

Vision and Language: A Primer

<https://elliottd.github.io/vlprimer/>



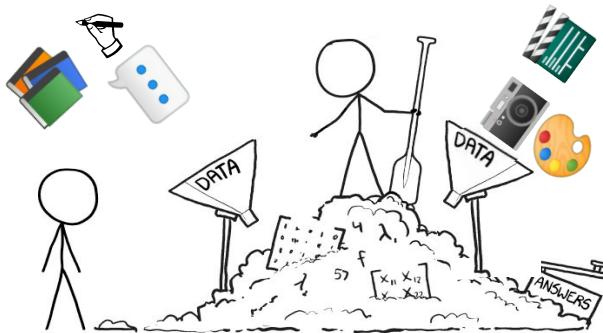
Desmond Elliott

Department of Computer Science
University of Copenhagen

 @delliott

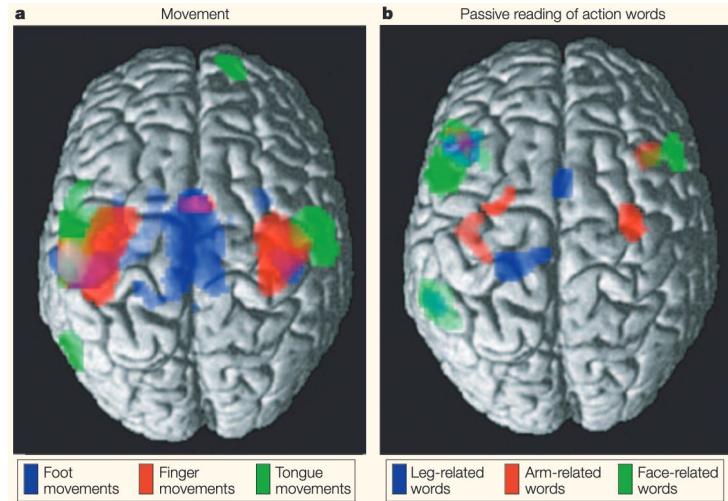
Working Definition

Multimodal models jointly processes information from two or more input modalities, e.g. images and text, speech and video, etc.



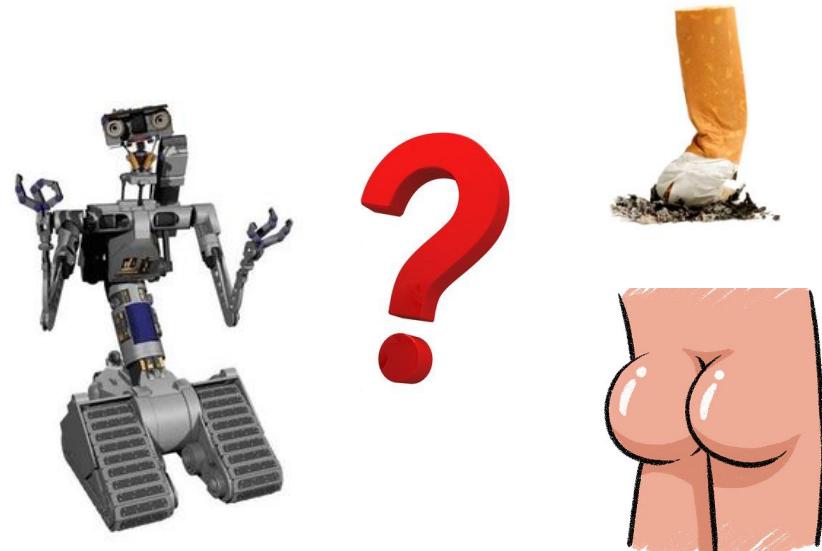
Why Vision and Language?

- Humans ground conceptual knowledge in modality processing systems in the brain
- Evidence that grounding activates similar brain regions for different input modalities



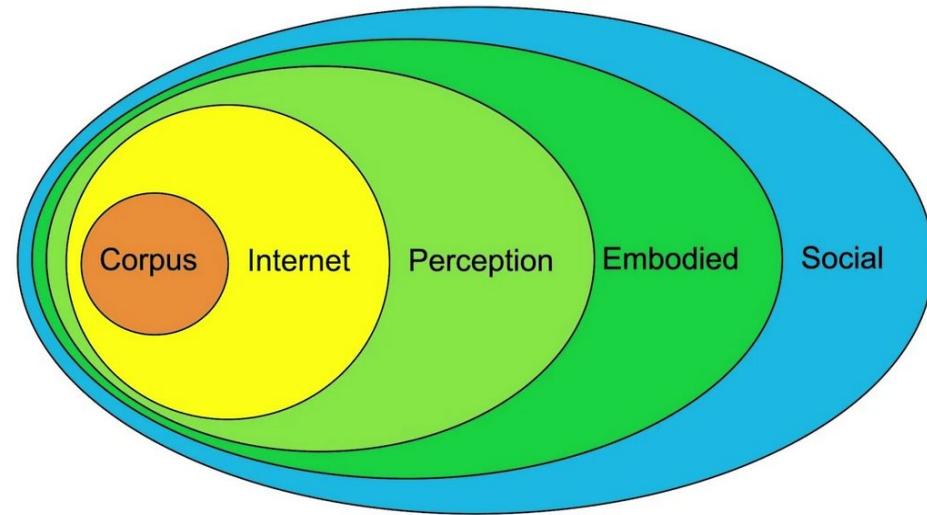
Barsalou et al. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2):84–91.
Pulvermüller. (2005). Brain mechanisms linking language and action. *Nature reviews neuroscience*, 6(7), 576–582.

Multimodality reduces ambiguity



You Cannot Learn Language From

- The radio without grounding
(lack perception)
- The television without actions
(lack embodiment)
- Without interacting with others
(lack social)

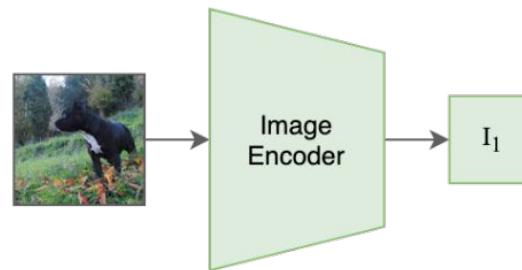
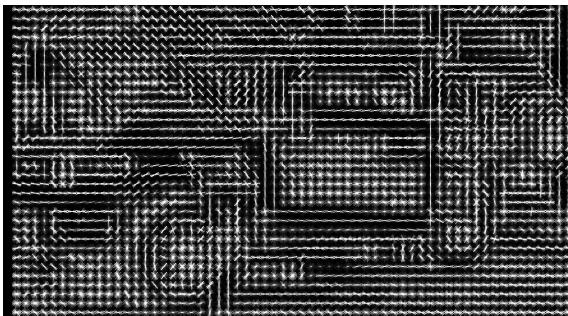


(At Least) Five Major Areas

- **Representation:** how to convert raw inputs into a usable format
- **Translation:** transform from one modality to another
- **Alignment:** predict relationships between elements across modalities
- **Fusion:** join features from modalities to support prediction
- **Co-learning:** transferring knowledge from one modality to another

Representation

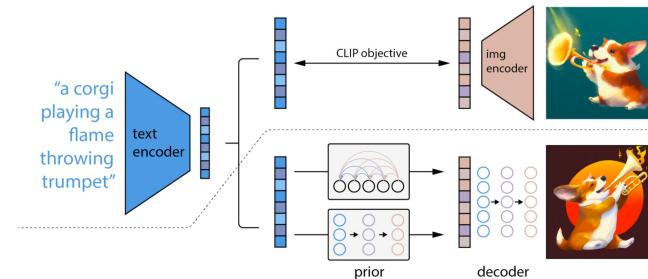
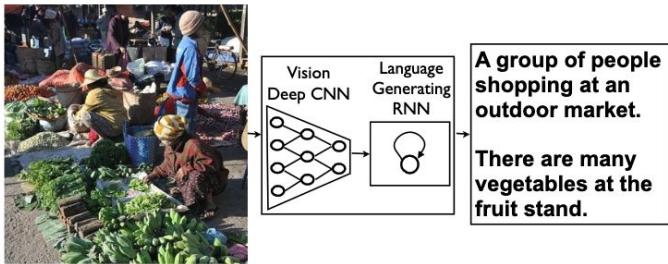
- Great deal of work over the last decade, from HOG features in the early 2000s to CLIP features in the 2020s.



Dalal & Triggs. (2005). Histograms of oriented gradients for human detection. CVPR
Radford et al. (2021) Learning transferable visual models from natural language supervision. ICML.

Translation

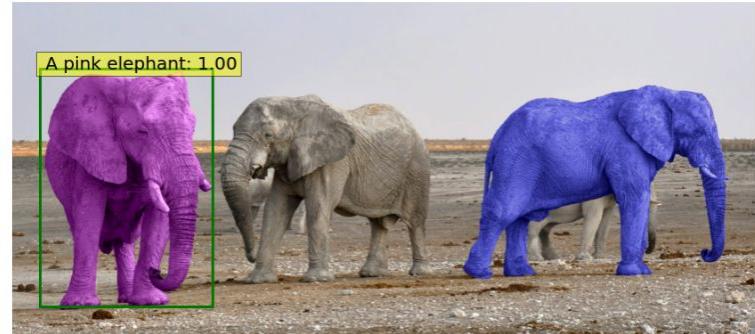
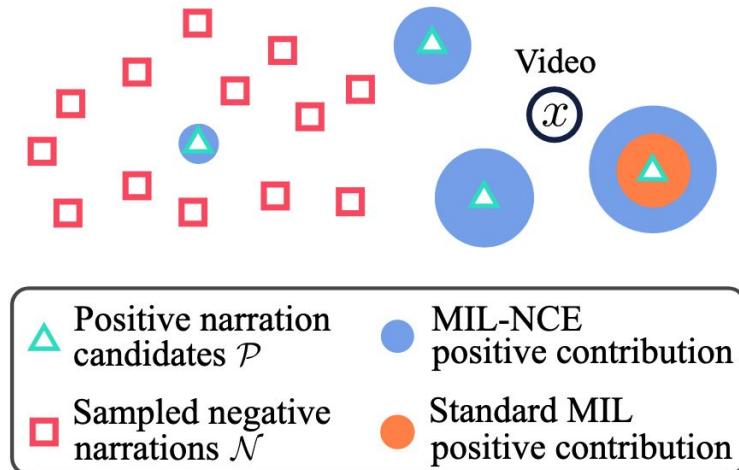
- Explosion of end-to-end neural network models since the mid 2010s



Vinyals et al. (2015). Show and tell: A neural image caption generator. CVPR.
Ramesh et al. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv.

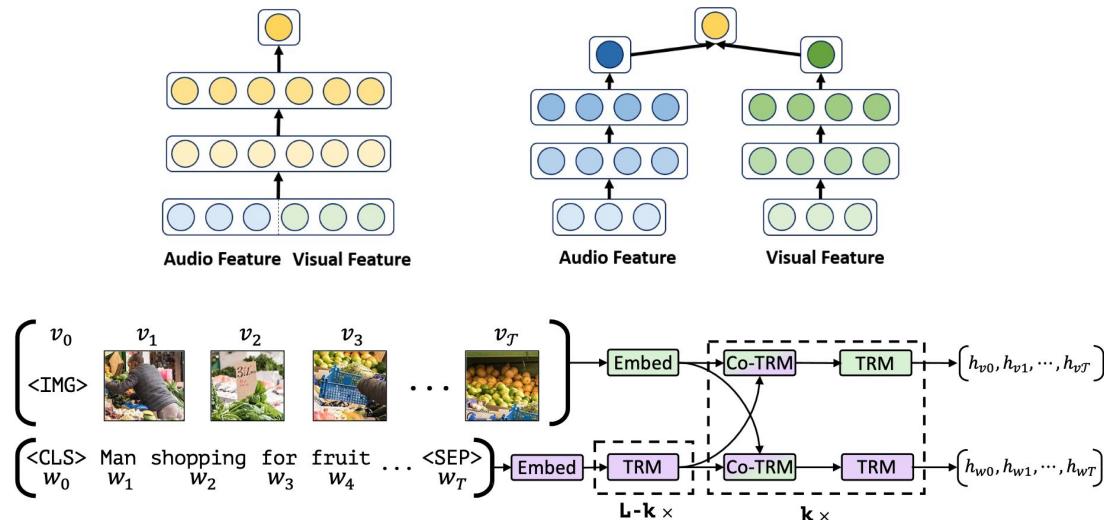
Alignment

- Important for self-supervised learning and also for phrase grounding



Fusion

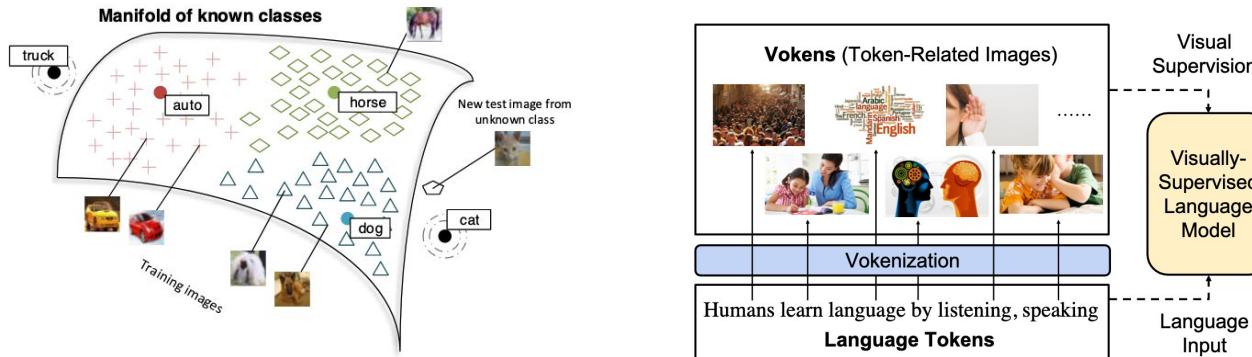
- Early work studied the differences between early and late fusion.
- Multi-head self-attention now provides model-based fusion.



Chen and Jin (2016). Multi-modal conditional attention fusion for dimensional emotion prediction. MM.
Lu et al. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NeurIPS.

Co-learning

- Zero-shot transfer across modalities, or using visual grounding to improve language models on text-only tasks.



Socher et al. (2013). Zero-shot learning through cross-modal transfer. NeurIPS.
Tan & Bansal. (2020). Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. EMNLP

Roadmap

1. Datasets for Multimodal Learning

 Visually Grounded Reasoning across Languages and Cultures (EMNLP 2021)

2. Data Representation

3. Modelling Techniques

 Retrieval-Augmentation in Image Captioning Image Captioning (EACL 2023, CVPR 2023, fACL 2023)

4. Understanding Multimodal Models

5. Future Directions

 Language Modelling with Pixels (ICLR 2023)

1. Datasets for Multimodal Learning

Anatomy of a Multimodal Dataset

- Datasets are typically either crowdsourced or harvested from the web
- Consist of two/three major parts:
 - Non-linguistic stimuli
 - Images or videos
 - Linguistic stimuli
 - Text or speech
 - Task labels (optional)



What color is the cat's leash?

purple

Two Types of Dataset

- **General-purpose:** visual data with descriptive annotations

- Conceptual Captions
- LAION-2/5B
- Speech-COCO



Blue Beach
Umbrellas, Point
Of Rocks, Crescent
Beach, Siesta Key -
Spiral Notebook

- **Task-specific:** visual data with e.g. classification labels

- Image / Video Captioning
- Visual Question Answering
- Visually Grounded Reasoning

What color is the cat's leash?
purple red



Degree of Multimodality

Social media platforms often form 'echo chambers' that encourage users to only read content that confirms beliefs they already hold (Getty)



Weak



A woman in a dark grey suit is giving a speech



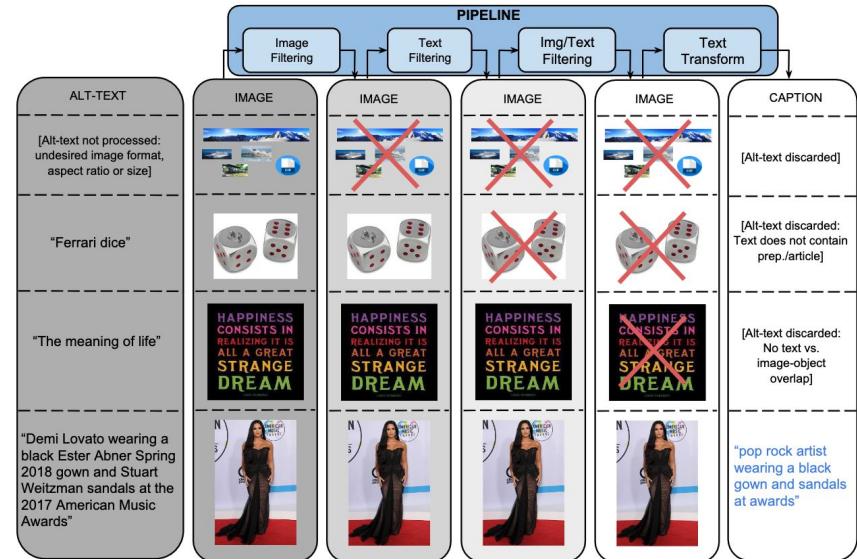
Strong

Panofsky. (1939). Studies in Iconology.

<https://www.independent.co.uk/news/angela-merkel-says-internet-search-engines-endangering-debate-algorithms-should-be-revealed-a7383811.html>

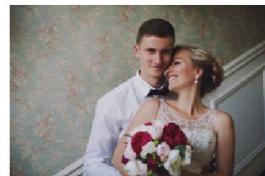
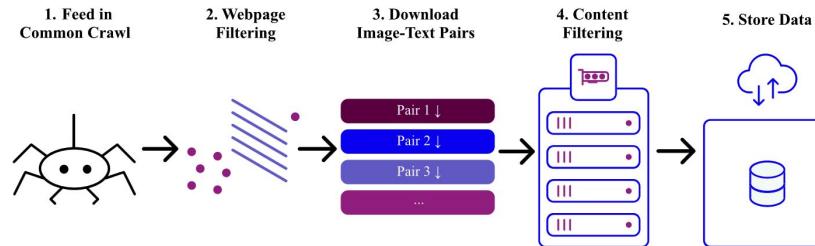
Conceptual Captions

- Used for pretraining
- 3M Images and *normalized* English captions.
- Normalization is not public.
- Due to *linkrot*, probably much less images still available.

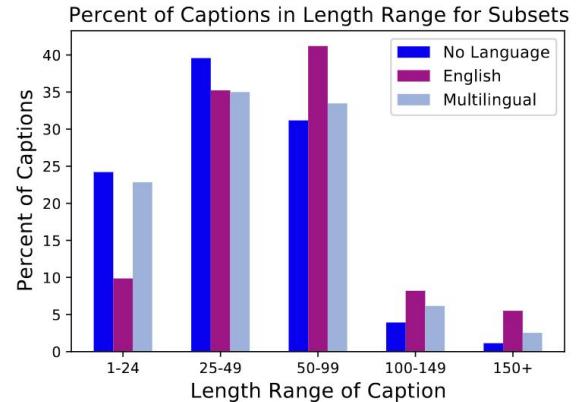


LAION

- Used for pretraining
- Image and multilingual *raw* captions harvested from within Common Crawl
- Data behind Stable Diffusion and OpenCLIP



Żeński i męski
portret Dama
outdoors i facet
Ślubna para [...]



COCO

- Used both a **general-purpose** and **task-specific** dataset
- Images covering 80 common objects in context with multiple human-authored captions.
- Object segmentation data too!

some sheep walking in the middle of a road
a herd of sheep with green markings walking down the road
a herd of sheep walking down a street next to a lush green grass covered hillside.
sheared sheep on roadway taken from vehicle, with green hillside in background.
a flock of freshly sheered sheep in the road.



VQAv2

- Answer questions about images
 - Task with multimodal inputs:
 - Image
 - Question
 - Commonly tackled as classification but increasing efforts as NLG
 - 1.1M image–question pairs with a careful effort to balance the distribution of answers

Who is wearing glasses?

man woman



Where is the child sitting?
fridge arms



NLVR2

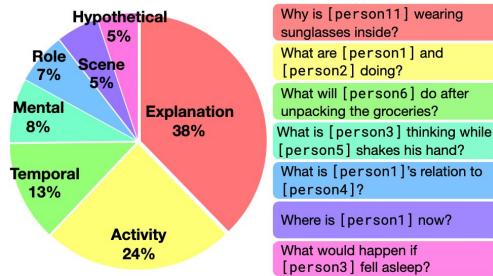
- Binary classification task that requires jointly reasoning over a pair of images and a sentence.
- Human-created hard negatives.
- 107K examples in total.



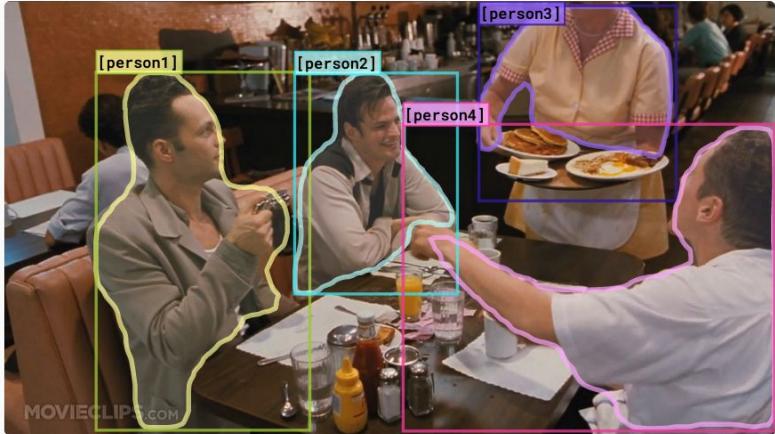
The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

Visual Commonsense Reasoning

- 290,000 multiple-choice VQA examples derived from movies.



- In addition to Question Answering, the dataset includes rationale selection too!



Why is [person4] pointing at [person1]?

- a) He is telling [person3] that [person1] ordered the pancakes.
- b) He just told a joke.
- c) He is feeling accusatory towards [person1].
- d) He is giving [person1] directions.

Rationale:

- a) [person1] has the pancakes in front of him.
- b) [person4] is taking everyone's order and asked for clarification.
- c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.
- d) [person3] is delivering food to the table, and she might not know whose order is whose.

Multi30K

- Multilingual aligned image–sentence dataset in many languages
 - English, German, French, Czech, Arabic, Japanese, Turkish, Ukrainian

A group of people are eating noodles.

Eine Gruppe von Leuten isst Nudeln.

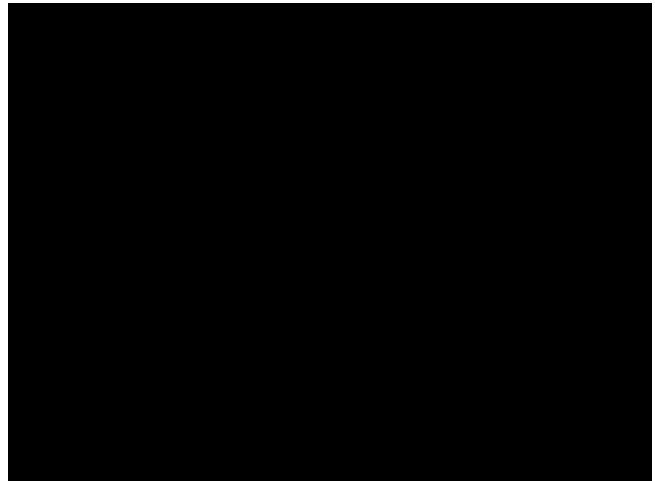
Un groupe de gens mangent des nouilles.

Skupina lidí jedí nudle.



BOBSL

- BBC-Oxford British Sign Language Dataset
- Sign spotting and sentence localization tasks
- 1,400 hours of signed shows
 - Factual, entertainment, drama, comedy, children's shows



Many Many More

- Visual Storytelling, e.g. VIST
- Grounded Referring Expression, e.g. Flickr30K Entities, Visual Genome
- Visual Entailment, e.g. SNLI-VE
- Vision & Language Navigation, e.g. RxR
- Visual Common Sense Reasoning: VCR
- Text-to-Image Generation, e.g. DALLEval
- Abstract reasoning, e.g. KiloGram, CRAFT
- Sign Language Processing, e.g. How2Sign
- *And more and more and more and more*

Ethical Issues

- Multimodal datasets are usually data scraped from the web with *unknown degrees of conformance*, or information about, licensing.



CC BY: This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use.

- As of 2022, there are an estimated 2.5B CC-licensed objects online.

Could people have reasonably expected that ***distribute*** or ***build upon*** included use for large-scale machine learning?



Image: © Neo Cali by Vektroid

A Problem with Scale

- Build multimodal systems that perpetuate harmful stereotypes



(Eileen Collins, American astronaut)

$$\cos(\cdot, \cdot)$$

0.276

This is a portrait of an astronaut
with the American flag

0.308

This is a photograph of a smiling
housewife in an orange jumpsuit with
the American flag

Pre-2022 web could be the last large-scale
source of non-AI generated multimodal data



Q: How can we collect multimodal data that better reflects the diversity of the world?

Visually Grounded Reasoning across Languages and Cultures

EMNLP 2021



F. Liu*



E. Bugliarello*



E.M. Ponti



S. Reddy



N. Collier



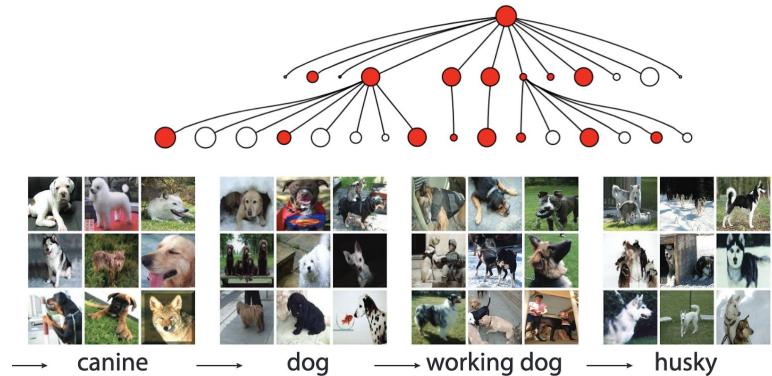
D. Elliott

Typical Vision and Language



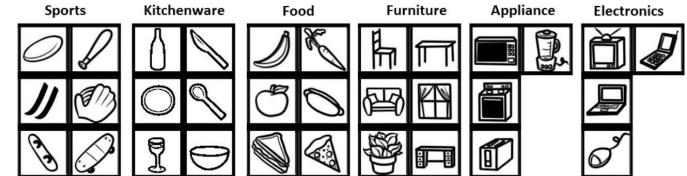
ImageNet (Deng et al. 2009)

- Train visual encoders
- Millions of labelled images
- Derived from the WordNet hierarchy



Common Objects in Context (Lin et al. 2014)

- Train and evaluate multimodal models
- 330K labelled images
 - 80 types of commonly occurring objects



Rethinking Vision and Language

DAT
A

Languages

- Mostly in English
- Or some Indo-European Languages



ENG: An **unusual** looking vehicle ...

NLD: Een mobiel **draaiorgel** ...

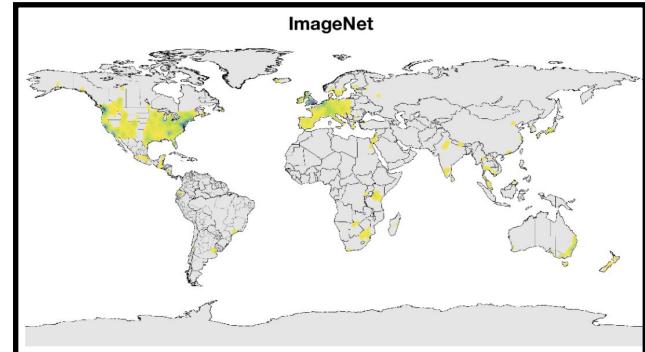
Example from [van Miltenburg+ 2017](#)

Image sources

- Mostly from ImageNet or COCO
- Reflecting North American and European cultures

Implications for V&L models

- Narrow linguistic/cultural domain
- No way to assess their real-world comprehension



Density map of geographical distribution of images in ImageNet ([DeVries+](#), 2019)

Concepts and Hierarchies

Category: objects with similar properties (Aristotle 40 BCE, ...)

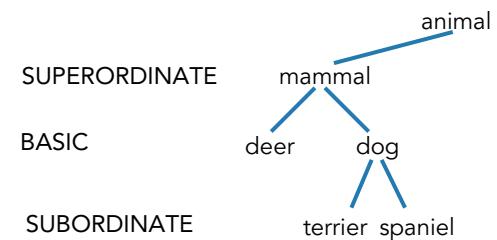
Concept: mental representation of a category (Rosch 1973)

Categories form a *hierarchy*

- Basic-level categories (Rosch 1976)

Somewhat universal

- Different cultures (Berlin 2014)
- Familiarity of individuals
(Wisniewski and Murphy, 1989)



Concrete Concepts in Cultural Context

- Some concepts are most immediately understood within a cultural background

Culture: The way of life of a collective of people that distinguishes them from other people (Mora, 2013; Shweder et al. 2007).



Pilota / Jai-alai



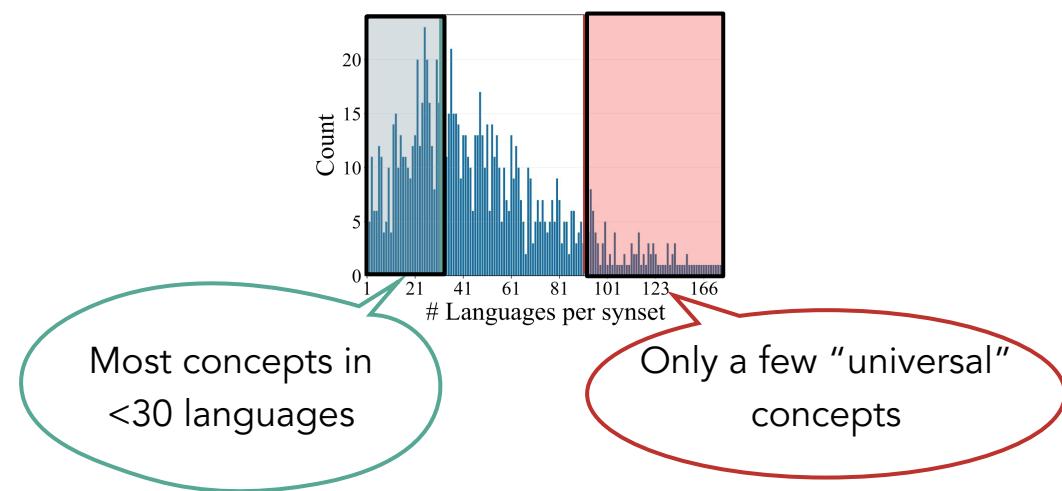
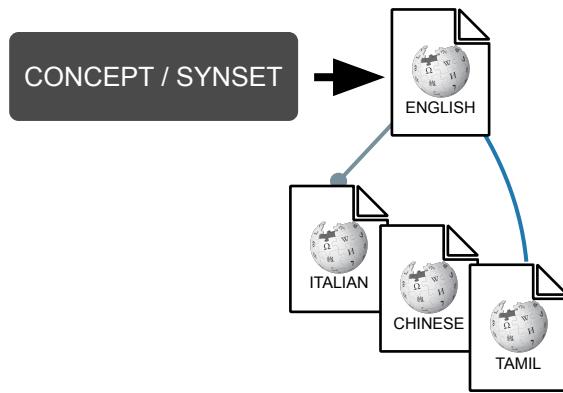
Sanxian / Shamisen



Clavie

Are ImageNet Concepts Cross-Lingual?

- The ImageNet, COCO and Visual Genome datasets use English WordNet concepts
- Idea: estimate cross-linguality using Wikipedia as a proxy





Representative of annotators' cultures



MaRVL-id Bola basket



MaRVL-sw Mpira wa kikapu



MaRVL-tr Basketbol



MaRVL-zh 篮球



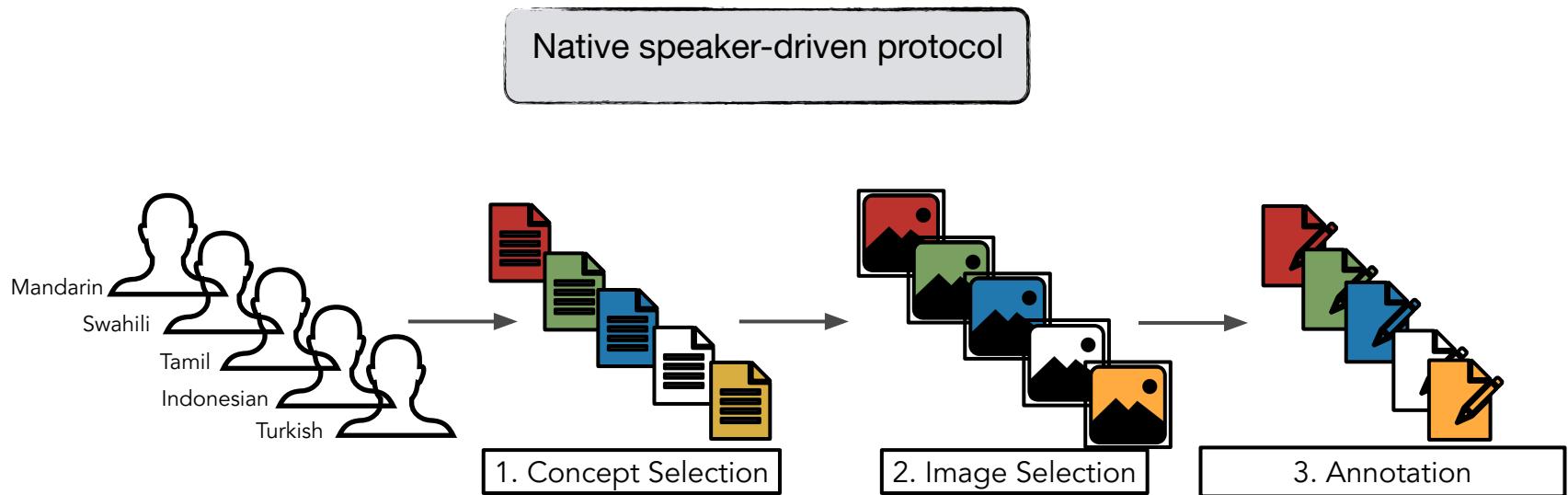
MaRVL-ta கூடைப்பந்தாட்டம்



5 typologically diverse languages

Independent, culture-specific annotations

Collecting MaRVL data



Visual Reasoning Task (Suhr et al. ACL 2019)

- **Datapoint:** two images (v_1 , v_2) paired with a sentence x
- **Task:** Predict whether x is a true description of the pair of images v_1 v_2



இரு படங்களில் ஒன்றில்
இரண்டிற்கும் மேற்பட்ட
மஞ்சள் சட்டை அணிந்த
வீரர்கள் காலையை அடக்கும்
பணியில் ஈடுப்பட்டிருப்பதை
காணமுடி.

True

X

Y

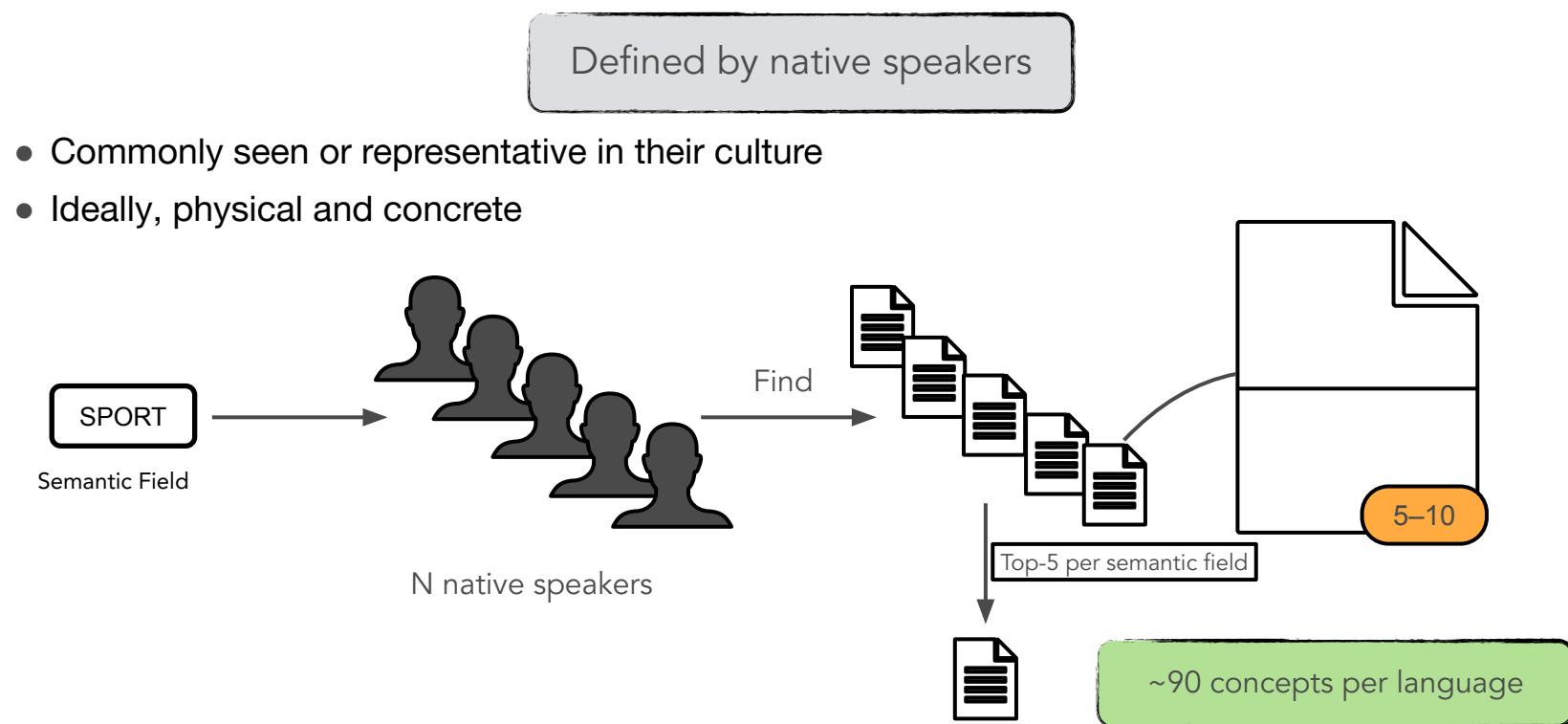
MaRVL is created from Universal Concepts

- Taken from the *Intercontinental Dictionary Series* ([Key & Comrie, 2015](#))
 - 18/22 chapters with concrete objects & events

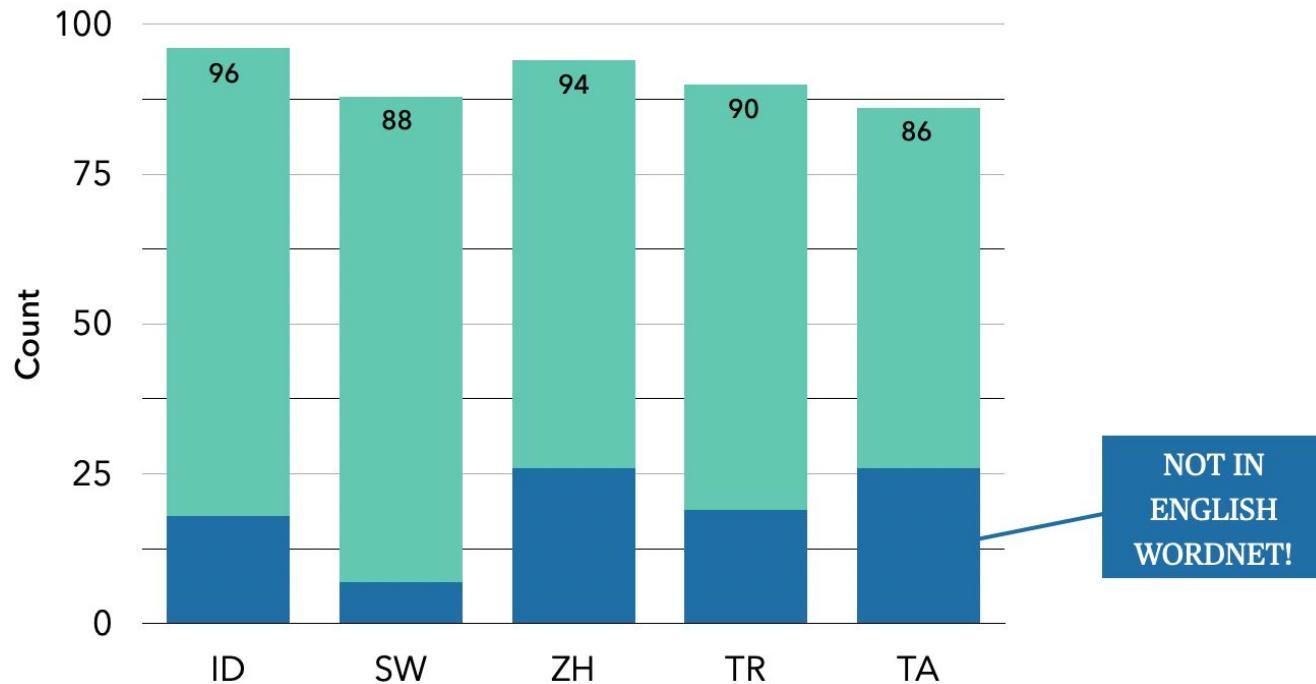
Chapter	Semantic Field
Animal	Bird, mammal
Food and Beverages	Food, Beverages
Clothing and grooming	Clothing
The house	Interior, exterior
Agriculture and vegetation	Flower, fruit, vegetable, agriculture
Basic actions and technology	Utensil/tool
Motion	Sport
Time	Celebrations
Cognition	Education
Speech and language	Music (instruments), visual arts
Religion and belief	Religion



Step 1. Language-Specific Concept Selection



Overview of Resulting Concepts



Step 2. Image Collection

Collected by native speakers

- Representative of the language population
- NLVR2 ([Suhr et al. ACL 2019](#)) requirements
 - 1. Contains more than one instance of a concept
 - 2. Shows an instance of the concept interacting with other objects
 - 3. Shows an instance of the concept performing an activity
 - 4. Displays a set of diverse objects or features



MaRVL-zh 花椰菜 (Cauliflower)



MaRVL-tr **Çılmış** (Buttermilk)



MaRVL-sw Jembe (Shovel)



MaRVL-tr Rakı (Rakı)

Step 3. Language Annotation

Written by native speakers



MATCH 4 PAIRS AT RANDOM



VALIDATE ANNOTATIONS



右图中的人在发球, 左图中的人在接球。



WRITE CAPTION TRUE ONLY FOR 2 PAIRS



右图中的人在发球, 左图中的人在接球。



FINAL VALIDATION



Fleiss' kappa:
93%

右图中的人在发球, 左图中的人在接球。

(The man in the right image is serving a ball while the man in the left image is returning a ball.)

Dataset Examples

MaRVL-tr Kanun (çalgı)



Görsellerden birinde dizlerinde kanun
bulunan birden çok insan var

(In one of the images, there are multiple
people with qanuns on their knees)

Label: True

MaRVL-ta தை (Vada)



இரண்டு படங்களிலும் நிறைய மசால் வடைகள்
உள்ளன

(Both images contain a lot of masala vadas)

Label: False

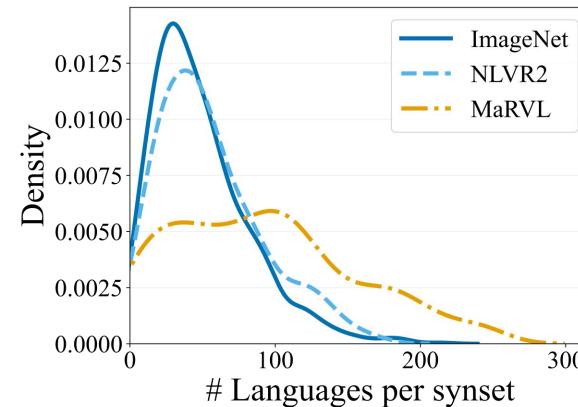
Summary Statistics

- 5560 data points across 5 languages
- 423 concepts (96 not in WordNet)
- 1390 unique captions

MaRVL covers more languages

MaRVL covers more language families

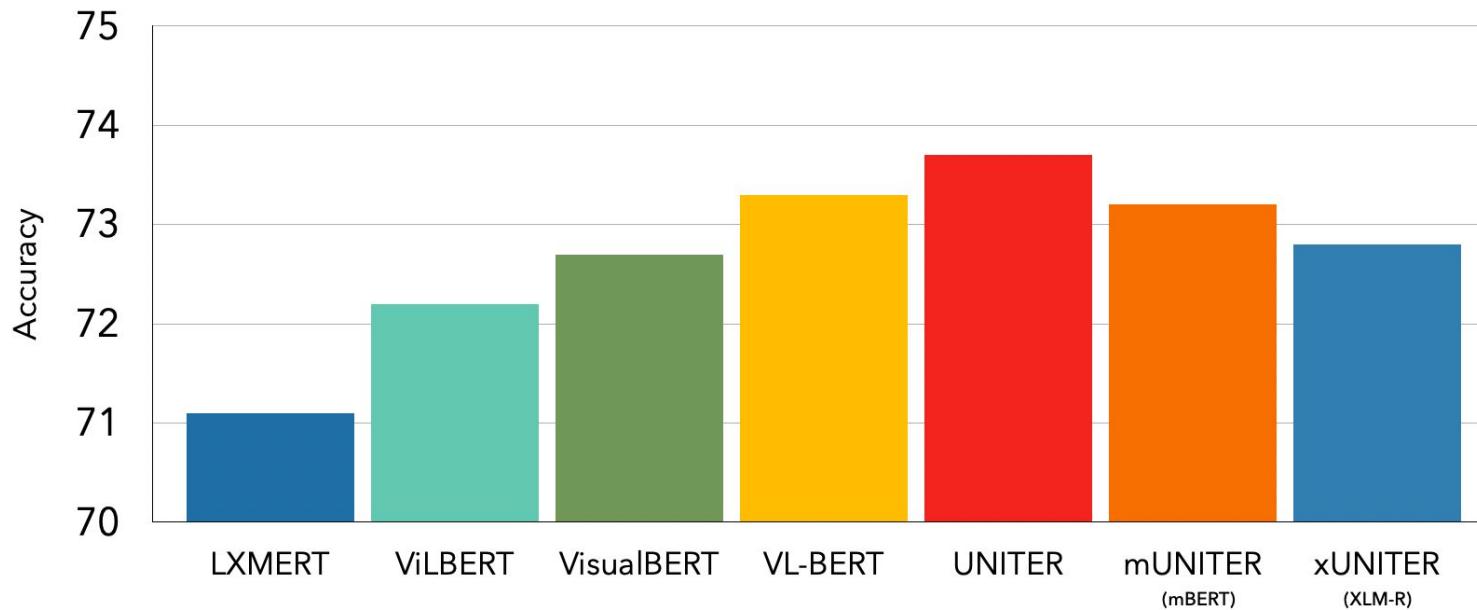
MaRVL covers more macroareas



Pretraining and Finetuning

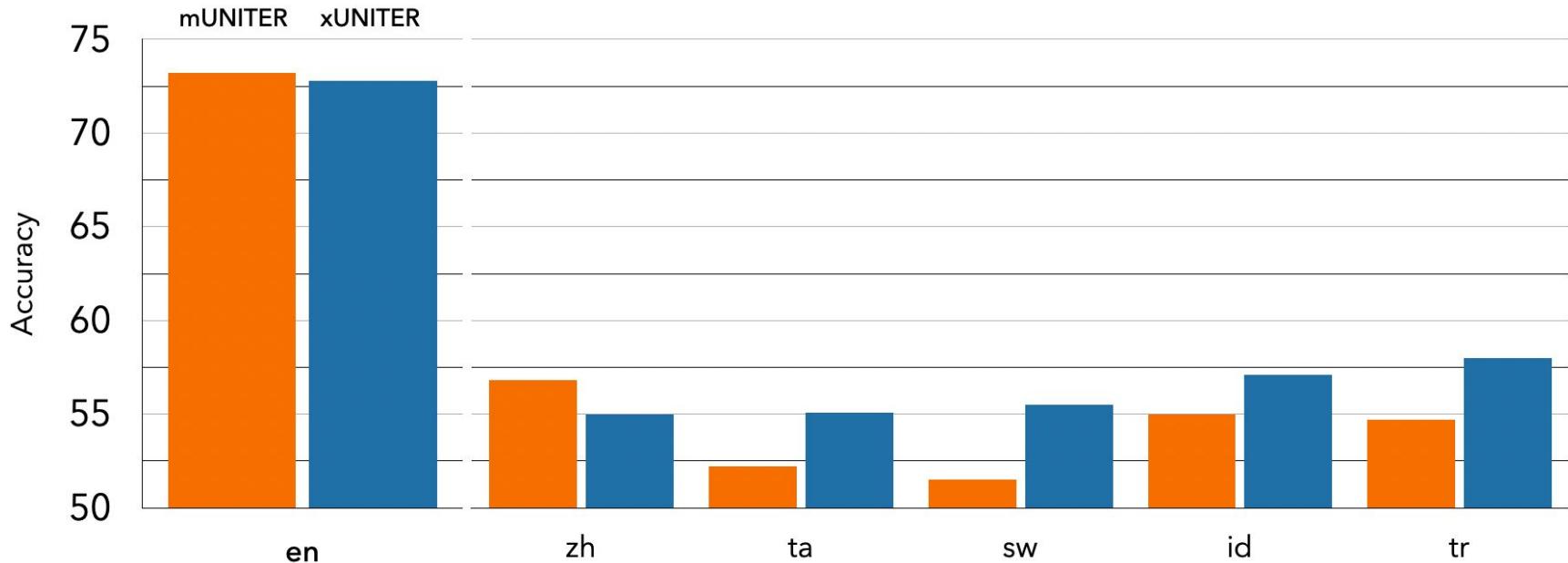
- Two new multilingual UNITER-based models
 - Pretrained on English Conceptual Captions + 104 languages Wikipedia
 - mUNITER: Initialised from mBERT
 - xUNITER: Initialised from XLM-R
- Finetune on 86,373 data points in **English NLVR2** ([Suhr+, 2019](#))
- Test on 5,560 datapoints in MaRVL (5,560 datapoints)
 - **Zero-shot**: Multilingual inputs directly in a cross-lingual approach
 - **Translate-test**: English models by machine translating language data

English NLVR2 Results (Sanity check)



m/xUNITER perform similarly to English-only models

MaRVL Zero-shot Results



Zero-shot transfer: substantial drop in performance

Conclusions

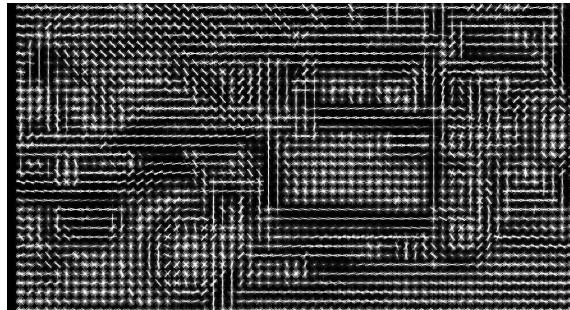
- Concepts and images in existing V&L datasets have an NA/EU bias
- Devise a new protocol for data creation driven by native speakers
- **MaRVL**: V&L reasoning dataset in 5 typologically diverse languages
- Implications beyond vision and language research
 - Multilingual datasets should not just be translations of English data

2. Data Representation

Three Levels of Representation

- Perceptual
- Pre-processed features
- Raw input

- Yellow
- Has wheels
- Metal
- Five-door
- Can transport
- ...



Perceptual Norms

- Ask people to write down the words that are triggered by textual stimuli.
- Stimuli: 541 noun concepts
- Norms are categorized into the likely knowledge source

Moose		
is large	27	visual-form and surface
has antlers	23	visual-form and surface
has legs	14	visual-form and surface
has four legs	12	visual-form and surface
has fur	7	visual-form and surface
has hair	5	visual-form and surface
has hooves	5	visual-form and surface
is brown	10	visual-color
hunted by people	17	function
eaten as meat	5	function
lives in woods	14	encyclopedic
lives in wilderness	8	encyclopedic
an animal	17	taxonomic
a mammal	9	taxonomic
an herbivore	8	taxonomic

Perceptual Norms: Pros / Cons

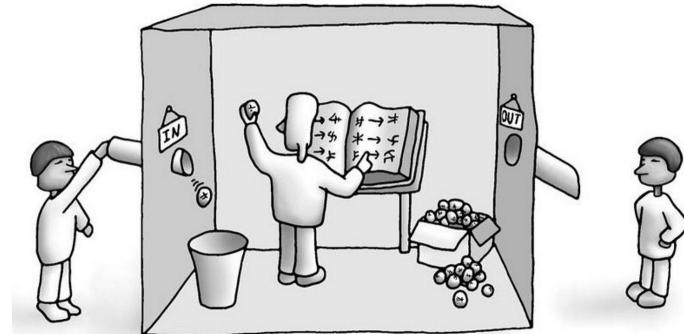
Pros

- Seemingly simple task
- Rich features

Cons

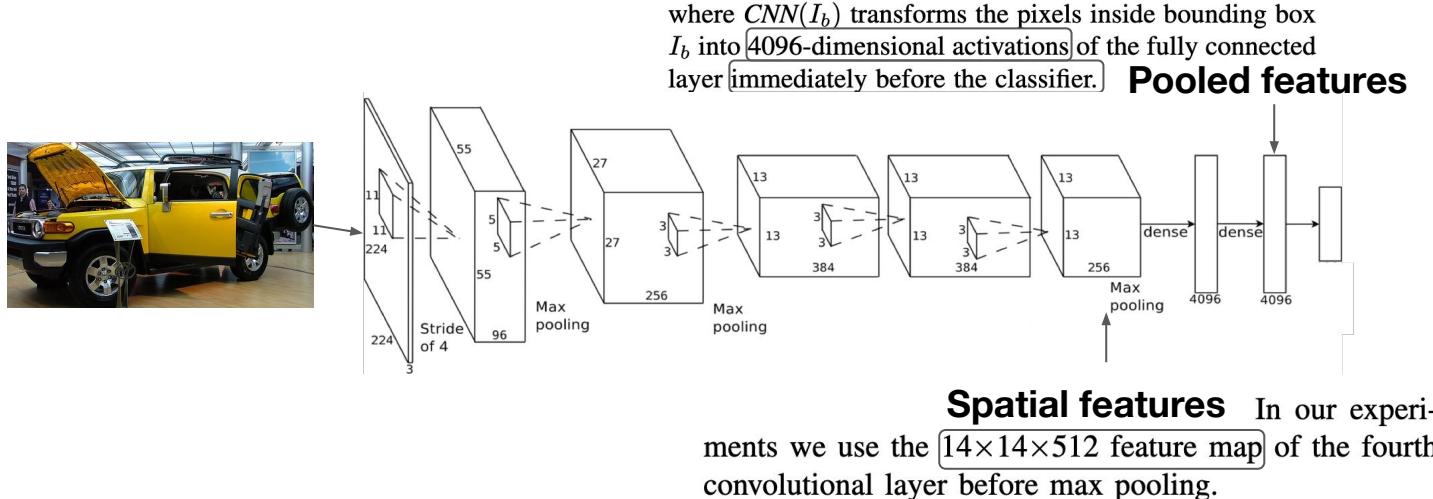
- Can it scale?
- Handling ambiguity

What does it mean to only understand symbols as defined by other symbols?



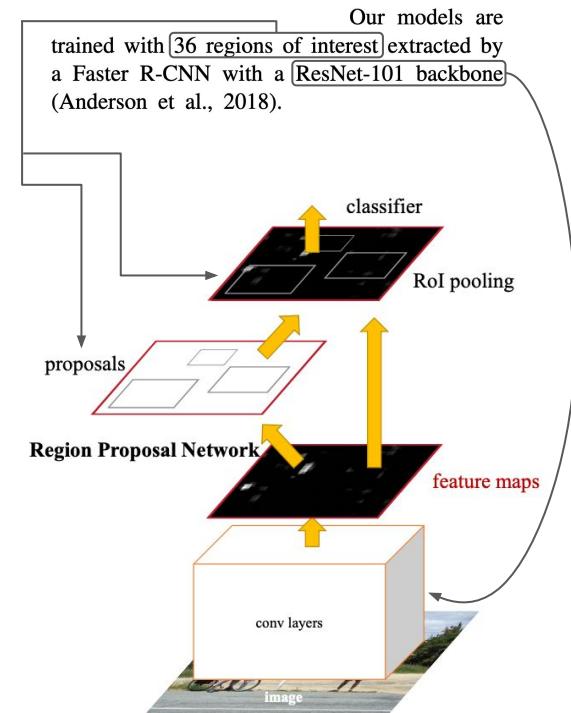
Spatial and Pooled Visual Features

- Earliest work in neural-network era used pooled or spatial preserving features from a pretrained Convolutional Neural Network.



Pre-processed Visual Features

- Faster R-CNN region-based feature vectors
 - Trained on the Visual Genome Dataset
 - The Region Proposal Network suggests the location of *regions of interest*.
 - RoI pooling performs spatial pooling in the final CNN layer to give a 2048D vector.



Pre-processed: Pros / Cons

Pros

- Long-established practice
- Usually an offline process:
do it once and forget

Cons

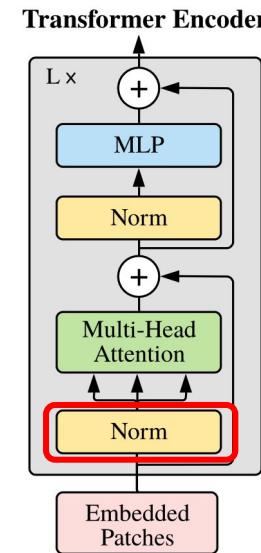
- Large datasets require specialized storage (e.g. *h5fs*)
- Not obvious how to randomly augment data
- Specialist knowledge can be opaque to newcomers

Raw Input

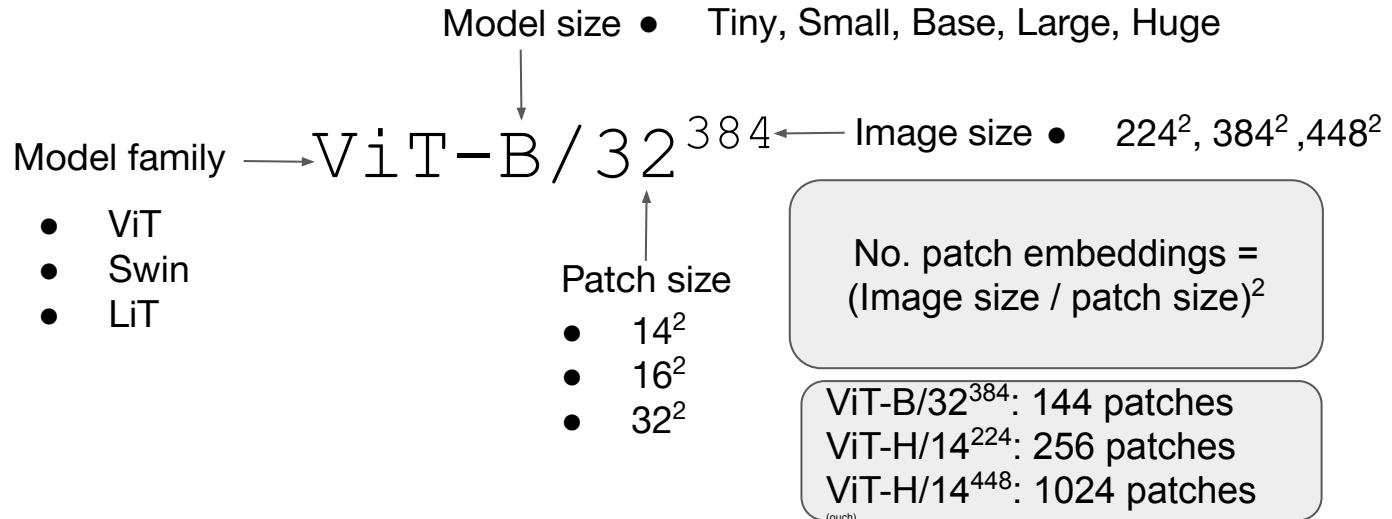
- Directly process data from the raw images or speech signal.
- Images:
 - Vision Transformer (ViT)
 - Swin Transformer
- Speech
 - Spectrogram Transformer
 - AudioMAE

Vision Transformer

- Good news! You are already almost an expert in how the Vision Transformer works
 - Split image into K patches
 - Embed each patch
 - Add position information
 - Encode using Transformer blocks that include an **extra pre-norm layer** for stability.



Nomenclature and Patch Count

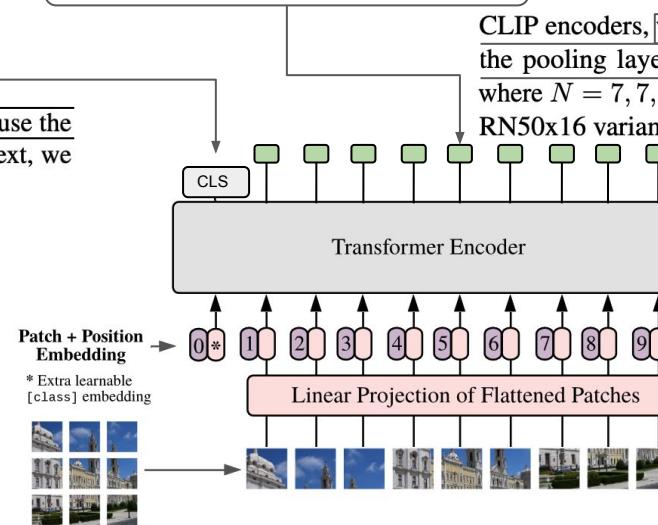


Extracting ViT Features

- Extract pooled features or patch-level features

To extract visual information from an image x^i , we use the visual encoder of a pre-trained CLIP [29] model. Next, we

For the CLIP encoders, we extract the feature grid before the pooling layers, resulting in an $N \times N$ grid, where $N = 7, 7, 12$ for the ViT-B/32, RN50x4 and RN50x16 variants of CLIP respectively.



Raw input: Pros / Cons

Pros

- Data augmentation is straightforward because you always have the raw input
- Fewer preprocessing steps means fewer creeping errors

Cons

- Smaller batches with an extra model on the GPU
- Potentially many inputs

Summary

- Many options for how to represent your multimodal inputs
 - Language-oriented
 - Object / stuff oriented
 - Raw inputs
- **No universally best option** but raw inputs are promising because the visual representation model can be fully differentiable.

3. Modelling

Two Main Approaches

Cross encoding models

aka Multimodal Transformer

Transformer

The red horse



Dual encoding models

aka Dual Tower Model

Language

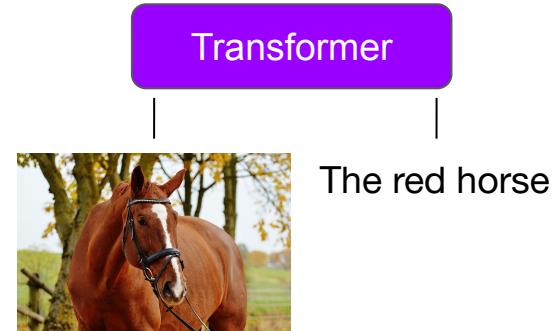
The red horse

Vision



Cross-encoding Models

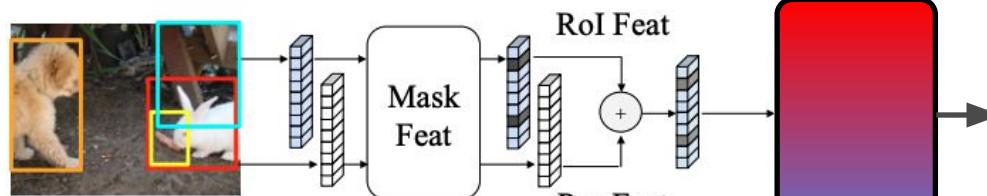
- Emerged as a key modelling approach in 2019 with a flurry of approaches to creating visually-grounded BERT models.
- This is a form of *model-based fusion*
- The backbone consists of two components:
 - language model
 - visual encoder



High-level Overview

Image:

Faster R-CNN, or
raw pixels + ViT



Language:

BERT tokens

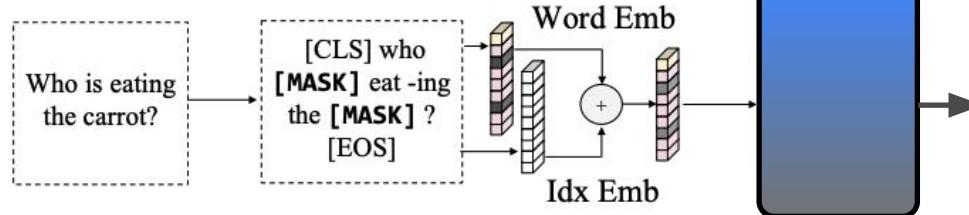
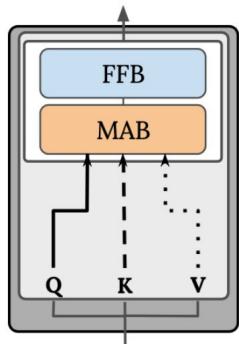


Figure: Tan & Bansal (2019). LXMERT: Learning Cross-Modality Encoder Representations from Transformers. EMNLP-IJCNLP

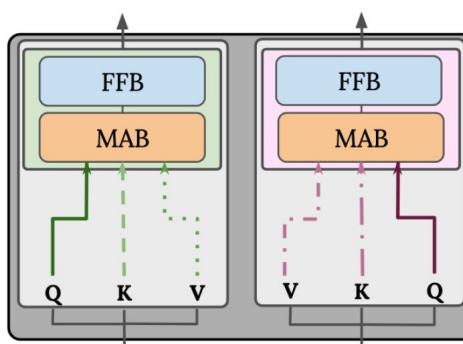
Building Blocks

- Single-stream or Dual-stream Transformer Blocks



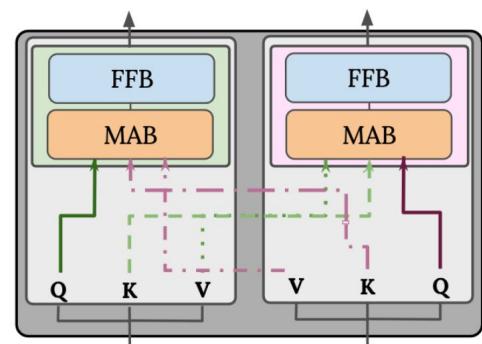
$\mathbf{w}_1 \dots \mathbf{w}_T$ $\mathbf{v}_1 \dots \mathbf{v}_K$

Single-stream



$\mathbf{w}_1 \dots \mathbf{w}_T$ $\mathbf{v}_1 \dots \mathbf{v}_K$

Dual-stream
intra-modal

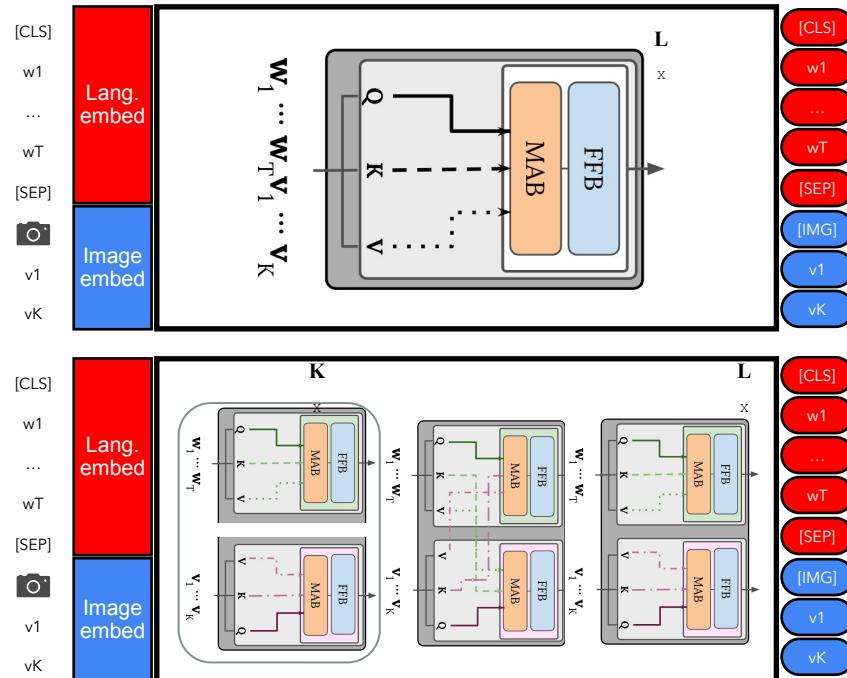


$\mathbf{w}_1 \dots \mathbf{w}_T$ $\mathbf{v}_1 \dots \mathbf{v}_K$

Dual-stream
inter-modal

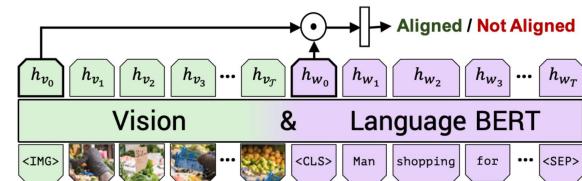
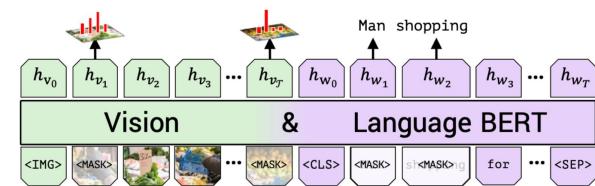
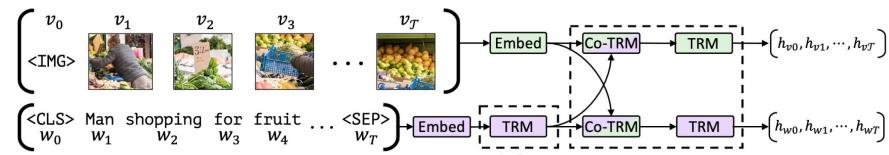
Single- & Dual-Stream Architectures

- Single-stream
 - Concatenate inputs into one sequence
- Dual-stream
 - Process modalities independently
 - Intra-modal
 - Inter-modal



2019: ViLBERT

- Dual-stream model
- Initialized from BERT
- Visual features extracted from 10-36 regions using Faster-RCNN
- Pretrained on Conceptual Captions
 - Masked Language Modelling
 - Masked Region Classification
 - Image-Text Matching

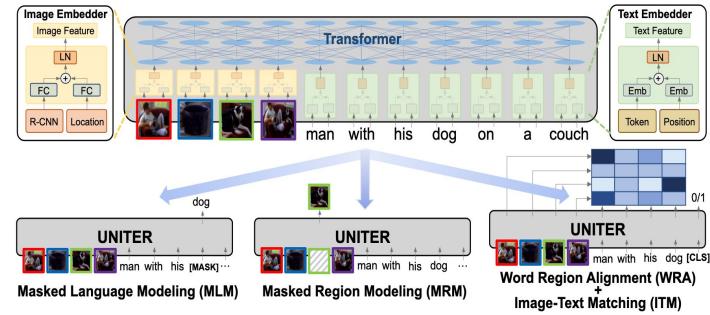


MLM, MRC, ITM

- Masked Language Modelling $\mathcal{L}_{\text{MLM}}(\theta) = -\mathbb{E}_{(\mathbf{w}, \mathbf{v}) \sim D} \log P_\theta(\mathbf{w}_m | \mathbf{w}_{\setminus m}, \mathbf{v})$
 - Same as BERT et al.
- Masked Region Classification $\mathcal{L}_{\text{MRM}}(\theta) = \mathbb{E}_{(\mathbf{w}, \mathbf{v}) \sim D} f_\theta(\mathbf{v}_m | \mathbf{v}_{\setminus m}, \mathbf{w})$
 - Mean Squared Error Regression over the 2048D feature vector; or
 - Predict the probability distribution over the 1600 Faster R-CNN classes
- Image-Text Matching $\mathcal{L}_{\text{ITM}}(\theta) = -\mathbb{E}_{(\mathbf{w}, \mathbf{v}) \sim D} [y \log s_\theta(\mathbf{w}, \mathbf{v}) + (1 - y) \log(1 - s_\theta(\mathbf{w}, \mathbf{v}))]$
 - 50% chance of randomly sampling a mis-matched sentence
 - Predict with a binary classifier (aka Next Sentence Prediction)
- Note: 15% masking usually spans both modalities

2020: UNITER

- Single-stream model
- Initialized from BERT
- Visual features from Faster-RCNN
- Pretrained on Conceptual Captions, Visual Genome, COCO, SBU Captions
 - Masked Language Modelling
 - Masked Region Classification
 - Image-Text Matching
 - Word-Region Alignment



WRA

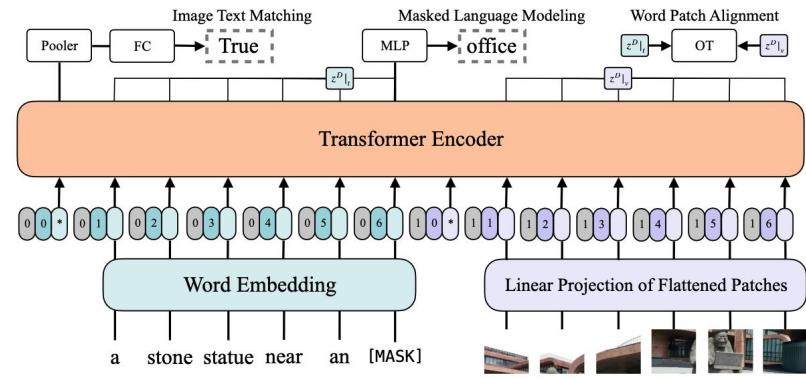
- Word-Region Alignment
 - Estimate an alignment between the text and image representations
 - Minimize the cost \mathbf{T} of transporting the embeddings from image regions to words in a sentence.
 - In other words, a fine-grained image-sentence matching

$$\mathcal{L}_{\text{WRA}}(\theta) = \mathcal{D}_{ot}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\mathbf{T} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i=1}^T \sum_{j=1}^K \mathbf{T}_{ij} \cdot c(\mathbf{w}_i, \mathbf{v}_j)$$

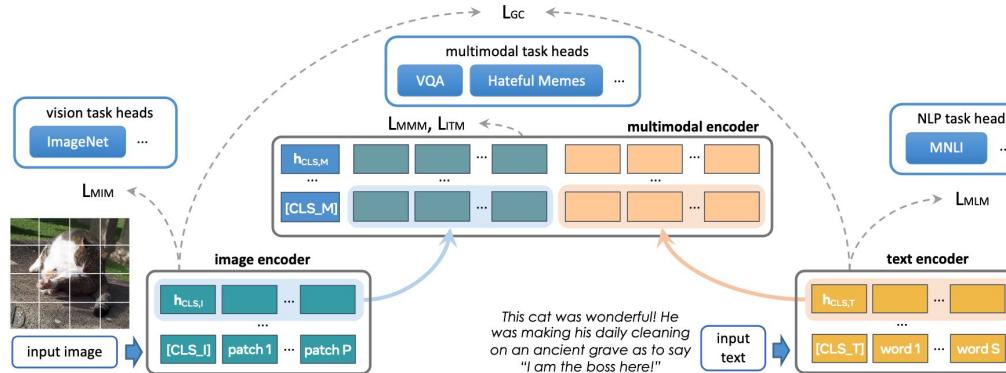
- $c(\cdot, \cdot)$ is cosine distance

2021: ViLT

- Single-stream model
- Initialized from BERT
- Visual features extracted from ViT-B/32
- Pretrained on Conceptual Captions, Visual Genome, COCO, SBU Captions
 - Masked Language Modelling
 - Image-Text Matching
 - Word-Patch Alignment



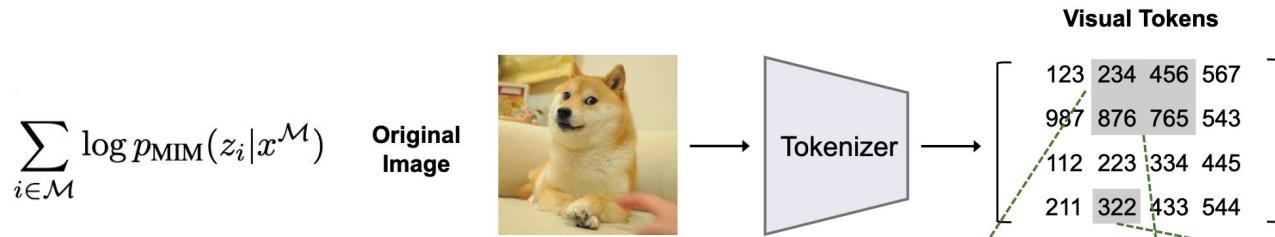
2022: FLAVA



- Dual-stream Visual features extracted from ViT-B/16
- Pretrained on PMD70M
 - Masked Language Modelling, Masking Image Modelling
 - Image-Text Matching, Masked Multimodal Modelling
 - Global Contrastive Matching

MIM and CL

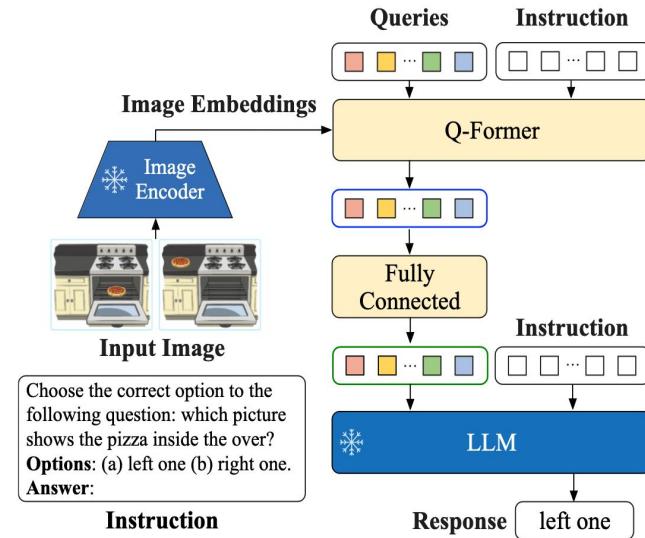
- Masked Image Modelling
 - Immediately after the image encoder and before multimodal encoding
 - Tokens from a discrete VAE (BEiT)



- Contrastive Loss
 - On the CLS embedding of each unimodal encoder $\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E} \left[\log \frac{f(\mathbf{t}, \mathbf{i})}{\sum_{\mathbf{t}' \in T} f(\mathbf{t}', \mathbf{i})} \right]$

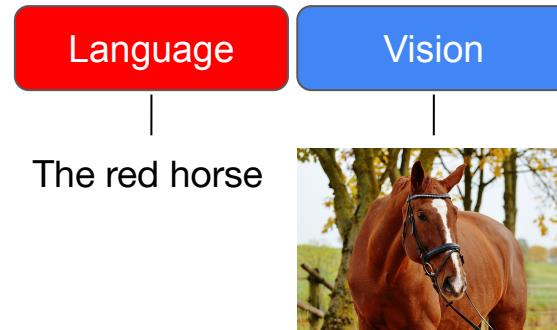
2023: InstructBLIP

- Initialized from BLIP-2
 - ViT & FlanT5_{XXL}
- Visual features from ViT-G/14
- Instruction-tuned on 13 datasets, each using 10-15 templates



Dual-encoding Models

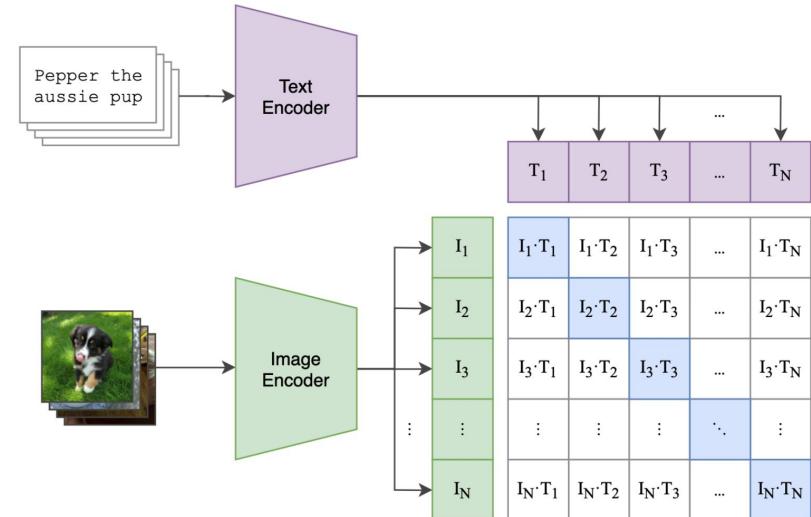
- Emerged as a sample-efficient alternative to cross-encoding.
- The backbone consists of two separate components:
 - language encoder
 - visual encoder



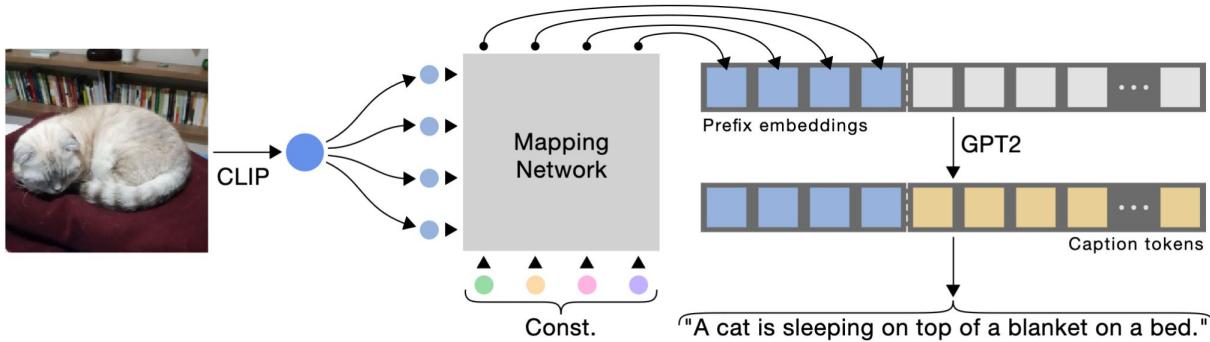
CLIP

- 12 Layer Transformer Encoder
- ViT or ResNet Visual Encoder
- Maximize the similarity of the embeddings of paired examples (I, T):

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E} \left[\log \frac{f(\mathbf{t}, \mathbf{i})}{\sum_{\mathbf{t}' \in T} f(\mathbf{t}', \mathbf{i})} \right]$$

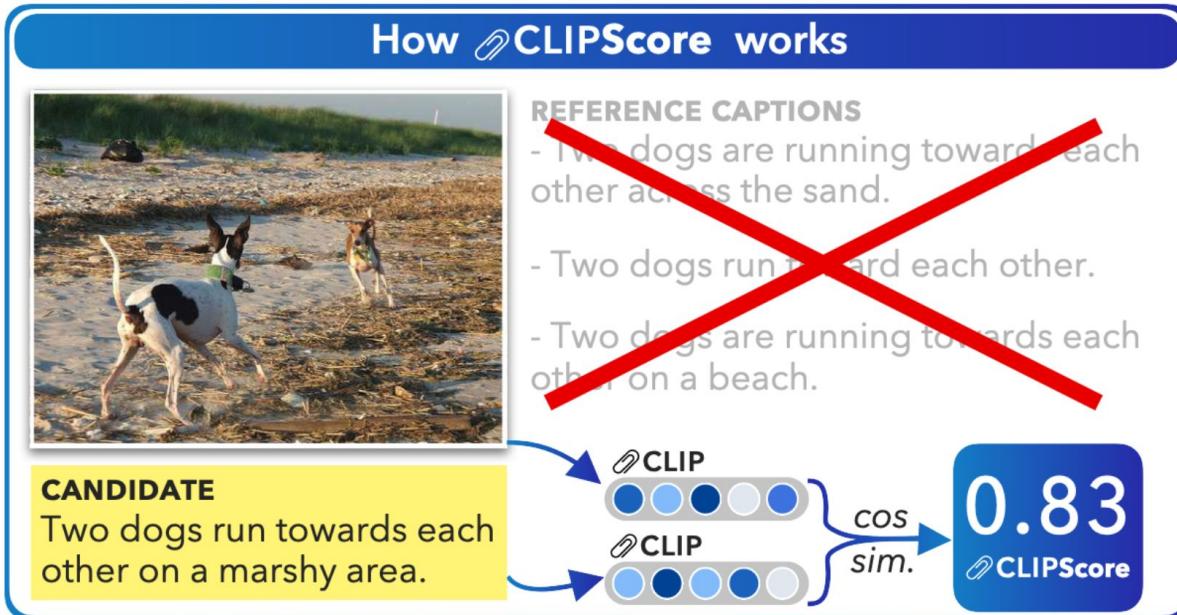


CLIP for Captioning



- Use CLIP as a feature extractor and GPT-2 as a language model.
 - Only train the mapping network to generate prefix embeddings
 - Lightweight system that exploits pretrained models

CLIP for Evaluation



Summary

- Cross-encoding:
 - Many advances in which parts of the input contribute to loss
 - Shift from regions-of-interest to Vision Transformers
- Dual-encoding:
 - Excellent cross-domain transfer to a wide range of problems
- Better performance: larger models pretrained with larger datasets.

Q: Can we offload some of the training budget by using retrieval augmentation?

Retrieval-Augmented Image Captioning

EACL 2023



R. Ramos



D. Elliott



B. Martins

Standard Approach to Image Captioning

Visual encoder



IMAGE



GENERATED CAPTION

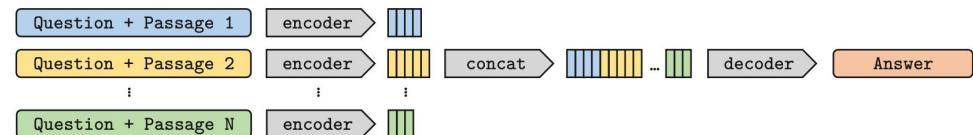
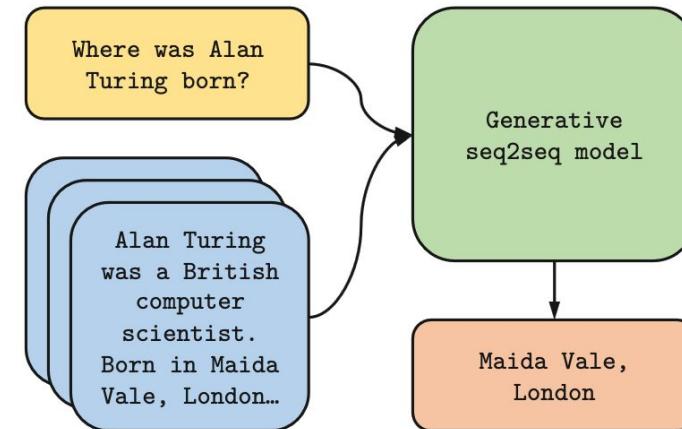
Proposed Approach

Visual and Language encoder

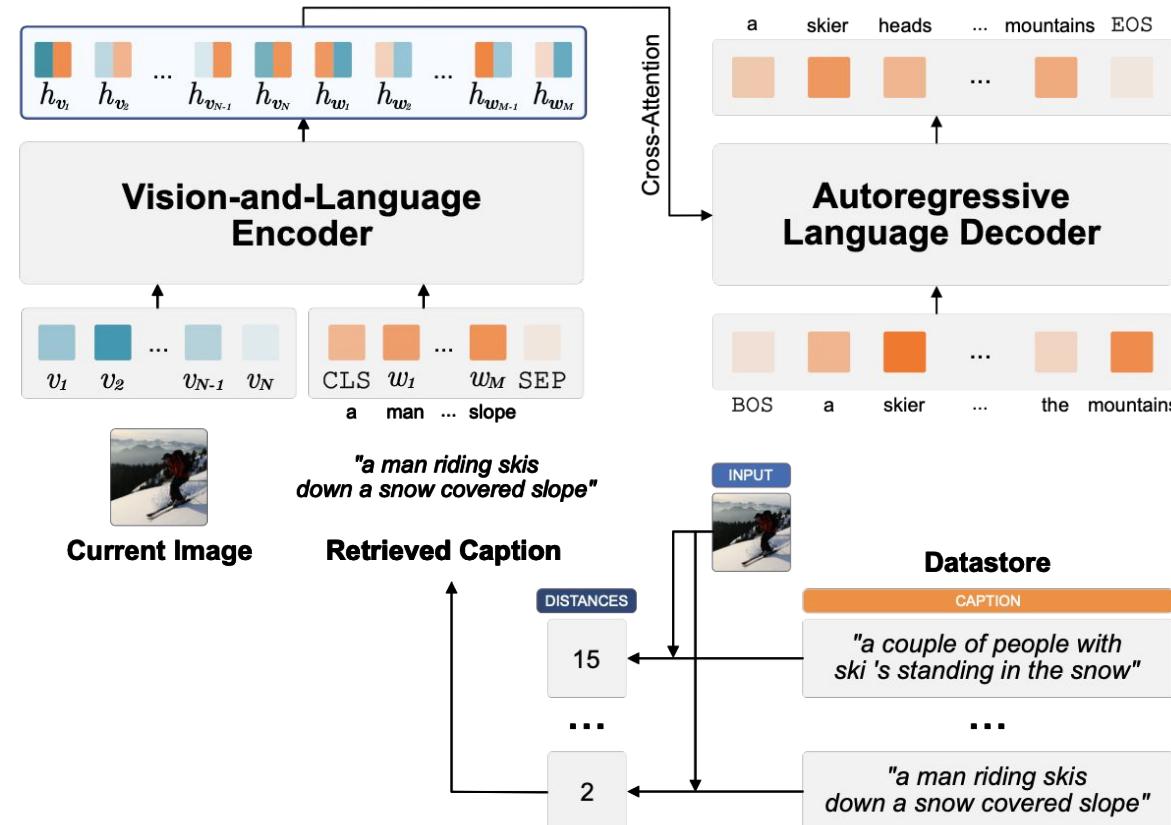


Retrieval-augmented Modelling

- Improve text generation quality by conditioning a model on relevant examples retrieved from a datastore

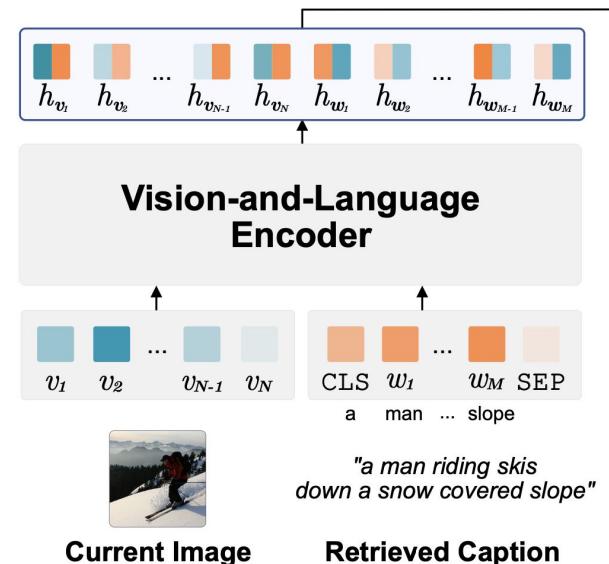


Encoder with Cross-modal representations Through Retrieval Augmentation



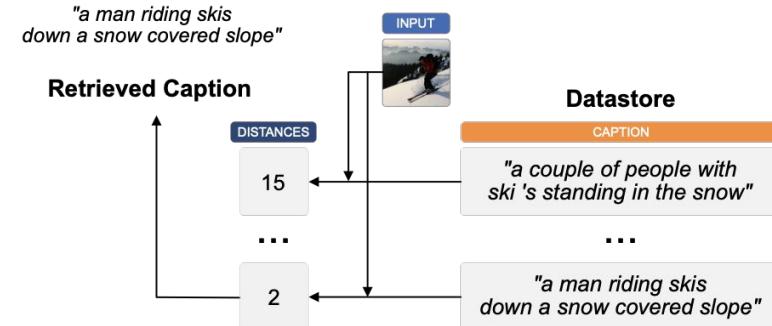
Joint Encoder

- Pretrained V&L encoder to jointly encode:
 - V - Visual input:
 - N = 36 Faster-RCNN regions-of-interest
 - L - Language Input
 - K retrieved sentences
 - M sub-words using the BERT tokenizer



Retrieval System

- Build a datastore: store high-dimensional vectors
 - FAISS: Facebook AI Similarity Search for nearest-neighbor search
 - Captions of images represented with CLIP embeddings
- Retrieve k nearest-neighbours captions from datastore
 - Image embedding compared against datastore caption vectors

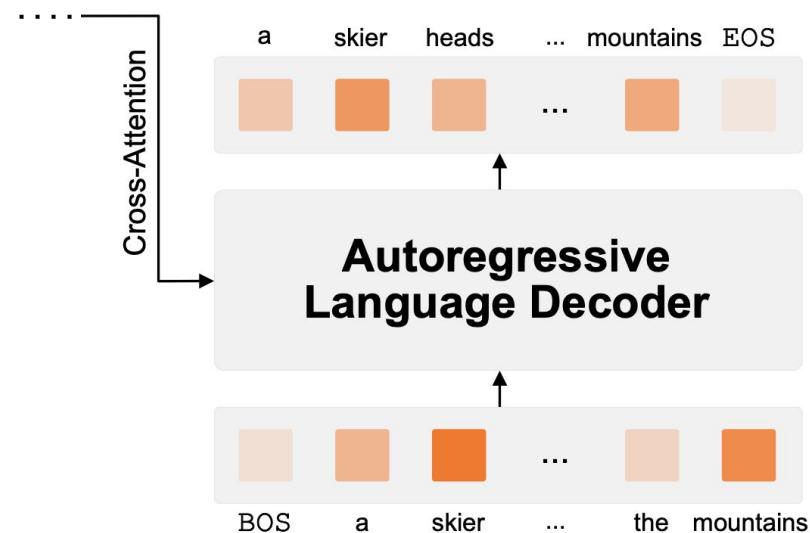


Decoder

GPT-2 style LM with additional learned cross-attention layers

The decoder predicts based on:

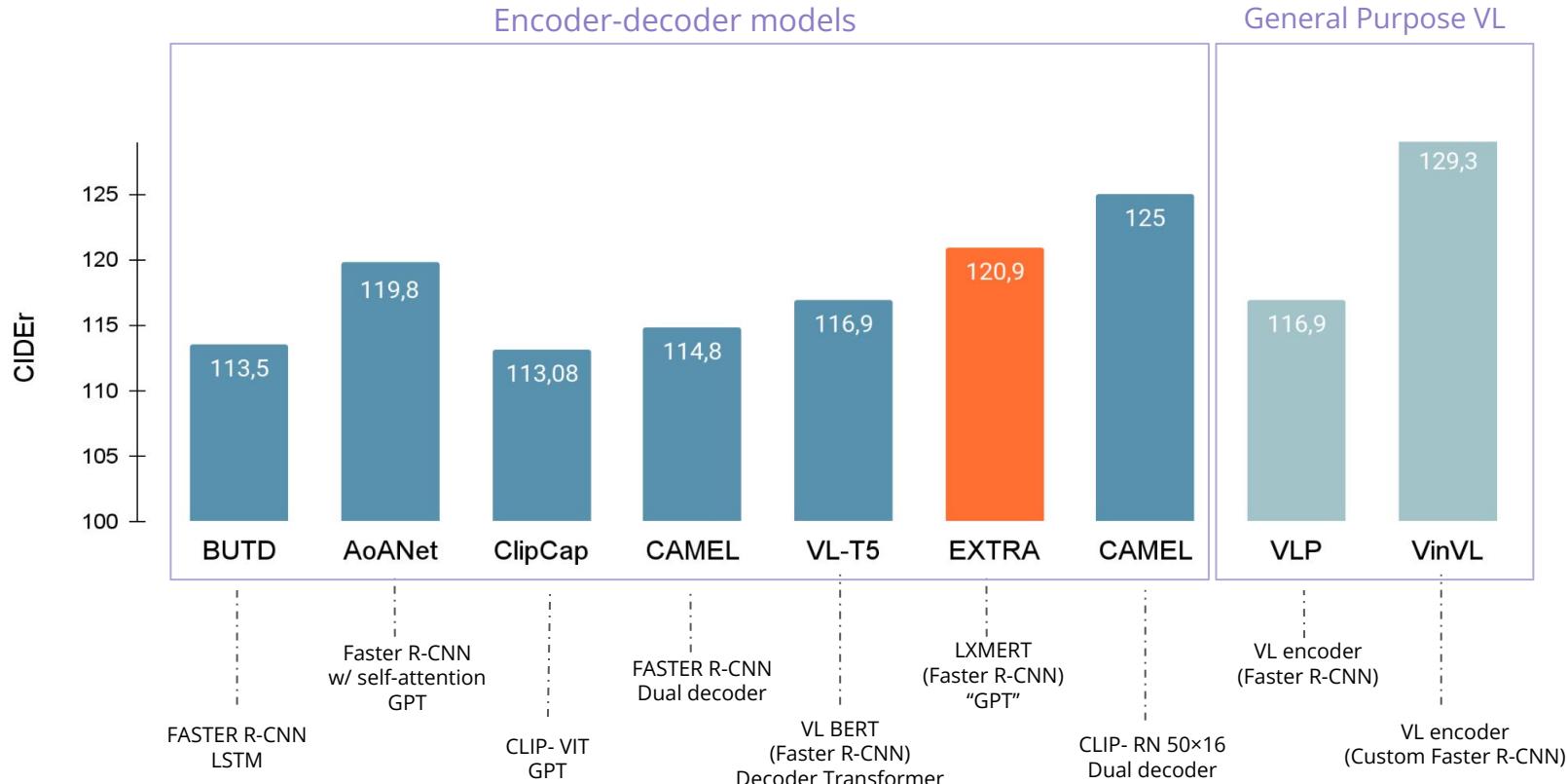
- previous tokens; and
- V&L encoder embeddings



Experimental Protocol

- Train and evaluate on COCO dataset.
 - 1 x 32GB NVIDIA V100S
 - Train the encoder and randomly initialized 4-layer decoder
- Retrieval augmentation is directly from the datastore. No gold data.
- Evaluate with BLEU, Meteor, SPICE, and **CIDEr**.
- Similar improvements with CE or additional SCST with CIDEr.

Results



Qualitative Examples



a vehicle parked by a parking meter in pile of snow

a vehicle parked next to a parking meter buried in snow near buildings

two parking meters nearly buried in heavy snow

a parking meter next to a car is surrounded by snow

a parking meter in front of a building piled in snow



EXTRA



a parking meter covered in snow next to a car



the kitten is stuck in the white basin

a very cute cat sitting in a sink drinking from a cup

a cat that is in a white sink

a cat sitting in a sink over a cup of something

a white cat is lying in a sink



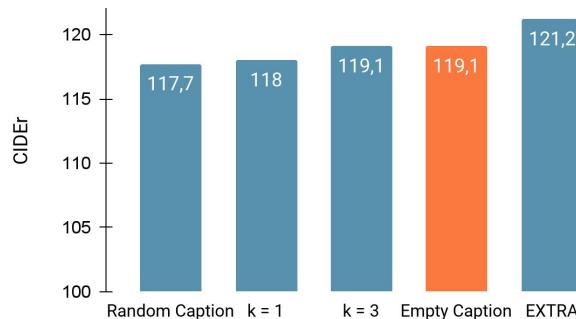
EXTRA



a cat sitting in a sink next to a bowl

Retrieval *Quantity* Matters

- Retrieve enough to handle mistakes
 - Encoding an empty caption is better than only retrieving one caption



EXTRA

a young boy holding an umbrella in the rain

Oracle Performance

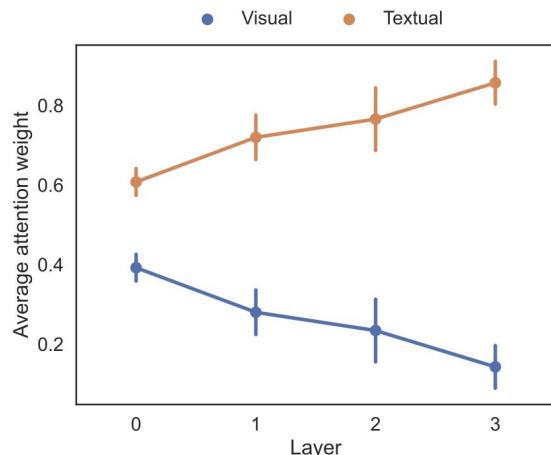
- EXTRA continues to improve when it encodes the other captions for a target image
 - Room for further improvement

	BLEU-4	CIDEr
k = 5 retrieved captions	38.3	121.2
5 references	40.9	129.5

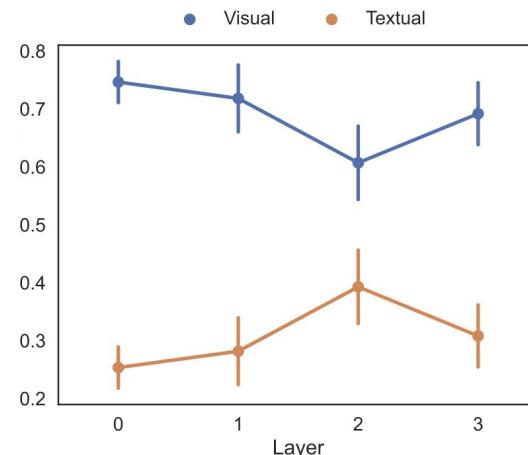
Oracle experiment after standard training

Vision First, Language Later

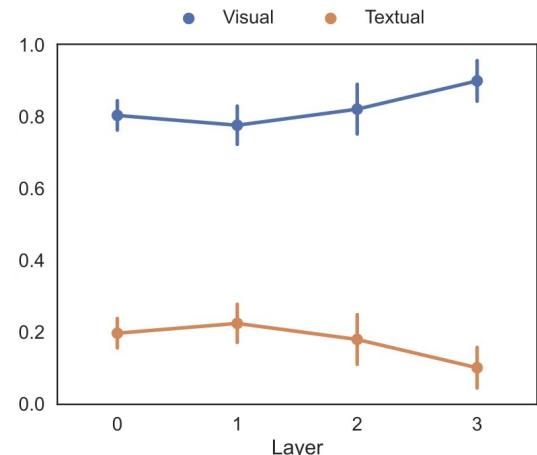
- Decoder attention shifts from the visual inputs to the textual inputs.



EXTRA



Empty Caption



Random Caption

Conclusions

- Image captioning addressed as language generation conditioned on *multimodal* inputs of the image and relevant retrieved sentences
- EXTRA outperforms similar models that only process the image
- Ablations and analyses reveal that even better performance is possible with better retrieval systems

Q: Can we train fewer parameters?

SmallCap: Lightweight Image Captioning Prompted with Retrieval Augmentation

CVPR 2023



R. Ramos



B. Martins



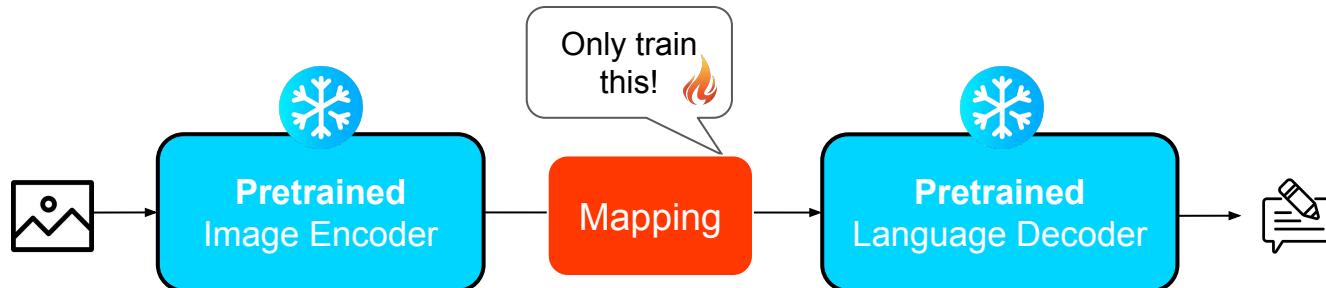
D. Elliott



Y. Kementchedjhieva

Motivation

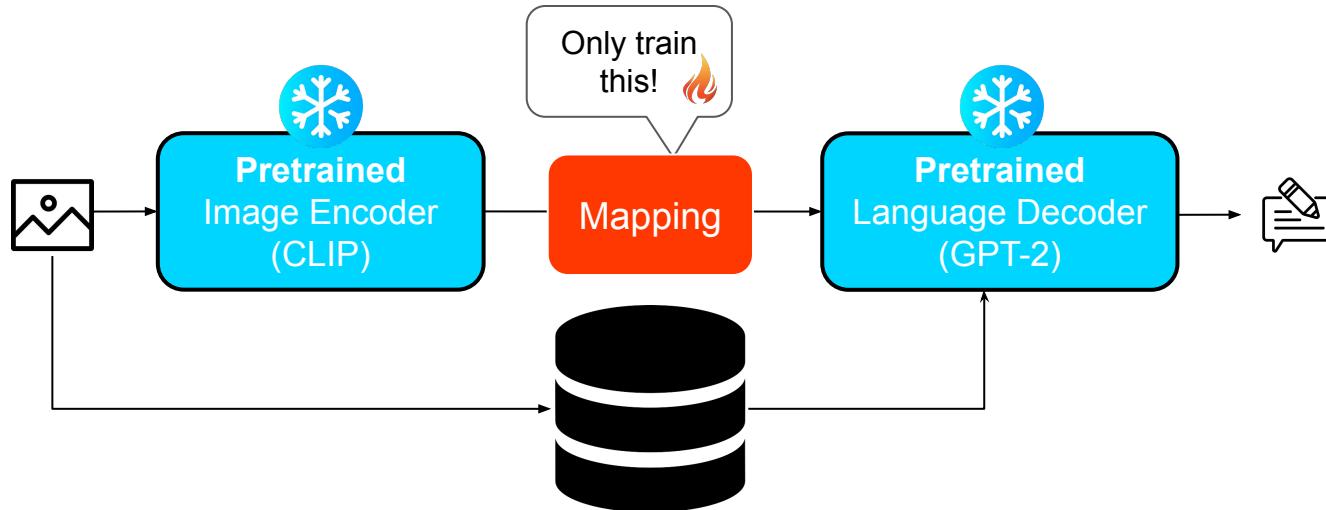
- Main trend in V&L is training bigger models on more data
- Alternative is emerging that re-uses independent backbone models
 - CLIPCap, I-Tuning



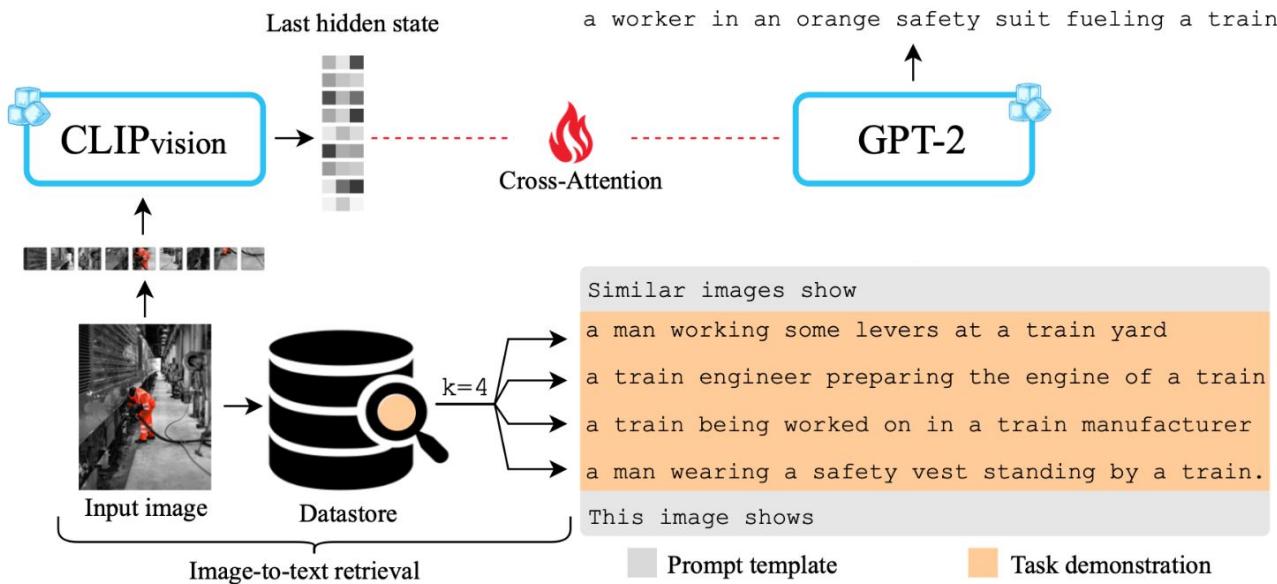
Mokady et al. (2021). ClipCap: CLIP Prefix for Image Captioning. arXiv.
Luo et al. (2023). I-Tuning: Tuning Frozen Language Models with Image for Lightweight Image Captioning. ICASSP.

Lightweight Training through Retrieval

- Given the success of multimodal retrieval augmentation, can we extend this to the lightweight training paradigm?



SmallCap Model



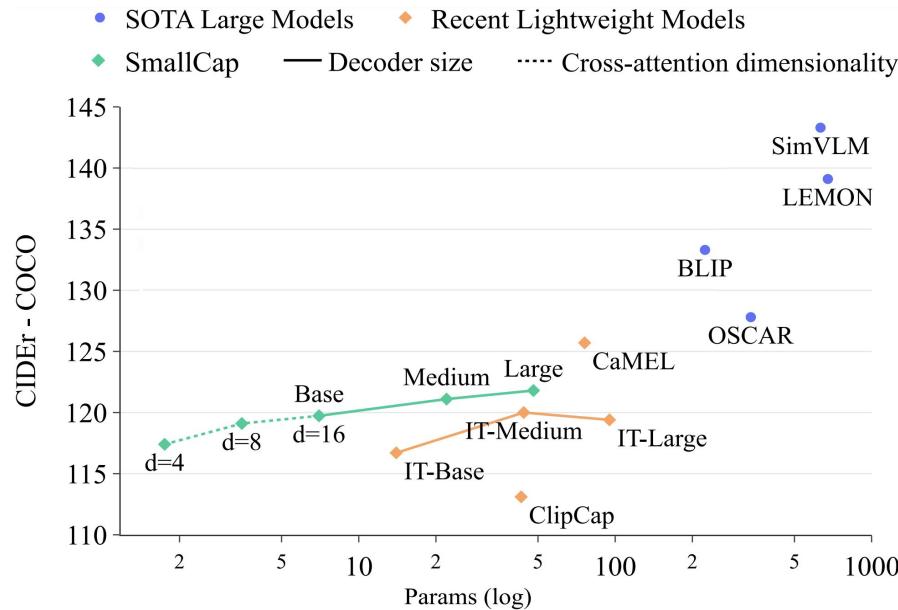
Experimental Setup

- Pretrained CLIP-ViT-B/32 and GPT/OPT backbone models
- Randomly initialize the cross-attention layer
- Train only on COCO in only 8 hours on 1 x 40GB NVIDIA A100 GPU

**Low-rank
cross-attention**
 $\text{Att}(\mathbf{QW}_i^Q, \mathbf{KW}_i^K, \mathbf{VW}_i^V)$
 W_i^K, W_i^Q
 $, W_i^V$
 $\in \mathbb{R}^{d_{\text{encoder}} \times d}$

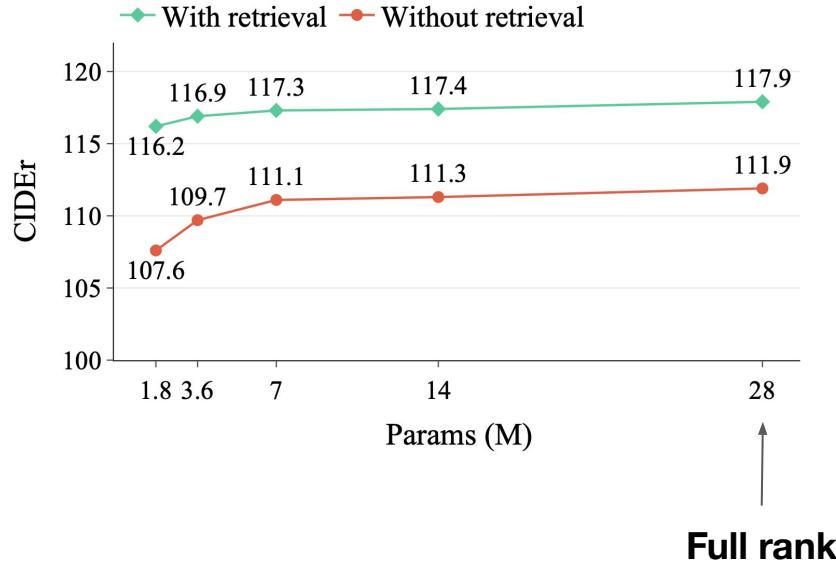
Attention rank	Params
d=64 (Full)	22M
d=16	7M
d=8	3.6M
d=4	1.8M

Results



- Outperform other lightweight approaches
- Effective with low-rank matrices: $4, 8, 16 \ll 64$
- Larger pretrained decoders further improve performance

Importance of Retrieval Augmentation



- With retrieval:
 - Performance is stable across the range of cross-attention sizes
- Without retrieval:
 - Drop in performance
 - SMALLCAP model performance degrades at a higher rate

Training-Free Domain Transfer

- SmallCap was trained on COCO but we can easily swap the datastore.
- Much stronger performance than other lightweight approaches

	Flickr30k	VizWiz	MSR-VTT
ClipCap	41.2	28.3	12.5
CaMEL	55.2	37.6	20.7
SMALLCAP	60.6	55.0	28.4

Qualitative Example on VizWiz



COCO



VizWiz

- some carrots potatoes garlic an onion and some chicken broth
- a selection of ingredients for soup includes carrots, meat, and prepackaged broth
- this is the makings of a meal with chicken and vegetables
- the meal has chicken, bread, and cole slaw

Generated caption:

a close up of a plate of food on a table

- a can of swanson fat free chicken broth
- a can of swanson brand chicken broth with less sodium
- a 14.5 ounce can of swanson branded chicken broth
- a can of swanson chicken broth on a table

Generated caption:

a can of swanson brand chicken broth on a table

Conclusions

- SmallCap:
 - light to train
 - easily transferred across domains without retraining
- Prompt-based conditioning method, wherein retrieved captions are used as a prompt to a generative language model
- Strong performance in out-of-domain settings

Q: Do you even train?

LMCap: Few-shot Multilingual Image Captioning by Retrieval Augmented Language Model Prompting

Findings of ACL 2023



R. Ramos



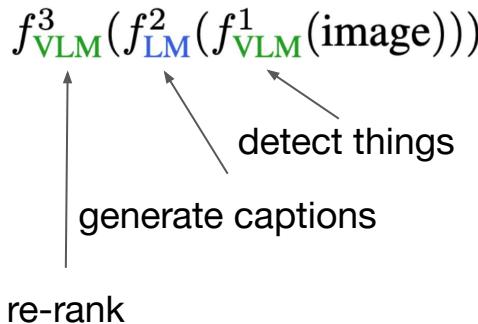
B. Martins



D. Elliott

Socratic Models

- Enable models to “communicate” with each other through their output labels, prompting, and ranking



I am an intelligent image captioning bot. This image is a {img_type}. There {num_people}. I think this photo was taken at a {place1}, {place2}, or {place3}. I think there might be a {object1}, {object2}, {object3},... in this {img_type}. A creative short caption I can generate to describe this image is:

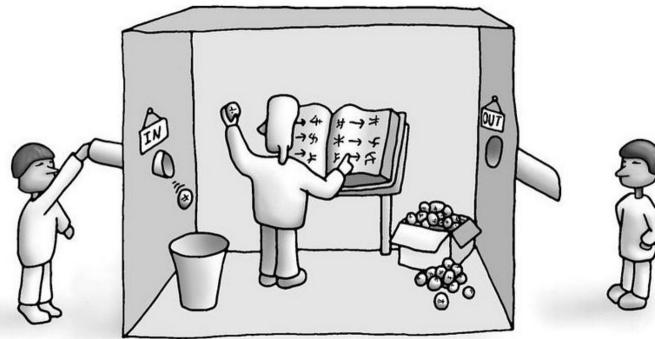


SM (ours): This image shows an inviting dining space with plenty of natural light.

ClipCap: A wooden table sitting in front of a window.

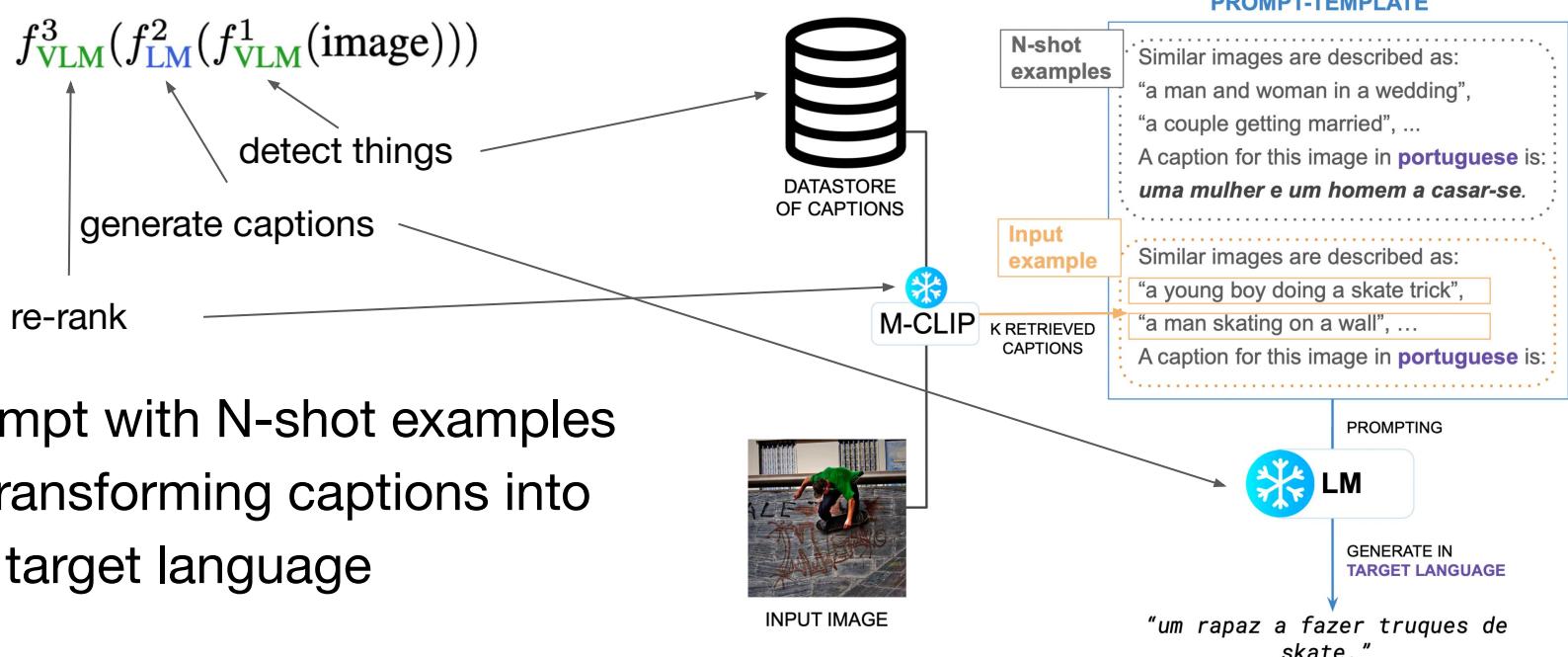


I am an intelligent image captioning bot. This image is a {img_type}. There {num_people}. I think this photo was taken at a {place1}, {place2}, or {place3}. I think there might be a {object1}, {object2}, {object3},... in this {img_type}. A creative short caption I can generate to describe this image is:



What does it mean to only understand symbols as defined by other symbols?

Multilingual Captioning with Retrieval Augmentation

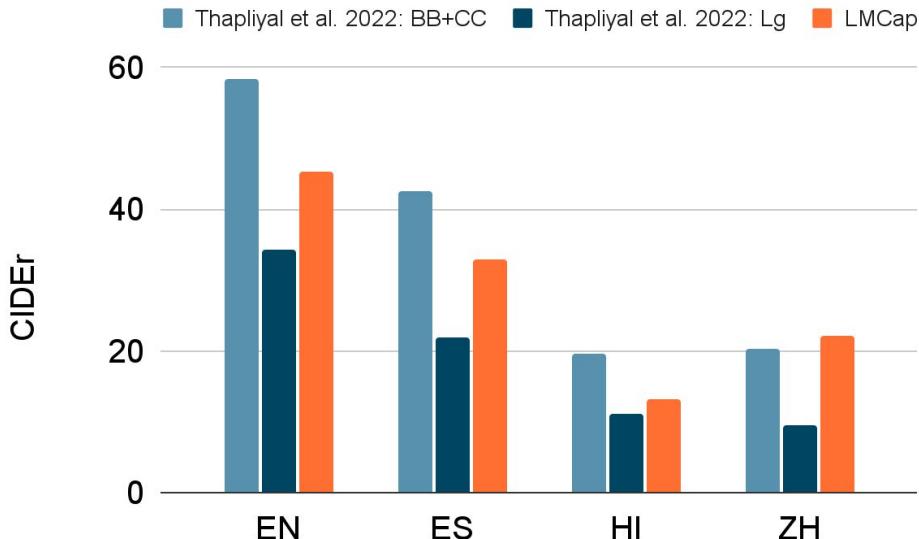


- Prompt with N-shot examples of transforming captions into the target language

Experimental Setup

- XGLM Language Model 564M - 7.6B params
- Multilingual CLIP (LAION)
- Experiments on XM3600
 - 100 images in 36 languages
- **No training or fine-tuning on any captioning data.**

Results



Competitive performance
compared to supervised models

Params	Config.	RAM	en	es	hi	zh
564M	K=4, N=3	6G	0.411	0.094	0.030	0.146
1.7B	K=4, N=3	12G	0.637	0.143	0.066	0.272
2.9B	K=4, N=3	16G	0.767	0.454	0.334	0.584
7.5B	K=4, N=3	22G	0.787	0.489	0.365	0.644

Need at least 2.9B parameter
decoder for multilingual generation

Qualitative Example

Retrieved Examples



two people and a kid skiing along a trail

an adult and two children are cross country skiing

two men and a little boy are skiing on a snowy spot

two adults on skis with a child on skis between them

Generated Captions

ENG: two people and a kid skiing along a trail

ESP: dos hombres y un niño esquiando en una pista de nieve

ZHO: 两个大人和一个小孩在雪地上滑雪

Conclusions

- Retrieval-augmentation is a powerful paradigm for V&L
 - Improve models with multimodal encoders
 - Improve lightweight trained models
 - Improve zero-training models
- Take advantage of in-domain resources and large pretrained models

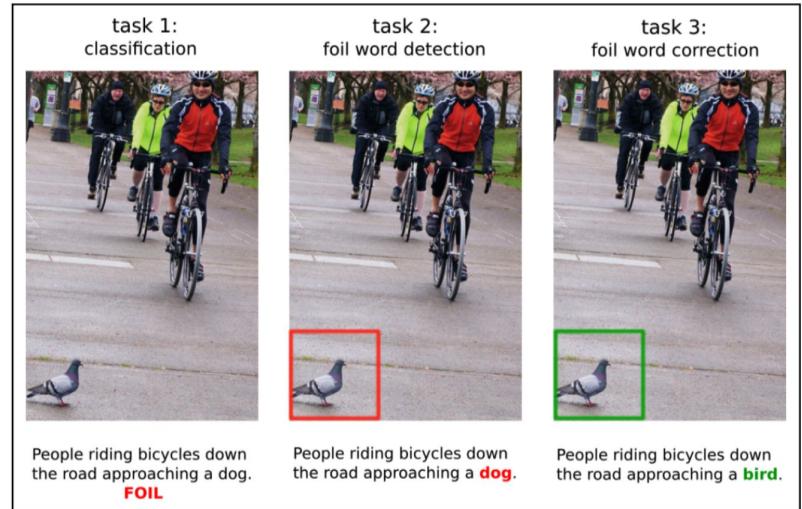
4. Understanding Multimodal Models

Beyond Benchmarking

- Many questions about what drives the success of these models?
 - Better contextualization: make better use of the multimodal inputs
 - Acquire certain “skills”, e.g. counting or localization
 - Understand linguistic structures
 - Something else?
- Model-internal behaviour
 - Attention mechanism patterns
- Probing
 - Tasks related to different skills

FOIL Captions

- Do V&L models really understand the relationship between words and images?
- Crowdsource datasets that contain contextually plausible but incorrect image–text pairs

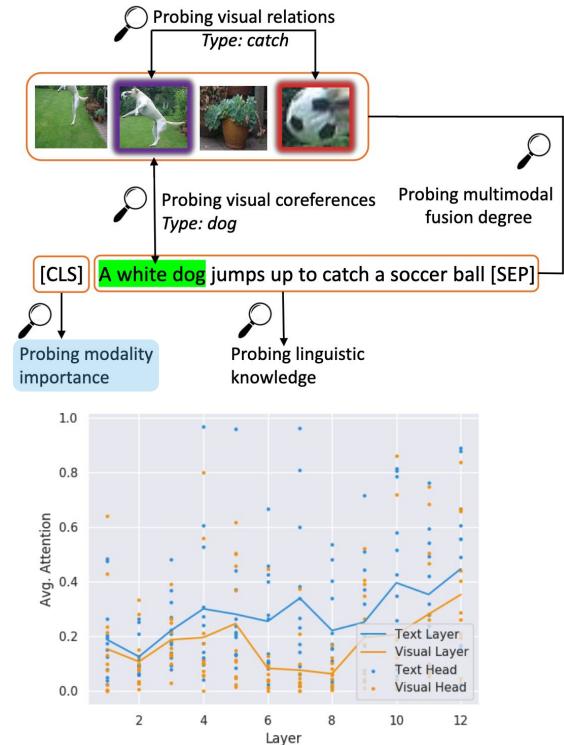


Vision and Language Understanding Evaluation

- Suite of five model probing tasks
- **Modality Influence:** Estimate the layer-wise contribution of each modality to the [CLS] embedding:

$$I_{M,j} = \sum_{i \in S} \mathbb{1}(i \in M) \cdot \alpha_{ij}$$

- The UNITER model relies more on textual features when fusing modalities throughout the model



VALSE Benchmark

- Test visio-linguistic capabilities with image-sentence foil pairs
- Image-sentence matching task
 - Existential quantifiers
 - Semantic number
 - Counting
 - Prepositional relations
 - Action replacement / swap
 - Co-reference



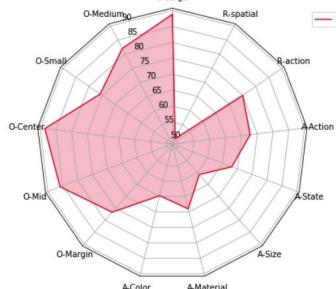
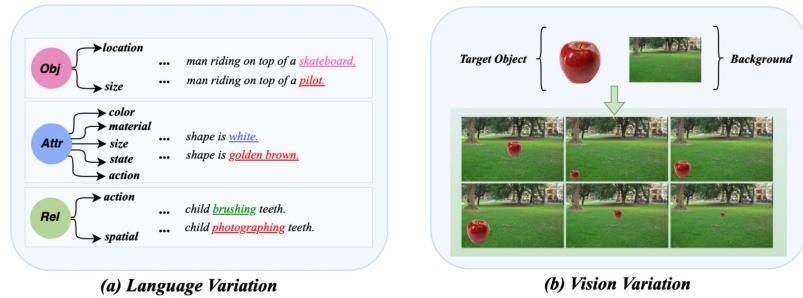
A small copper vase with **some flowers** / **exactly one flower** in it.

Metric	Model	Avg.
	Random	50.0
	GPT1*	60.7
	GPT2*	60.1
	CLIP	64.0
acc_r	LXMERT	59.6
	ViLBERT	63.7
	12-in-1	75.1
	VisualBERT	<u>46.4</u>

$$p(caption, img) > p(foil, img)$$

VL-CheckList

- Evaluate V&L models based on automatic manipulations to vision and language data.
- Image-Sentence matching task
- Radar chart overviews based on object / attribute / relationship variations



Subject-Verb-Object Probes

- Large-scale dataset with SVO triplets mined from Conceptual Captions and 14K images and with crowdsourced captions
- Foil detection formulation



WinoGround

- 1,600 text-image pairs to evaluate compositional understanding



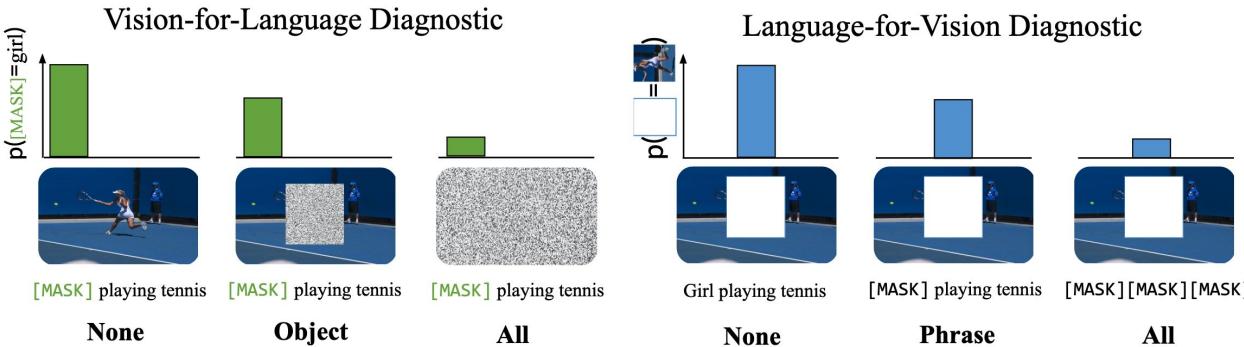
some plants
surrounding a
lightbulb



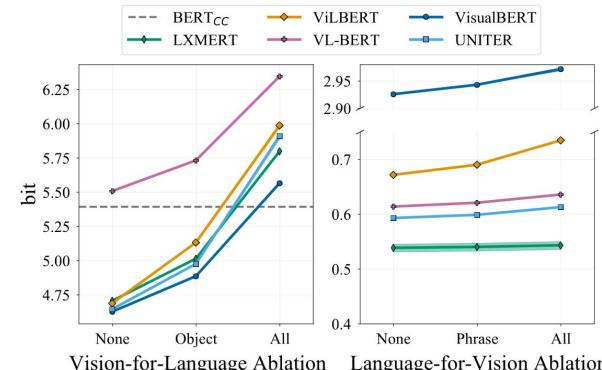
a lightbulb
surrounding
some plants

- Images sourced **with permission** from Getty.
- Differences are categorised into: swap dependent, swap-independent, and visual differences

Vision-for-Language?



- Pair of diagnostic evaluations that can be applied to any model that makes MLM and MRC predictions.



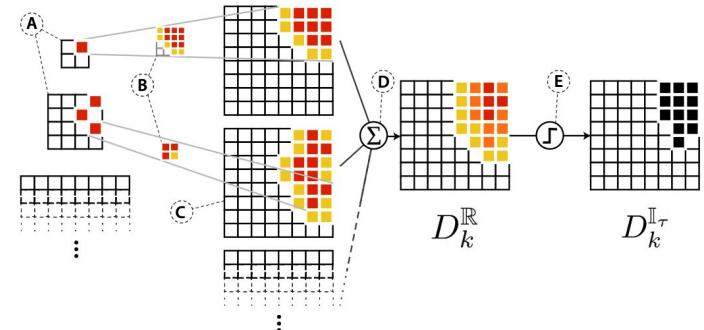
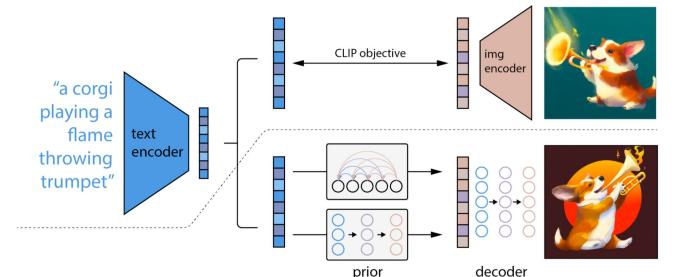
Summary

- Understanding and analysis is a vibrant area of research
- Foil detection is the most popular methodology
- Witnessing a methodological shift
 - attention analyses to linguistically-informed analyses
 - hand-crafted datasets
 - simpler accuracy-based metrics

5. Future Directions

Text-to-Image Generation

- Two main approaches: VQVAE and Conditional diffusion
- Big questions:
 - Do these models produce verifiably correct outputs?
 - How can they deal with cross-cultural generation?



Humour

- Could humour a new frontier in multimodal understanding?
 - Non-literal understanding
 - Deeper multimodal interaction
 - Social / world-level knowledge



Can you please pass the cow?

Task 1 Match the Caption + Cartoon

- I'd kill for some cream cheese
vs.
 - Can you please pass the cow?

Task 2 Rank the Finalist

- ## Welcome to Insomniacs Anonymous

Task

Explanation Generation

Human-authored

When drinking coffee or tea, people often add cream, and may ask others to pass it if it's on the other side of a table. But here, the mugs are huge, so instead of asking for a small cup of cream, they are asking for the entire cow, which is the appropriately-sized cream dispenser for these huge drinks.

From Pixels (OFA + T5-11B)

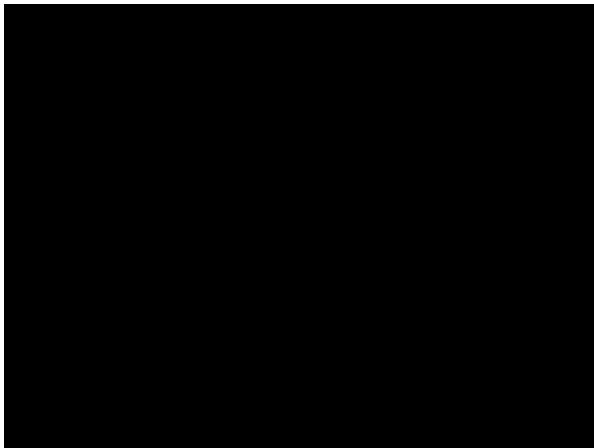
The joke is that the meeting participants are actually sitting on chairs made out of coffee mugs, which is an unlikely location for the discussion. Instead of asking for another mug of coffee, the person at the head of the table simply asks for "the cow", or a coffee machine.

From Description (5-shot GPT 3.5):

"Pass the cow" is an example of a non sequitur, something that looks like a logical thing to say, but doesn't make sense in context. The humor in this cartoon comes from the large size of the coffee mugs: they are so large that they resemble buckets rather than mugs, thus making the request to "pass the cow" almost reasonable.

Physical Understanding

- Predicting and explaining physical actions in the world will become of increasing importance as we create embodied agents



Q: How many objects are prevented by the tiny green triangle from falling into the basket?

Q: What is the color of the last object that collided with the tiny red circle?

Q: If any of the other objects are removed, will the tiny green circle end up in the basket?

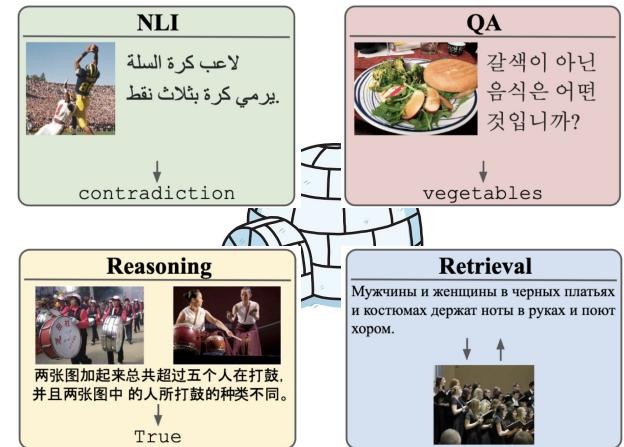
Multimodality and Interaction

- Learning to act in procedurally-generated video game environments with rich contexts, action spaces, and long-term rewards



Multilinguality

- The majority of Vision and Language research is in English
- We need resources, models, and evaluations to create useful multilingual multimodal models
- High-quality data requires:
 - time
 - money
 - community engagement



Q: What if we treated language as vision?

Language Modelling with Pixels

ICLR 2023



P. Rust



J. F. Lotz



E. Bugliarello



E. Salesky



M. de Lhoneux

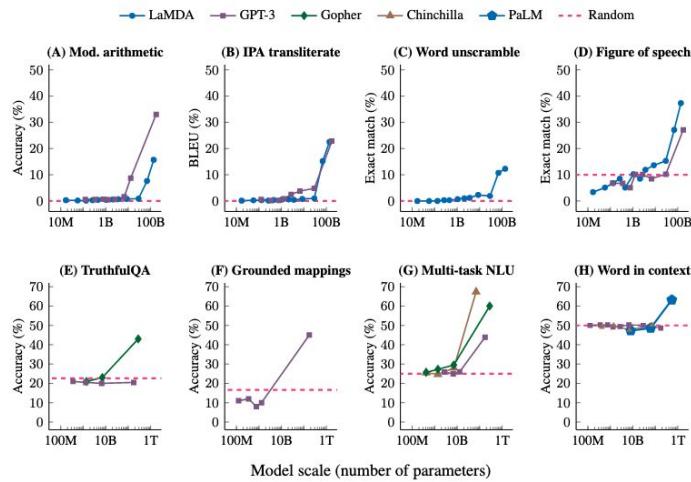
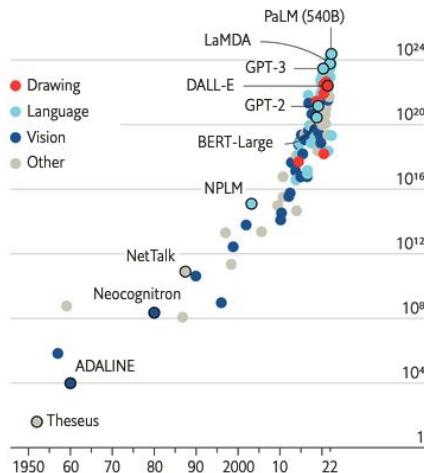


D. Elliott

NLP in the Era of Scale

The blessings of scale

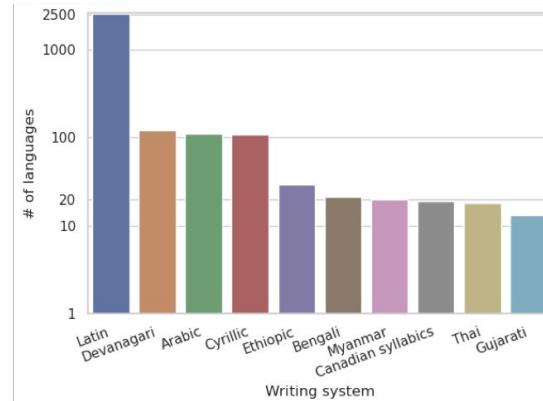
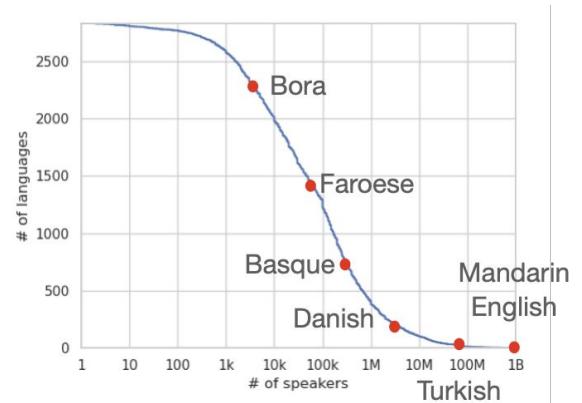
AI training runs, estimated computing resources used
Floating-point operations, selected systems, by type, log scale



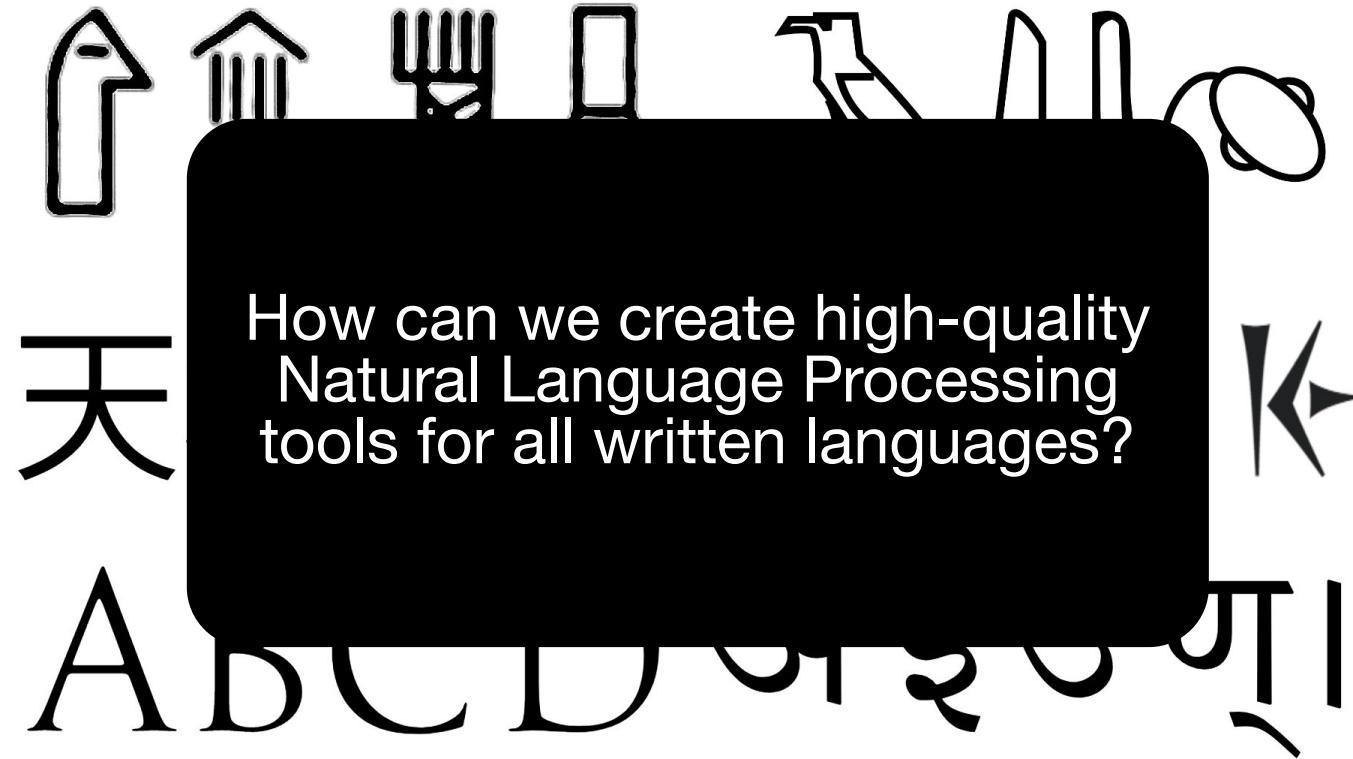
Source: www.economist.com/interactive/briefing/2022/06/11/huge-foundation-models-are-turbo-charging-ai-progress
Emergent Abilities of Large Language Models. (Wei+ TMLR)

What's Left? NLP for All Written Languages

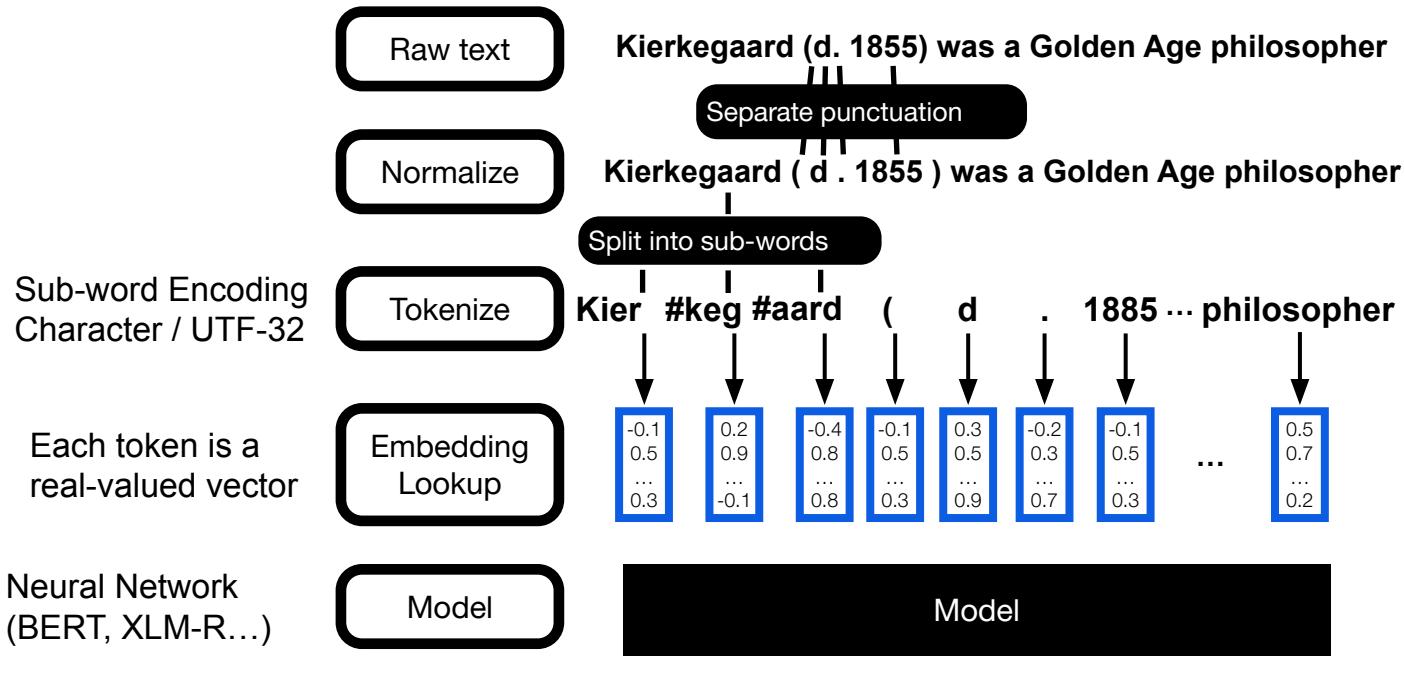
- There are 7,000 spoken languages, of which 3,000 are written
 - There is at least 400 languages with >1M speakers
- But NLP only covers 100 languages ([van Esch+ LREC22](#))
 - Lack of technological inclusion for billions of people



Slide credit: Sebastian Ruder



Recap: Language Processing is a pipeline

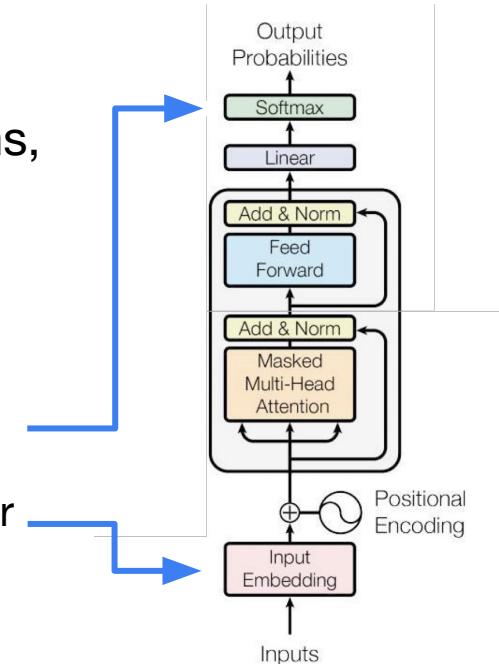


Q: What's Stopping Us?

- NLP is an **open vocabulary problem** and the ability of a model to process unseen words is determined by its vocabulary.
 1. “Trained” over a corpus: Byte-Pair Encoding ([Sennrich+ ACL16](#))
 - Unseen tokens not in the vocabulary unless there is a byte-level backoff
 2. Corpus independent: characters ([Clark+ TACL22](#)) / bytes ([Xue+ ACL22](#))
 - Need to deal with longer sequence lengths

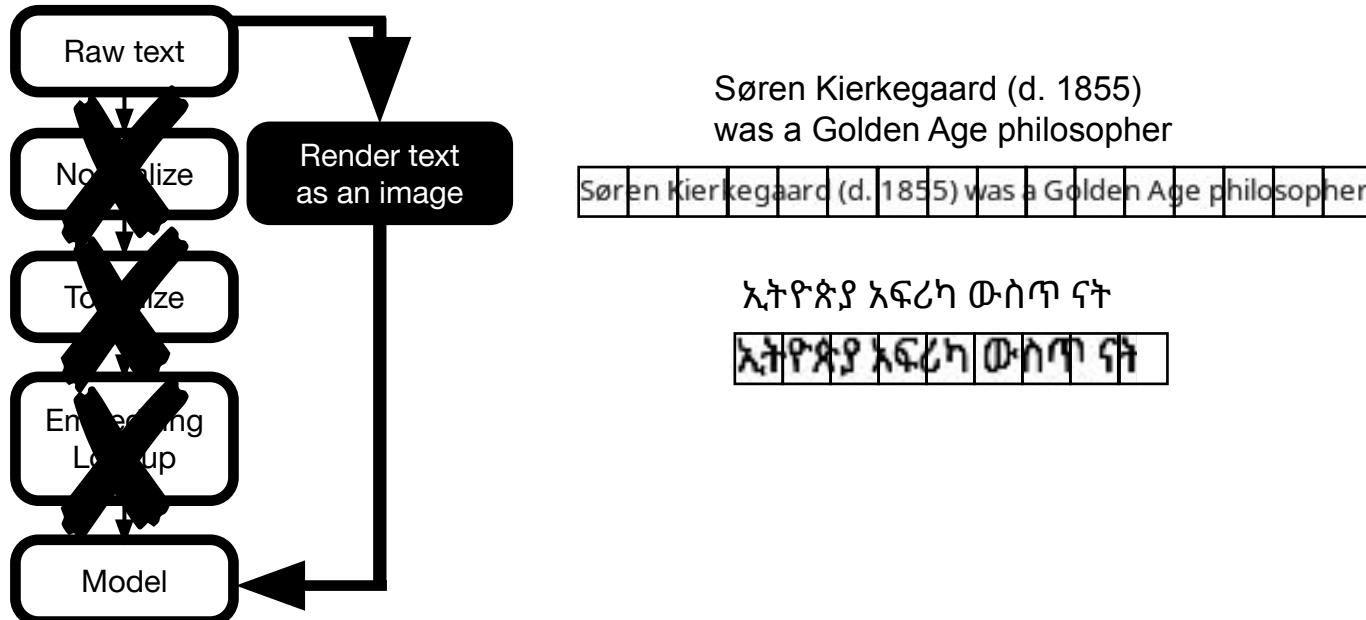
A: The Vocabulary Bottleneck

- Language models have discrete input and output vocabularies expressed over a finite inventory of tokens, characters, words, sub-words, etc.
 - This creates a bottleneck in two places:
 1. *Computational bottleneck* in the Output layer
 2. *Representational bottleneck* in the Embedding layer

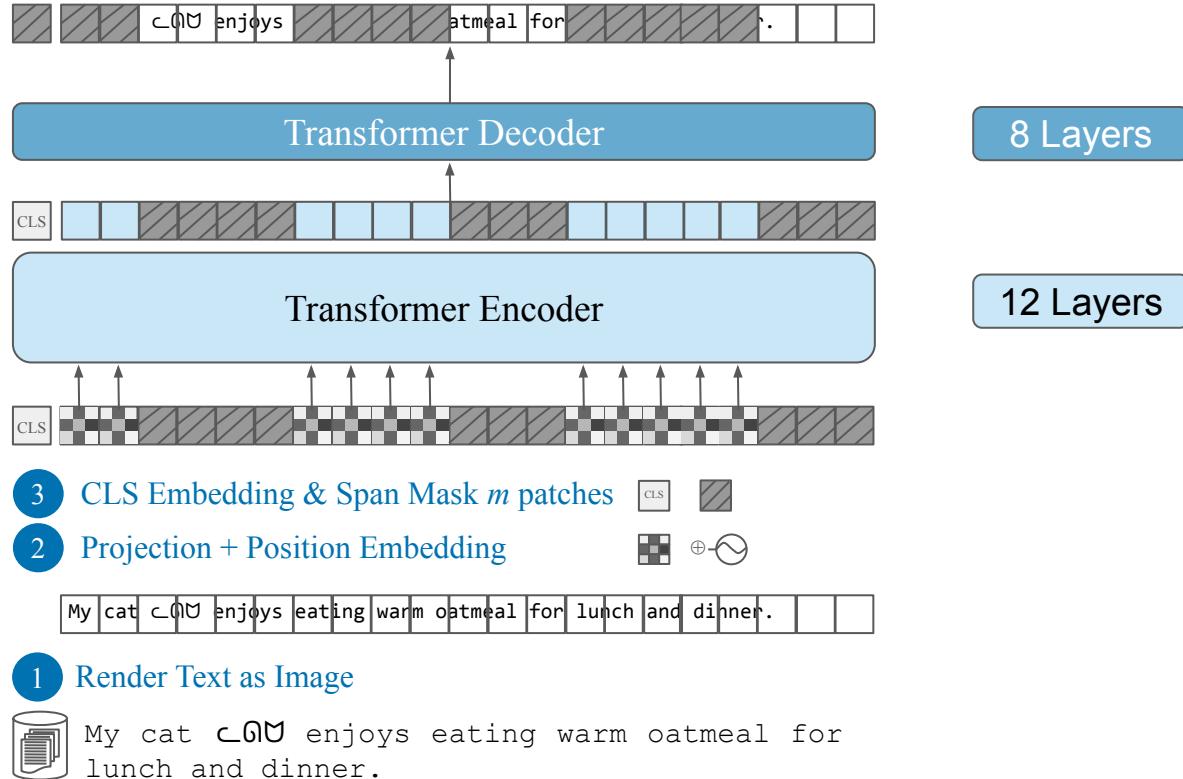


Pixel-based Language Modelling

- Key insight: treat language processing as visual processing



The PIXEL Model



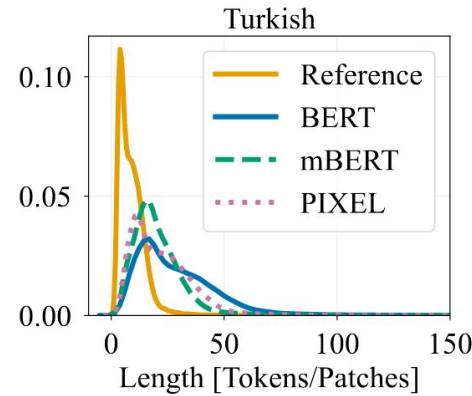
16pixel x 16pixel patch

Google Noto Fonts

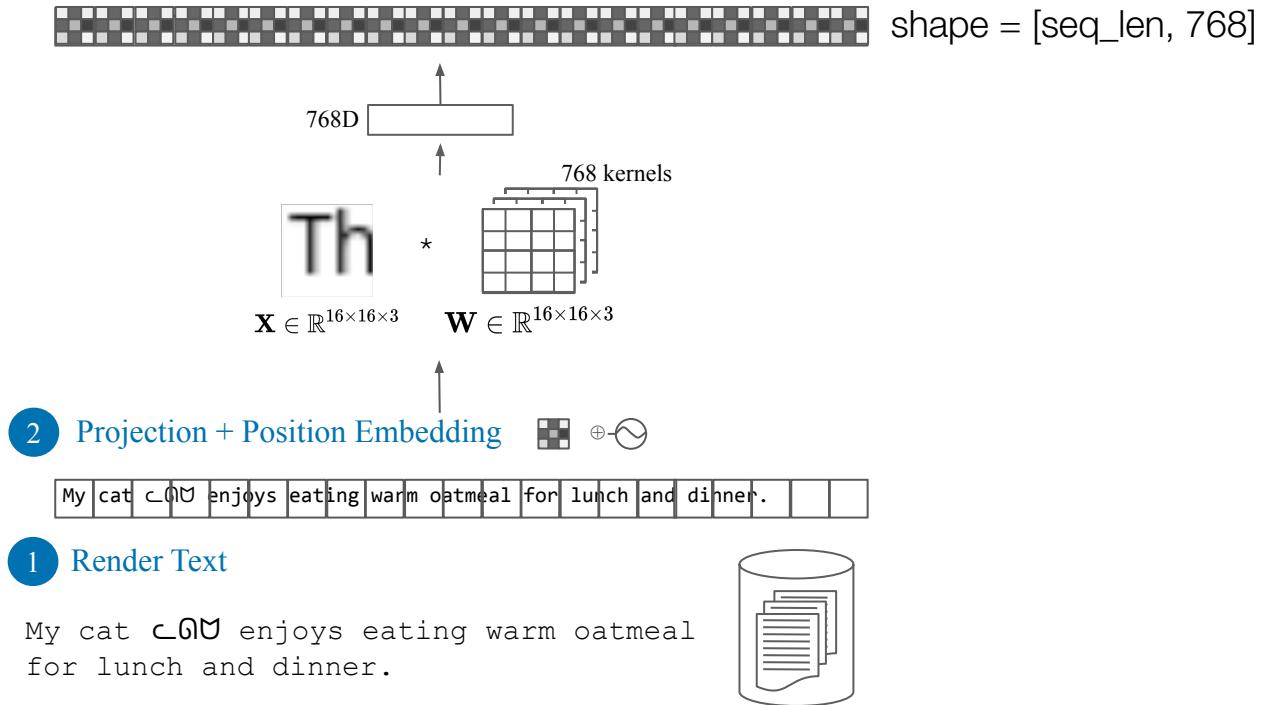
PyGame / PangoCairo

Rendered Text is Compact

- Proportion of text that can be encoded in k subwords / patches.
- PIXEL encoding produces sequence lengths that are at least as long as as BERT.
 - No length penalty.

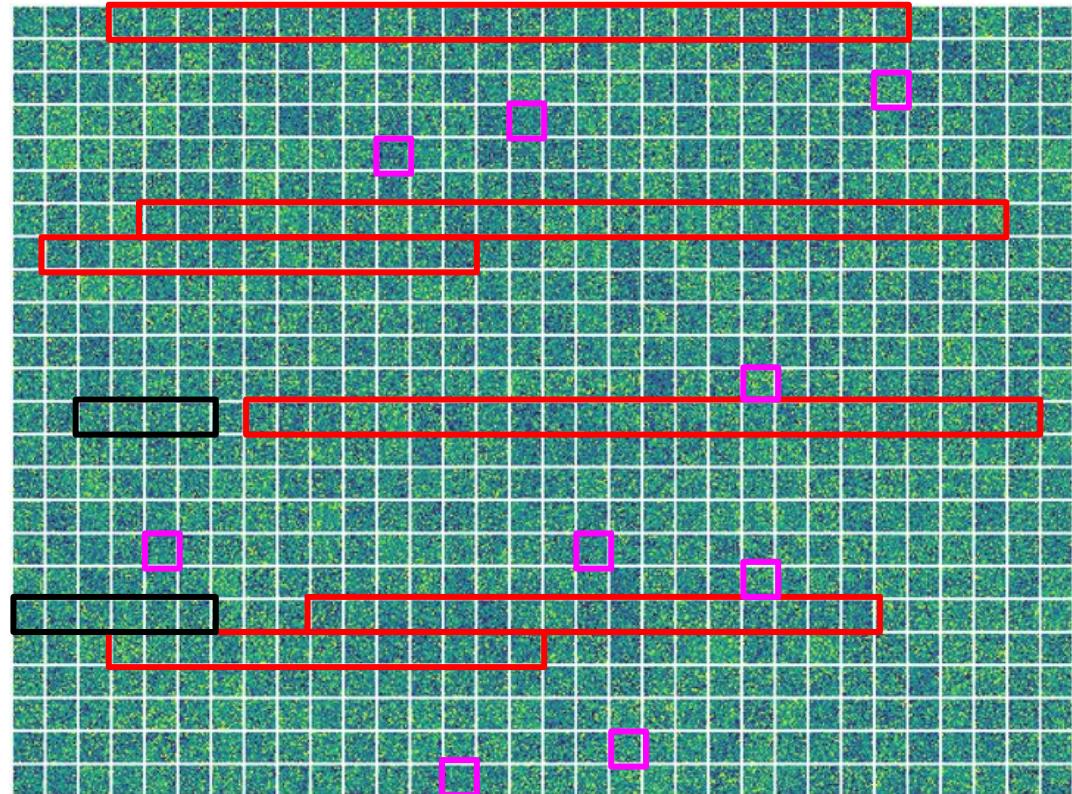


Detail: “Embedding” Layer



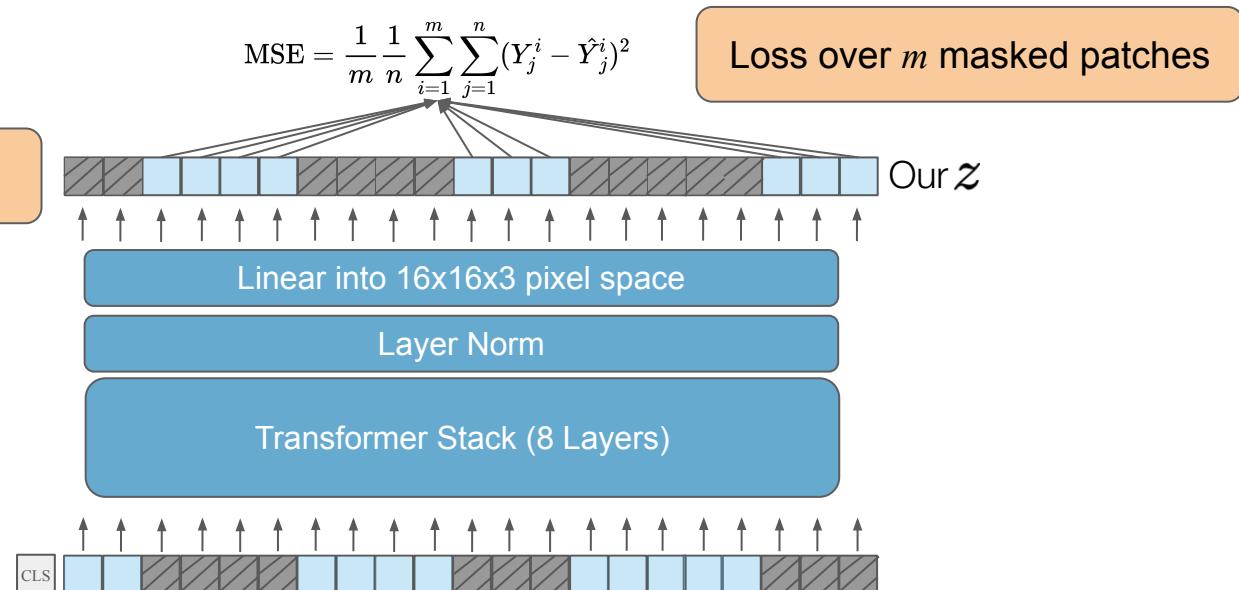
Visualization of Convolution Kernels

1. Some kernels learn about the presence / absence of any pixels.
2. Many kernels capture horizontal strokes
3. Only a few kernels capture curved shapes (*likely due to letters rendered across patch boundaries*)



Evolution of Conv2D weights during pretraining step 10K–1M

Detail: Pixel Reconstruction



A new type of generative model

Penguins are designed to be streamlined and hydrodynamic, so having long legs would add expanding. Having short legs with webbed feet to act like runners, helps to give them that the do-like figure. If we compare bird anatomy with humans, we would see something peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones are close to ours. What most people mistake for knees are actually the ankles of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the boxes of the bird! So how does this look inside of a penguin? In the images below, you can see boxes surrounding the penguins' knees.

Penguins are designed to be streamlined and hydrodynamic, so having long legs would add expanding. Having short legs with webbed feet to act like runners, helps to give them that the do-like figure. If we compare bird anatomy with humans, we would see something peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones are close to ours. What most people mistake for knees are actually the ankles of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the boxes of the bird! So how does this look inside of a penguin? In the images below, you can see boxes surrounding the penguins' knees.

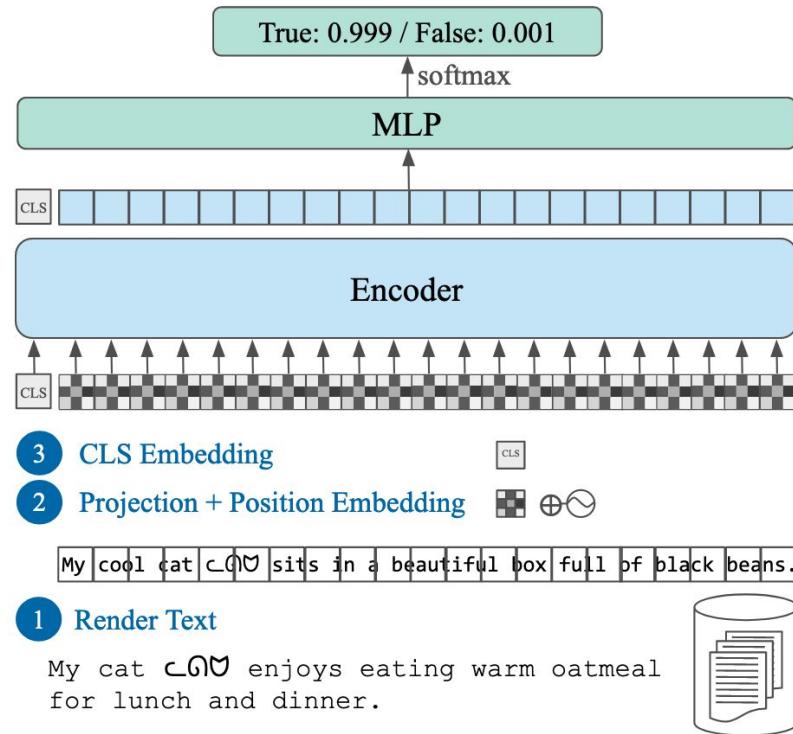
Penguins are designed to be streamlined and hydrodynamic, so having long legs would add expanding. Having short legs with webbed feet to act like runners, helps to give them that the do-like figure. If we compare bird anatomy with humans, we would see something peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones are close to ours. What most people mistake for knees are actually the ankles of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the boxes of the bird! So how does this look inside of a penguin? In the images below, you can see boxes surrounding the penguins' knees.

100K steps

500K steps

1M steps

Adapting to Downstream Tasks



The Benefits of Pixels

- **PIXEL can process anything that can be rendered**
 - Open vocabulary which is easily extensible to unseen text
 - Support all written languages
 - Greater flexibility to process written language in different forms
 - PDFs, scanned newspapers, etc.
- **Complete parameter sharing** from the input representation
 - Unlike separate-but-related subwords in an embedding matrix

Pretraining

- **English Dataset:** English Wikipedia and Books Corpus
- **Masking:** 25% Span Masking
- **Maximum sequence length:** 529 patches (16×8464 pixels)
- **Compute:** 8 x 40GB A100 GPUs for 8 days
- **Parameters:** 86M encoder + 26M decoder

There is only 0.05% non-English text in our pretraining data (estimated by Blevins and Zettlemoyer 2022)

The **Great Wall of China** (traditional Chinese: 萬里長城; simplified Chinese: 万里长城; pinyin: Wàn lǐ Chángchéng)

Downstream Tasks

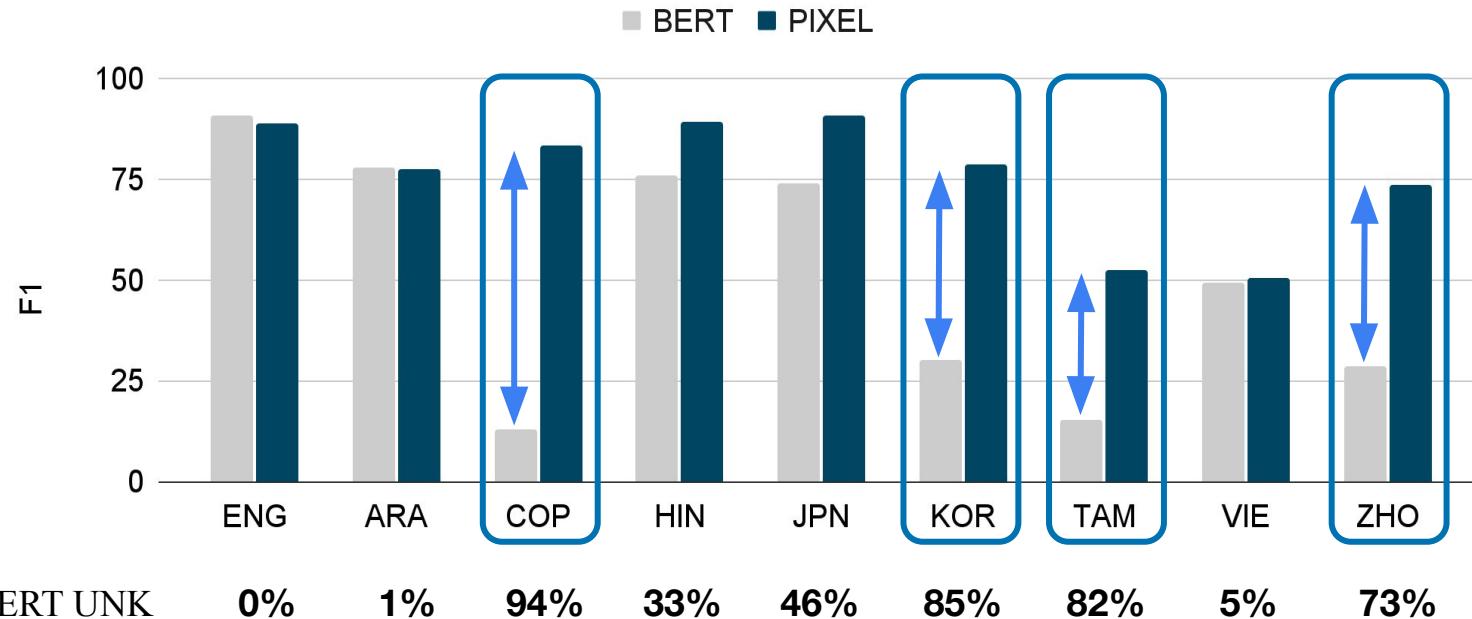
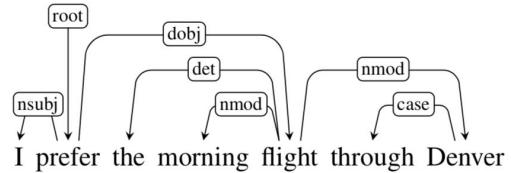
- **Datasets:** Universal Dependencies, MasakhaNER, GLUE, Zeroé
- **Models:**

	Parameters	Pretraining Data
PIXEL _{BASE}	86M	English Wikipedia + Bookcorpus
BERT _{BASE}	110M	—
CANINE-C	127M	104-languages from Wikipedia

Similar pretraining setup

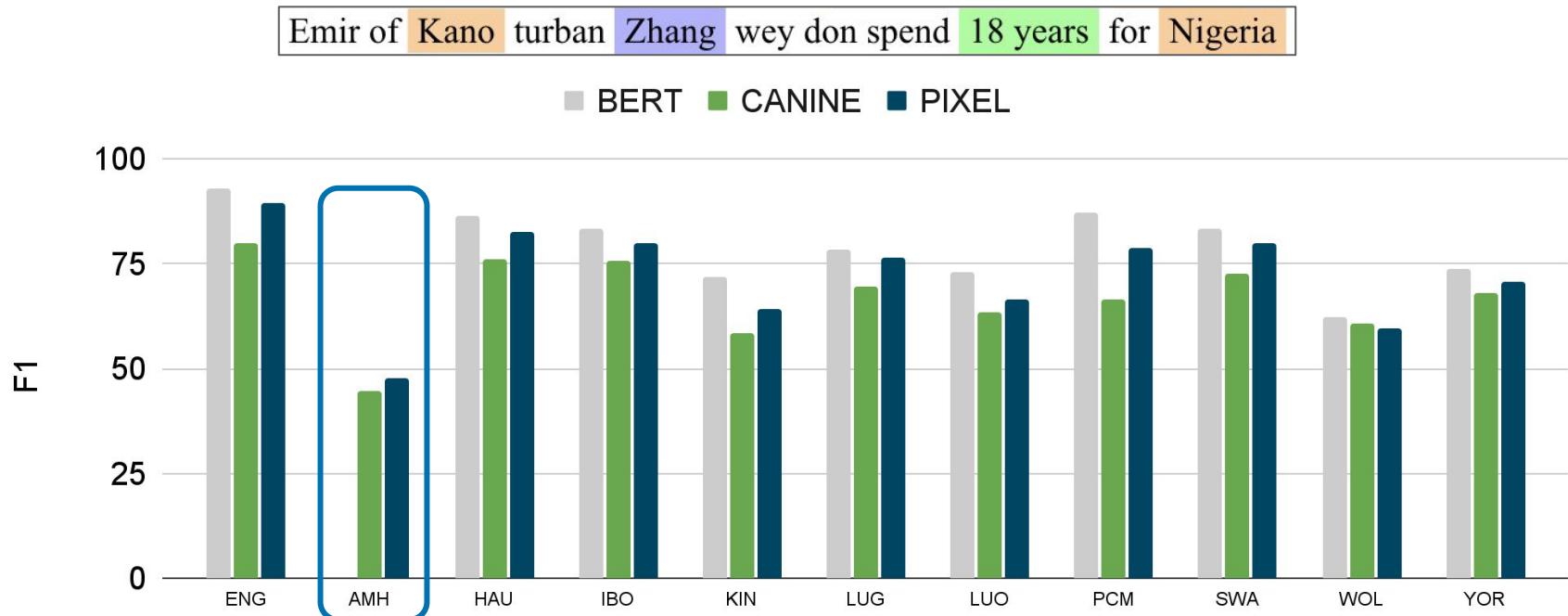
Tries to solve the same
problem using UTF-32

Dependency Parsing Results



PIXEL vastly outperforms BERT on unseen scripts

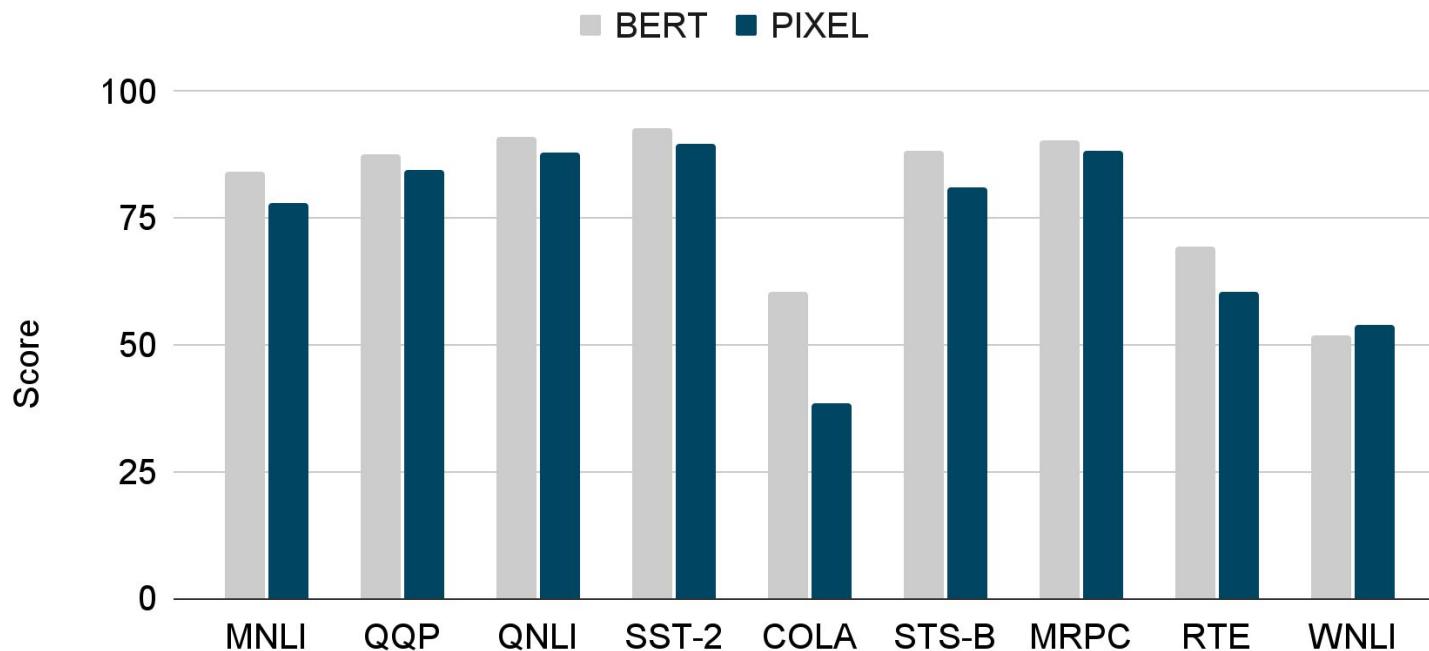
Named Entity Recognition in African Languages



PIXEL outperforms BERT
on the non-Latin script

PIXEL outperforms the
multilingually pretrained CANINE-C

GLUE: Sentence-level Understanding



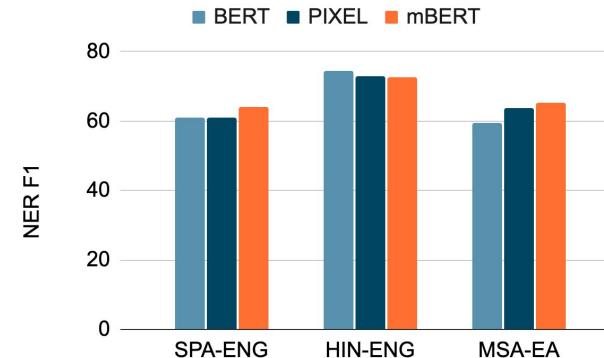
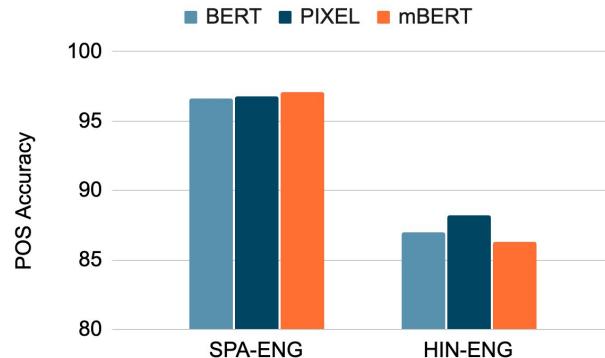
BERT outperforms PIXEL on English sentence-level tasks

Code-switching

- Grammatically consistent switching between two or more languages in a single utterance ([Joshi 1982](#))

“Michael Jackson PER revivió en los **Billboard 2014** EVENT

- NER and POS tagging evaluation ([Aguilar et al., 2020](#))



PIXEL adapts to new fonts

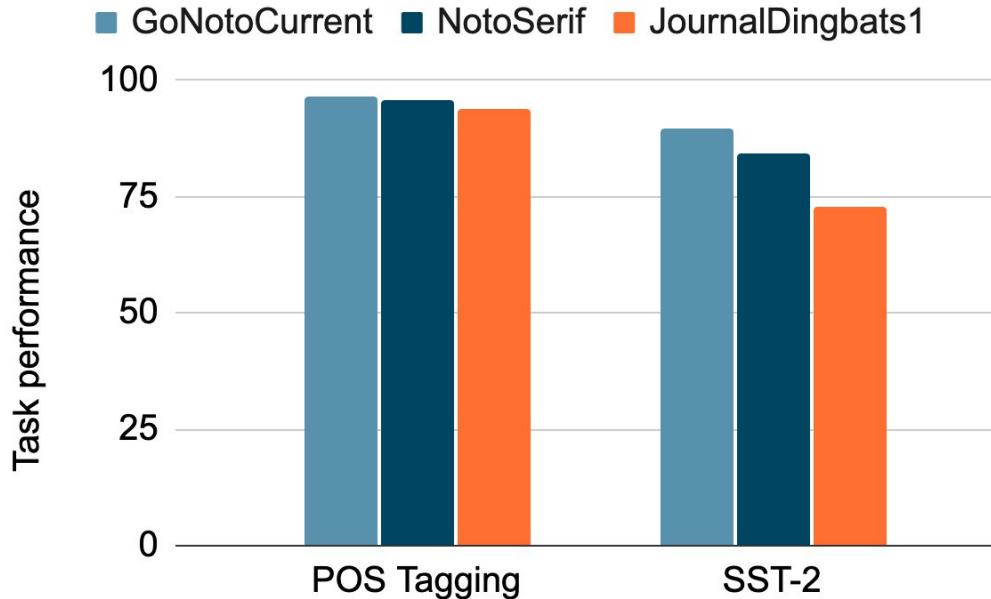
GoToCurrent

My cat loves oatmeal and pancakes. ■

NotoSerif-Regular

My cat loves oatmeal and pancakes. ■

JournalDingbats1



Crucial importance for
processing new
scripts, e.g.
handwritten text

Big Question

- Masked Language Modelling is **classic distributional semantics**: model the identity of a (masked) word, given the unmasked context

$$-\sum_{m \in M} \log p(x_m | \mathbf{x}_{\setminus m})$$

BERT: Masked Language Modelling

$$\frac{1}{M} \sum_{m \in M} (Y_m - \hat{Y}_m | \mathbf{x}_{\setminus m})^2$$

PIXEL: Masked Autoencoding

- Why is it possible to learn a good encoder by predicting masked pixels?

Conclusions

- PIXEL is a new type of language model that tackles the open vocabulary problem using visually rendered text.
 1. This enables high-quality transfer to unseen scripts.
 2. Robustness to orthographic attacks and code-switching.
- *My opinion:* Language is special but its computer format should be as flexible and expressive as possible.

Questions?

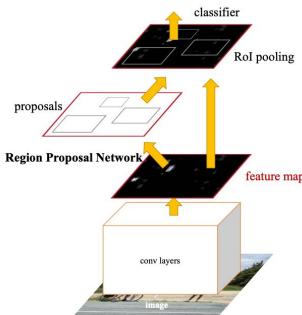
Wrap-up

1. Datasets

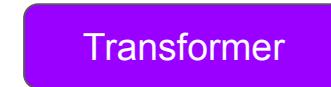
some sheep walking in the middle of a road
a herd of sheep with green markings walking down the road
a herd of sheep walking down a street next to a lush green grass covered hillside.
sheared sheep on roadway taken from vehicle, with green hillside in background.
a flock of freshly sheered sheep in the road.



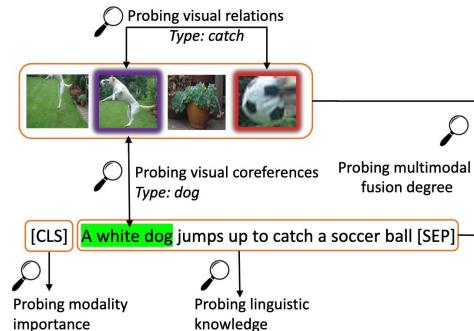
2. Representation



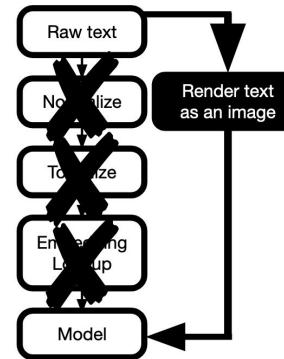
3. Modelling



The red horse



4. Understanding



5. New Ideas

Where to find more research?

CVPR

NeurIPS

ACL

NAACL

LREC

ICML

EACL

ECCV

ICCV

IJCAI

ICLR

COLING

arXiv

...



DALL-E 2: “digital art of someone drinking from arxiv firehose every morning”

Surveys

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 41, NO. 2, FEBRUARY 2019

Multimodal Machine Learning: A Survey and Taxonomy

Tadas Baltrušaitis , Chaitanya Ahn

Journal of Artificial Intelligence Research 71 (2021) 1183 - 1317

Submitted 09/2019; published 08/2021

Abstract—Our experience of the world is multimodal - we see objects, hear sounds, smell scents, taste flavors, and feel textures. Modality refers to the way in which something happens or is experienced. It includes multiple such modalities. In order for Artificial Intelligence to be able to interpret such multimodal signals together, *Multimodal learning* is needed to extract information from multiple modalities. It is a vibrant multi-disciplinary field. Instead of focusing on specific multimodal applications, this paper surveys and presents them in a common taxonomy. We go beyond the typical challenges that are faced by multimodal machine learning, namely: This new taxonomy will enable researchers to better understand the

Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods

Aditya Mogadala
Marimuthu Kalimuthu
Dietrich Klakow
Spoken Language Systems (LSV)
Saarland Informatics Campus
Saarland University
66123 Saarbrücken, Germany

Multimodal Learning with Transformers: A Survey

Peng Xu, Xiatian Zhu, and David A. Clifton

Abstract—Transformer is a promising neural network learner, and has achieved great success in various machine learning tasks. Thanks to the recent prevalence of multimodal applications and big data, Transformer-based multimodal learning has become a hot topic in AI research. This paper presents a comprehensive survey of Transformer techniques oriented at multimodal data. The main contents of this survey include: (1) a background of multimodal learning, Transformer ecosystem, and the multimodal big data era, (2) a systematic review of *Vanilla Transformer*, *Vision Transformer*, and *multimodal Transformers*, from a geometrically topological perspective, (3) a review of multimodal Transformer applications, via two important paradigms, i.e., for multimodal pretraining and for specific multimodal tasks, (4) a summary of the common challenges and designs shared by the multimodal Transformer models and applications, and (5) a discussion of open problems and potential research directions for the community.

Index Terms—Multimodal Learning, Transformer, Introductory, Taxonomy, Deep Learning, Machine Learning.

Predictions & Speculations

- Increasing societal impact of V&L models
 - Both for entertainment and for misinformation
- Shift in focus to zero-shot instruction-based models
 - Fine-tuning is too expensive for each task
- Concentrated focus on understanding how models work
 - Bigger and better datasets will continue to be major contributions
- Big challenge to evaluate bidirectional generative models

Reflection

- Multimodality is an exciting area of research
 - Big leaps in recent years powered by pretraining and bigger data
- Some friction between public/non-public models and data
 - Opportunity to make an impact at your own scale
- This area is still in its infancy
 - Huge opportunities for your ideas to make an impact in every area
 - Modelling, task creation and development, evaluation, understanding