Deep learning for NLP

Eneko Agirre, Gorka Azkune
Ander Barrena, Oier Lopez de Lacalle

@eagirre @gazkune @4nderB @oierldl #dl4nlp

http://ixa2.si.ehu.eus/eneko/dl4nlp

Session 7: Bridging the gap between natural languages and the visual world

Plan for the course

- Introduction: machine learning and NLP
- Multilayer perceptron
- Word representation and
  Recurrent neural networks (RNN)
- Sequence-to-Sequence (seq2seq) and
  Machine Translation
- Attention, transformers and
  Natural language inference
- Pre-trained transformers, BERTology and final words
- Bridging the gap between natural languages
  and the visual world

Plan for this session

- NLU and grounding
- Grounding and multimodal tasks:
  - Exemplary discriminative tasks
  - Exemplary generative tasks
  - More tasks?
- Multimodal systems
- Multimodal systems for NLU
- Conclusions

NLU and GROUNDING

NLU and grounding

Large LMs are great!

NLU and grounding

Large LMs are great! QA

SQUAD2.0

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**

SQUAD1.1

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Jul 24, 2021	{ANNA} (single model) *LG AI Research*	**90.622**	**95.719**

NLU and grounding

Large LMs are great! GLUE

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE
1	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4
2	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1
3	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6
4	DIRL Team	DeBERTa + CLEVER		91.0	74.5	97.5	93.3/91.0	93.4/93.1	76.4/90.9	92.1	91.8	96.7	93.1
5	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5
6	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2
20	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6

NLU and grounding

Large LMs are great! SuperGLUE

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	Liam Fedus	SS-MoE		91.0	92.3	96.9/98.0	99.2	89.2/65.2	95.0/94.2	93.5	77.4	96.6	72.3	96.1/94.1
	2	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
	3	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
+	4	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+	5	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
	6	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	7	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9

NLU and grounding

- Superhuman performance on NLU benchmarks.
- LMs are also very close to humans on more difficult tasks like commonsense knowledge.
- Then, do LMs **understand** natural language?

NLU and grounding

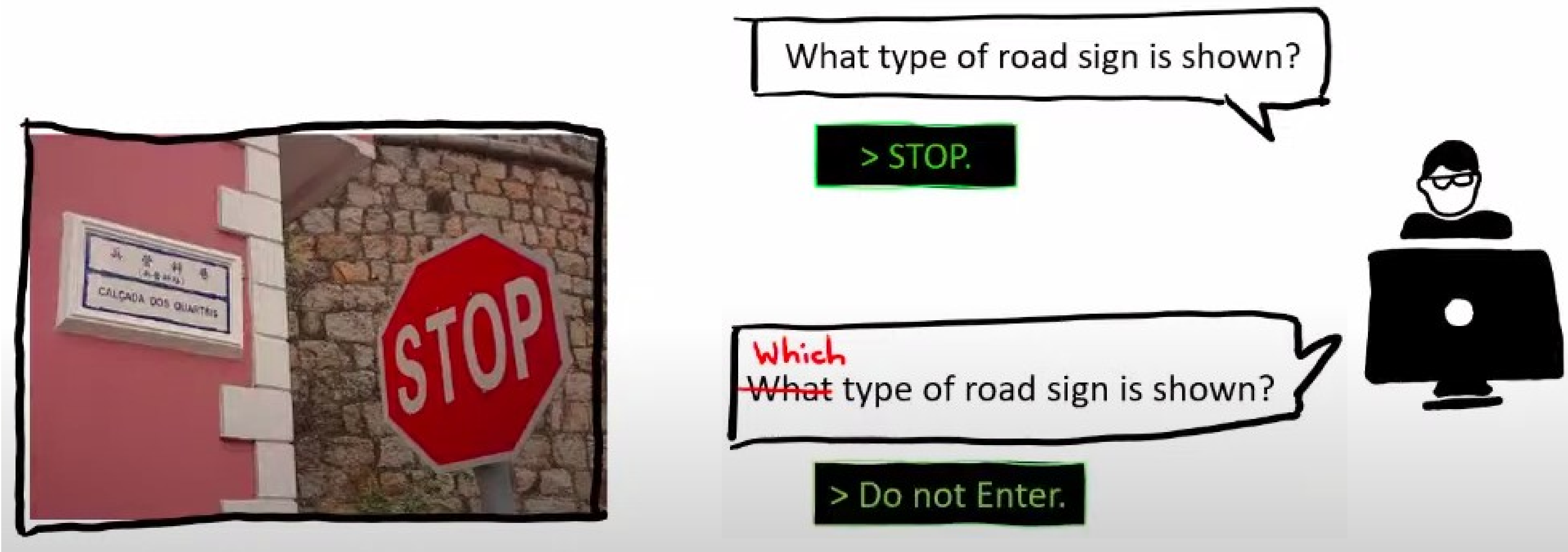
If we scratch the surface:

- LMs are sensitive to phrasing.

NLU and grounding

If we scratch the surface:

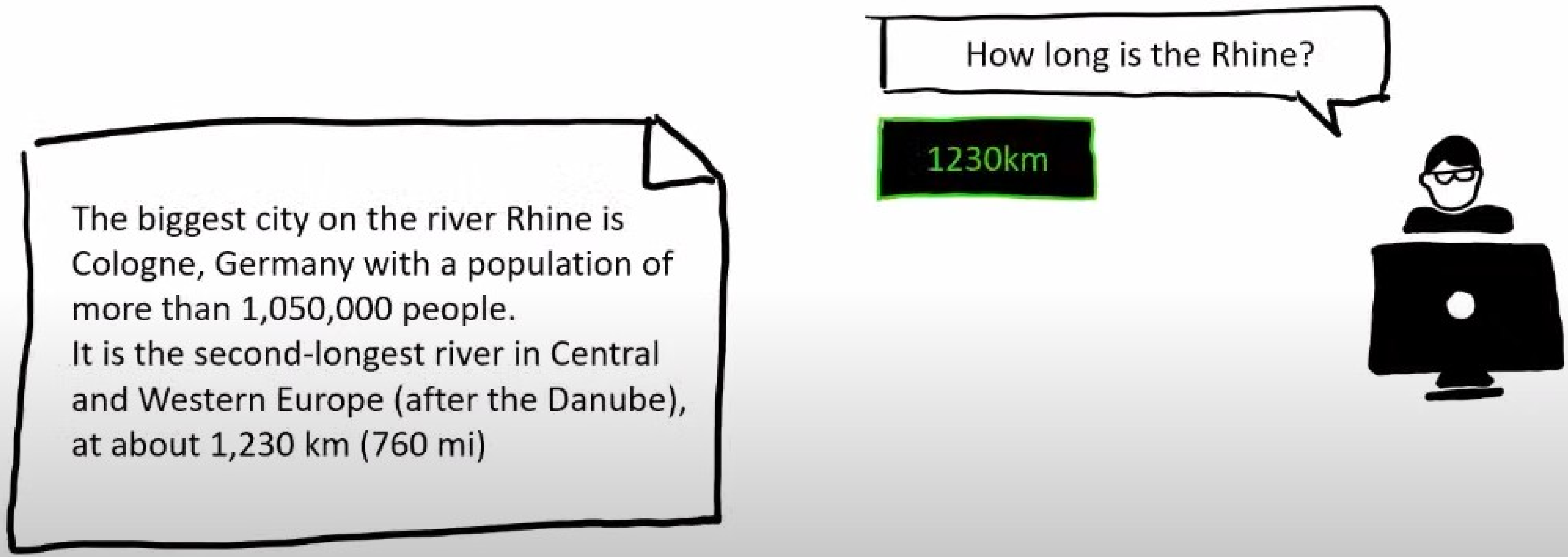
- LMs are sensitive to phrasing.



NLU and grounding

If we scratch the surface:

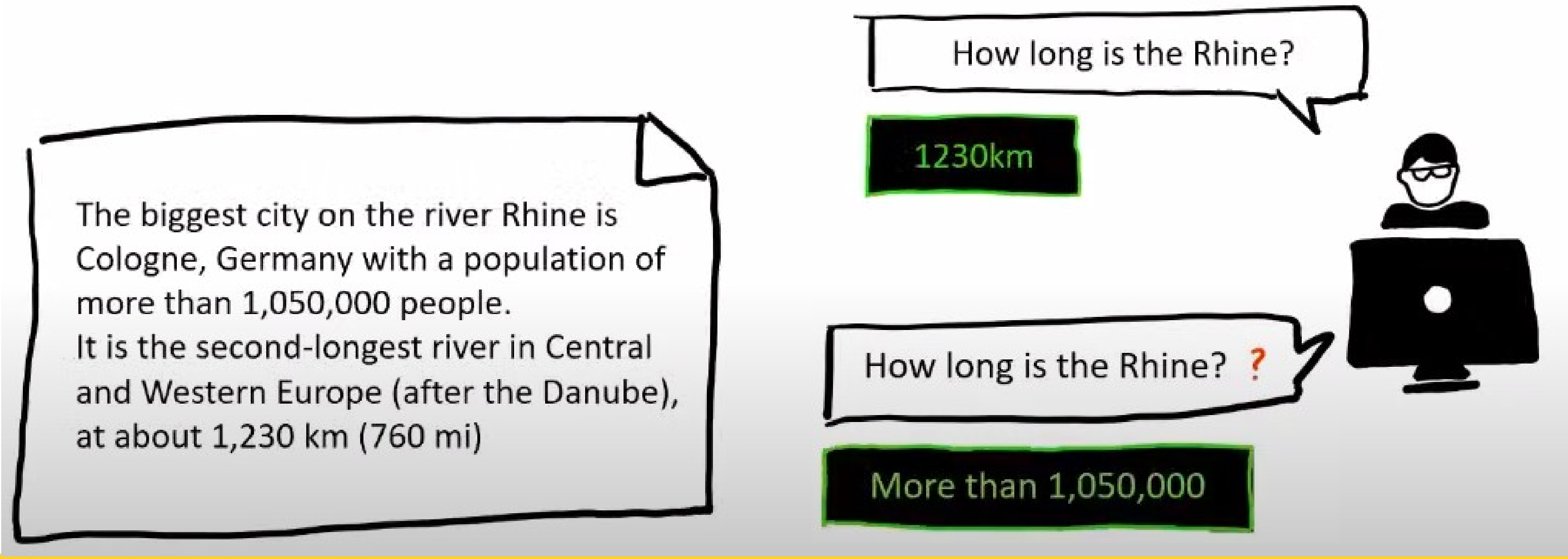
- LMs are sensitive to typos.



NLU and grounding

If we scratch the surface:

- LMs are sensitive to typos.



NLU and grounding

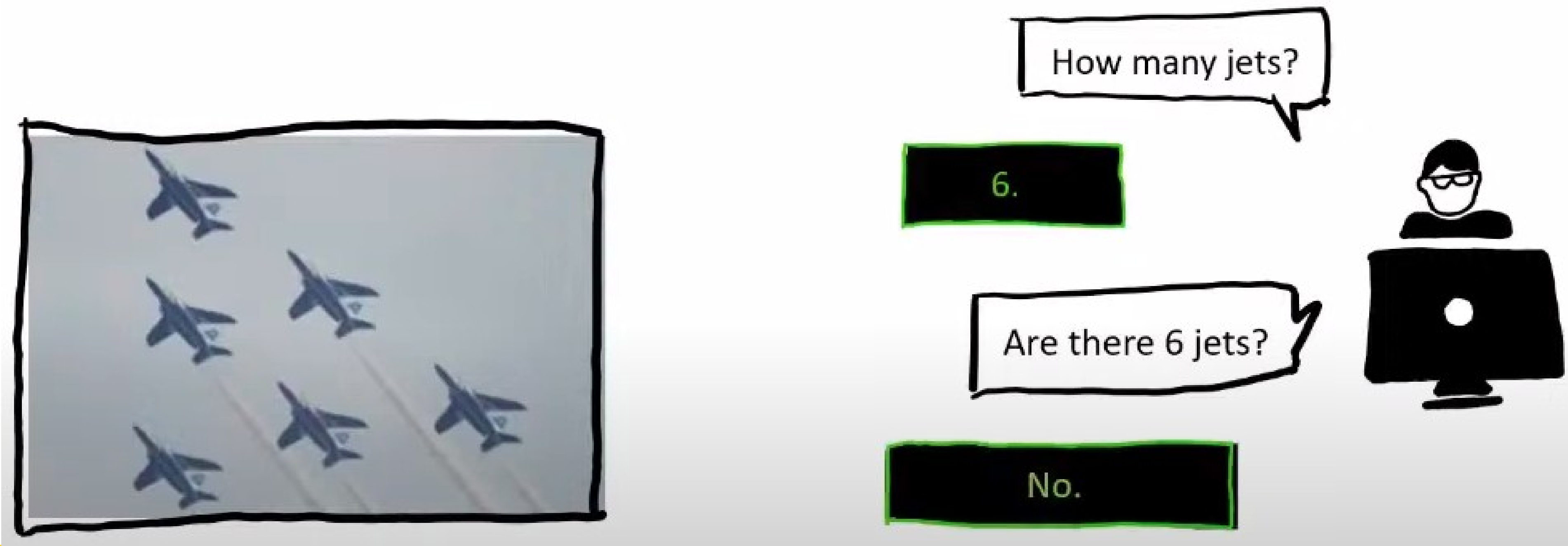
If we scratch the surface:

- LMs are not consistent with themselves.

NLU and grounding

If we scratch the surface:

- LMs are not consistent with themselves.



NLU and grounding

Two main conclusions from those research works:

- We have a problem with leaderboards and automatic evaluation metrics.
  - Dataset artifacts.
  - A single metric is not enough.
- LMs do not understand as we do.

NLU and grounding

Two main conclusions from those research works:

- ~~We have a problem with leaderboards and automatic evaluation metrics.~~
  - ~~Dataset artifacts.~~
  - ~~A single metric is not enough.~~
- LMs do not understand as we do.

NLU and grounding

Two main conclusions from those research works:

- ~~We have a problem with leaderboards and automatic evaluation metrics.~~
  - ~~Dataset artifacts.~~
  - ~~A single metric is not enough.~~
- **LMs do not understand as we do.**

NLU and grounding

Do LMs understand natural language or learn its meaning?

- Bender et al. argue that the language modeling task, using only **form** as training data, cannot lead to learning of meaning.
- Human-analogous NLU is a grand challenge of artificial intelligence, which involves mastery of the structure and use of language and the ability to **ground** it in the world.

Source: Bender et al. Climbing towards NLU: On meaning, form, and understanding in the age of data. 2020

NLU and grounding

What is meaning?

- Form: any observable realization of language: marks on a page, pixels or bytes in a digital representation of text, or movements of the articulators.
- Meaning: the relation between the form and something **external to language.**

NLU and grounding

Grounding distributional representations in the real world is challenging. Approaches:

- Train distributional models on corpora **augmented with perceptual data**, such as photos or other modalities.
- Look to **interaction data**, e.g. a dialogue corpus with success annotations, low-level success signals such as emotional stress or eye gaze... a signal about the felicitous uses of forms.

NLU and grounding

Grounding distributional representations in the real world is challenging. Approaches:

- Train distributional models on corpora **augmented with perceptual data**, such as photos or other modalities.
- ~~Look to **interaction data**, e.g. a dialogue corpus with success annotations, low-level success signals such as emotional stress or eye gaze... a signal about the felicitous uses of forms.~~

NLU and grounding

Where can we ground language?

- Visual information: images and videos.
- Knowledge graphs.
- Audio data.
- …

NLU and grounding

Where can we ground language?

- **Visual information: images and videos.**
- Knowledge graphs.
- Audio data.
- …

NLU and grounding

Why visual information?

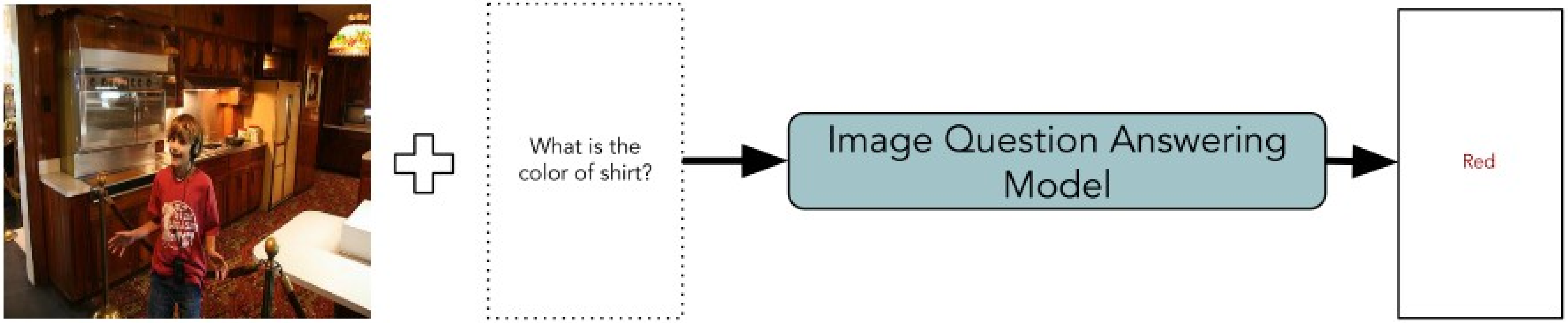
- There are many and large resources.
- Videos and images contain many relevant information about our world.
- KGs are expensive and limited.
- Audio is also present at videos (can be included easily).

GROUNDING and MULTIMODAL TASKS

Grounding and multimodal tasks

Exemplary discriminative task:

- Visual Question Answering (VQA): given an image and a question about that image, produce the answer

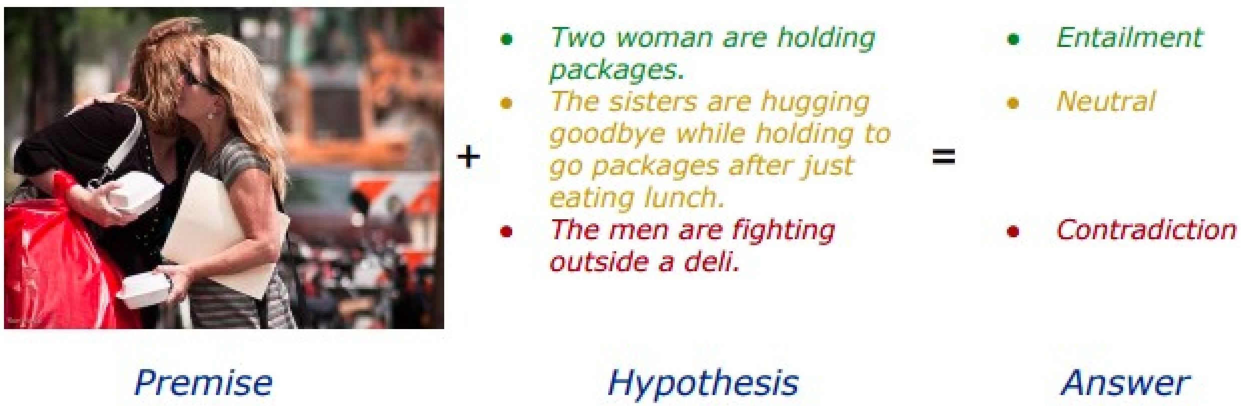


Source: Mogadala et al. Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods. 2019

Grounding and multimodal tasks

Exemplary discriminative task:

- Visual entailment: given an image and a sentence, decide whether the sentence is an entailment, neutral or contradiction.

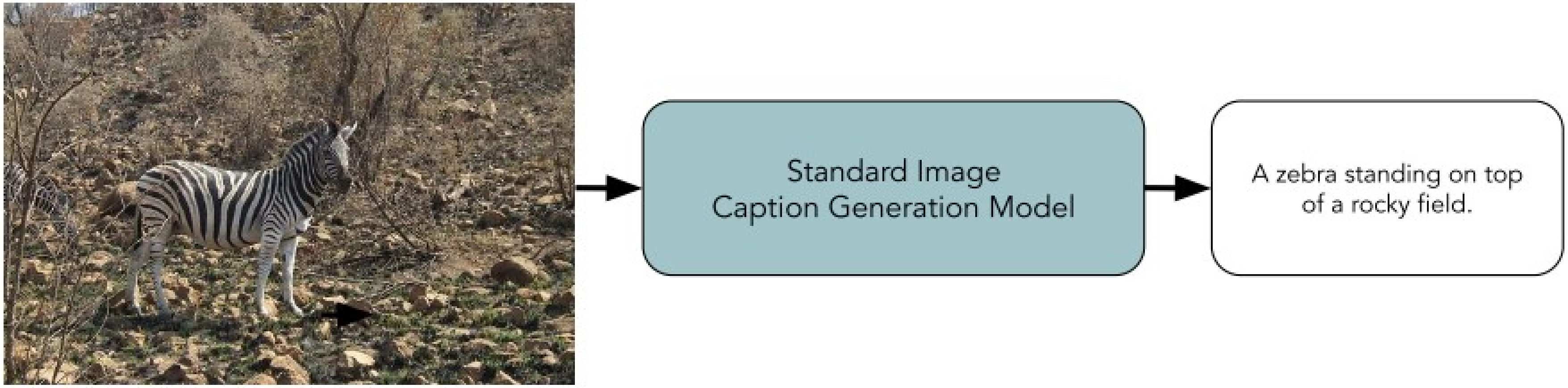


Source: Xie et al. Visual Entailment: A Novel Task for Fine-grained Image Understanding. 2018

Grounding and multimodal tasks

Exemplary generative task:

- Image captioning: describe the contents of an image using text

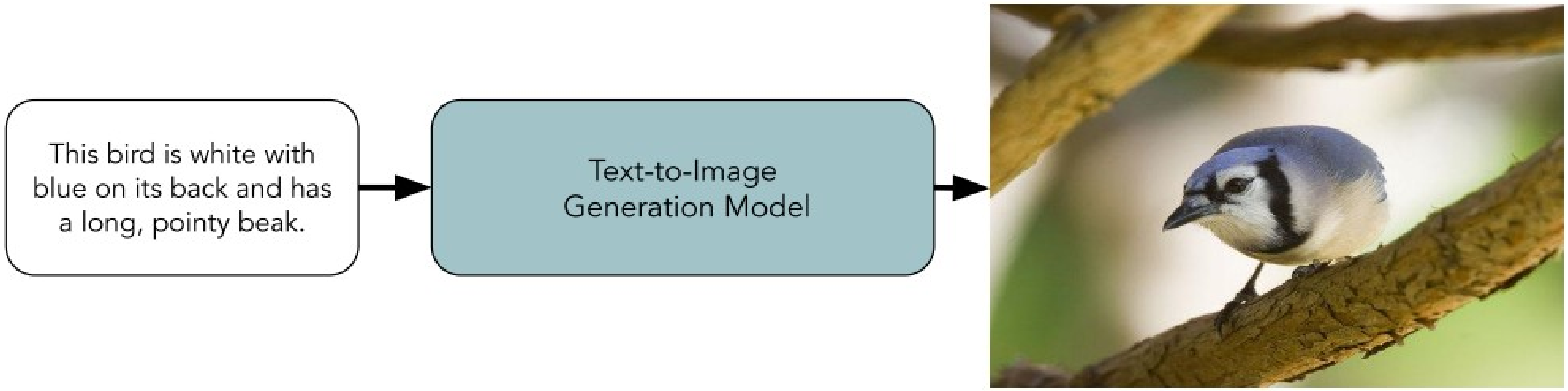


Source: Mogadala et al. Trends in Integration of Vision and Language
Research: A Survey of Tasks, Datasets, and Methods. 2019

Grounding and multimodal tasks

Exemplary generative task:

- Text to image: given a textual description generate the described image.



Source: Mogadala et al. Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods. 2019

Grounding and multimodal tasks

More tasks?

- Text-image retrieval.
- Referring expresions.
- Semantic similarity.
- Text-guided visual navigation.
- Machine translation.
- Videogame playing (reinforcement learning)
- ...

MULTIMODAL SYSTEMS

Multimodal systems

To understand current multimodal systems we need:

- Some image basics.
- Neural network architectures for CV.
- Different image representation approaches.

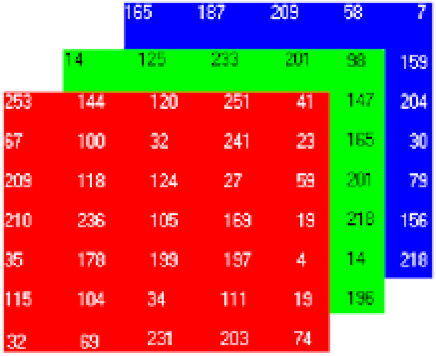
Computer vision basics

What is a (digital) image?

- A 3 dimensional tensor: Height x Width x Channels.
  - Channels usually refer to Red, Green and Blue intensities (RGB).
  - Values are usually between 0 and 255.
  - A pixel = 1x3 vector; values in [0, 255]

Computer vision basics

What we see



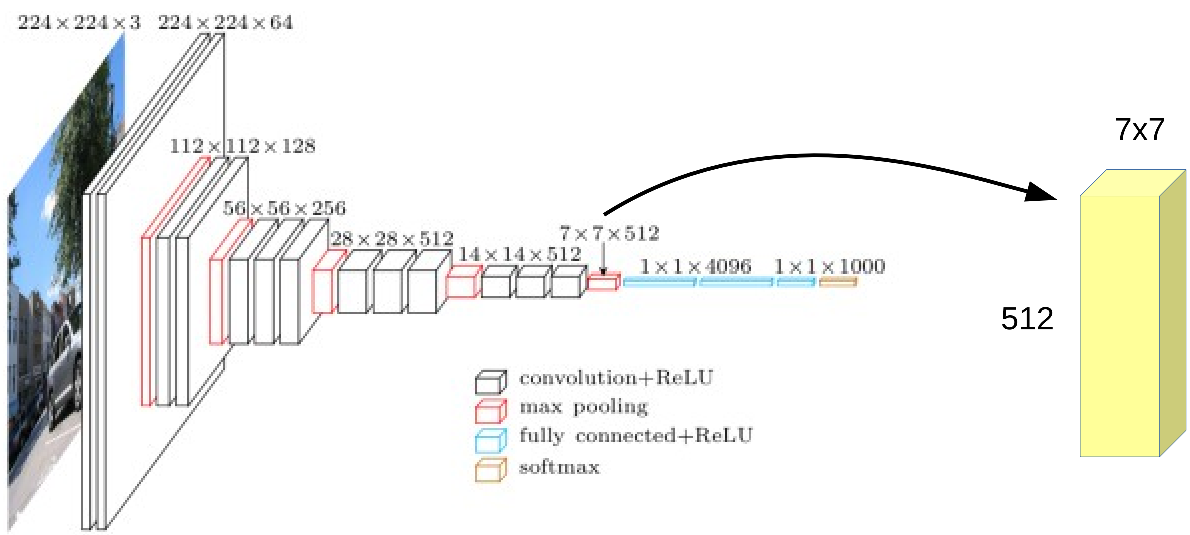
What the computer sees

Computer vision basics

- Neural network architectures for CV:
  - Convolutional Neural Networks.
  - Transformers.
- Visual representation or embeddings
  - Grid-based region embeddings.
  - Object-based region embeddings.

Computer vision basics

Example: VGG16 (Simonyan and Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. 2015)

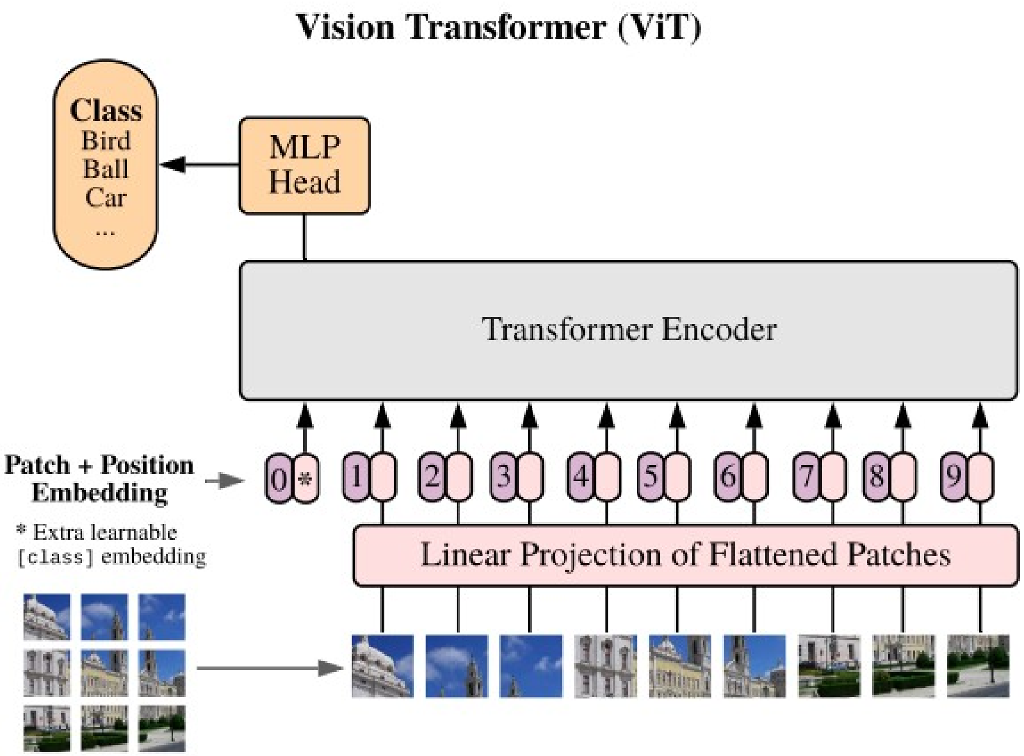


Each cell represents
a region of 32x32
in the original image

Source: https://towardsdatascience.com/step-by-step-vgg16-implementation-in-keras-for-beginners-a833c686ae6c

Computer vision basics

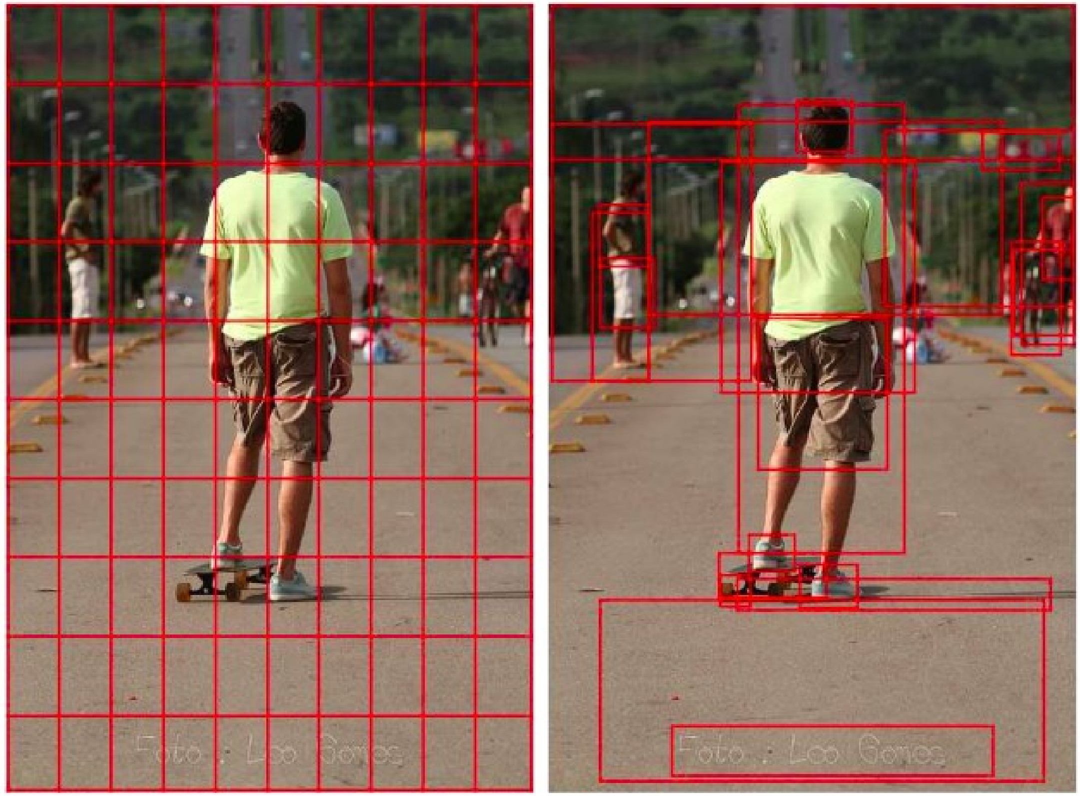
Vision transformers



Source: Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2021

Computer vision basics

Grid-based vs object-based embeddings

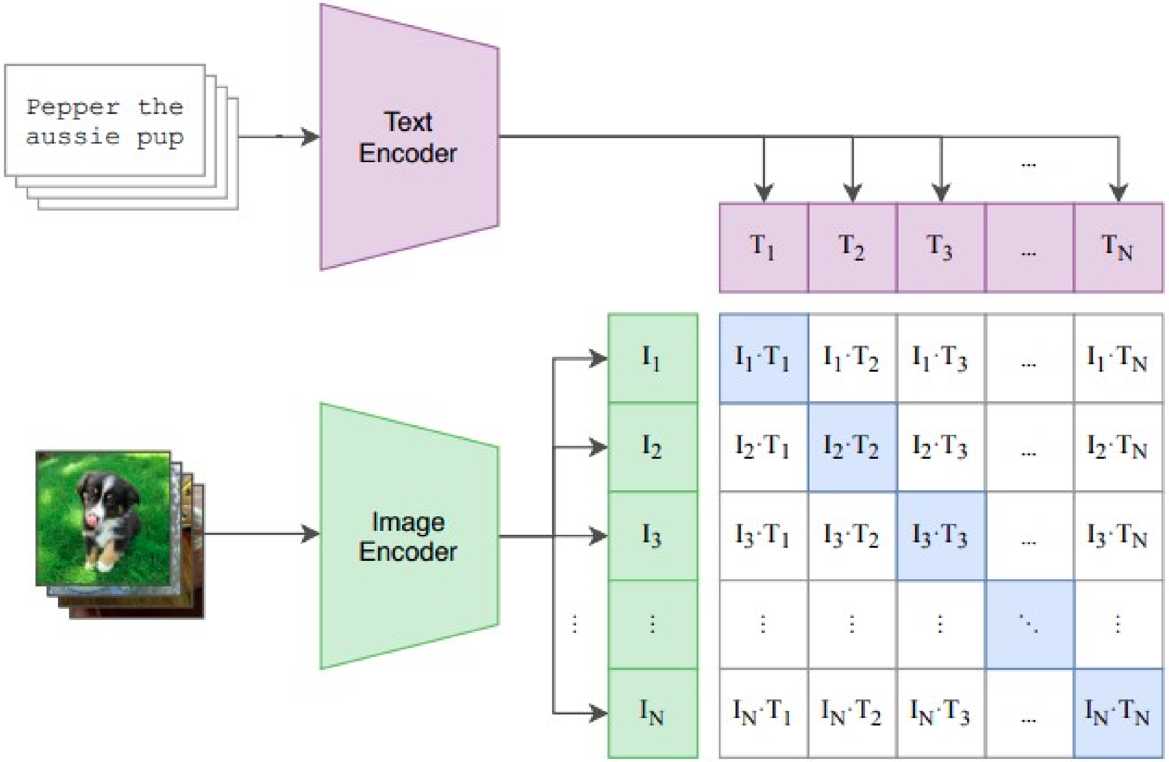


Multimodal systems

Three main trends:

- Contrastive learning for visual and textual representations.
- Multimodal transformers:
  - Encoder-only systems.
  - Encoder-decoder systems.
  - Decoder-only systems.
- Problem-specific solutions.

Contrastive learning



$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Source: Raford et al. Learning Transferable Visual Models From Natural Language Supervision. 2021

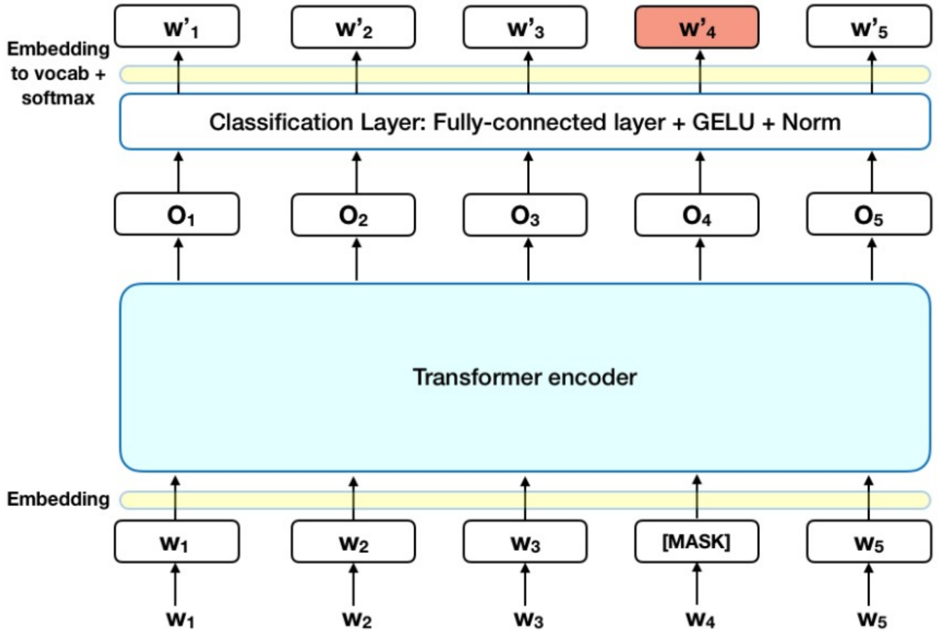
Contrastive learning

- Requires large datasets of aligned image-text pairs.
  - Nowadays about 1.8B noisy image-text pairs (private) or 400M (public).
- Specially well-suited for image-text and text-image retrieval.
- Provides strong models for visual tasks.

Multimodal transformers

Transformer encoder revisited (BERT)

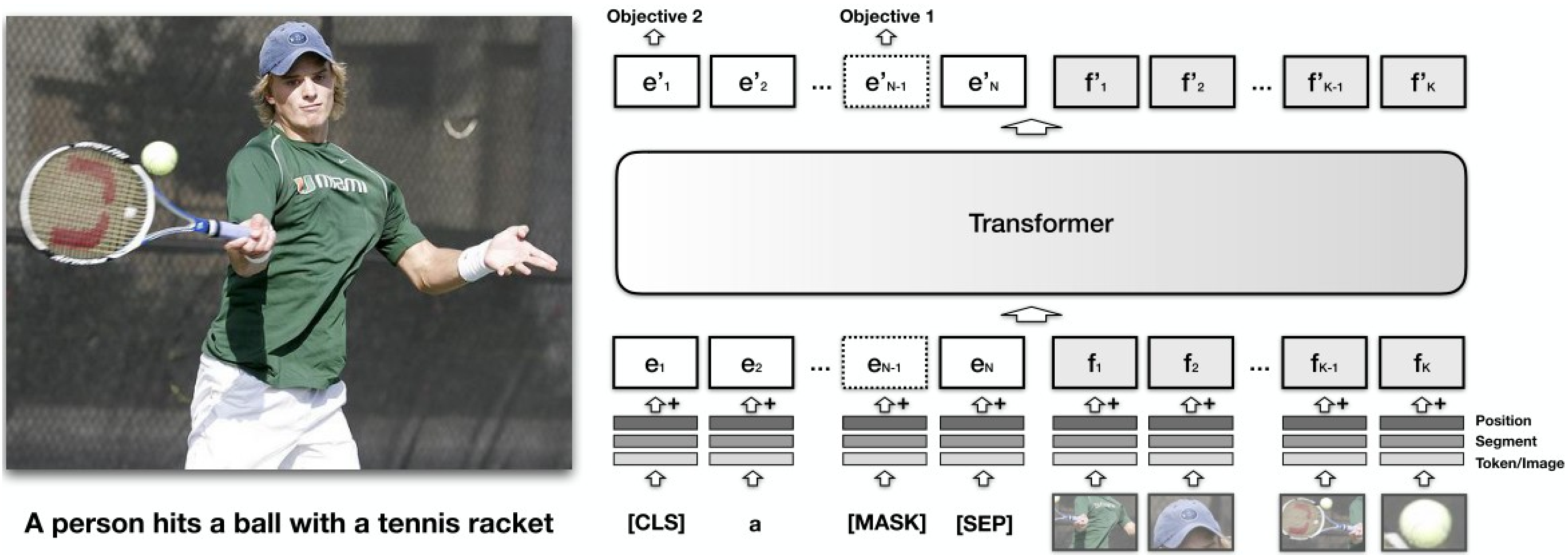
- Transformer encoder + pre-training (MLM)



Source: http://jalammar.github.io/illustrated-bert/

Multimodal systems

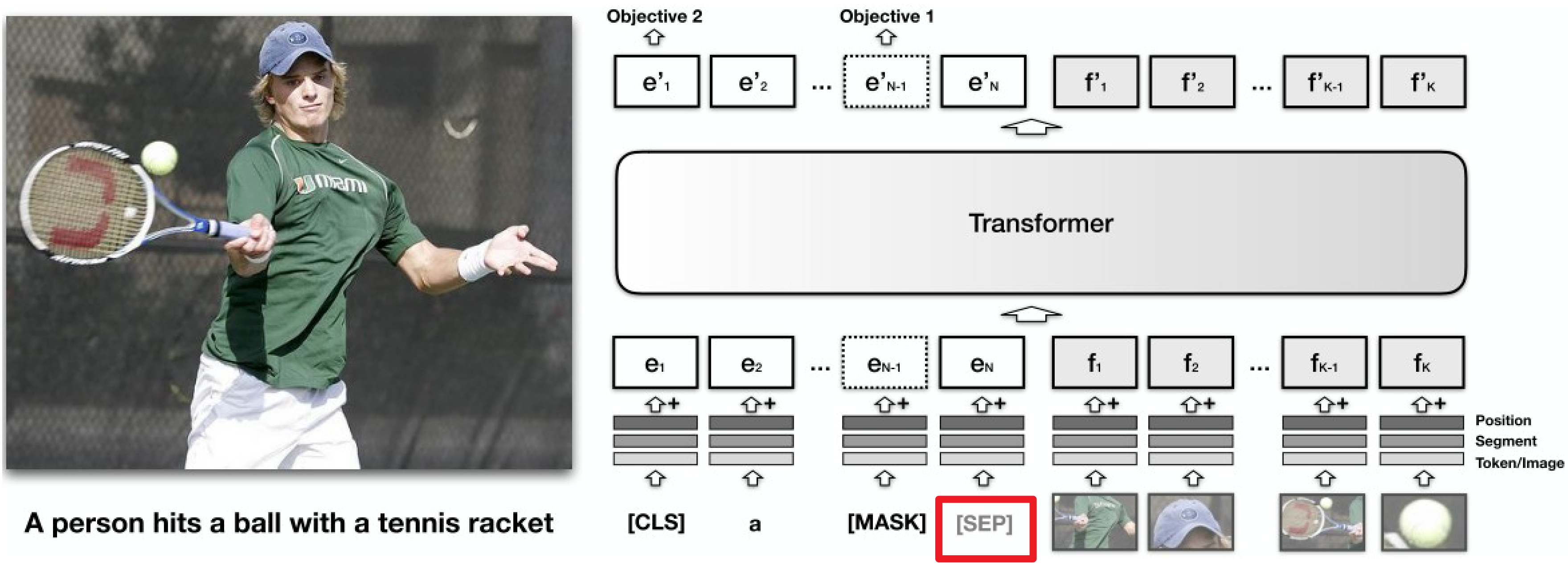
A “simple” multimodal transformer: VisualBERT



Source: Li et al. VisualBERT: A Simple and Performant Baseline for Vision and Language. 2019

Multimodal systems

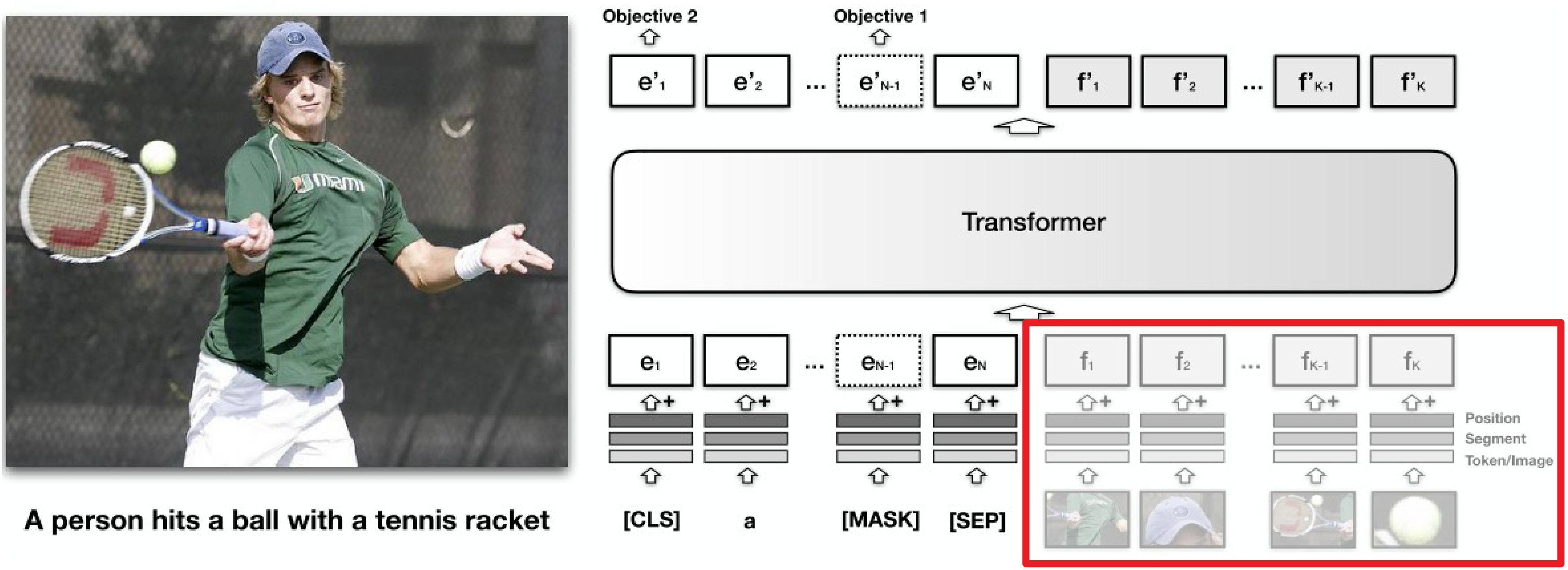
A “simple” multimodal transformer: VisualBERT



Source: Li et al. VisualBERT: A Simple and Performant Baseline for Vision and Language. 2019

Multimodal systems

A “simple” multimodal transformer: VisualBERT



Source: Li et al. VisualBERT: A Simple and Performant Baseline for Vision and Language. 2019

VISUAL EMBEDDINGS

Multimodal transformers

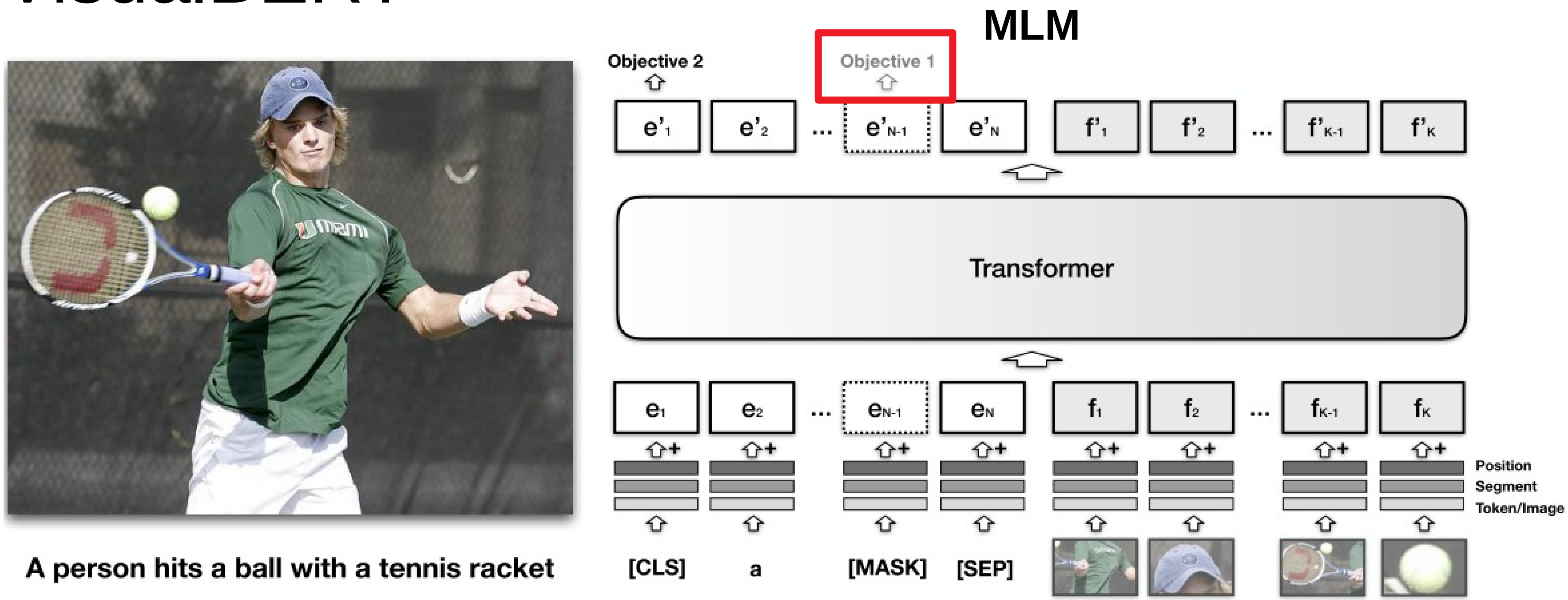
A set of visual embeddings *F*.

- Each f in F corresponds to a region in the image derived from an object detector.
- $f = f_0 + f_s + f_p$
  - f_0: visual feature representation of the img region.
  - f_s: image embedding or text embedding?
  - f_p: position embedding

Source: Mogadala et al. Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods. 2019

Multimodal transformers

A “simple” multimodal transformer: VisualBERT

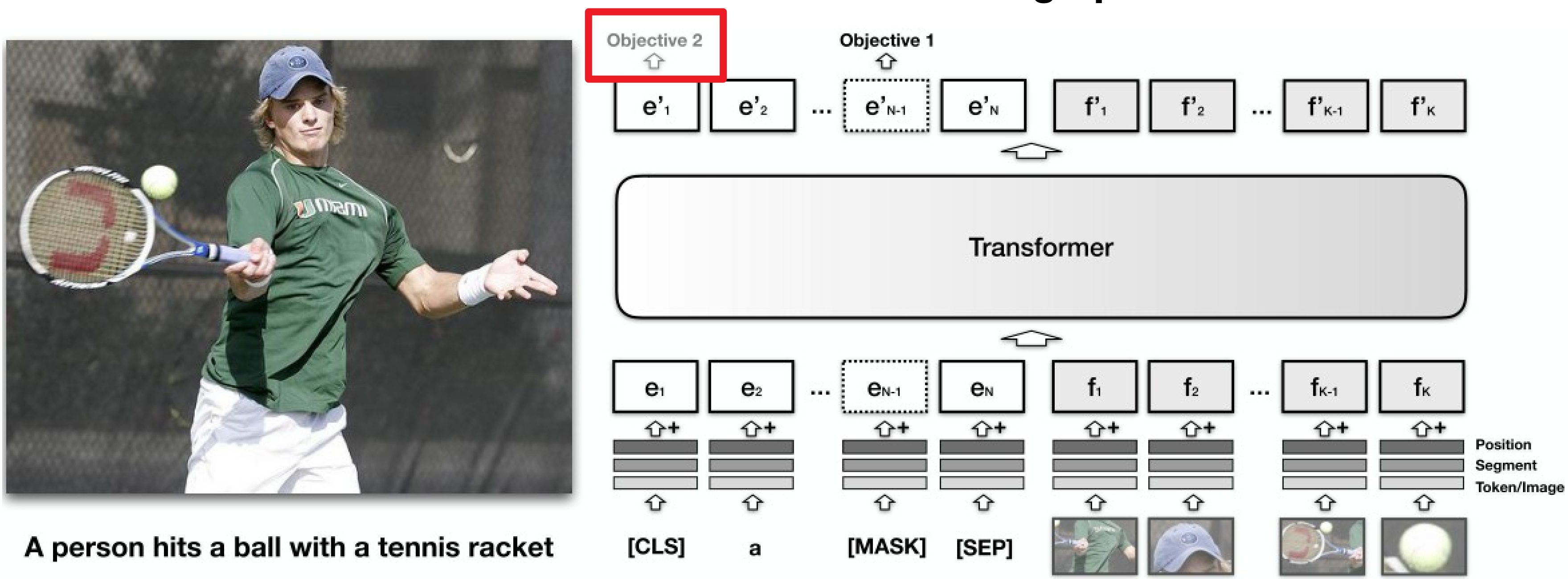


Source: Li et al. VisualBERT: A Simple and Performant Baseline for Vision and Language. 2019

Multimodal transformers

A “simple” multimodal transformer: VisualBERT

Sentence-image prediction



Source: Li et al. VisualBERT: A Simple and Performant Baseline for Vision and Language. 2019

Multimodal transformers

Training VisualBERT

- Task-agnostic pre-training:
  - Use COCO dataset.
  - MLM (Objective 1): mask text tokens but not image tokens.
  - Sentence-image prediction (Objective 2)
- Task-specific pre-training:
  - Use the target dataset with Objective 1 and 2.
- Fine-tuning:
  - Task specific input, output and objective.

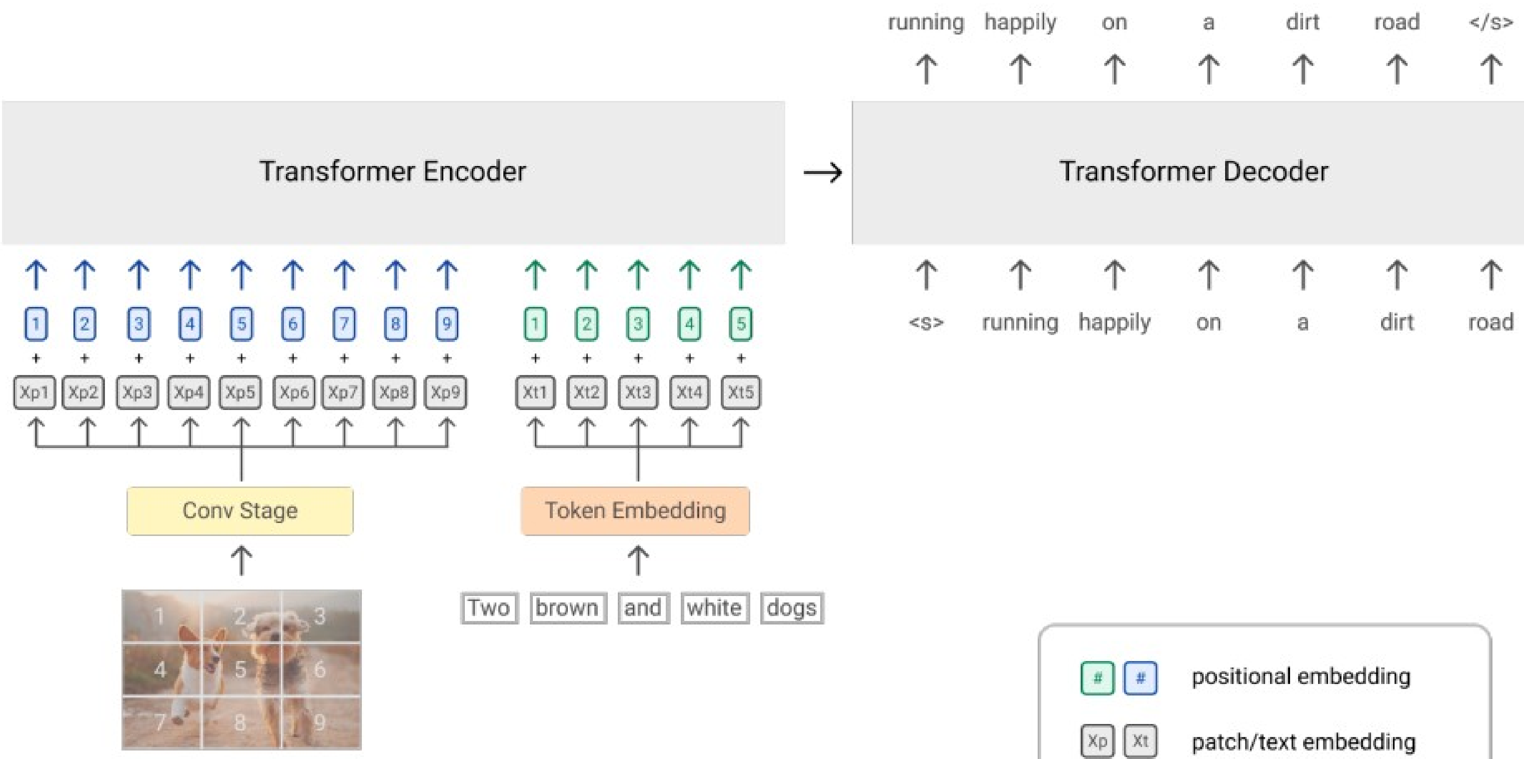
Multimodal transformers

Tons of different multimodal transformers:

- Architectural design: encoder-only, encoder-decoder, decoder-only.
- Pre-training strategies.
- Inputs: grid region embeddings, object region embeddings...
- Target tasks.

Multimodal transformers

SimVL: encoder-decoder example



Source: Wang et al. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. 2021

Multimodal transformers

- SimVL training recipe:
  - LM loss with visual+textual prefix and text generation.
  - Uses grid-based region features (simpler than object-based ones).
  - Pre-trains on a huge dataset of aligned image-text pairs (1.8B pairs).
  - Fine-tune on the specific task.
- Achieves SOTA results in 6 VL benchmarks: VQA, Visual entailment, Image captioning…
  - VisualBERT achieved ~71% in VQA.
  - SimVL achieved ~80% in VQA!
- Zero-shot capabilities.

Multimodal transformers

Zero-shot image captioning

Multimodal transformers

Zero-shot VQA

Multimodal transformers

Zero-shot visual text completion

Multimodal transformers

DALL·E: a decoder-only multimodal transformer for text2image

- Two-stage training process:
  - Stage 1: learn the visual codebook (dVAE)
  - Stage 2: learn the transformer (12B parameters)
    - 256 textual tokens max with vocab size 16,384.
    - 1024 visual tokens with vocab size 8,192.
    - 3 self-attention masks: text-to-text, image-to-text, image-to-image.
    - Pre-training task: next visual token prediction.

Source: Ramesh et al. Zero-shot text-to-image generation. 2021

Multimodal transformers

DALL·E: a decoder-only multimodal transformer for text2image

- Sample generation process:
  - Get N samples from the transformer.
  - Rerank using CLIP (pretrained contrastive model).
- Dataset:
  - 250 million text-images pairs from the internet.

Source: Ramesh et al. Zero-shot text-to-image generation. 2021

Multimodal transformers

DALL·E:

examples

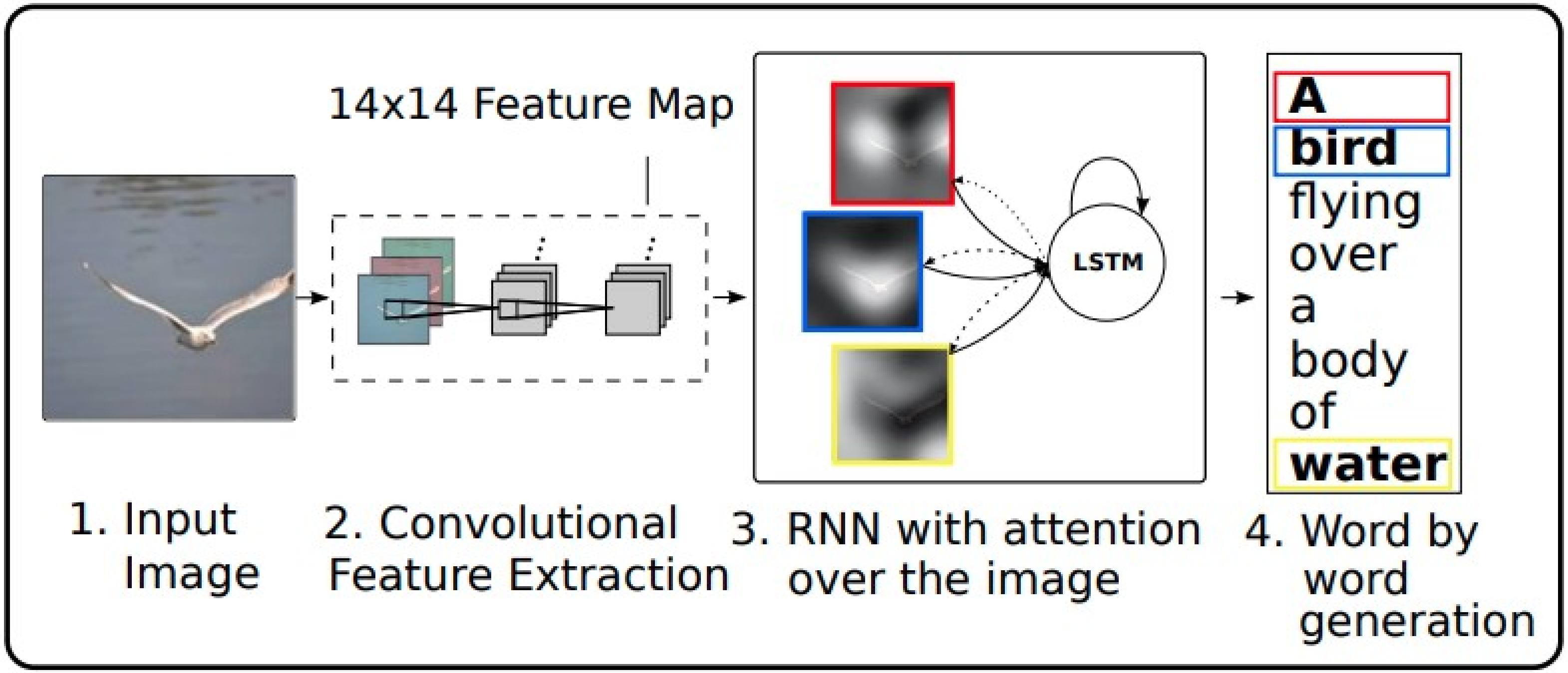
Multimodal transformers

DALL·E:

examples

Specific solutions

Image captioning (NOTE: far from SOTA)



Source: Xu et al. Show, attend and tell: neural image caption generation with visual attention. 2015

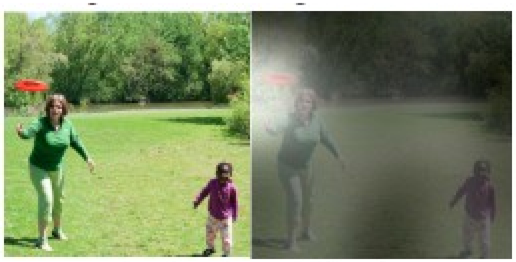
Specific solutions

Image captioning

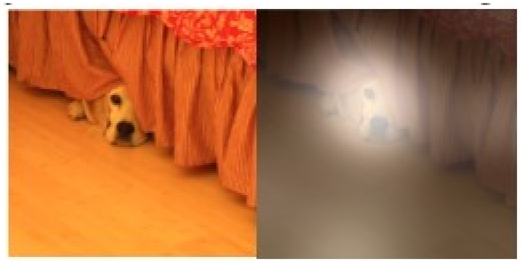
- Encoder-decoder architecture with attention.
  - The encoder is a CNN.
  - The decoder is an LSTM (language model).
  - Soft-attention on the input image is used to generate each word.
- End-to-end training.
  - Input: image.
  - Output: caption (text).

Specific solutions

Image captioning: Some examples

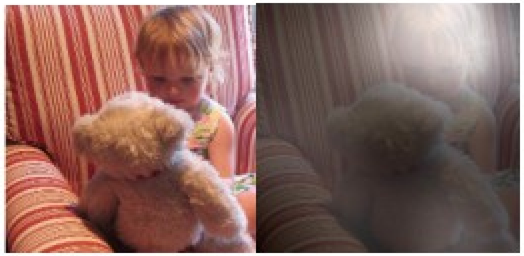


A woman is throwing a frisbee in a park.



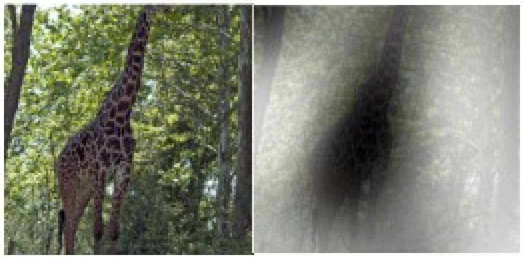
A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.



A giraffe standing in a forest with trees in the background.

Source: Xu et al. Show, attend and tell: neural image caption generation with visual attention. 2015

MULTIMODAL SYSTEMS FOR NLU

Multimodal systems for NLU

- Do those multimodal systems really help for NLU?
  - Caution! Remember evaluation problems for NLU.
- There are not many studies in this line.
  - The community is more focused on multimodal tasks.
  - We will review a couple of interesting papers.

Multimodal systems for NLU

MM transformers in NLU benchmarks

Model	Init. with BERT?	Diff. to BERT Weight	SST-2	QNLI	QQP	MNLI
ViLBERT (Lu et al., 2019)	Yes	0.0e-3	90.3	89.6	88.4	82.4
VL-BERT (Su et al., 2020)	Yes	6.4e-3	90.1	89.5	88.6	82.9
VisualBERT (Li et al., 2019)	Yes	6.5e-3	90.3	88.9	88.4	82.4
Oscar (Li et al., 2020a)	Yes	41.6e-3	87.3	50.5	86.6	77.3
LXMERT (Tan and Bansal, 2019)	No	42.0e-3	82.4	50.5	79.8	31.8
$\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019)	-	0.0e-3	90.3	89.6	88.4	82.4
$\text{BERT}_{\text{BASE}}$ + Weight Noise	-	6.5e-3	89.9	89.9	88.4	82.3

Source: Tan and Bansal. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. 2020

Multimodal systems for NLU

MM transformers in NLU benchmarks

Model	Init. with BERT?	Diff. to BERT Weight	SST-2	QNLI	QQP	MNLI
ViLBERT (Lu et al., 2019)	Yes	0.0e-3	90.3	89.6	88.4	82.4
VL-BERT (Su et al., 2020)	Yes	6.4e-3	90.1	89.5	88.6	82.9
VisualBERT (Li et al., 2019)	Yes	6.5e-3	90.3	88.9	88.4	82.4
Oscar (Li et al., 2020a)	Yes	41.6e-3	87.3	50.5	86.6	77.3
LXMERT (Tan and Bansal, 2019)	No	42.0e-3	82.4	50.5	79.8	31.8
$\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019)	-	0.0e-3	90.3	89.6	88.4	82.4
$\text{BERT}_{\text{BASE}}$ + Weight Noise	-	6.5e-3	89.9	89.9	88.4	82.3

No gains over BERT!

Multimodal systems for NLU

Some reasons for those results:

- Large discrepancy between visually-grounded language and other types of natural language.
  - 120M tokens in VL datasets VS 220B in C4 corpus.
  - Short and instructive descriptions in VL datasets.
- Most of the words in natural language are not visually grounded.
  - The ratio of grounded tokens is only about 28% in English Wikipedia (approximate estimation).

Multimodal systems for NLU

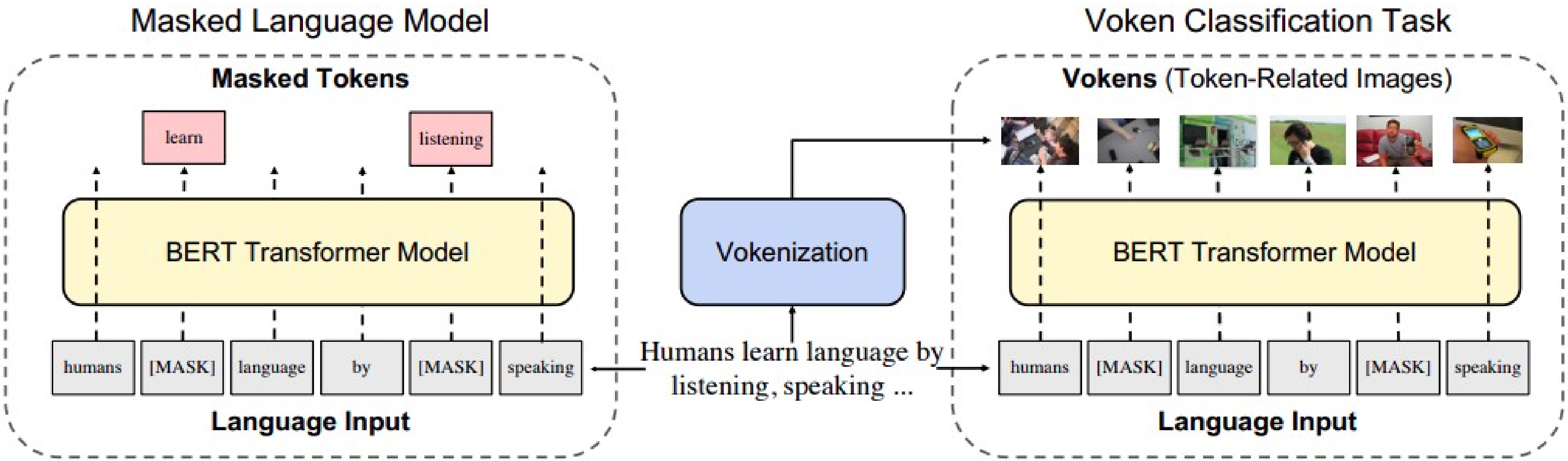
Statistics of image-captioning dataset and other natural language corpora

Dataset	# of Tokens	# of Sents	Vocab. Size	Tokens #/ Sent.	1-Gram JSD	2-Gram JSD	Grounding Ratio
MS COCO	7.0M	0.6M	9K	11.8	0.15	0.27	54.8%
VG	29.2M	5.3M	13K	5.5	0.16	0.28	57.6%
CC	29.9M	2.8M	17K	10.7	0.09	0.20	41.7%
Wiki103	111M	4.2M	29K	26.5	0.01	0.05	26.6%
Eng Wiki	2889M	120M	29K	24.1	0.00	0.00	27.7%
CNN/DM	294M	10.9M	28K	26.9	0.04	0.10	28.3%

Source: Tan and Bansal. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. 2020

Multimodal systems for NLU

Vokenization approach:



- Vokens are generated from VL datasets, using a contrastive model.
- Vokens generated for purely textual corpora.

Multimodal systems for NLU

Vokenization results:

Method	SST-2	QNLI	QQP	MNLI	SQuAD v1.1	SQuAD v2.0	SWAG	Avg.
$\text{BERT}_{6L/512H}$	88.0	85.2	87.1	77.9	71.3/80.2	57.2/60.8	56.2	75.6
$\text{BERT}_{6L/512H}$ + Voken-cls	89.7	85.0	87.3	78.6	71.5/80.2	61.3/64.6	58.2	76.8
$\text{BERT}_{12L/768H}$	89.3	87.9	83.2	79.4	77.0/85.3	67.7/71.1	65.7	79.4
$\text{BERT}_{12L/768H}$ + Voken-cls	**92.2**	**88.6**	**88.6**	**82.6**	**78.8/86.7**	68.1/71.2	**70.6**	**82.1**
$\text{RoBERTa}_{6L/512H}$	87.8	82.4	85.2	73.1	50.9/61.9	49.6/52.7	55.1	70.2
$\text{RoBERTa}_{6L/512H}$ + Voken-cls	87.8	85.1	85.3	76.5	55.0/66.4	50.9/54.1	60.0	72.6
$\text{RoBERTa}_{12L/768H}$	89.2	87.5	86.2	79.0	70.2/79.9	59.2/63.1	65.2	77.6
$\text{RoBERTa}_{12L/768H}$ + Voken-cls	**90.5**	**89.2**	**87.8**	**81.0**	**73.0/82.5**	**65.9/69.3**	**70.4**	**80.6**

Multimodal systems for NLU

Vokenization limitations:

- Approximation error of using finite image labels.
- Lack of vocabulary diversity of a small image-text dataset.

Multimodal systems for NLU

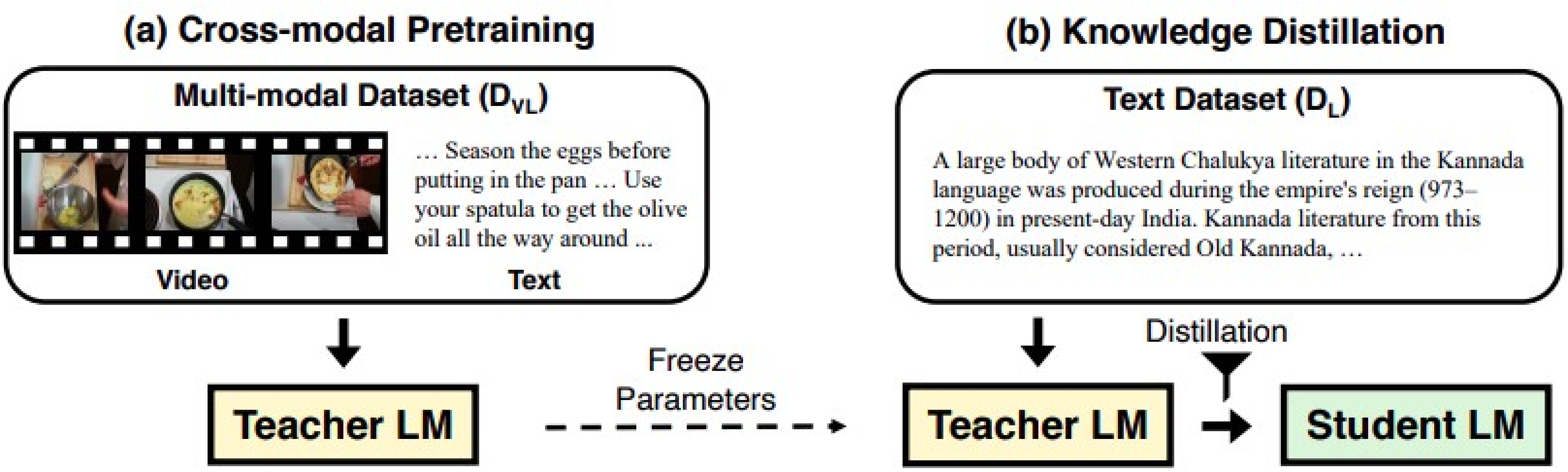
VidLanKD: a video-language knowledge distillation method for improving NLU.

- Train a multi-modal teacher model on a video-text dataset.
- Transfer its knowledge to a student language model with a text dataset.
- Results on GLUE, SquAD, SWAG, GLUE-diagnostics, PIQA and Tracie.

Source: Tang et al. VIDLANKD: Improving Language Understanding via Video-Distilled Knowledge Transfer. 2021

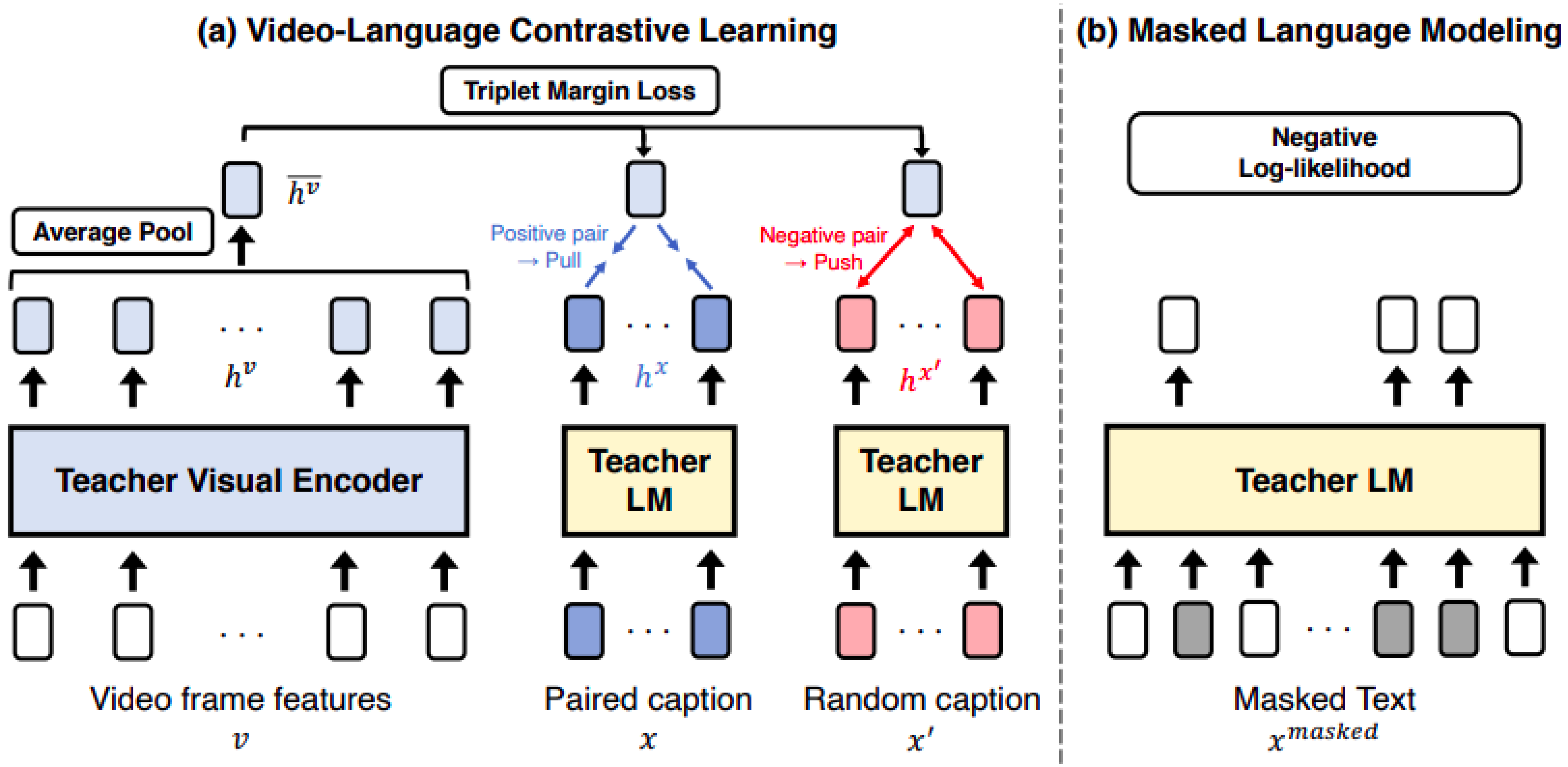
Multimodal systems for NLU

VidLanKD overview:



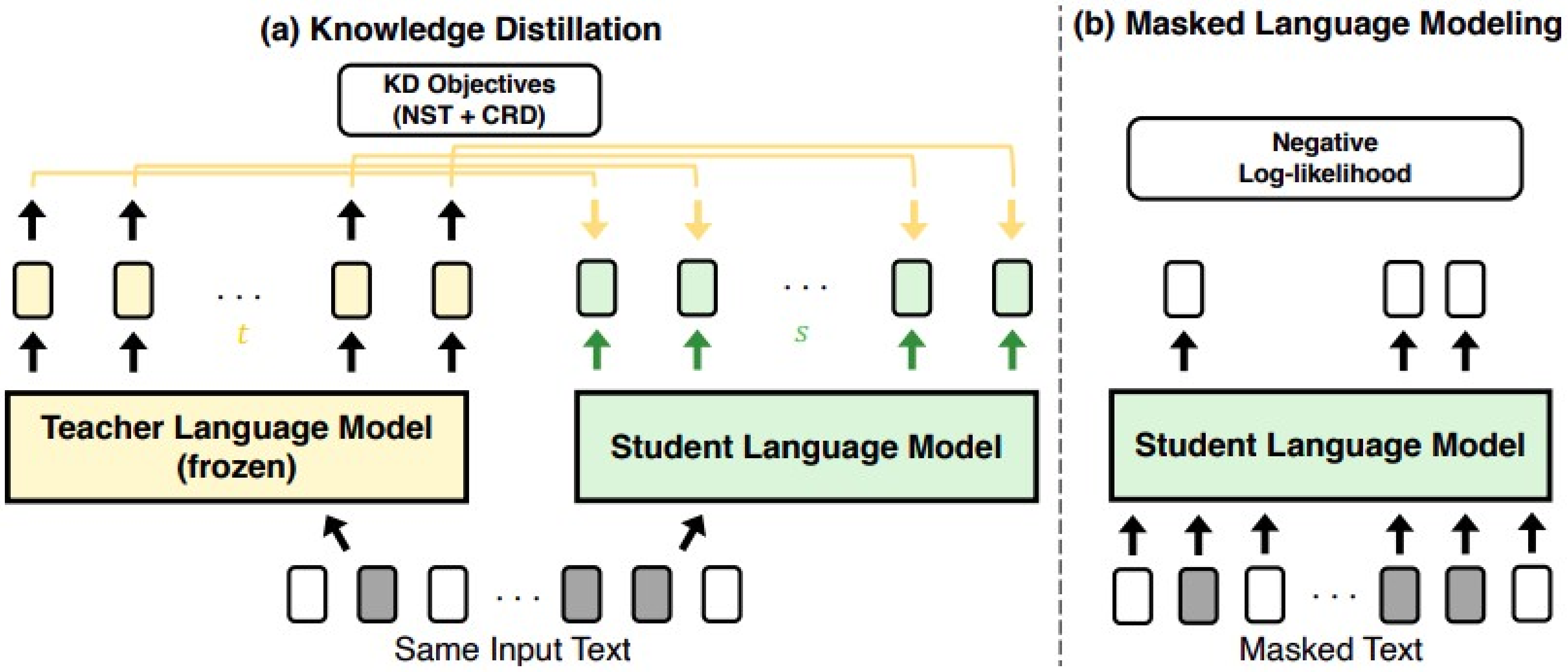
Multimodal systems for NLU

VidLanKD cross-modal training:



Multimodal systems for NLU

VidLanKD teacher-student distillation:



Multimodal systems for NLU

VidLanKD datasets:

- Video-text: HowTo100M.
  - 136M video clips.
  - 138M captions.
  - 568M tokens.
- Text pretraining: English Wikipedia.
  - 2.9B tokens.
  - 120M sentences.

Multimodal systems for NLU

VidLanKD results:

	SST-2 Acc	QNLI Acc	QQP Acc	MNLI Acc	SQuAD v1.1 EM†	SQuAD v2.0 EM	SWAG Acc	Avg.
BERT$_{12L/768H}$ [68]	89.3	87.9	83.2	79.4	77.0	67.7	65.7	78.6
+ KD (Img-Voken) [68]	92.2	88.6	88.6	82.6	78.8	68.1	70.6	81.4
BERT$_{12L/768H}$	89.0	88.0	86.2	79.2	77.2	68.0	65.0	78.9
+ KD (Vid-Voken) w/ ResNet	93.4	89.2	88.7	83.0	78.9	68.7	70.0	81.7
+ KD (Vid-Voken) w/ CLIP	94.1	**89.8**	89.0	83.9	79.2	68.6	71.6	82.3
+ KD (NST+CRD) w/ ResNet	94.2	89.3	89.7	84.0	79.0	**68.9**	71.8	82.4
+ KD (NST+CRD) w/ CLIP	**94.5**	89.6	**89.8**	**84.2**	**79.6**	68.7	**72.0**	**82.6**

	GLUE diagnostics				PIQA	TRACIE
	Lexicon	Predicate	Logic	Knowledge		
BERT$_{6L/512H}$	53.0	64.2	44.5	44.0	56.9	63.4
+ KD-NST	53.3 (+0.3)	63.7 (-0.5)	44.8 (+0.3)	48.6 (**+4.6**)	60.0 (**+3.1**)	66.7 (**+3.3**)

CONCLUSIONS

Conclusions

- Multimodal systems show some promise, but still lag behind text-only systems.
  - Scale seems to be the main driver (again).
  - Multimodally grounded large and diverse resources are needed.
- Evaluation is a major concern for NLU.
- Grounding in different modalities seems important.
  - Perceptual data: audio, tactile, smell…
  - Knowledge graphs.
  - Interaction data.

Conclusions

Closing words from Sam Altman (CEO OpenAI)

The text-encoding part of DALL·E probably can't beat pure text models yet. But I would be very surprised if multimodal models do not start outperforming pure text models in the next few years.

THANKS!