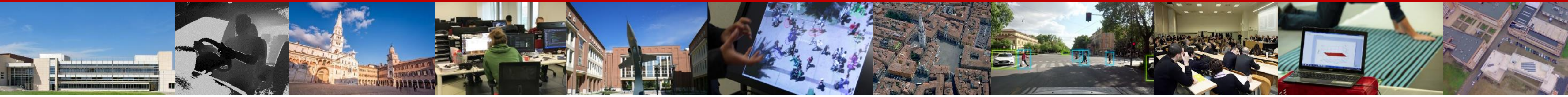
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImageLab

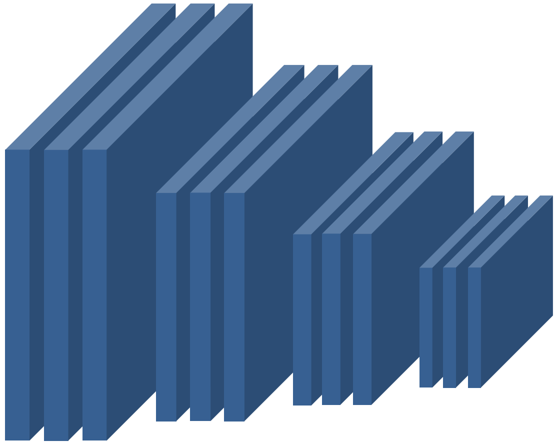
From Images to Text: New forms of Human-AI Interaction



Lorenzo Baraldi

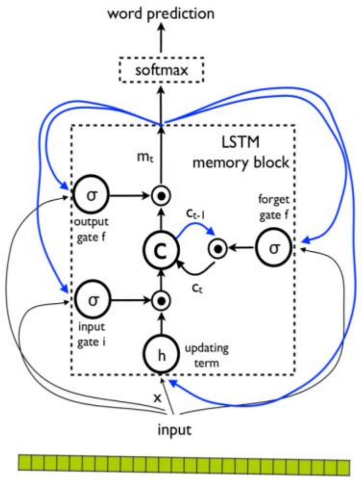
VISMAC

Describing images in Natural Language



Visual feature extractor

+



Language model

..a white shark swims in the ocean water..

Goal: describe a visual input in natural language.

Base technical idea: Combine visual feature extractors with language models

1. Karpathy, A., & Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *CVPR 2015*.
2. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. Show and tell: A neural image caption generator. In *CVPR 2015*.
3. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR 2015*.

The Image Captioning Journey

Image

TRAINING STRATEGIES

1. **Cross Entropy Loss**
2. **Masked Language Model**
3. **Reinforcement Learning**
4. **VL Pre-Training**

VISUAL ENCODING

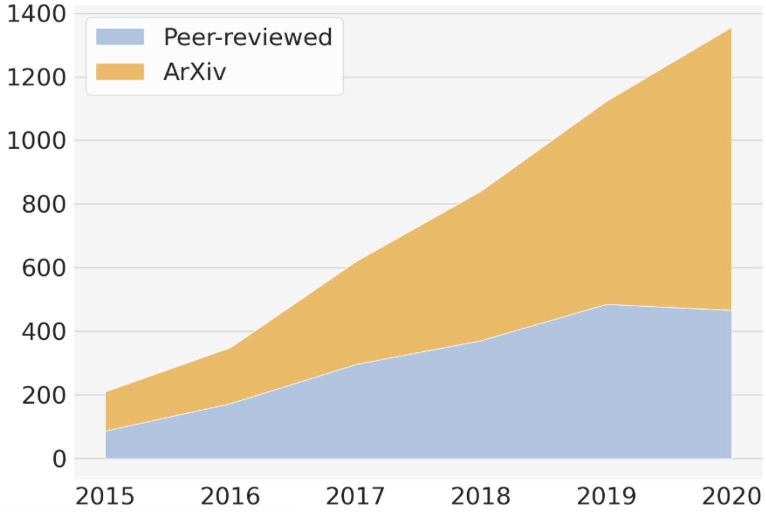
1. **Non-Attentive**
   (Global CNN Features)
2. **Additive Attention**:
   - **Grid-based**
   - **Region-based**
3. **Graph-based Attention**
4. **Self-Attention:**
   - **Region-based**
   - **Patch-based**
   - Image-Text Early Fusion

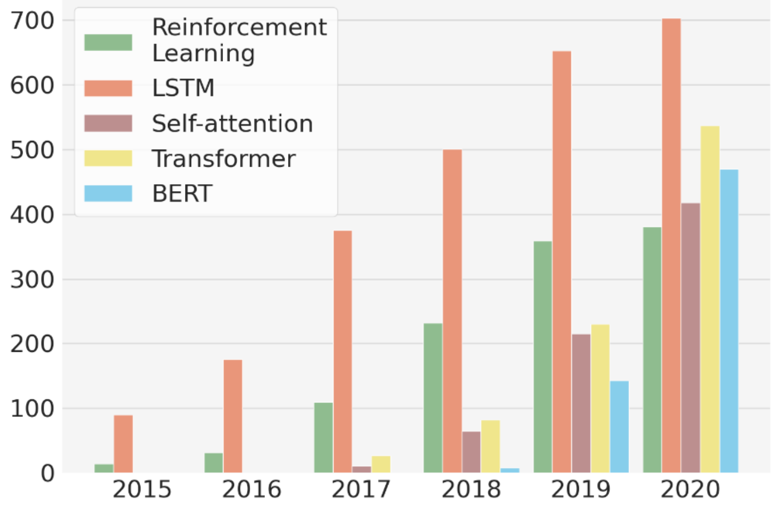
LANGUAGE MODELS

1. **LSTM-based:**
   - **Single-layer**
   - **Two-layer**
2. **CNN-based**
3. **Transformer-based**
4. **Image-Text Early Fusion**
   (BERT-like)

A herd of zebras grazing with a rainbow behind.

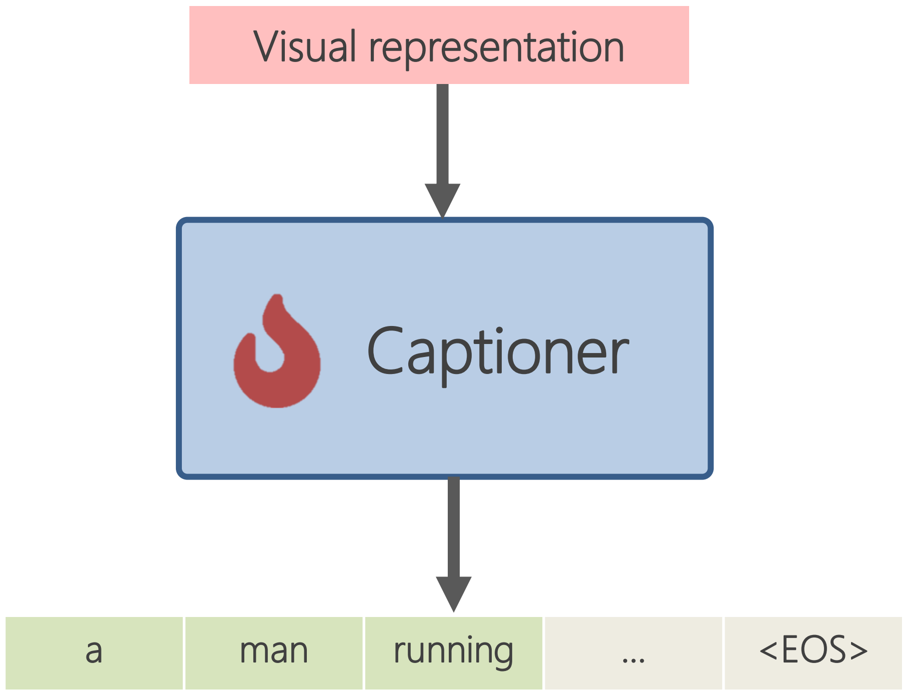
"Image Captioning" and related keywords in the text of recent papers:





Anatomy of a Captioning System

Base technical idea:
combine visual feature extractors with language models.

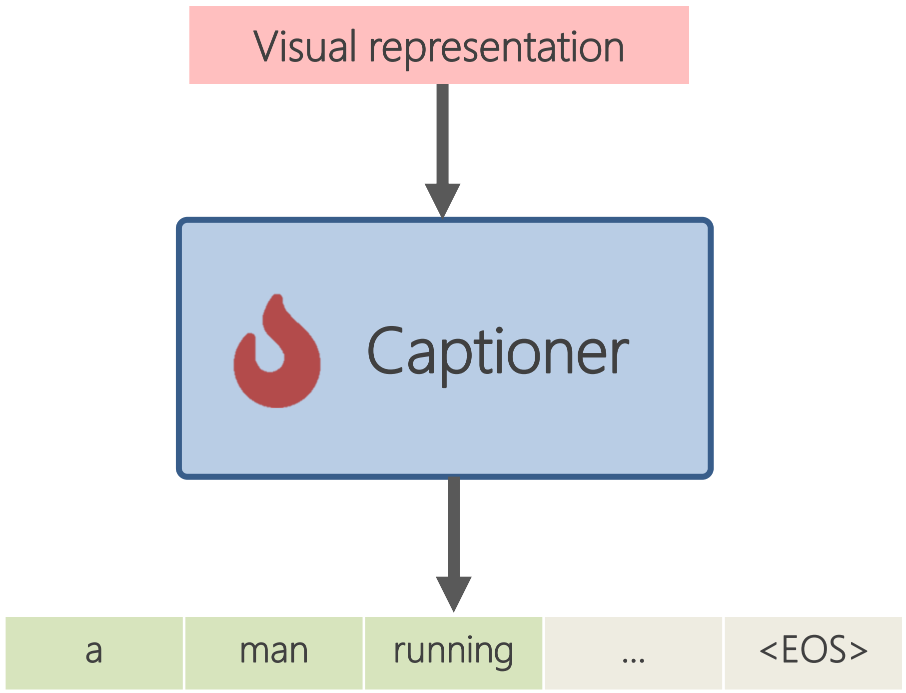


Anatomy of a Captioning System

Base technical idea: combine visual feature extractors with language models.

Many possibilities

- **Language model**:
  - RNN and variants (LSTM)
  - 1-d CNN
  - Transformer-based (recently)

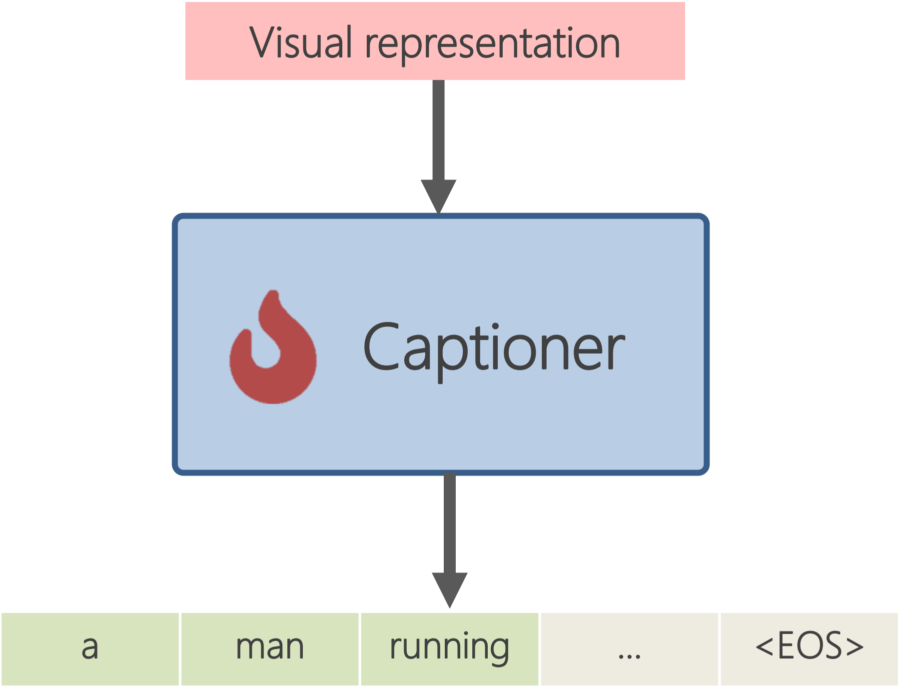


Anatomy of a Captioning System

Base technical idea: combine visual feature extractors with language models.

Many possibilities

- **Language model**:
  - RNN and variants (LSTM)
  - 1-d CNN
  - Transformer-based (recently)
- Conditioning on the **visual input:**
  - Single feature (e.g. pooling)
  - Sequence of features (e.g. video captioning)
  - Set of features (models based on attention)

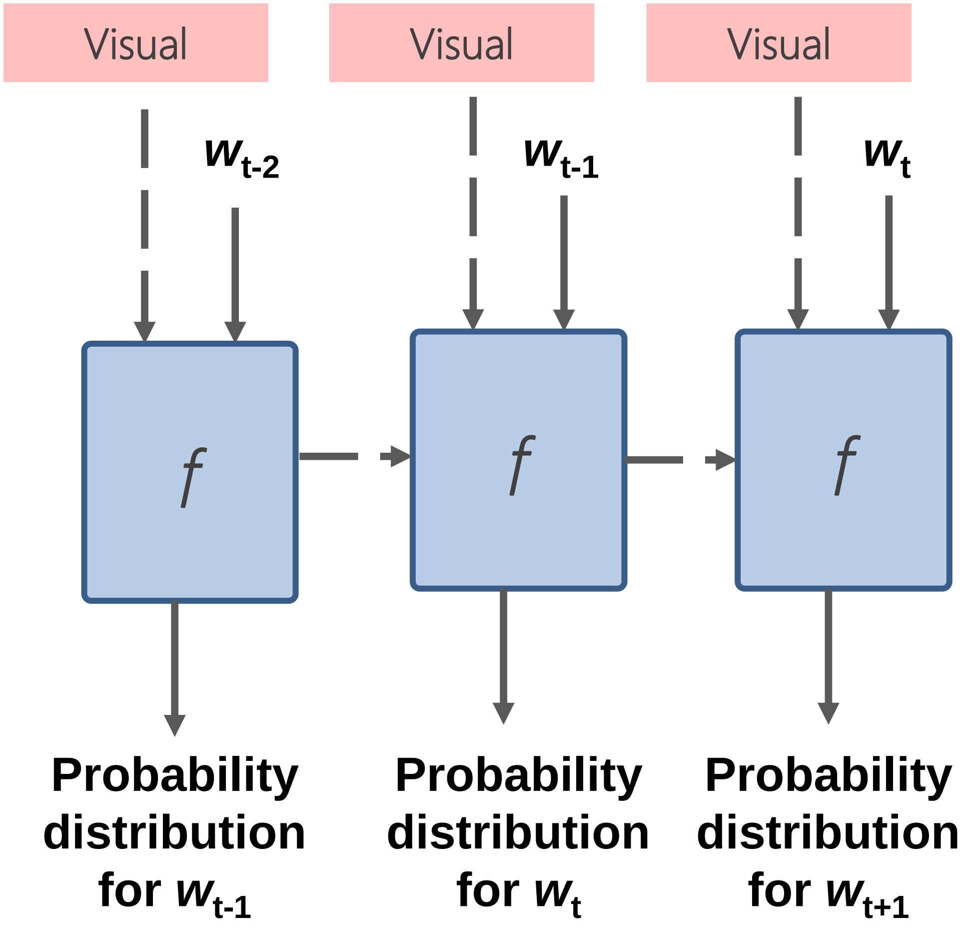


Anatomy of a Captioning System

Language model

- Prediction process is always **sequential**, i.e. we model the probability of outputting a word given previous words in the sentence.

- The probability distribution for w_t is conditioned on w_{t-1} , w_{t-2} , … w_0



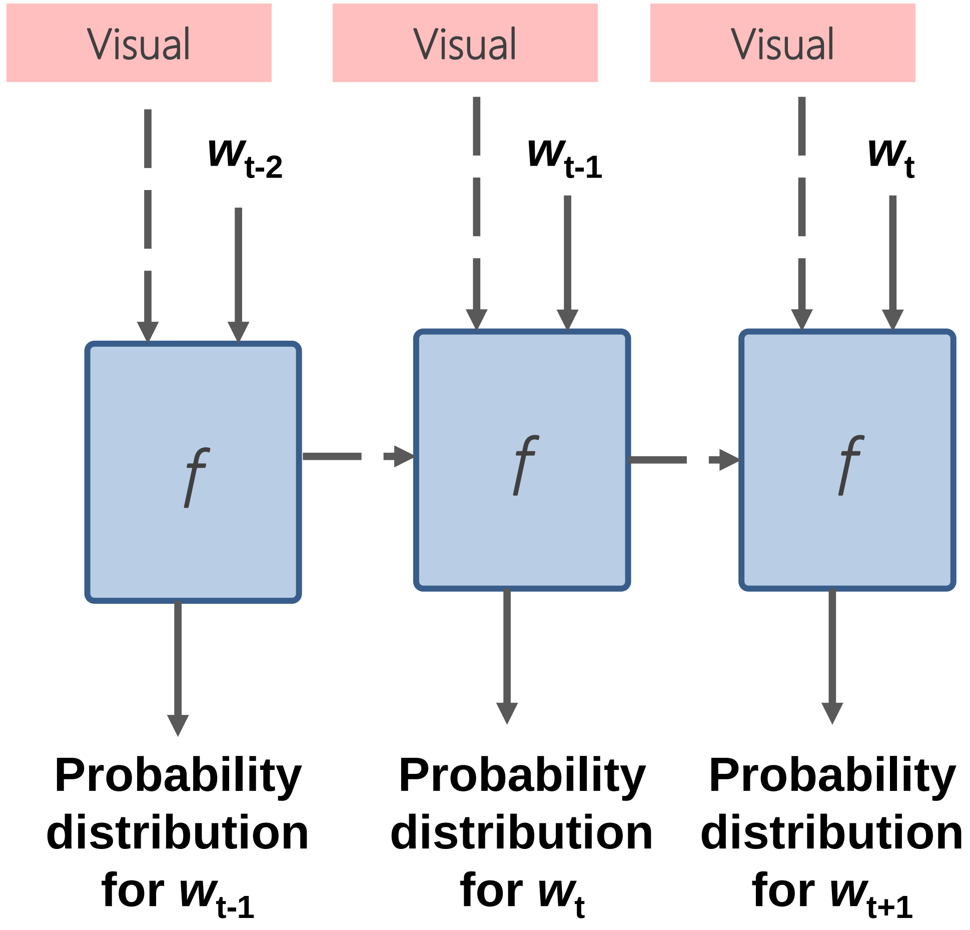
Anatomy of a Captioning System

Language model

- Prediction process is always **sequential**, i.e. we model the probability of outputting a word given previous words in the sentence.

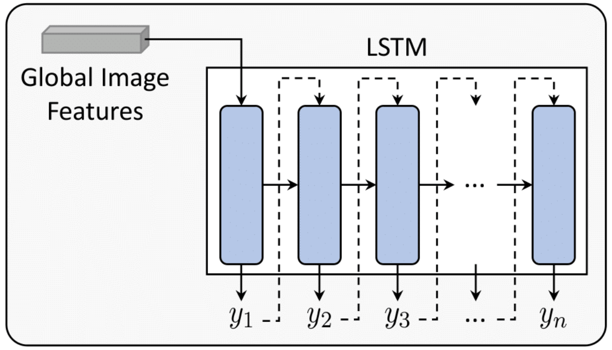
- The probability distribution for w_t is conditioned on w_{t-1} , w_{t-2} , … w_0

- A function *f* models the computational graph for predicting the word at each step (the "**step function**").
    - *Any of {RNN, CNN, Transformer, …}*

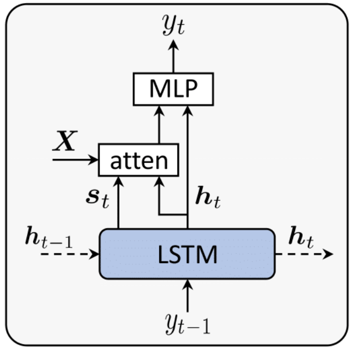


Language Models for Image Captioning

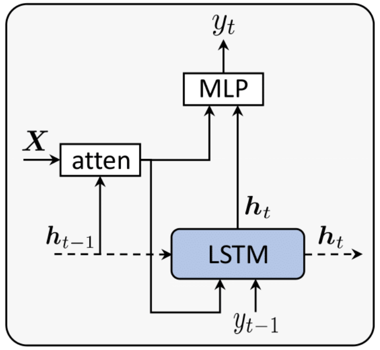
Single-Layer LSTM



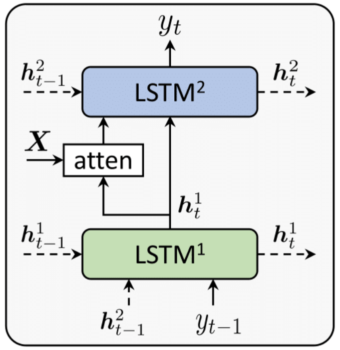
Single-Layer LSTM with Visual Sentinel



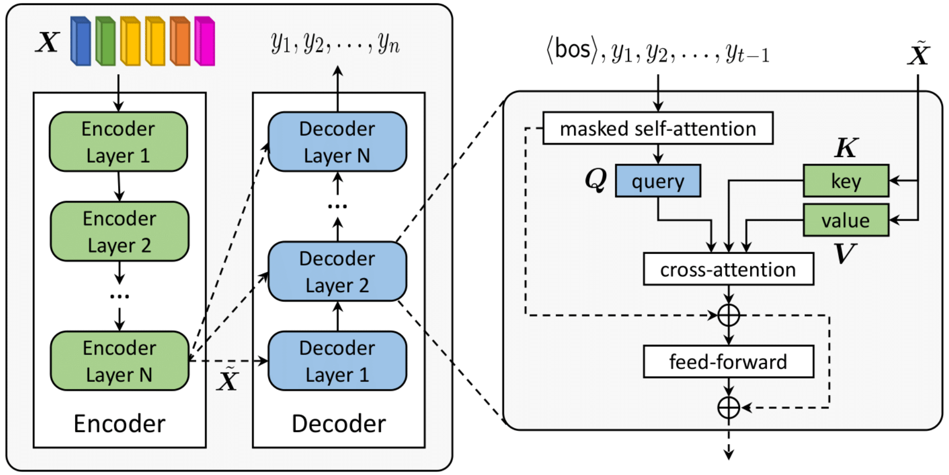
Single-Layer LSTM with Attention



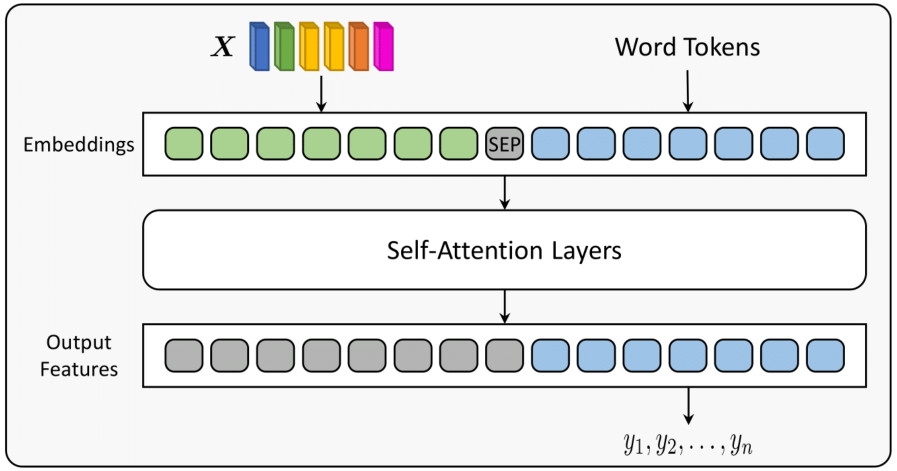
Two-Layer LSTM with Attention



Transformer



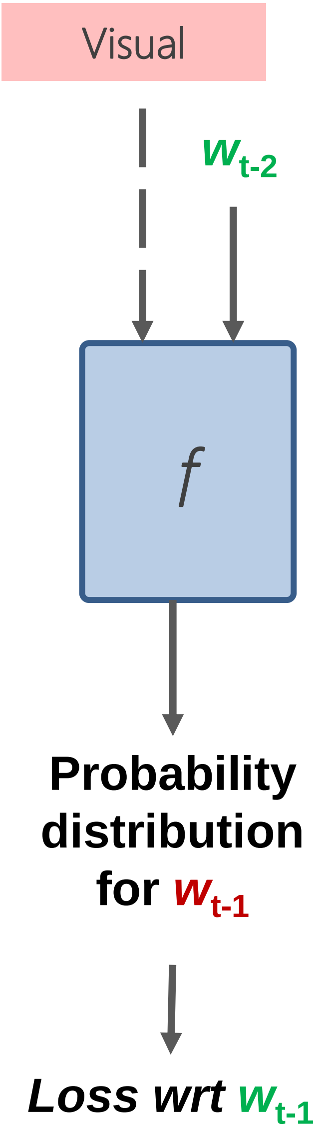
BERT-like



Anatomy of a Captioning System

At training time

- *Train the model to predict a word given the previous ground-truth words.*

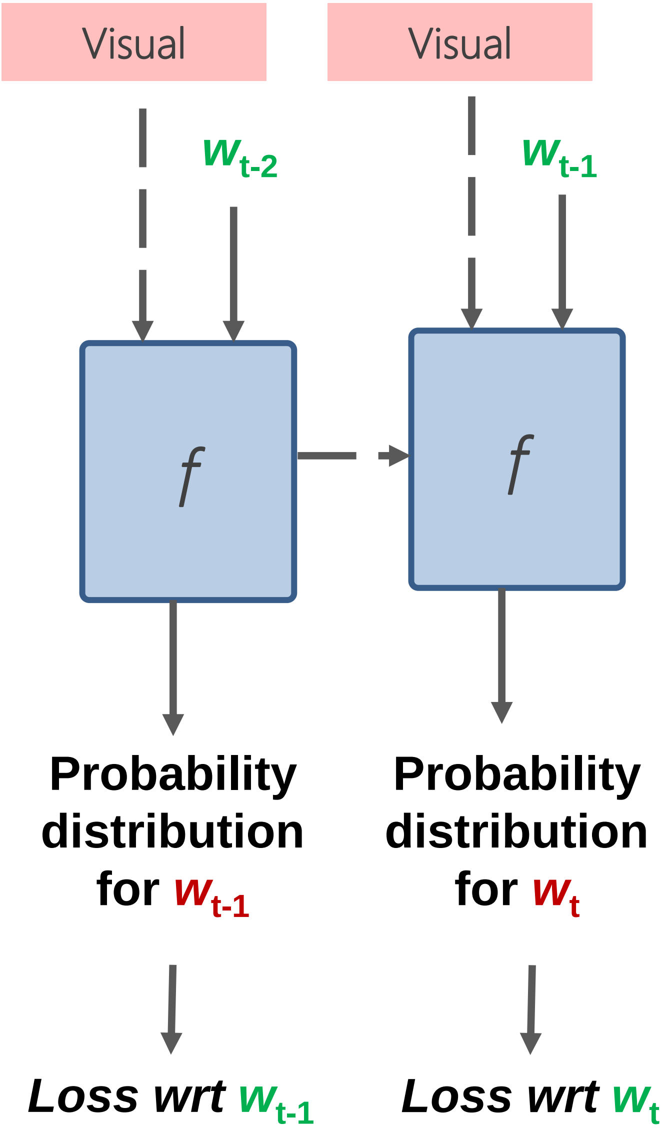


w_t :ground-truth words

Anatomy of a Captioning System

At training time

- *Train the model to predict a word given the previous ground-truth words.*

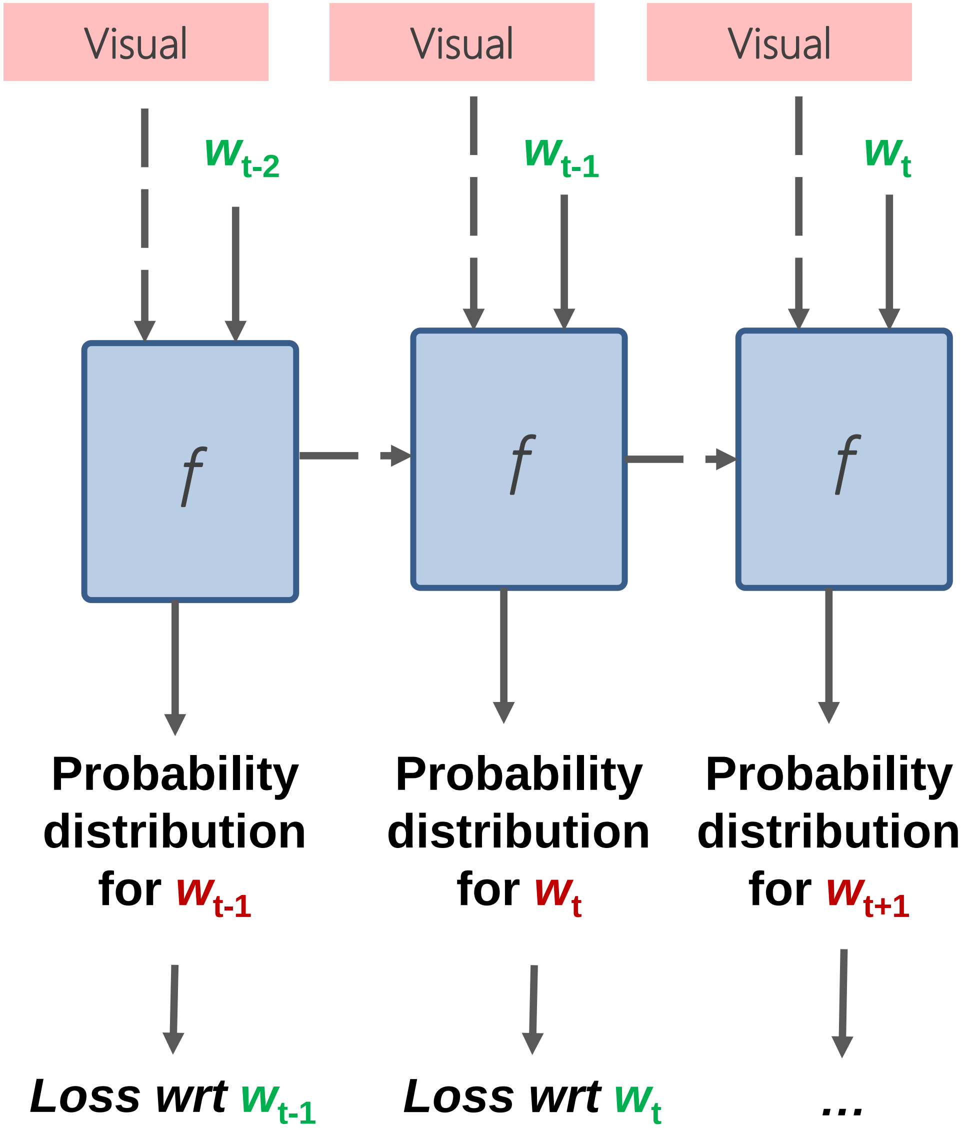


w_t :ground-truth words

Anatomy of a Captioning System

At training time

- *Train the model to predict a word given the previous ground-truth words.*



w_t :ground-truth words

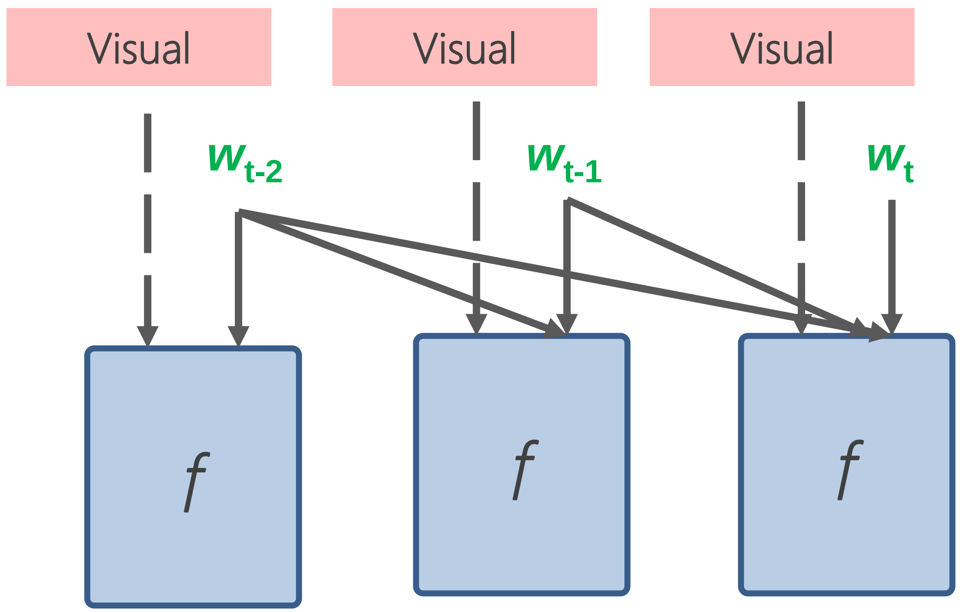
Anatomy of a Captioning System

At training time

- *Train the model to predict a word given the previous ground-truth words.*

If the step function does not depend on its own output at previous timesteps:

- We can parallelize over the t axis.
  - → Training time reduction
  - E.g. Conv1D, Transformer



w_t :ground-truth words

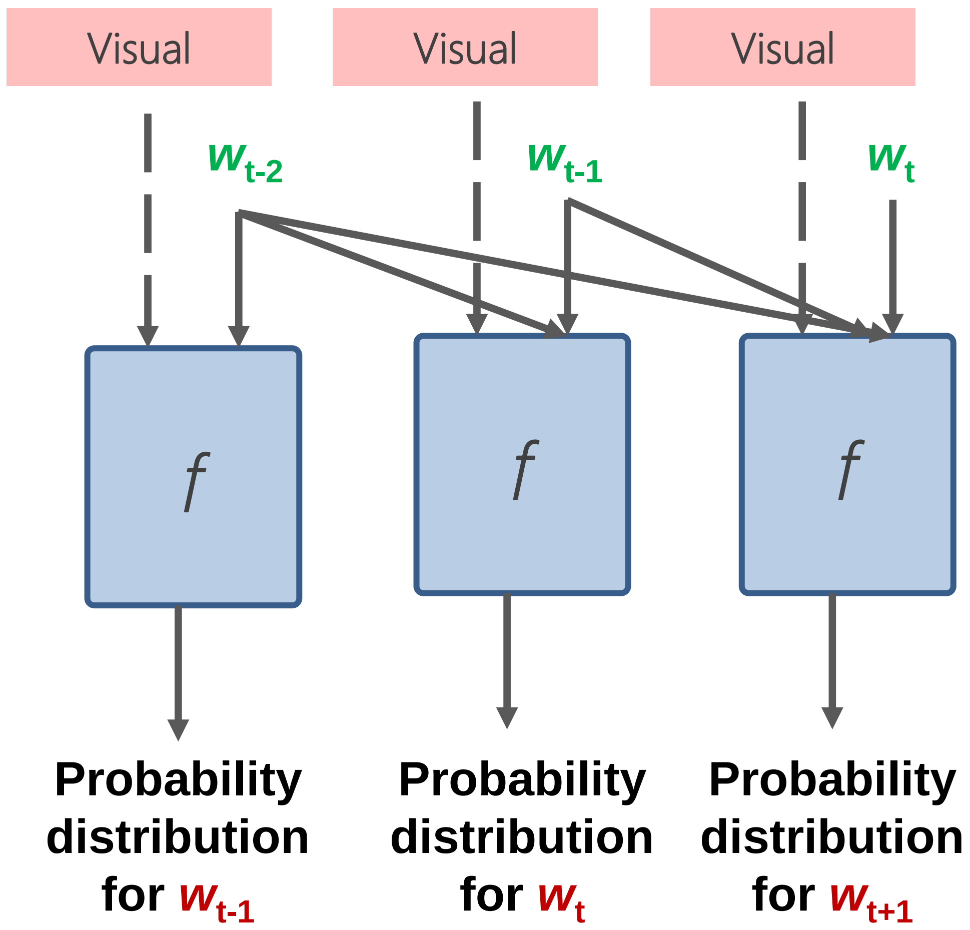
Anatomy of a Captioning System

At training time

- *Train the model to predict a word given the previous ground-truth words.*

If the step function does not depend on its own output at previous timesteps:

- We can parallelize over the t axis.
  - → Training time reduction
  - E.g. Conv1D, Transformer



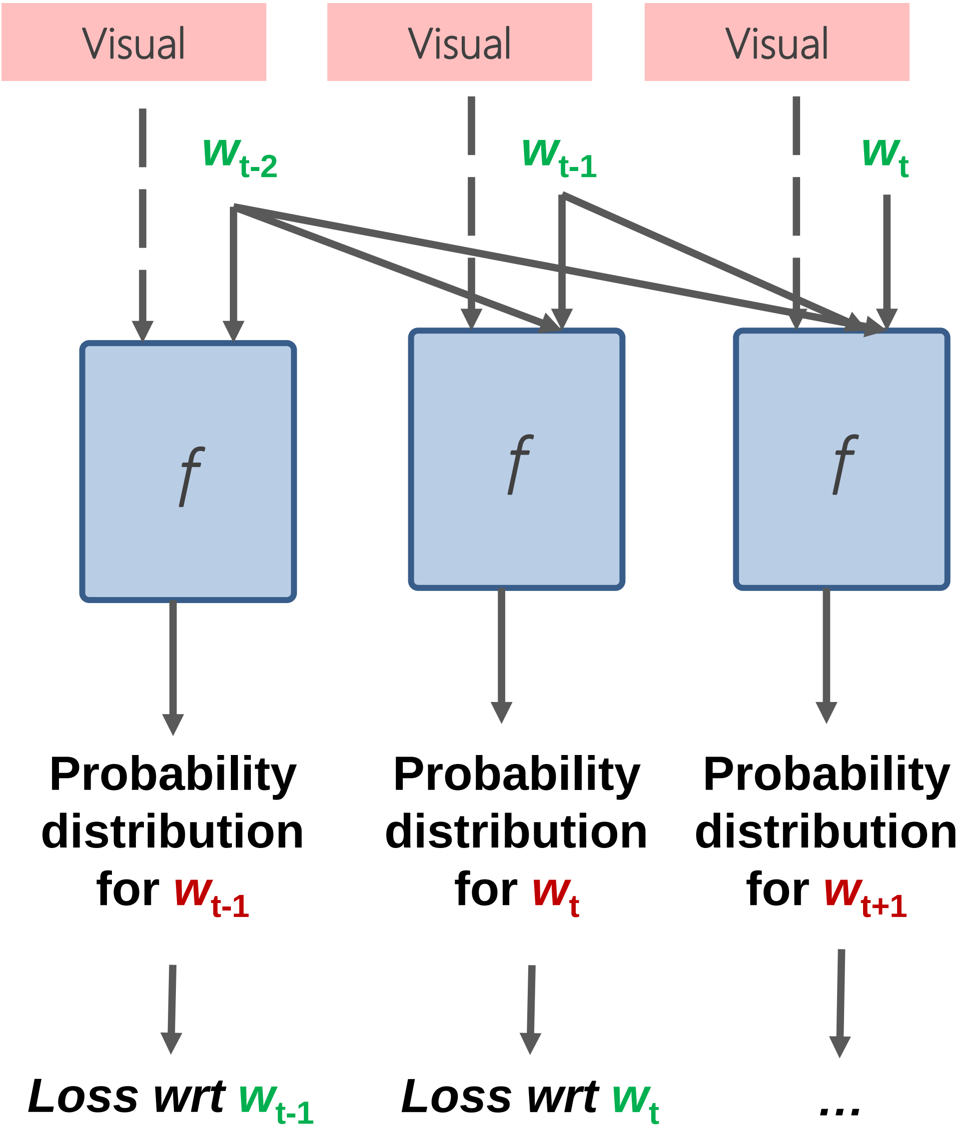
w_t :ground-truth words

At training time

- *Train the model to predict a word given the previous ground-truth words.*

If the step function does not depend on its own output at previous timesteps:

- We can parallelize over the t axis.
  - → Training time reduction
  - E.g. Conv1D, Transformer

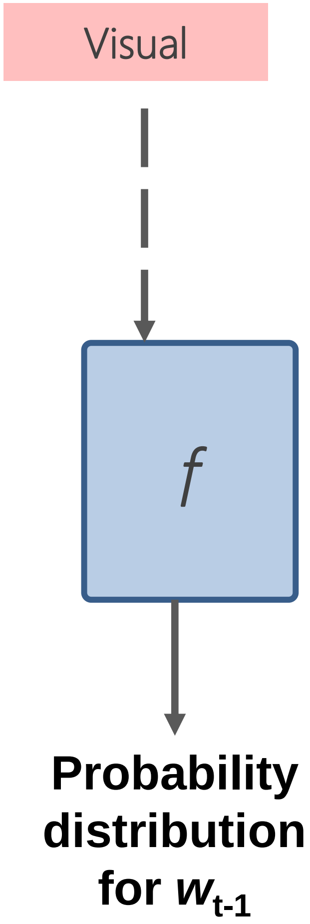


w_t :ground-truth words

Anatomy of a Captioning System

At prediction time (sampling)

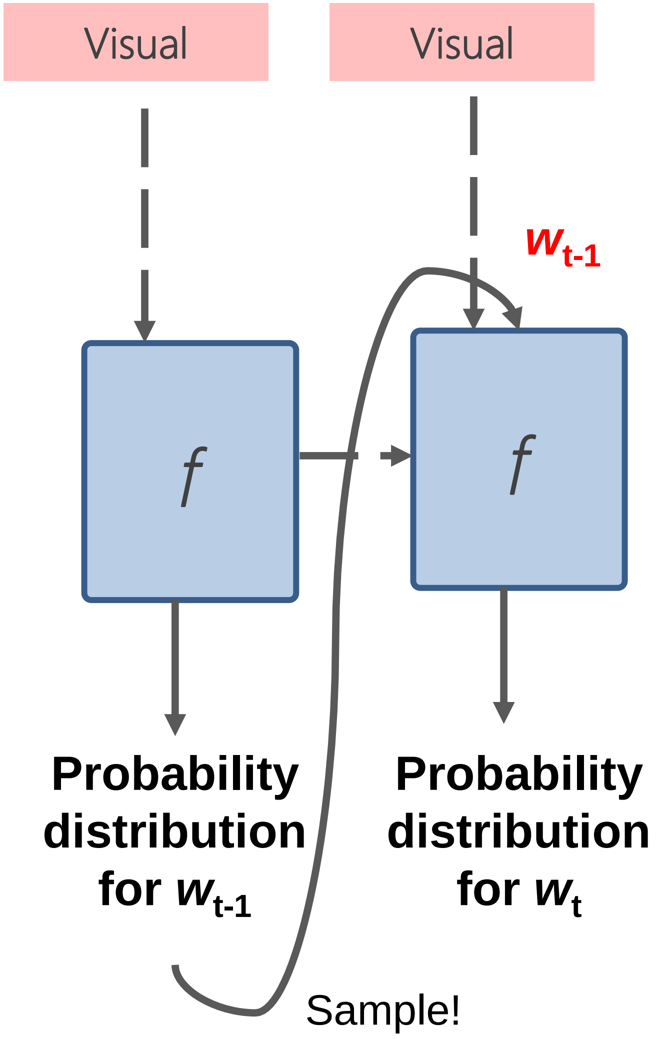
- We sample one word from the previous output, and use this as an input.



w_t : sampled words

At prediction time (sampling)

- We sample one word from the previous output, and use this as an input.

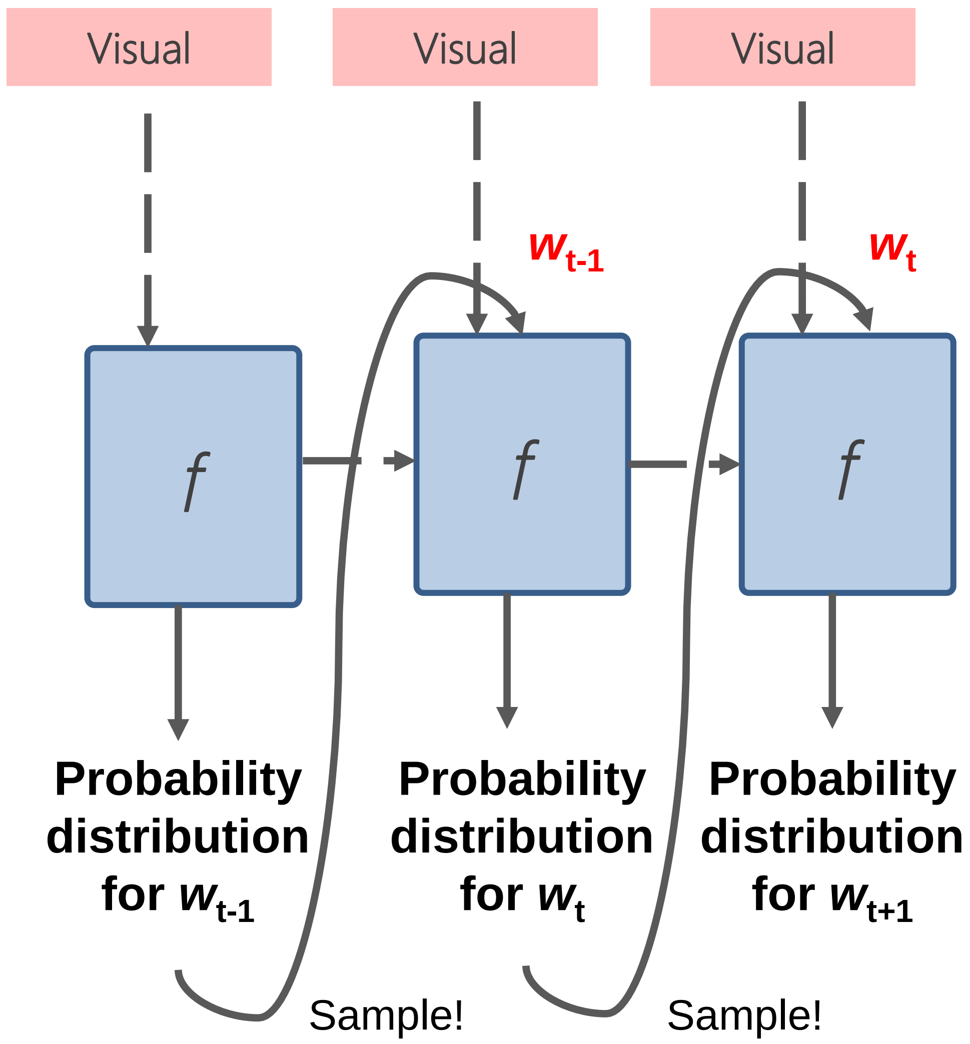


w_t : sampled words

Anatomy of a Captioning System

At prediction time (sampling)

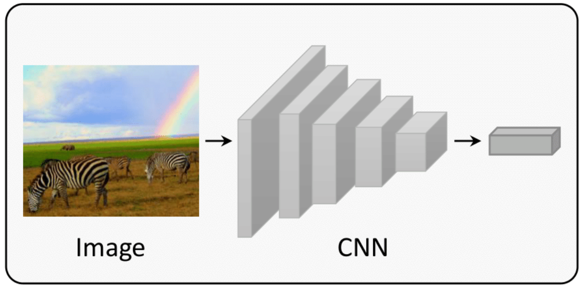
- We sample one word from the previous output, and use this as an input.
- Possible strategies:
  - Always sample the most probable word
  - Build a tree of possible choices, then select the chain of predictions with maximum probability (*beam search*)



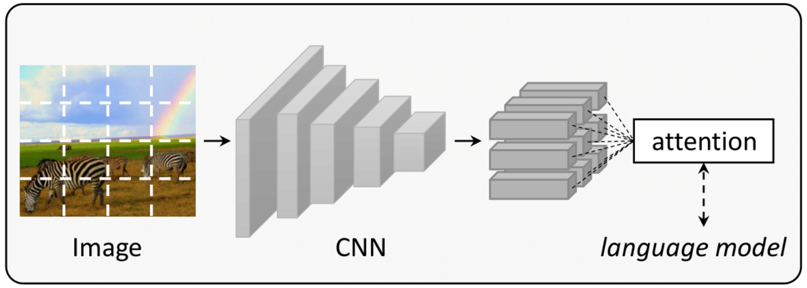
w_t : sampled words

What About Image Encoding

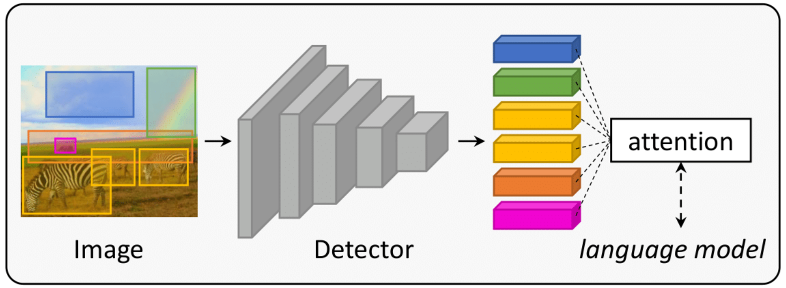
Global CNN Features



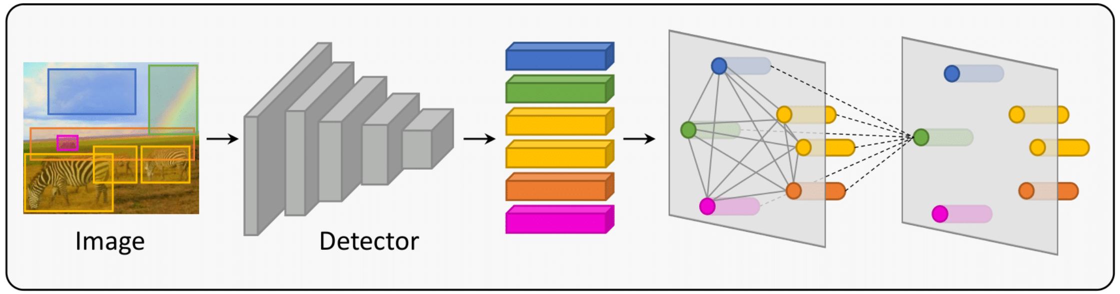
Attention Over Grid of CNN Features



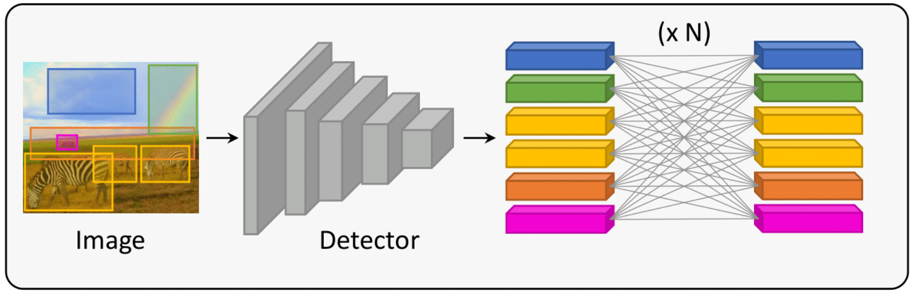
Attention Over Visual Regions



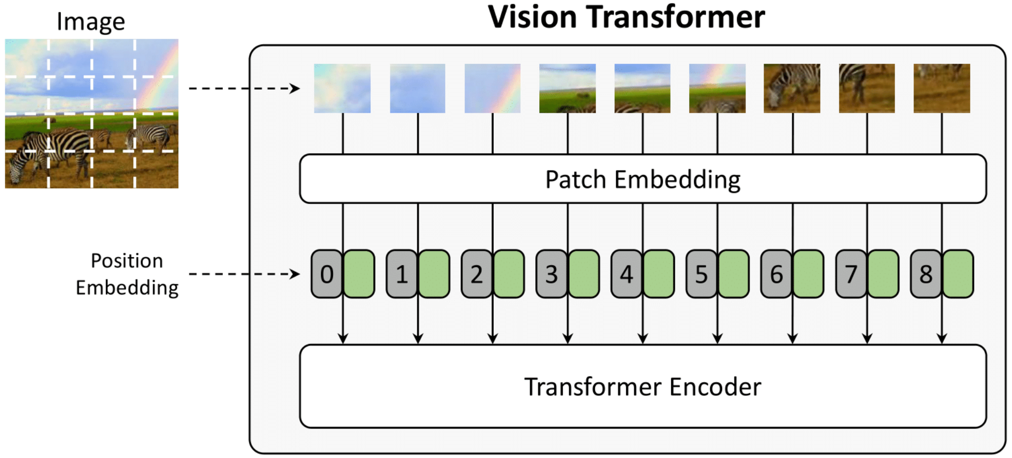
Graph-based Encoding



Self-Attention Encoding



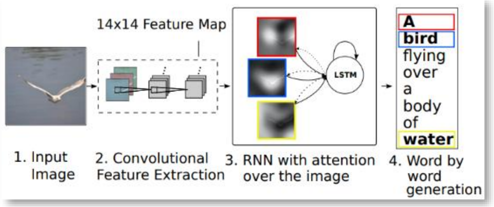
Vision Transformer



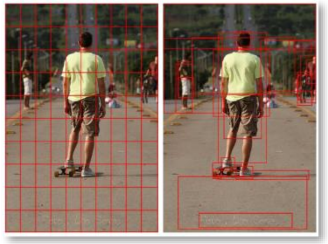
AImage Lab

Describing Images in Natural Language

- **RNN-based Captioning** — Vinyals, *et al*. Show and Tell - *CVPR, 2015*; Karpathy, *et al*. Deep Visual-Semantic Alignments - *CVPR, 2015*

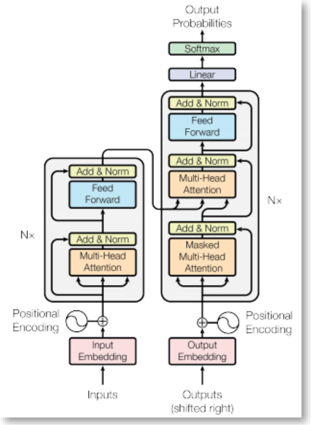


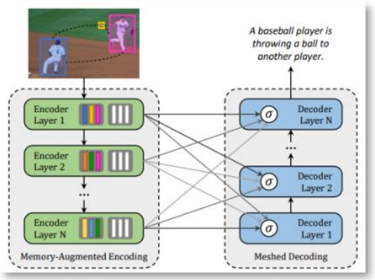
  - Attentive models
    - → attention on spatial features — Xu, *et al*. Show, Attend and Tell - *ICML, 2015*
    - → saliency-driven attention — Cornia, *et al*. Paying More Attention on Saliency - *TOMM, 2018*
    - → attention on image regions — Anderson, *et al*. Bottom-up Top-down Attention - *CVPR, 2018*



  - Training strategies
    - → cross-entropy loss
    - → reinforcement learning — Rennie, *et al*. Self-Critical Sequence Training - *CVPR, 2017*

- **Transformer-based Captioning** — Huang, *et al*. Attention on Attention - *ICCV, 2019*; Li, *et al*. Entangled Transformer - *ICCV, 2019*; Herdade, *et al*. Transforming Objects into Words - *NeurIPS, 2019*; Cornia, *et al.* M^2 Transformer - *CVPR, 2020*





- …

Datasets for Image Captioning

- **Standard datasets** (e.g., Microsoft COCO, FLickr8k, Flickr30k)
- **Pre-training datasets** (e.g., SBU Captions, Conceptual Captions 3M/12M)
- **Domain-specific datasets** (e.g., VizWiz Captions, CUB-200, Oxford-102, Fashion Captioning, Breaking News, GoodNews, TextCaps Localized Narratives)

	Domain	Nb. Images	Nb. Caps (per Image)	Vocab Size	Nb. Words (per Cap.)
MS COCO	Generic	132K	5	27K (10K)	10.5
Flickr30K	Generic	31K	5	18K (7K)	12.4
Flickr8K	Generic	8K	5	8K (3K)	10.9
CC3M	Generic	3.3M	1	48K (25K)	10.3
CC12M	Generic	12.4M	1	523K (163K)	20.0
SBU Captions	Generic	1M	1	238K (46K)	12.1
VizWiz	Assistive	70K	5	20K (8K)	13.0
CUB-200	Birds	12K	10	6K (2K)	15.2
Oxford-102	Flowers	8K	10	5K (2K)	14.1
Fashion Cap.	Fashion	130K	1	17K (16K)	21.0
BreakingNews	News	115K	1	85K (10K)	28.1
GoodNews	News	466K	1	192K (54K)	18.2
TextCaps	OCR	28K	5/6	44K (13K)	12.4
Loc. Narratives	Generic	849K	1/5	16K (7K)	41.8

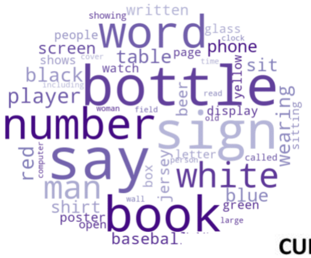
COCO

Woman on a horse jumping over a pole jump.

A glass bowl contains peeled tangerines and cut strawberries.

TextCaps

The billboard displays 'Welcome to Yakima The Palm Springs of Washington'.

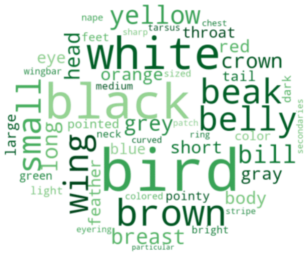


Fashion Captioning

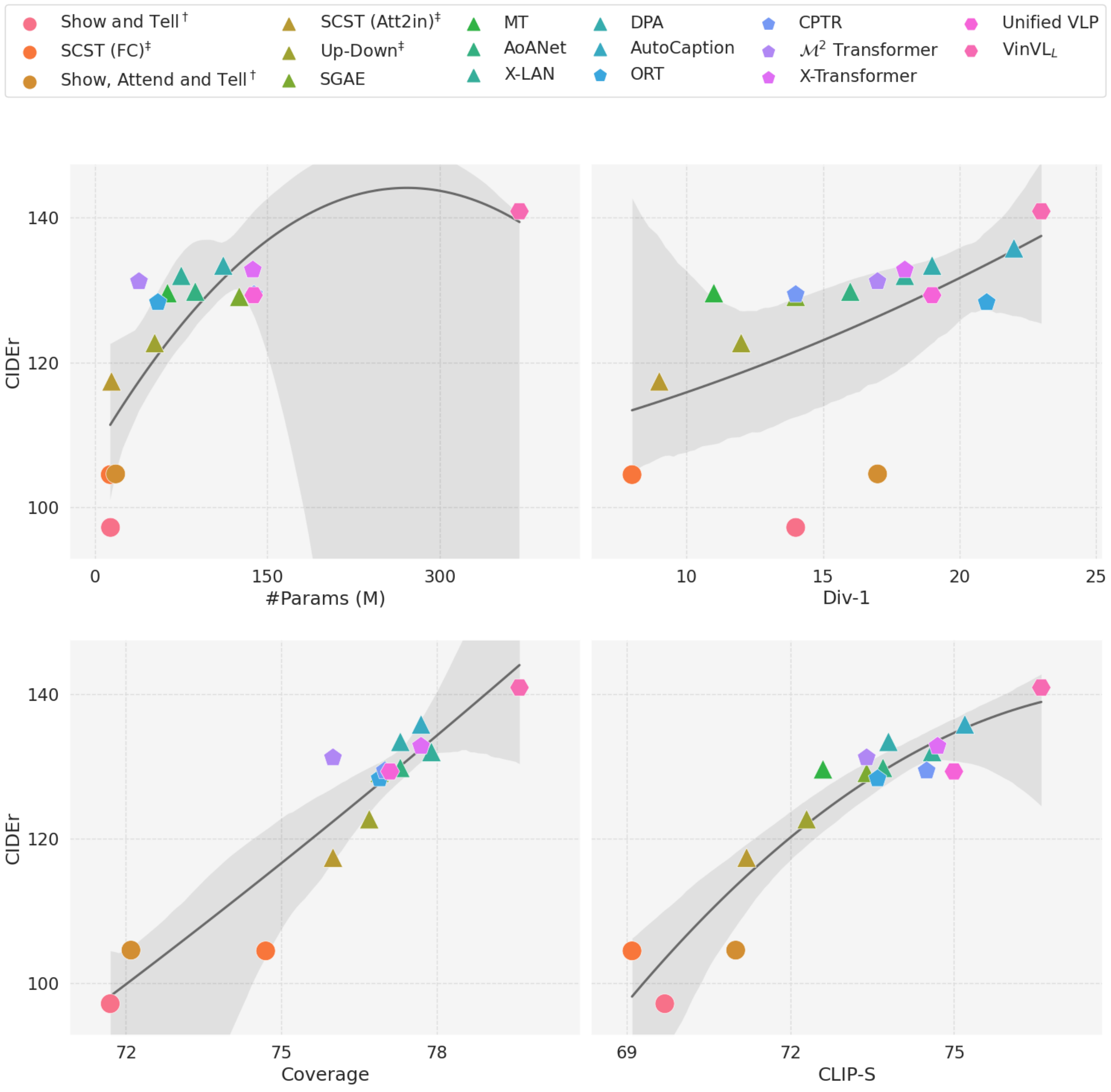
A decorative leather padlock on a compact bag with croc embossed leather.

CUB-200

This bird is blue with white on its chest and has a very short beak.



Evaluation for Image Captioning



			Inputs		
		Original Task	**Pred**	**Refs**	**Image**
Standard	BLEU	Translation	✓	✓	
	METEOR	Translation	✓	✓	
	ROUGE	Summarization	✓	✓	
	CIDEr	Captioning	✓	✓	
	SPICE	Captioning	✓	(✓)	(✓)
Diversity	Div-1/2	Captioning	✓		
	Vocab. Size	Captioning	✓		
	%Novel	Captioning	✓		
Embedding-based	WMD	Doc. Dissimilarity	✓	✓	
	Alignment	Captioning	✓	✓	
	Coverage	Captioning	✓	(✓)	(✓)
Learning-based	TIGEr	Captioning	✓	✓	✓
	BERT-S	Text Similarity	✓	✓	
	CLIP-S	Captioning	✓	(✓)	✓

But looking at the captions is key…

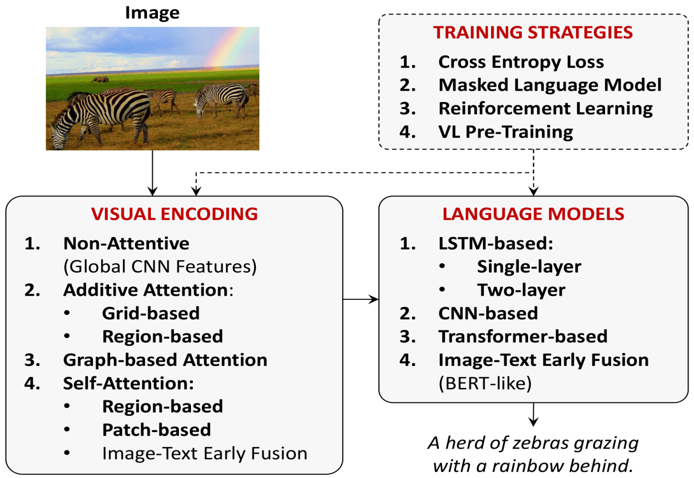
Want to know more?

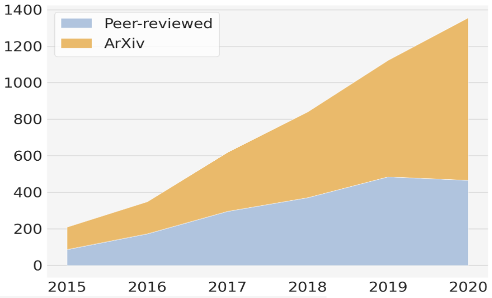
From Show to Tell: A survey on Image Captioning

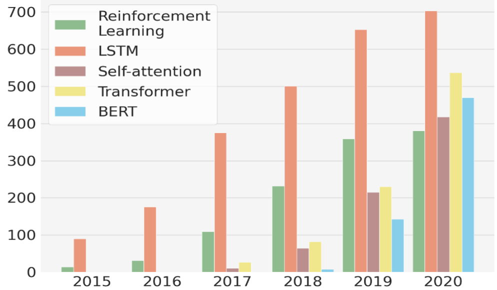
M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, R. Cucchiara

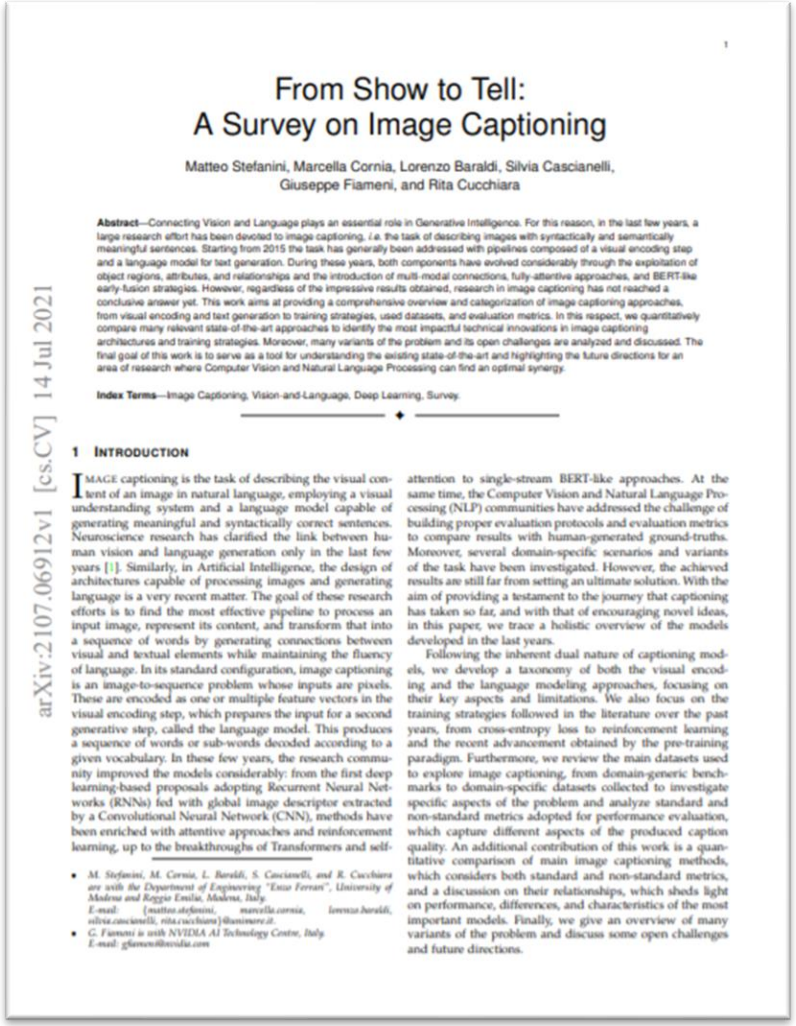
TPAMI 2023 - https://arxiv.org/pdf/2107.06912.pdf

Covers all visual and textual encoding modalities, training strategies, datasets, evaluation metrics and variants, over **more than 177 captioning papers**!





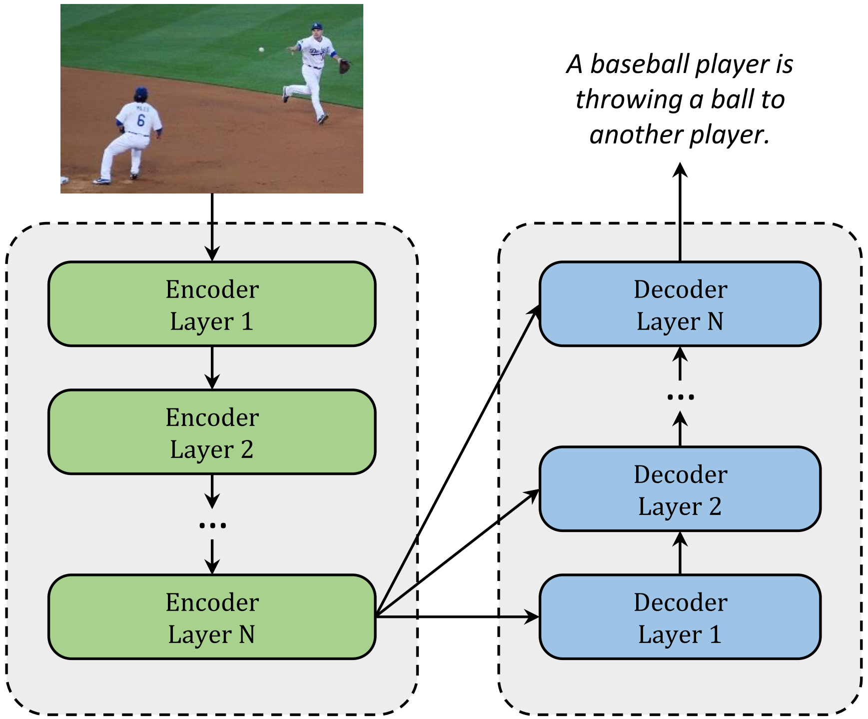




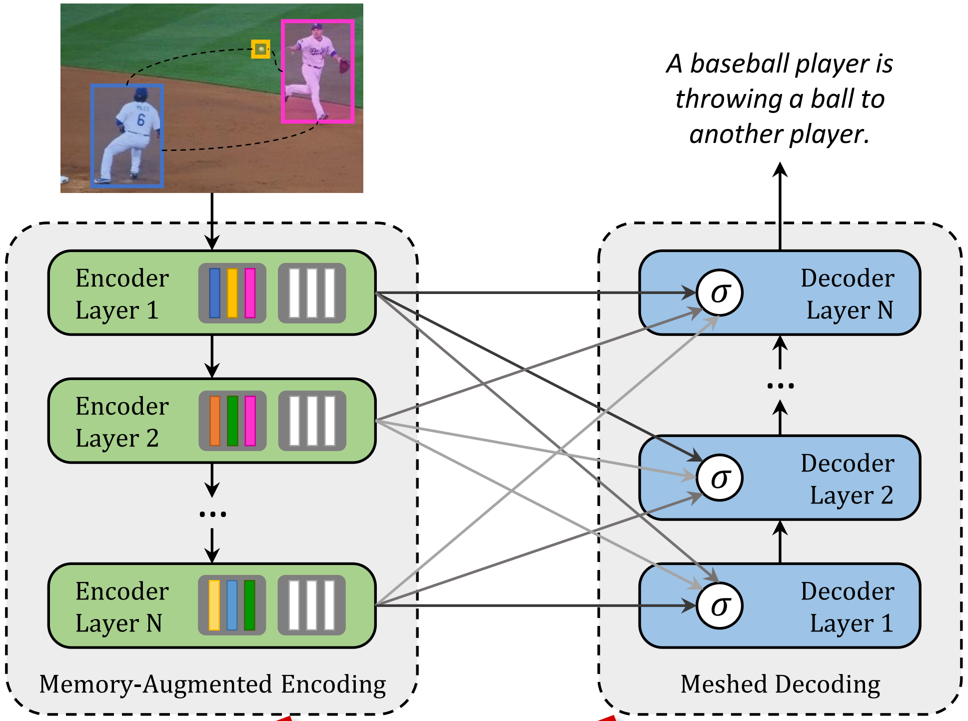
Meshed-Memory Transformer

Meshed-Memory Transformer

Original Transformer



M^2 Transformer



Relationships between image regions are modeled via **persistent memory vectors**.

Encoder and decoder layers are connected in **a mesh-like structure**.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. "Meshed-Memory Transformer for Image Captioning." CVPR 2020.

Meshed-Memory Transformer

- Our model is divided into an **encoder** and a **decoder** module, both made of stacks of attentive layers.
- All intra-modality and cross-modality interactions between word and image features are modeled via **scaled dot-product attention**, without using recurrence.

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V}$$

Self-Attention

- Queries, keys, and values come from the same modality.
- They are extracted from image features in the encoder, and from words in the decoder.

Cross-Attention (decoder only)

- Queries are extracted from words.
- Keys and values are extracted from image features coming from the encoder layers.

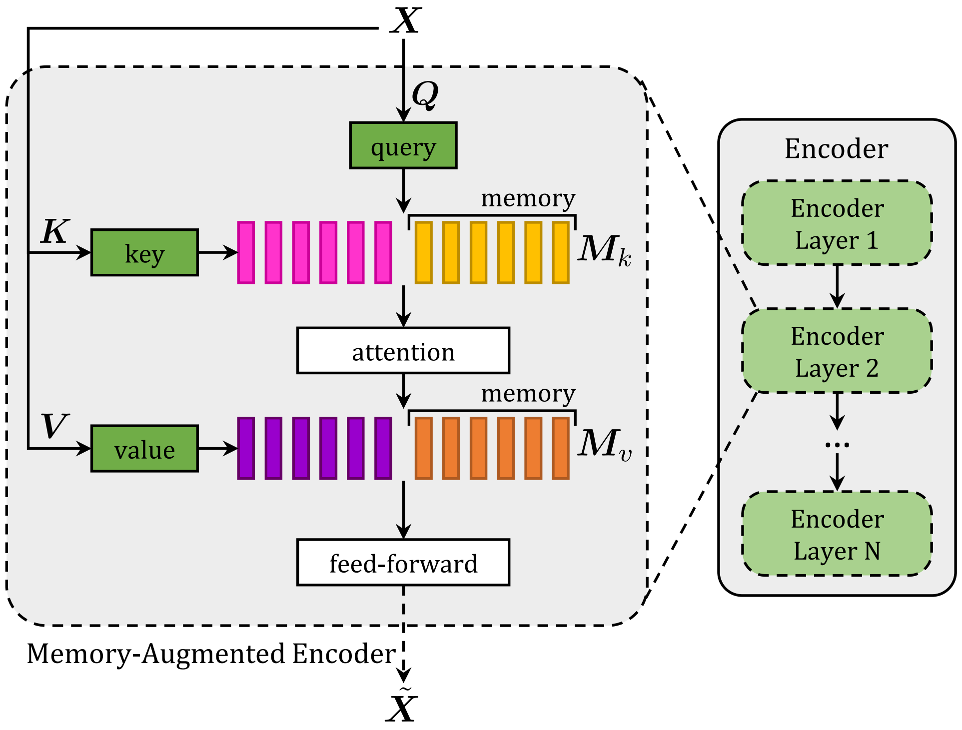
Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. "Meshed-Memory Transformer for Image Captioning." CVPR 2020.

Meshed-Memory Transformer

- The set of keys and values used in an encoder layer is extended with vectors which can encode **a priori information**.
- The additional keys and values are implemented as plain **learnable vectors** which can be directly updated via SGD.
- The operator is defined as:

$$\mathcal{M}_{\text{mem}}(\boldsymbol{X}) = \text{Attention}(W_q\boldsymbol{X}, \boldsymbol{K}, \boldsymbol{V})$$
$$\boldsymbol{K} = [W_k\boldsymbol{X}, \boldsymbol{M}_k]$$
$$\boldsymbol{V} = [W_v\boldsymbol{X}, \boldsymbol{M}_v],$$

- We can learn a multi-level representation of the relationships between image regions integrating learned a priori knowledge.



Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. “Meshed-Memory Transformer for Image Captioning." CVPR 2020.

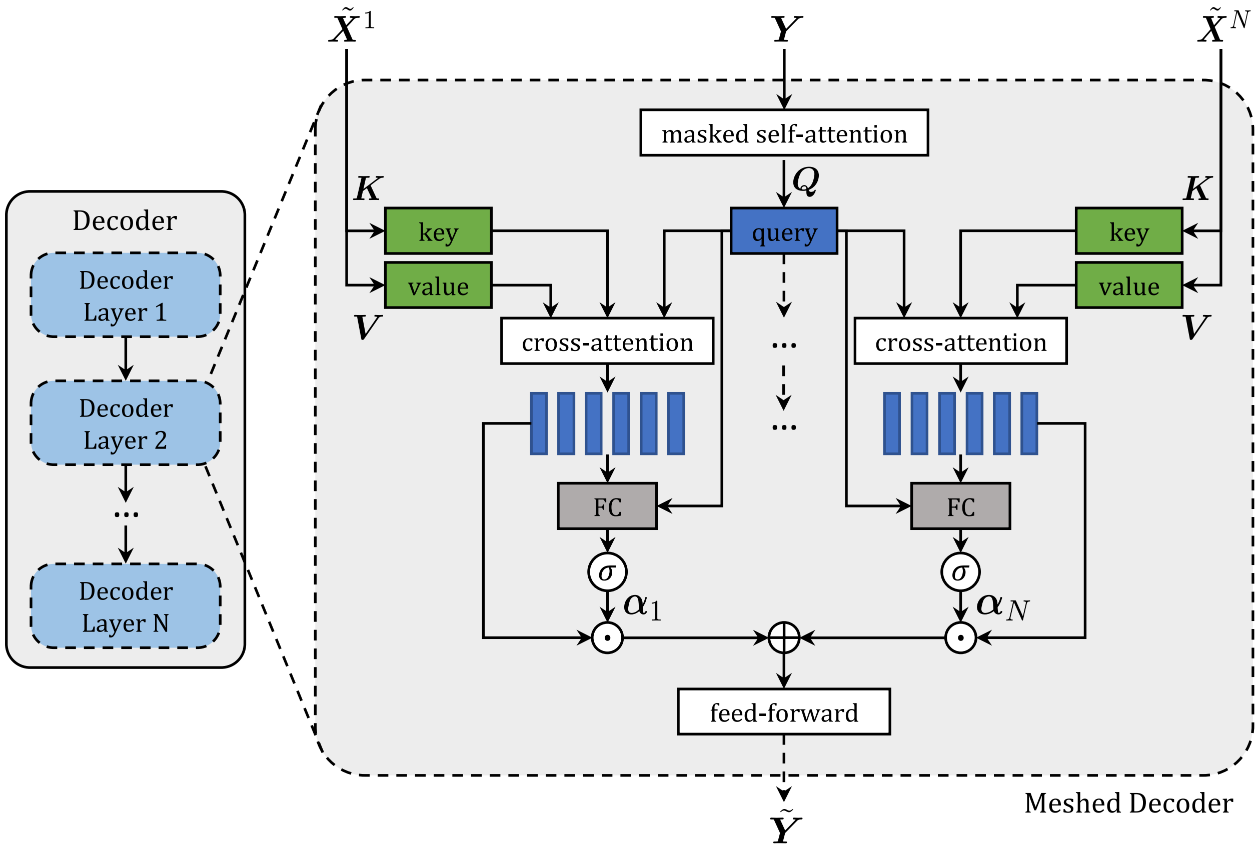
Meshed-Memory Transformer

- We perform **a cross-attention with all encoding layers**.
- Our **meshed attention operator** is defined as

$$\mathcal{M}_{\text{mesh}}(\tilde{\mathcal{X}}, \boldsymbol{Y}) = \sum_{i=1}^{N} \boldsymbol{\alpha}_i \odot \mathcal{C}(\tilde{\boldsymbol{X}}^i, \boldsymbol{Y})$$

where $\mathcal{C}(\cdot,\cdot)$ is the cross-attention computed using queries from the decoder and keys and values from the encoder:

$$\mathcal{C}(\tilde{\boldsymbol{X}}^i, \boldsymbol{Y}) = \text{Attention}(W_q \boldsymbol{Y}, W_k \tilde{\boldsymbol{X}}^i, W_v \tilde{\boldsymbol{X}}^i)$$



- Weights in $\boldsymbol{\alpha}_i$ modulate the contribution of each encoding layer and the relative importance between different layers.

$$\boldsymbol{\alpha}_i = \sigma\left(W_i\left[\boldsymbol{Y}, \mathcal{C}(\tilde{\boldsymbol{X}}^i, \boldsymbol{Y})\right] + b_i\right)$$

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. "Meshed-Memory Transformer for Image Captioning." CVPR 2020.

M² Transformer: Results

		BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE		CIDEr	
		c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
CVPR 2017	SCST [33]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
CVPR 2018	Up-Down [4]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
ICCV 2019	RDN [18]	80.2	95.3	-	-	-	-	37.3	69.5	28.1	37.8	57.4	73.3	121.2	125.2
ECCV 2018	RFNet [15]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
ECCV 2018	GCN-LSTM [48]	80.8	95.9	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
CVPR 2019	SGAE [46]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
ICCV 2019	ETA [24]	81.2	95.0	65.5	89.0	50.9	80.4	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
ICCV 2019	AoANet [14]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
ICCV 2019	GCN-LSTM+HIP [49]	**81.6**	95.9	66.2	90.4	51.5	81.6	39.3	71.0	28.8	38.1	59.0	74.1	127.9	130.2
	$\mathcal{M}^2$ Transformer	**81.6**	**96.0**	**66.4**	**90.8**	**51.8**	**82.7**	**39.7**	**72.8**	**29.4**	**39.0**	**59.2**	**74.8**	**129.3**	**132.1**

→ At the beginning of 2020, our model reached the **first place in the COCO leaderboard.**

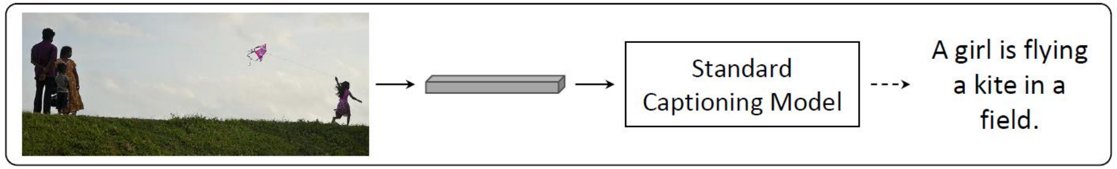
Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. "Meshed-Memory Transformer for Image Captioning." CVPR 2020.

Controllable Captioning

Controllable Image Captioning

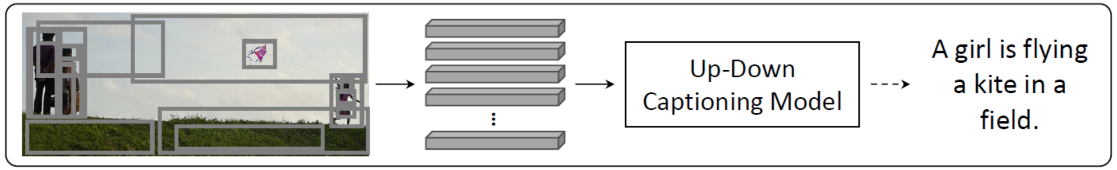
Early captioning approaches:

- Global image feature vector



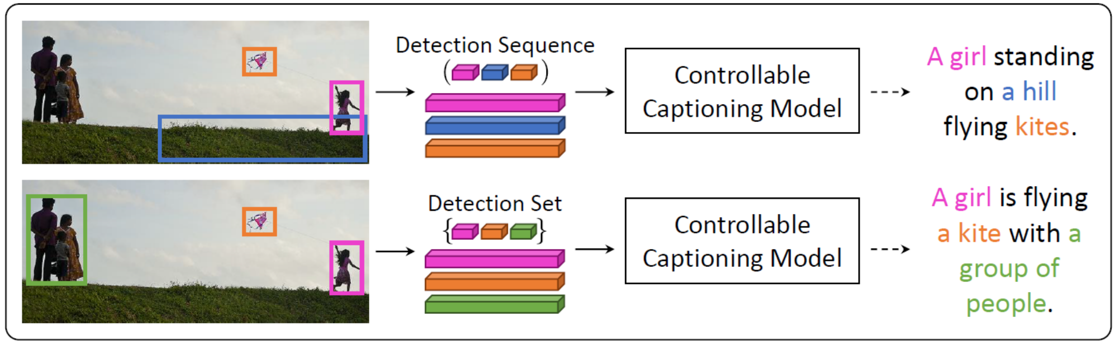
Attention-based approaches:

- Weakly interpretable (through attention)
- Not controllable.
  - We can't decide which regions get processed
  - No control over the generation process.



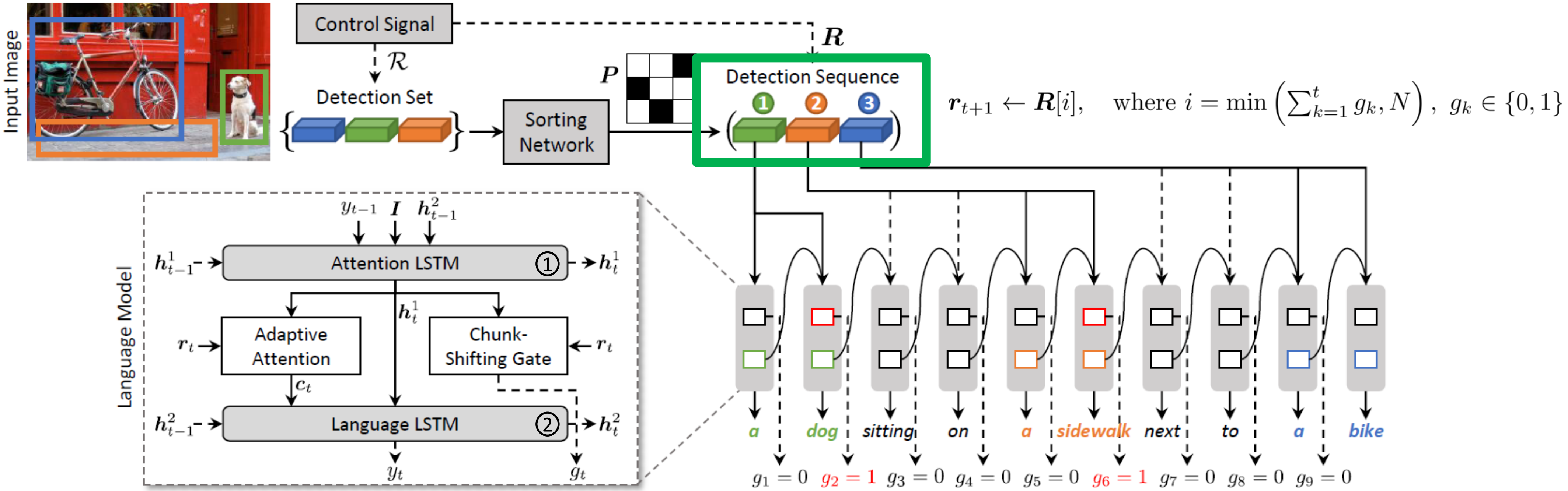
Show, control and tell

- Controllable via regions
  - A sequence (ordered)
  - A set (unordered)



Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

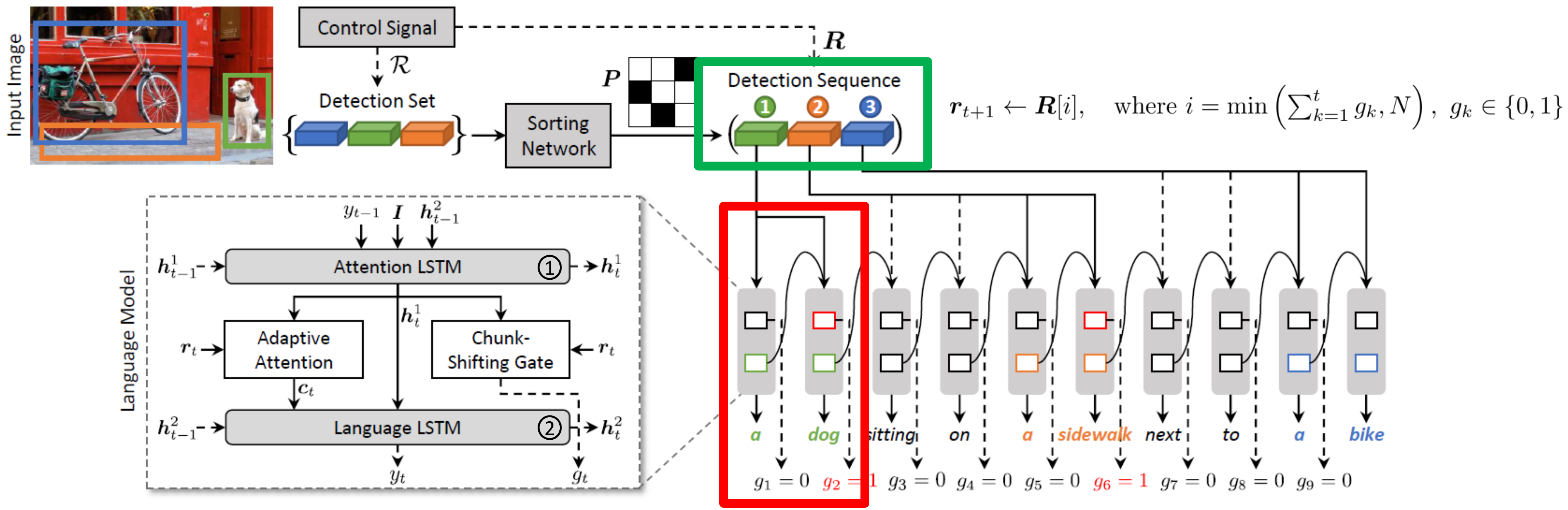
Generating Controllable Captions



- Language model takes as input **a sequence of regions**
- Switches between one region and the next one via a learned chunk-shifting gate
  - When it's done with the generation of chunk, it moves to the next region in the sequence

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

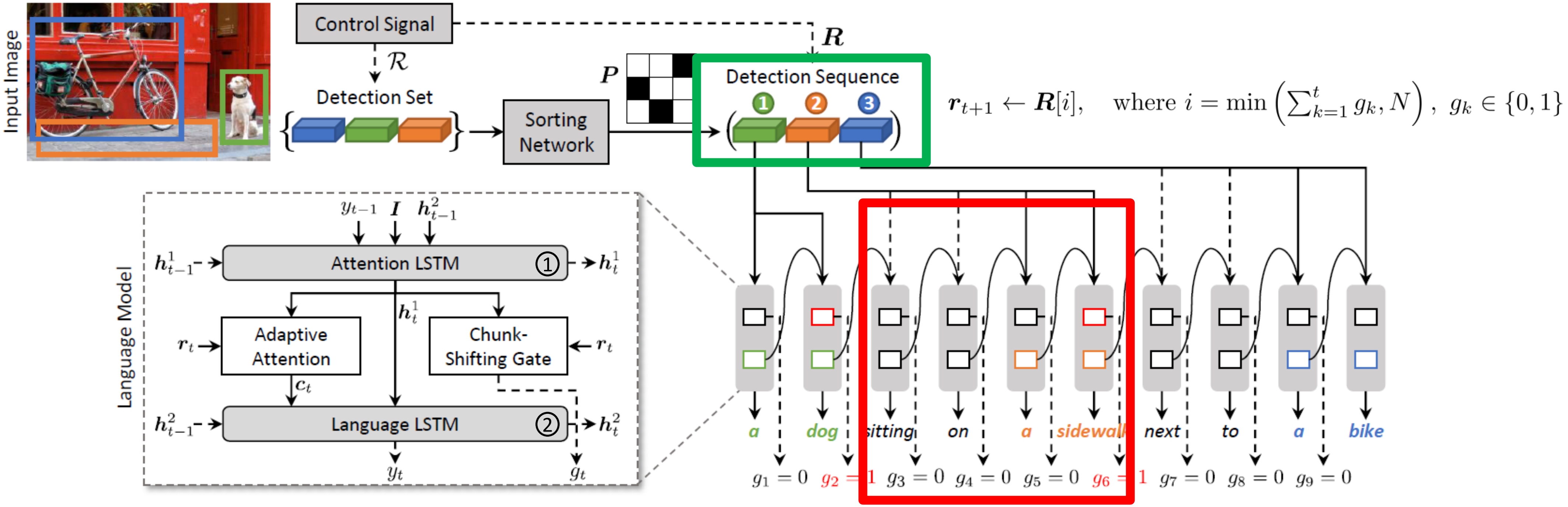
Generating Controllable Captions



- Language model takes as input a sequence of regions
- **Switches between one region and the next one via a learned chunk-shifting gate**
  - When it's done with the generation of chunk, it moves to the next region in the sequence

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

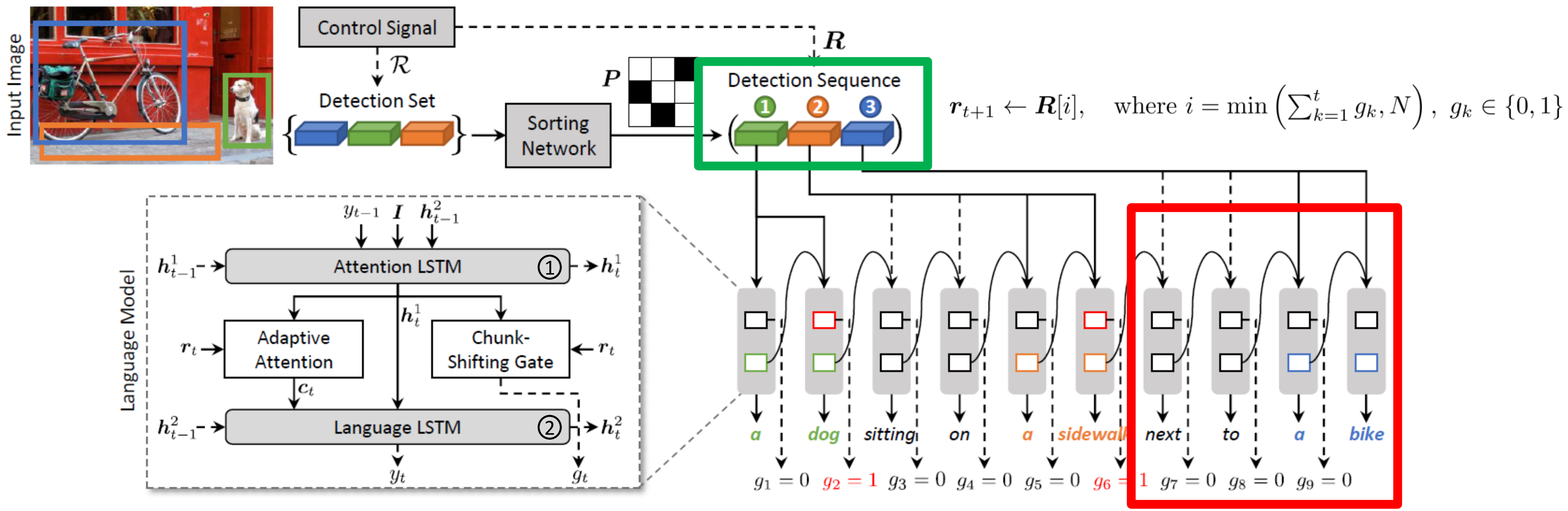
Generating Controllable Captions



- Language model takes as input a sequence of regions
- **Switches between one region and the next one via a learned chunk-shifting gate**
  - When it’s done with the generation of chunk, it moves to the next region in the sequence

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

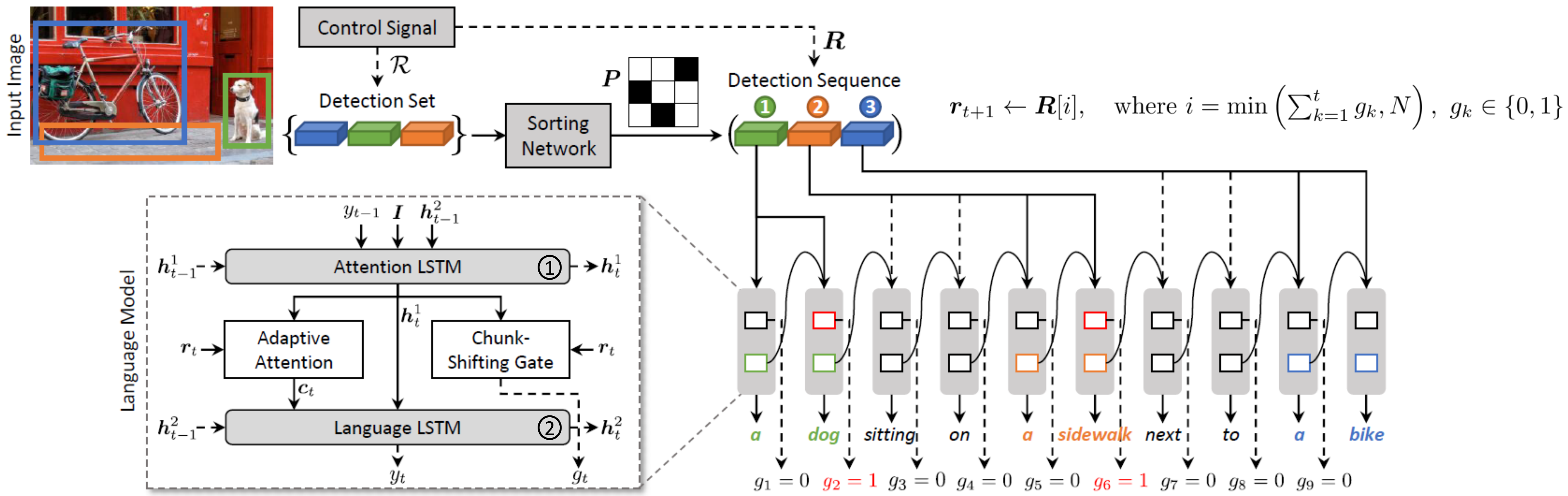
Generating Controllable Captions



- Language model takes as input a sequence of regions
- **Switches between one region and the next one via a learned chunk-shifting gate**
  - When it's done with the generation of chunk, it moves to the next region in the sequence

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Generating Controllable Captions

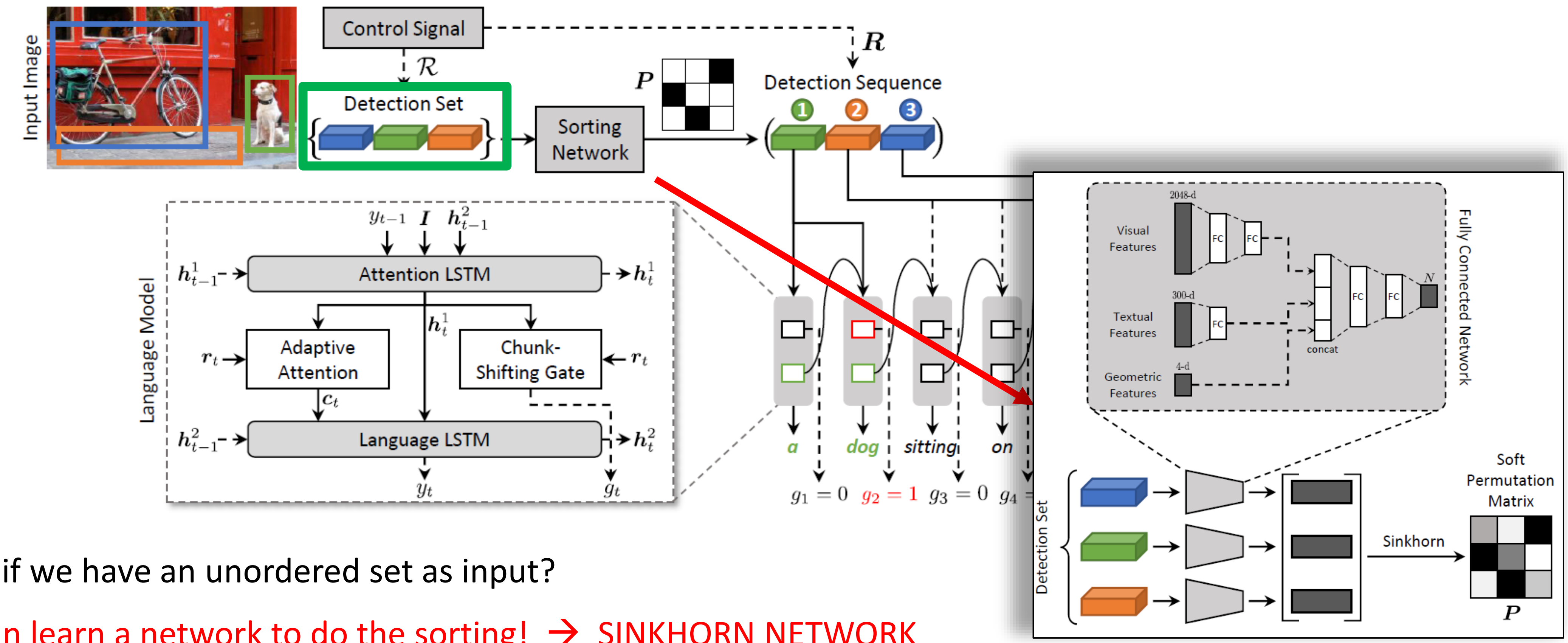


- Train on GT words and shifting gate values (obtained via NLP)

$$L(\theta) = -\sum_{t=1}^{T} \Big(\log \overbrace{p(y^*_t | \boldsymbol{r}^*_{1:t}, \boldsymbol{y}^*_{1:t-1})}^{\text{Word-level probability}} + \\ + g^*_t \log p(g_t = 1 | \boldsymbol{r}^*_{1:t}, \boldsymbol{y}^*_{1:t-1}) + \\ + (1 - g^*_t) \log (1 - \underbrace{p(g_t = 1 | \boldsymbol{r}^*_{1:t}, \boldsymbol{y}^*_{1:t-1})}_{\text{Chunk-level probability}}) \Big)$$

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

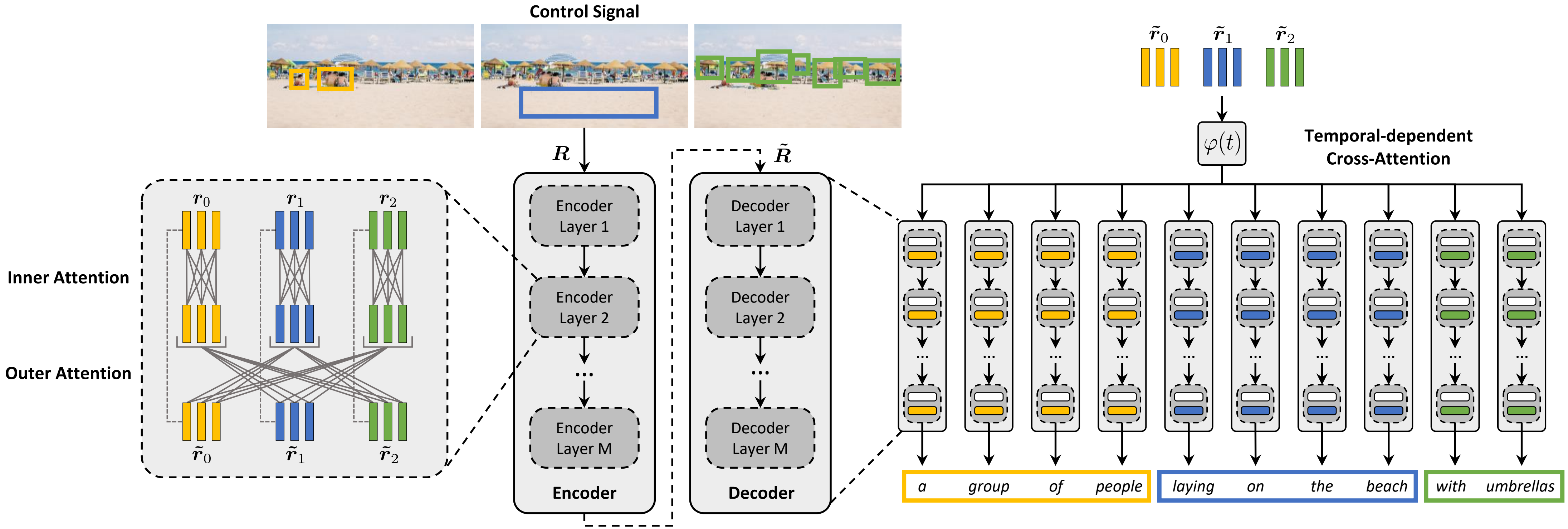
Sorting Network



- What if we have an unordered set as input?
- We can learn a network to do the sorting! → SINKHORN NETWORK
  - Approximates a derivable permutation matrix
  - Train on real data, then use the Hungarian to get the true permutation matrix.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

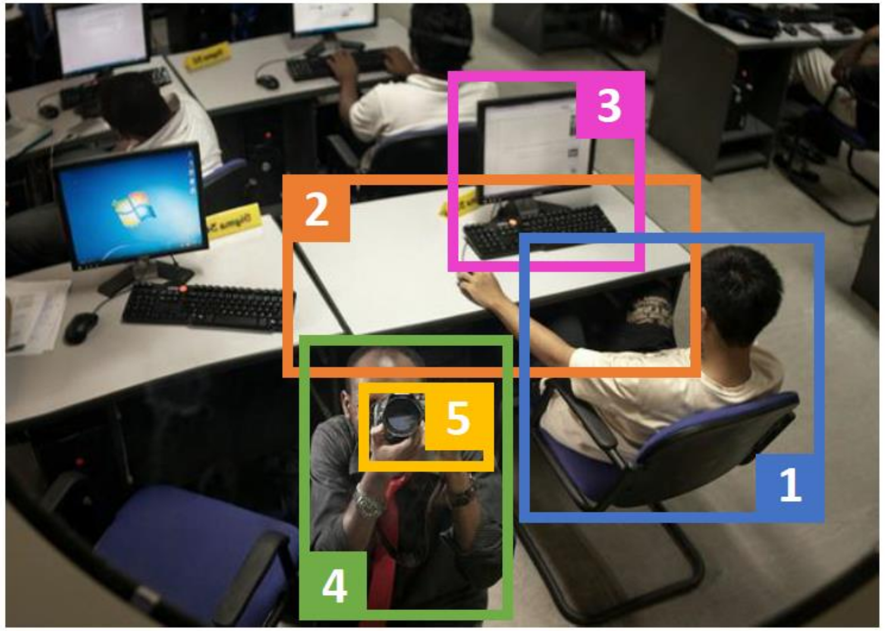
Transformer-based Show, Control and Tell



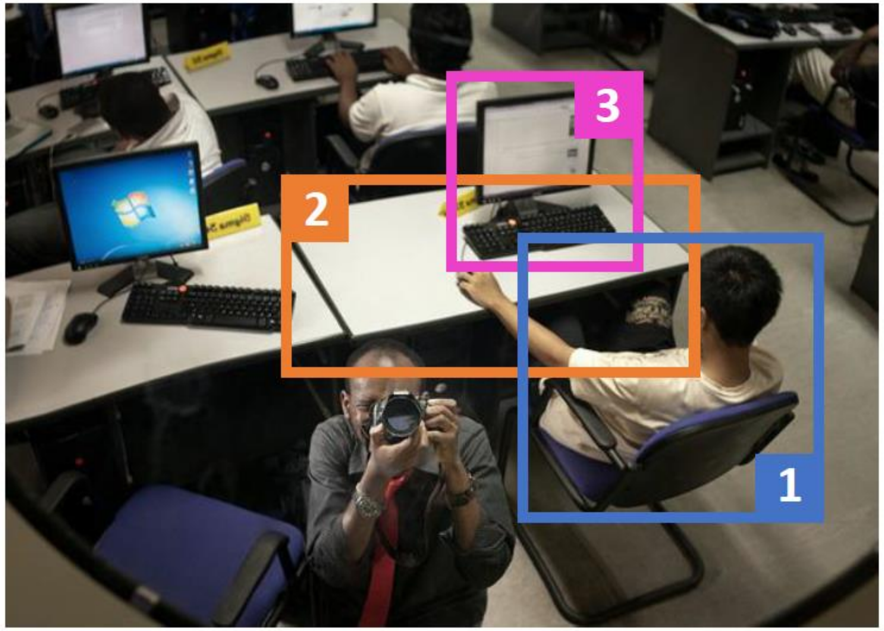
Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Fully-Attentive Iterative Networks for Region-Controlled Image and Video Captioning." Under Review.

Controllability via a Sequence of Regions

Results when controlling with a sequence of regions



A man sitting at a desk with a computer and a man holding a camera.



A man sitting at a desk with a computer.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Fully-Attentive Iterative Networks for Region-Controlled Image and Video Captioning." Under Review.
Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Results when controlling with a set of regions

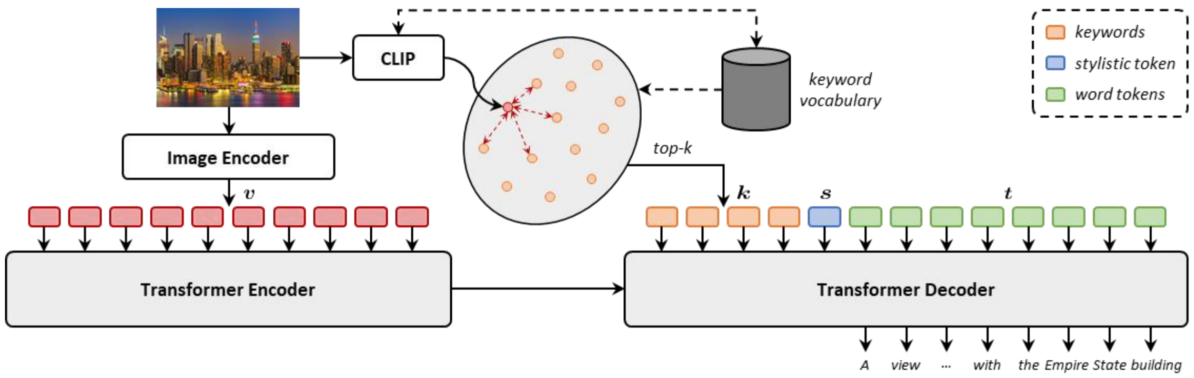
A man in a black jacket skiing down a hill. A man on skis down a snow covered slope.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Fully-Attentive Iterative Networks for Region-Controlled Image and Video Captioning." Under Review.
Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Universal Captioner

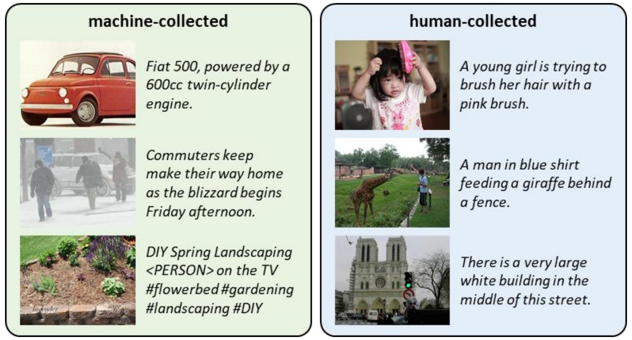
Universal Captioner

- Current captioning models do not cover the entire long-tail distribution of real-world concepts.
- We address the task of generating human-like descriptions with in-the-wild concepts:
  - training on web-scale automatically collected datasets;
  - while maintaining the descriptive style of traditional human-annotated datasets like COCO.



Inputs

- **CNN feature extractors** which can directly take raw pixels as input and avoid the need of using object detectors;
- **Textual keywords** extracted with large-scale cross-modal models;
- **Stylistic tokens** to separate hand-collected and web-based image-caption pairs.



Architecture

- **Fully-attentive encoder-decoder** that jointly encodes keywords, style, and text.

Data

- Training on hand-collected and web-scale datasets, for a total of **36.4 million image-text pairs**.

Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. "Universal Captioner: Long-Tail Vision-and-Language Model Training through Content-Style Separation". *Under Review*

Main results:

- State-of-the-art results on COCO;
- State-of-the-art results on nocaps when using external data;
- Zero-shot generalization to other datasets;
- Capability to name long-tail concepts (*e.g.* proper nouns of places, famous people, brands).

Results on COCO

	B-4	M	R	C	S
$\mathcal{M}^2$ Transformer	39.1	29.2	58.6	131.2	22.6
X-Transformer	39.7	29.5	59.1	132.8	23.4
AutoCaption	40.2	29.9	59.5	135.8	23.8
OSCAR$^{\text{base}}$	40.5	29.7	-	137.6	22.8
VinVL$^{\text{base}}$	40.9	**30.9**	-	140.6	**25.1**
UniversalCap$^{\text{tiny}}$	40.8	29.9	59.9	140.4	23.4
UniversalCap$^{\text{small}}$	**41.2**	30.4	**60.2**	143.0	24.1
UniversalCap$^{\text{base}}$	40.8	30.4	**60.2**	**143.4**	24.2

Results on Open Images

	CLIP-S	# Long-tail Words	# Proper Nouns
VinVL$^{\text{base}}$	0.708	149	55
VinVL$^{\text{large}}$	0.715	186	60
UniversalCap$^{\text{tiny}}$	0.728	821	410
UniversalCap$^{\text{small}}$	0.732	866	432
UniversalCap$^{\text{base}}$	**0.739**	**1,071**	**469**

Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. “Universal Captioner: Long-Tail Vision-and-Language Model Training through Content-Style Separation”. *Under Review*

AImage Lab

Universal Captioner

Standard Captioner: *A large building with a statue on the front.*

Universal Captioner: *The **Arc de Triomphe** in **Paris** with a blue sky.*

Standard Captioner: *A president speaking at a podium in front of a flag.*

Universal Captioner: ***President Obama** giving a speech in front of an American flag.*

Standard Captioner: *Two plates of pancakes with syrup on a table.*

Universal Captioner: *A plate of pancakes and **a jar of Nutella** on a table.*

Standard Captioner: *A red truck driving down a highway.*

Universal Captioner: *A red **Coca-Cola truck** driving down the highway.*

Standard Captioner: *A man standing in front of an apple screen.*

Universal Captioner: ***Steve Jobs** standing in front of an **Apple logo**.*

Standard Captioner: *A castle with flowers in the middle of a body of water.*

Universal Captioner: *A view of the **Taj Mahal** reflecting in the water.*

Standard Captioner: *A woman with blonde hair is posing for a picture.*

Universal Captioner: *A picture of **Marilyn Monroe** with a red lipstick.*

Standard Captioner: *A person holding a cellphone in their hand.*

Universal Captioner: *A person holding a cellphone with a **Facebook logo** on it.*

Standard Captioner: *A picture of a bridge over a body of water.*

Universal Captioner: *A picture of the **Golden Gate bridge** in **San Francisco**.*

Standard Captioner: *A crowd of people standing in front of a tall tower.*

Universal Captioner: *A group of people standing near the **leaning tower of Pisa**.*

Standard Captioner: *There is a clown mask on top of a store.*

Universal Captioner: *A statue of **Ronald McDonald** in front of a **McDonald's**.*

Standard Captioner: *A poster of a young boy with two children.*

Universal Captioner: *A **Harry Potter and the Prisoner of Azkaban** concert poster.*

Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. "Universal Captioner: Long-Tail Vision-and-Language Model Training through Content-Style Separation". *Under Review*

Evaluation Metrics

A New Evaluation metric

PAC-S: A new metric for evaluating image-text correspondence

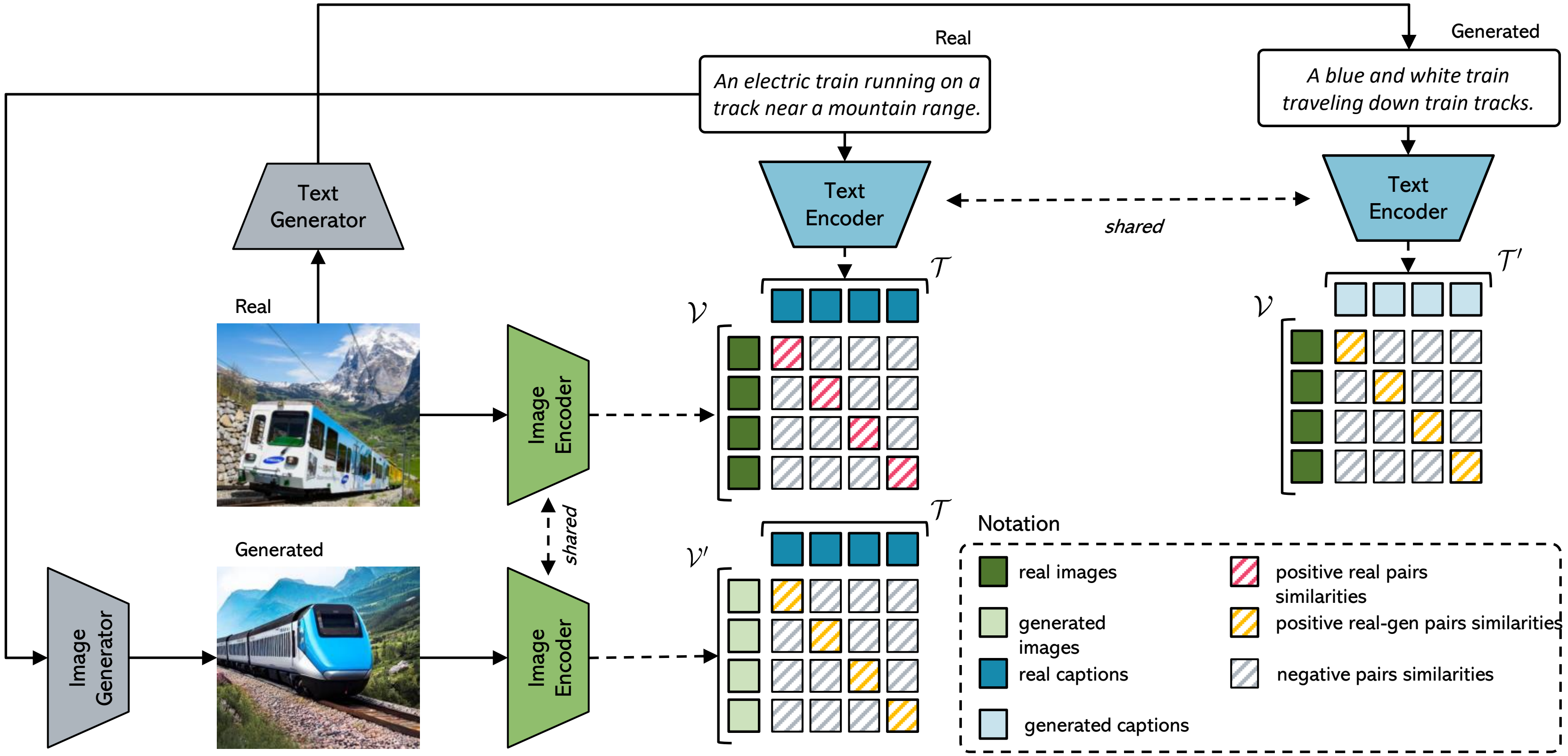
- Existing metrics for image-text correspondence are either only based on **(few) human references** or multi-modal embeddings trained on **noisy data**.
- We propose a **learnable metric** for video and image captioning, which employs both pre-training on **web-collected data**, **generated data for data augmentation** and the power of **human annotations**.
- Based on a ***positive-augmented training*** of a multimodal embedding space.
- Our metric outperforms previous reference-free and reference-based metrics in terms of ***correlation with human judgment***.

Image	Candidate Captions	METEOR	CIDEr	CLIP-S	PAC-S
	*A black cow by a person.*	**9.67**	14.9	**0.766**	0.676
	A cow walking through a field.	15.0	17.2	0.754	**0.775**
	*A silver bicycle is parked in a living room.*	23.1	**68.6**	**0.686**	0.853
	A silver bicycle leaning up against a kitchen table and chairs.	32.4	63.7	0.637	**0.862**
	*A yellow bus passes through an intersection.*	**42.7**	**167.0**	**0.816**	0.836
	A yellow bus is traveling down a city street just past an intersection.	33.9	94.5	0.813	**0.844**

S. Sarto, M. Barraco, M. Cornia, L. Baraldi, R. Cucchiara "Positive-Augmented Constrastive Learning for Image and Video Captioning Evaluation«, **CVPR 2023 Highlight**

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Positive-Augmented Contrastive Learning



- ***Dual-encoder architecture*** comparing the visual and textual inputs via cosine similarity
- Usage of ***synthetic generators*** of both visual and textual data (Stable Diffusion[1] and BLIP[2], respectively)

Fine-tuning on human annotated data by taking into account ***contrastive relationship*** between real and generated matching image-caption pairs.

1. *Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022.*
2. *Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In ICML, 2022.*

Image Captioning Correlation with Human Judgment

PAC score achieves the **best correlation with human judgment** and accuracy on all the considered image datasets, demonstrating its ***effectiveness*** compared to previously proposed metrics.

	Flickr8k-Expert		Flickr8k-CF	
	Kendall τ_b	Kendall τ_c	Kendall τ_b	Kendall τ_c
BLEU-1	32.2	32.3	17.9	9.3
BLEU-4	30.6	30.8	16.9	8.7
ROUGE	31.1	32.3	19.9	10.3
METEOR	41.5	41.8	22.2	11.5
CIDEr	43.6	43.9	24.6	12.7
SPICE	51.7	44.9	24.4	12.0
BERT-S	-	39.2	22.8	-
LEIC	46.6	-	29.5	-
BERT-S++	-	46.7	-	-
UMIC	-	46.8	-	-
TIGEr	-	49.3	-	-
ViLBERTScore	-	50.1	-	-
MID	-	54.9	37.3	-
CLIP-S	51.1	51.2	34.4	17.7
PAC-S	**53.9** (+2.8)	**54.3** (+3.1)	**36.0** (+1.6)	**18.6** (+0.9)
RefCLIP-S	52.6	53.0	36.4	18.8
RefPAC-S	55.4 (+2.8)	55.8 (+2.8)	37.6 (+1.2)	19.5 (+0.7)

	Composite	
	Kendall τ_b	Kendall τ_c
BLEU-1	29.0	31.3
BLEU-4	28.3	30.6
ROUGE	30.0	32.4
METEOR	36.0	38.9
CIDEr	34.9	37.7
SPICE	38.8	40.3
BERT-S	-	30.1
BERT-S++	-	44.9
TIGEr	-	45.4
ViLBERTScore	-	52.4
FAIEr	-	51.4
CLIP-S	49.8	53.8
PAC-S	**51.5** (+1.7)	**55.7** (+1.9)
RefCLIP-S	51.2	55.4
RefPAC-S	52.8 (+1.6)	57.1 (+1.7)

	Pascal-50S				
	HC	HI	HM	MM	Mean
length	51.7	52.3	63.6	49.6	54.3
BLEU-1	64.6	95.2	91.2	60.7	77.9
BLEU-4	60.3	93.1	85.7	57.0	74.0
ROUGE	63.9	95.0	92.3	60.9	78.0
METEOR	66.0	97.7	94.0	66.6	81.1
CIDEr	66.5	97.9	90.7	65.2	80.1
BERT-S	65.4	96.2	93.3	61.4	79.1
BERT-S++	65.4	98.1	96.4	60.3	80.1
TIGEr	56.0	99.8	92.8	74.2	80.7
ViLBERTScore	49.9	99.6	93.1	75.8	79.6
FAIEr	59.7	99.9	92.7	73.4	81.4
MID	67.0	99.7	97.4	76.8	85.2
CLIP-S	55.9	**99.3**	96.5	72.0	80.9
PAC-S	**60.6** (+4.7)	**99.3** (+0.0)	**96.9** (+0.4)	**72.9** (+0.9)	**82.4** (+1.5)
RefCLIP-S	64.9	**99.5**	95.5	73.3	83.3
RefPAC-S	68.2 (+3.3)	**99.5** (+0.0)	**95.6** (+0.1)	**75.9** (+2.6)	**84.8** (+1.5)

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. JAIR, 47:853–899, 2013

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge

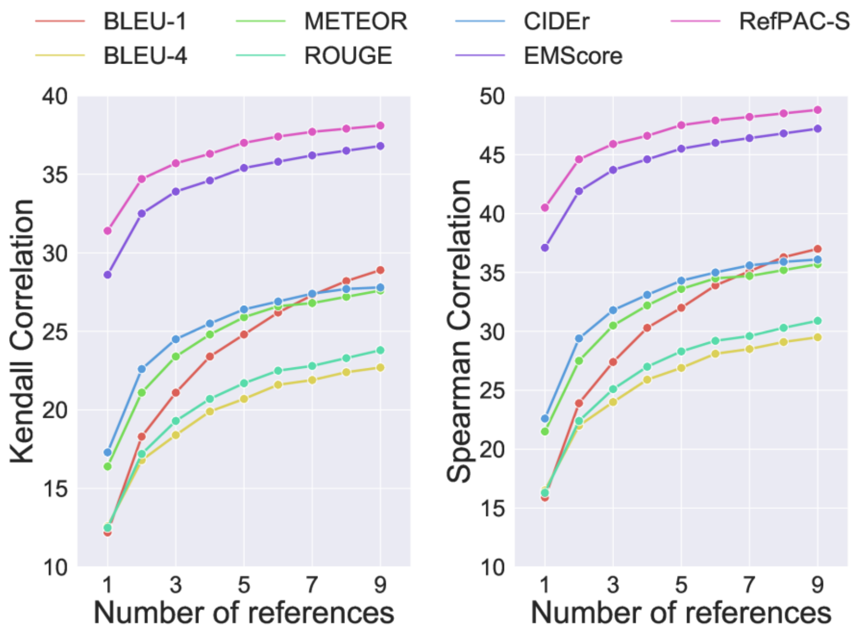
Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In CVPR, 2015

Video Captioning Correlation with Human Judgment

It works well on videos too!

	No Ref		1 Ref		9 Refs	
	Kendall τ_b	Spearman ρ	Kendall τ_b	Spearman ρ	Kendall τ_b	Spearman ρ
BLEU-1	-	-	12.2	15.9	28.9	37.0
BLEU-4	-	-	12.6	16.4	22.4	29.5
ROUGE	-	-	12.5	16.3	23.8	30.9
METEOR	-	-	16.4	21.5	27.6	35.7
CIDEr	-	-	17.3	22.6	27.8	36.1
BERT-S	-	-	18.2	23.7	29.3	37.8
BERT-S++	-	-	15.2	19.8	24.4	31.7
EMScore	23.2	30.3	28.6	37.1	36.8	47.2
PAC-S / RefPAC-S	**25.1** (+1.9)	**32.6** (+2.3)	**31.4** (+2.8)	**40.5** (+3.4)	**38.1** (+1.3)	**48.8** (+1.6)

Human judgment correlation scores on the VATEX-EVAL[1] dataset. We show Kendall τ_B correlation score at varying of the number of reference captions.



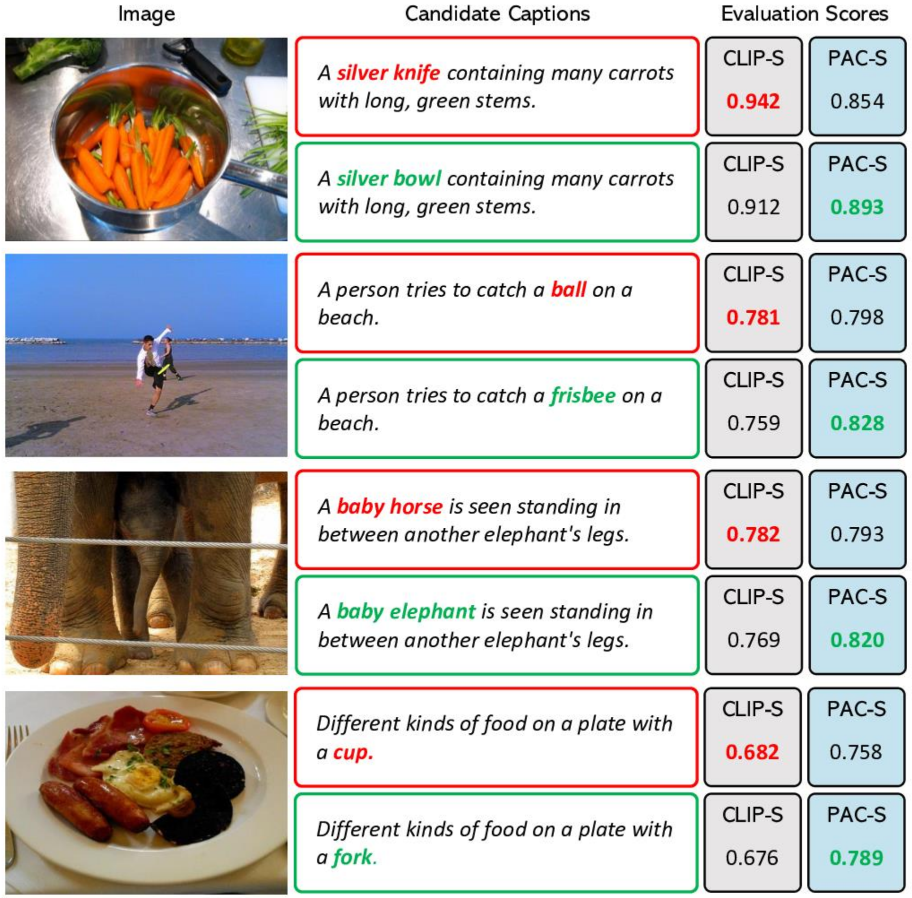
Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embedding Matching. In CVPR, 2022

S. Sarto, M. Barraco, M. Cornia, L. Baraldi, R. Cucchiara "Positive-Augmented Constrastive Learning for Image and Video Captioning Evaluation«, **CVPR 2023 Highlight**

And it hallucinates less than previous metrics ☺

	FOIL		ActivityNet-FOIL
	Acc. (1 Ref)	Acc. (4 Refs)	Accuracy
BLEU-1	65.7	85.4	60.1
BLEU-4	66.2	87.0	66.1
ROUGE	54.6	70.4	56.7
METEOR	70.1	82.0	72.9
CIDEr	85.7	94.1	77.9
MID	90.5	90.5	-
CLIP-S	87.2	87.2	-
EMScore	-	-	89.5
PAC-S	**89.9** (+2.7)	**89.9** (+2.7)	**90.1** (+0.6)
RefCLIP-S	91.0	92.6	-
EMScoreRef	-	-	92.4
RefPAC-S	**93.8** (+2.8)	**95.2** (+2.6)	**93.5** (+1.1)

We extend our analysis to two datasets for detecting hallucinations in textual sentences, namely FOIL[2] and ActivityNet[1].



Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embedding Matching. In CVPR, 2022

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aur´elie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. FOIL it! Find One mismatch between Image and Language caption. In ACL, 2017.

Different Cross-Modal Features

PAC-S achieves the best results across ***all cross-modal backbones*** and almost all datasets, overcoming correlation and accuracy scores of other metrics by a large margin.

		Flickr8k-Expert		Flickr8k-CF		VATEX-EVAL		Pascal-50S	FOIL	ActivityNet-FOIL
		Kendall τ_b	Kendall τ_c	Kendall τ_b	Kendall τ_c	Kendall τ_b	Spearman ρ	Accuracy	Accuracy	Accuracy
CLIP ViT-B/16	CLIP-S	51.7	52.1	34.9	18.0	-	-	81.1	90.6	-
	EMScore	-	-	-	-	24 1	31.4	-	-	90.0
	PAC-S	**54.5** (+2.8)	**54.9** (+2.8)	**35.9** (+1.0)	**18.5** (+0.5)	**26.8** (+2.7)	**34.7** (+3.3)	**82.9** (+1.8)	**91.1** (+0.5)	**90.7** (+0.7)
CLIP ViT-L/14	CLIP-S	52.6	53.0	35.2	18.2	-	-	81.7	90.9	-
	EMScore	-	-	-	-	26.7	34.7	-	-	89.0
	PAC-S	**55.4** (+2.8)	**55.8** (+2.8)	**36.8** (+1.6)	**19.0** (+0.8)	**28.9** (+2.2)	**37.4** (+2.7)	**82.0** (+0.3)	**91.9** (+1.0)	**91.2** (+2.2)
OpenCLIP ViT-B/32	CLIP-S	52.3	52.6	35.4	18.3	-	-	81.2	88.9	-
	EMScore	-	-	-	-	24.8	32.2	-	-	88.2
	PAC-S	**53.6** (+1.3)	**53.9** (+1.3)	**36.1** (+0.7)	**18.6** (+0.3)	**25.4** (+0.6)	**33.1** (+0.9)	**82.4** (+1.2)	**90.1** (+1.2)	**89.5** (+1.3)
OpenCLIP ViT-L/14	CLIP-S	54.4	54.5	36.6	18.9	-	-	82.5	92.2	-
	EMScore	-	-	-	-	27.0	35.0	-	-	90.7
	PAC-S	**55.3** (+0.9)	**55.7** (+1.2)	**37.0** (+0.4)	**19.1** (+0.2)	**27.8** (+0.8)	**36.1** (+1.1)	**82.7** (+0.2)	**93.1** (+0.9)	**91.2** (+0.5)

S. Sarto, M. Barraco, M. Cornia, L. Baraldi, R. Cucchiara "Positive-Augmented Constrastive Learning for Image and Video Captioning Evaluation», CVPR 2023 Highlight

Qualitative Results

Image	Candidate Captions	CLIP-S	PAC-S
	*Two white dogs running.*	**0.530**	0.500
	A man riding a motorbike kicks up dirt.	0.486	**0.542**
	*Little girl in bare feet sitting in a circle.*	**0.524**	0.431
	A white dog runs in the grass.	0.426	**0.456**
	*Four woman wearing formal gowns pose together and smile.*	**0.700**	0.730
	A man in a wetsuit surfs.	0.613	**0.762**
	*Boy with a red crown in a shopping cart.*	**0.385**	0.467
	People stand outside near a concrete wall and a window.	0.359	**0.509**

Image	Candidate Captions	CLIP-S	PAC-S
	*A man and young girl eat a meal on a city street .*	**0.769**	0.764
	A small brown and white dog running through tall grass.	0.752	**0.820**
	*A man jumps while snow skiing.*	**0.512**	0.503
	A man is hiking on a snow-covered trail.	0.464	**0.567**
	*Two girls walking down the street.*	**0.583**	0.556
	A dog lies down on a cobblestone street.	0.550	**0.562**
	*A woman is signaling is to traffic , as seen from behind.*	**0.753**	0.767
	A man rides a bike through a course.	0.714	**0.800**

S. Sarto, M. Barraco, M. Cornia, L. Baraldi, R. Cucchiara "Positive-Augmented Constrastive Learning for Image and Video Captioning Evaluation», CVPR 2023 Highlight

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

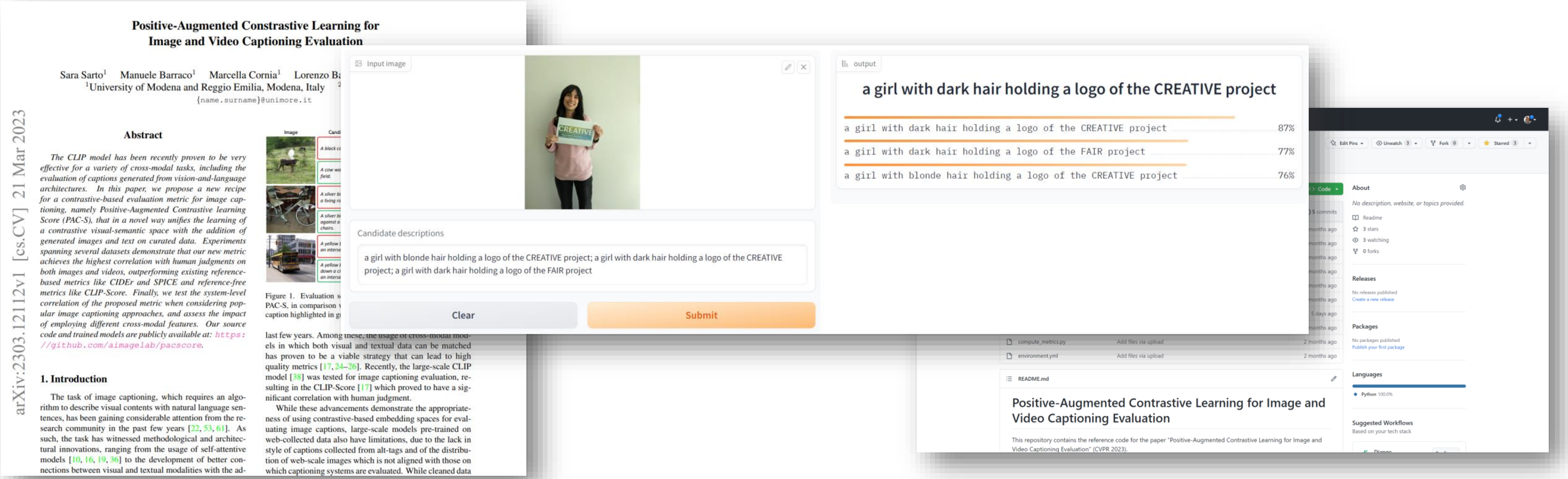
Qualitative Results

Image	Candidate Captions	METEOR	CIDEr	CLIP-S	PAC-S
	*A blue bird being held by a handler.*	35.2	96.3	80.1	80.0
	A blue bird perched on a gloved hand.	18.6	39.0	76.1	82.1
	*A black boxer dog with a white underbelly and brown collar looks at the camera.*	35.1	26.6	77.5	82.3
	A close up of a black pug.	11.6	21.1	71.0	83.5
	*Trains amble by the rail yard.*	26.2	68.8	81.9	75.4
	The red train and the yellow train on on the tracks.	14.7	28.3	79.8	81.6

Image	Candidate Captions	METEOR	CIDEr	CLIP-S	PAC-S
	*A passenger train in the snow.*	26.8	89.7	83.5	83.1
	A red train driving through a snow covered city.	27.2	72.6	81.4	85.7
	*A dog pokes it's head out from under a pile of stuff.*	25.8	60.5	67,5	75.6
	A dog underneath a wooden beam.	22.0	38.9	63.9	81.6
	*A large green coach with a bridge in the background*	28.3	32.0	87.1	76.7
	Green bus and tan truck on a city street with a man waiting to cross the street.	34.0	17.8	79.2	79.4

S. Sarto, M. Barraco, M. Cornia, L. Baraldi, R. Cucchiara "Positive-Augmented Constrastive Learning for Image and Video Captioning Evaluation», **CVPR 2023 Highlight**

Want to know more?



Read the paper

Use it in your projects ☺

https://arxiv.org/abs/2303.12112

https://github.com/aimagelab/pacscore

S. Sarto, M. Barraco, M. Cornia, L. Baraldi, R. Cucchiara "Positive-Augmented Constrastive Learning for Image and Video Captioning Evaluation», **CVPR 2023 Highlight**

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Thank you!

Manuele Barraco

Sara Sarto

Nicholas Moratelli

Marcella Cornia

Lorenzo Baraldi

Rita Cucchiara

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImageLab