

Multimodal Machine Learning

Visual Language Models: Text & Image

Steffen Jauch-Walser

21st of July, 2022

Multimodality Seminar - LMU Munich

Table of contents

1. Motivation
2. Data2Vec
3. ViLBert
4. Flamingo
5. Outlook
6. References

- Topic 4 Image to Text
- Topic 5 Text to Image
- Topic 6 Image Supporting Language Models
- Topic 7 Text Supporting CV Models
- **Topic 8 Text + Image**
- Chapter 3 Further topics

Challenges in AI

- Data
 - need more data, but labelling data prohibitively costly
 - model training takes too much time
- Too many models
- Knowledge transfer across tasks

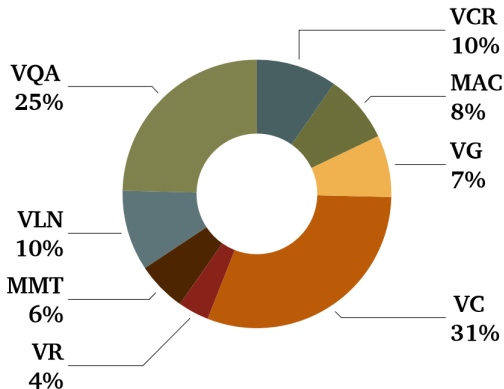
Is there a general model?

Visual Language Tasks

- Visual Question Answering (VQA)
- Visual Captioning (VC)
- Visual Common Sense Reasoning (VCR)
- Visual Language Navigation (VLN)
- Multimodal Affective Computing (MAC)
- Multimodal Machine Translation (MMT)

Visual-Language Tasks Trends

Trends in VisLang Research



Uppal et al. (2022): VisLang Paper Trends (previous 2 years)

- developed by Meta AI
- one baseline model that is able to deal with three modalities: speech, vision, language
- no cross modal interactions
- self-supervised transformer model with modality specific encoding
- borrows from earlier literature
- continuous and contextualized representations as prediction targets

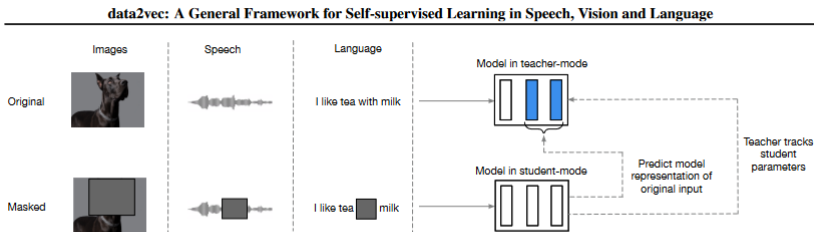


Figure 1. Illustration of how data2vec follows the same learning process for different modalities. The model first produces representations of the original input example (teacher mode) which are then regressed by the same model based on a masked version of the input. The teacher parameters are an exponentially moving average of the student weights. The student predicts the average of K network layers of the teacher (shaded in blue).

Baevski et al. (2022)

Data2Vec: Prediction Targets

Denote by a_t^l the output of block l at timestep t , then the training targets y_t are given by

$$y_t = \frac{1}{K} \sum_{l=L-K+1}^L \hat{a}_t^l$$

- normalized outputs \hat{a}_t^l averaged across the top K blocks
- contextualized targets based on complete input data

$$L(y_t, f_t(x)) = \begin{cases} \frac{(y_t - f_t(x))^2}{\beta} & \text{if } |(y_t - f_t(x))| \leq \beta \\ |y_t - f_t(x)| - \frac{\beta}{2} & \text{otherwise} \end{cases}$$

- smooth L_1 loss
- less sensitive to outliers, but β needs tuning

Data2Vec: Masking and Encoding

- student model only sees masked version of the input and predicts y_t
- the masking is modality specific and learned
- teacher model is updated frequently at the start and slowly later on (exponentially moving average)

- built on the BERT re-implementation RoBERTa
- byte-pair encodings used to tokenize input as sub-words (50K types)
- 15% of uniformly selected tokens:
 - 80% are replaced by a learned mask token,
 - 10% are left unchanged and
 - 10% are replaced by randomly selected vocabulary token

Data2Vec: Masking and Encoding — Vision

- images of 224x224 pixels embedded as patches of 16x16 pixels
- linearly transformed into a sequence of 196 representations
- they mask blocks of multiple adjacent patches which contain at least 16 patches with random aspect ratio
- 60% of patches mask
- image crops, horizontal flipping, and color jittering

1. modality specific encoding (features and position)
2. modality specific masking and prediction target creation (self-supervised)
3. modality agnostic student model training process

The underlying architecture follows a standard transformer architecture approach using self-attention.

Data2Vec: Transformers

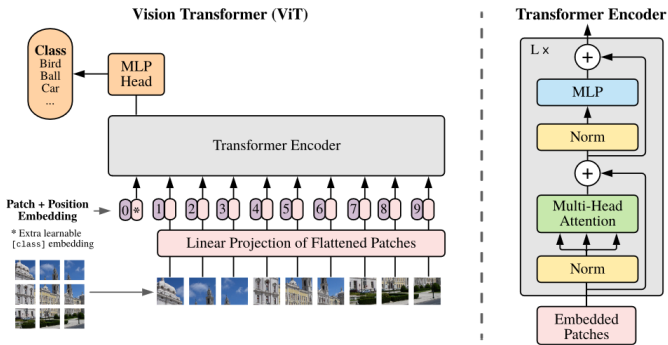


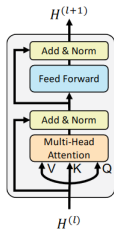
Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable “classification token” to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

- state-of-the-art performance on common benchmarks
- Is data2vec really a multimodal model?

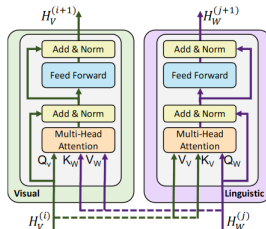
Can we find ways to leverage
modality agnostic encoding or cross modality inputs?

- popular, earlier model
- cross modal version of Bert
- introduces co-attention

Should we use a single stream to operate on both visual and language input or two separate streams?



(a) Standard encoder transformer block



(b) Our co-attention transformer layer

Figure 2: We introduce a novel co-attention mechanism based on the transformer architecture. By exchanging key-value pairs in multi-headed attention, this structure enables vision-attended language features to be incorporated into visual representations (and vice versa).

Lu et al. (2019)

- co-attentional transformer layers
- keys, queries, values are input into the other modality's multi-head attention block
- attention-pooled features for each modality conditioned on the other
- image-conditioned language attention
- language-conditioned image attention

- masking similar to Bert (15%)
- training: masked multi-modal modelling and alignment
- model predicts distribution over semantic classes for corresponding image regions

Table 1: Transfer task results for our ViLBERT model compared with existing state-of-the-art and sensible architectural ablations. [†] indicates models without pretraining on Conceptual Captions. For VCR and VQA which have private test sets, we report test results (in parentheses) only for our full model. Our full ViLBERT model outperforms task-specific state-of-the-art models across all tasks.

Method	VQA [8]	VCR [25]			RefCOCO+ [39]			Image Retrieval [26]			ZS Image Retrieval		
	test-dev (test-std)	Q→A	QA→R	Q→AR	val	testA	testB	R1	R5	R10	R1	R5	R10
SOTA	DFAF [36]	70.22 (70.34)	-	-	-	-	-	-	-	-	-	-	-
	R2C [25]	-	63.8 (65.1)	67.2 (67.3)	43.1 (44.0)	-	-	-	-	-	-	-	-
	MAttNet [33]	-	-	-	-	65.33	71.62	56.02	-	-	-	-	-
	SCAN [35]	-	-	-	-	-	-	-	48.60	77.70	85.20	-	-
Ours	Single-Stream [†]	65.90	68.15	68.89	47.27	65.64	72.02	56.04	-	-	-	-	-
	Single-Stream	68.85	71.09	73.93	52.73	69.21	75.32	61.02	-	-	-	-	-
	ViLBERT [†]	68.93	69.26	71.01	49.48	68.61	75.97	58.44	45.50	76.78	85.02	0.00	0.00
	ViLBERT	70.55 (70.92)	72.42 (73.3)	74.47 (74.6)	54.04 (54.8)	72.34	78.52	62.61	58.20	84.90	91.52	31.86	61.12

Lu et al. (2019)


- pretraining and finetuning helps, more pretraining helps more
- handling vislang inputs separately improves over single streams
- mixed results on the effects of model depth
 - VQA seems to prefer a layer depth of 6
 - zero shot image retrieval prefers greater depth
 - VCR and RefCoCo seem to prefer shallower networks

Can we do better? Less finetuning? More open-ended tasks?

- developed by Google (DeepMind)
- few shot learning model: how can we quickly adapt to new tasks?
- state-of-the-art on open-ended tasks by prompting the model with task-specific examples
- architectural innovation
 - bridging vision- and language-only models
 - handle sequences of arbitrarily interleaved visual and textual data
- 80 billion parameters

Flamingo

	pandas: 3		dogs: 2		→	giraffes: 4	
I like reading			, my favourite play is Hamlet. I also like			→	Dreams from my Father.
						→	he falls down.
What happens to the man after hitting the ball? Answer:							



This is a picture of two teddy bears on the moon.

What are they doing?

They are having a conversation.

What object are they using?




It looks like a computer.

Is this surprising?

Yes, it is surprising.

Why is this picture surprising to you?

I think it is surprising because teddy bears are not usually found on the moon.






What is the common thing about these three images?

They are all flamingos.

What is the difference between these three images?

The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.



This is an apple with a sticker on it.

What does the sticker say?

The sticker says "iPod".

Where is the photo taken?

It looks like it's taken in a backyard.

Do you think it is printed or handwritten?

It looks like it's handwritten.

What color is the sticker?

It's white.

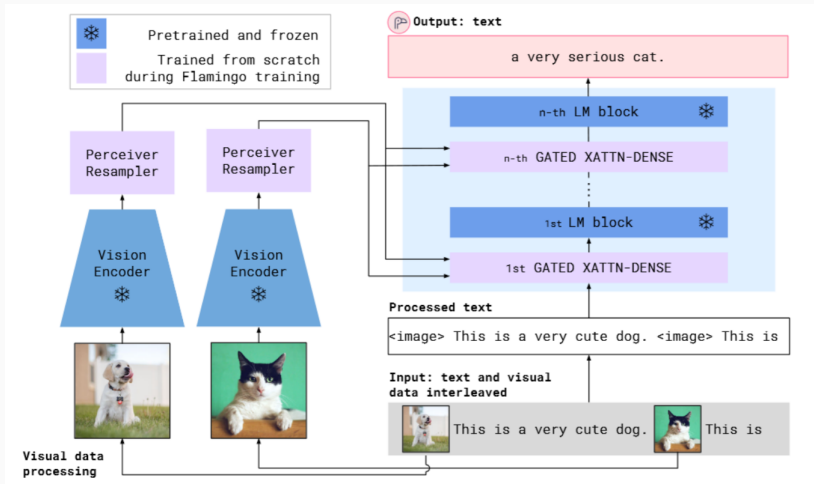
Idea:

- large language models (LM) are great at generating expressive language
- can we combine them with a visual model without huge training effort?

Solution:

- pre-train both models and freeze them
- add a perceiver resampler and cross-attention layers as bridge

Flamingo



Flamingo Architecture: Alayrac et al. (2022)

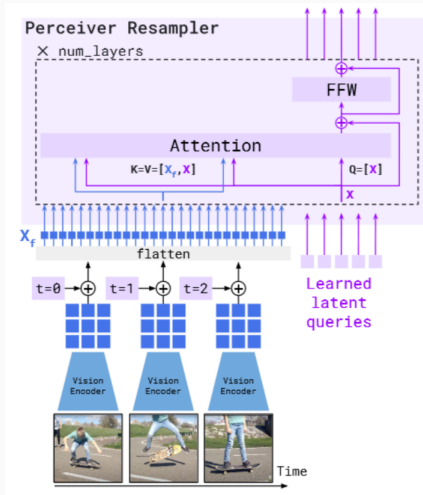
Flamingo: Approach

The vision encoder

- is pretrained similar to the Clip model
- i.e. uses a contrastive text-image approach
- aims to extract colour, shape, nature, position of objects, ...
(queries)

The language model gives Flamingo rich text generation capabilities.
Both models are frozen to avoid retraining and to ensure functionality.

Flamingo: Resampler



Alayrac et al. (2022)

Flamingo: Gated Cross-Attention

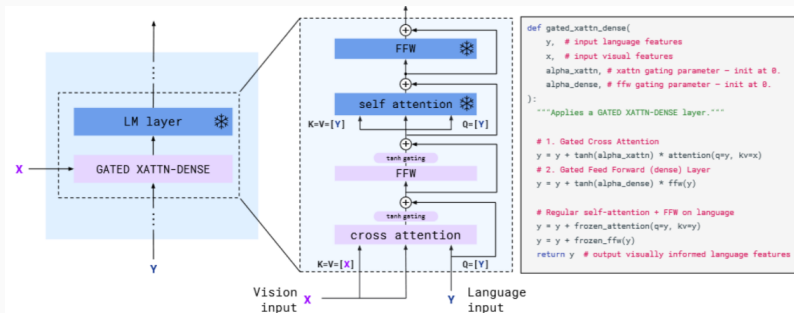


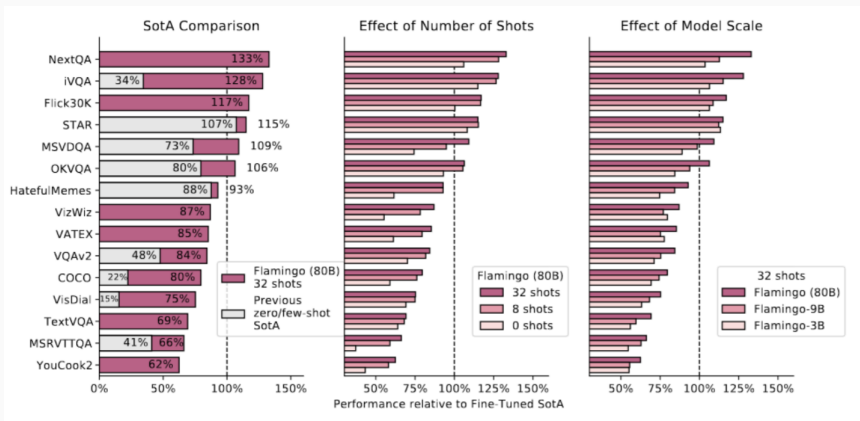
Figure 5 | **GATED XATTN-DENSE layers**. We insert new cross-attention layers, whose keys and values are obtained from the vision features while using language queries, followed by dense feed forward layers in between existing pretrained and frozen LM layers in order to condition the LM on visual inputs. These layers are *gated* so that the LM is kept intact at initialization for improved stability and performance.

XATTN at certain depth frequencies. Alayrac et al. (2022)

Flamingo: Datasets

- trained solely on task agnostic web scraped data
- three types of training data
- text-image interleaved, text-image pairs, video-image pairs
- no training sets specifically designed for machine learning
- data from 43 Million webpages
- importantly, the weigh the loss from different datasets
- gradient accumulation over datasets for high performance

Flamingo: Performance



Influence of Model Scale and Number of Shots. Alayrac et al. (2022)

Important Concepts:

- contextualization
- diverse attention mechanisms
- pretraining and model freezing
- data set weighting
- perceiver resampler

- bigger, more general and adaptive models
- communication between pretrained models rather than complexity
- leveraging huge data through easy training
- exciting tasks
- prohibitive resource constraints
- two big players (Meta and Google)

References

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Baevski, Alexei, et al. "Data2vec: A general framework for self-supervised learning in speech, vision and language." arXiv preprint arXiv:2202.03555 (2022).

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." Advances in neural information processing systems 32 (2019).

Uppal, Shagun, et al. "Multimodal research in vision and language: A review of current and emerging trends." Information Fusion 77 (2022): 149-171.

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).