearing a hat in the desert."

Thanks to my (PhD) students Asya Grechka, Mustafa Shukor, Hugo Laurençon, Guillaume Couairon, Alexandre Ramé, Corentin Dancette, Hugo Touvron, Arthur Douillard,Théophane Vallaeys

Matthieu Cord
Prof. at Sorbonne University, ISIR lab.
Scientific Director of Valeo.ai

scai
SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

valeo.ai

SCIENCES SORBONNE UNIVERSITÉ