



**UNIMORE**  
UNIVERSITÀ DEGLI STUDI DI  
MODENA E REGGIO EMILIA

AI Image<sup>Lab</sup>

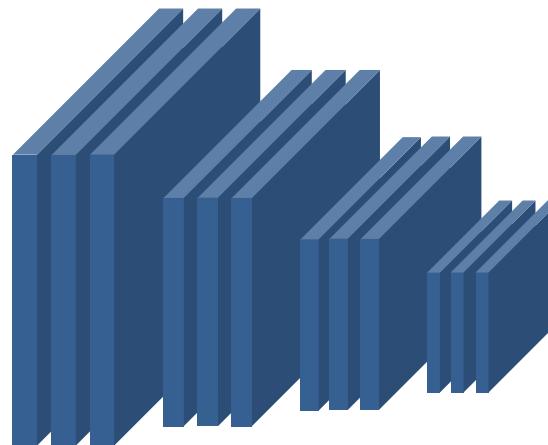
# From Images to Text: New forms of Human-AI Interaction



Lorenzo Baraldi

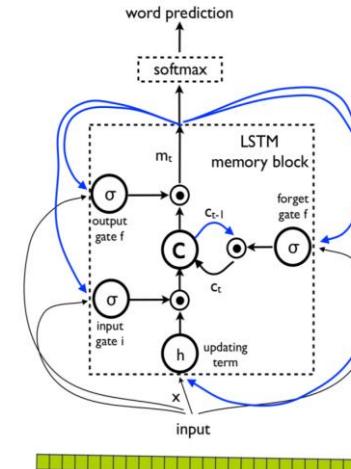
VISMAC

# Describing images in Natural Language



Visual feature extractor

+



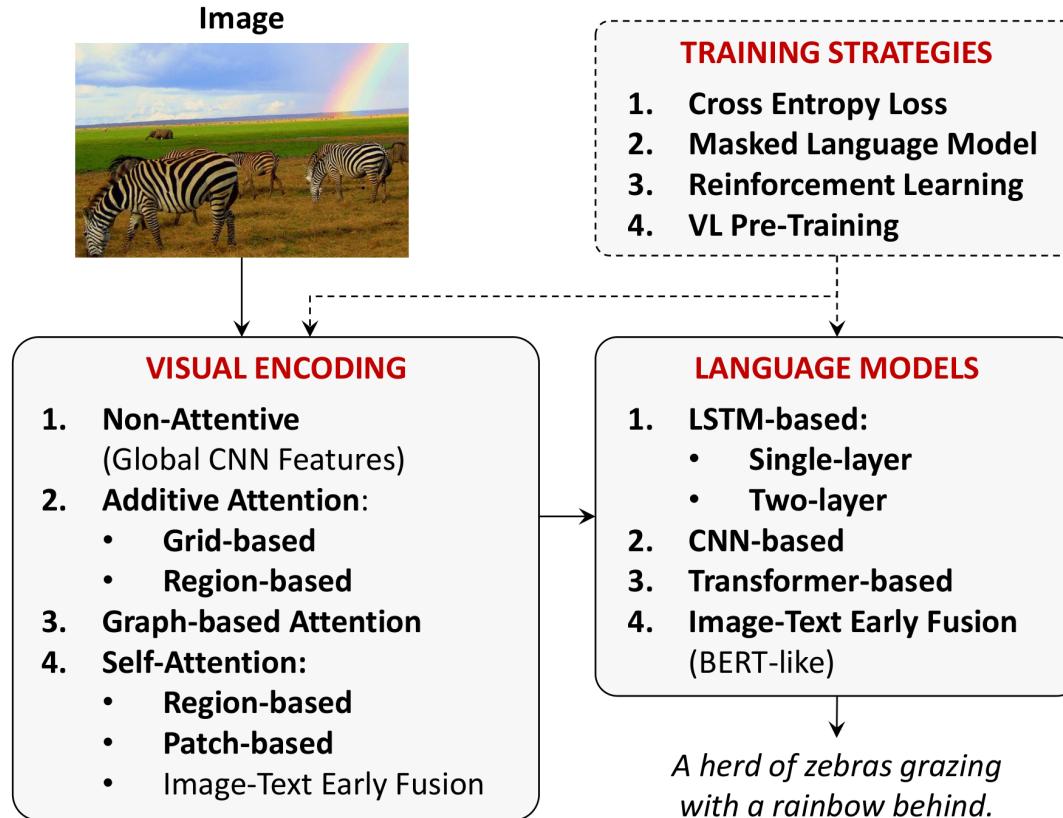
Language model

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

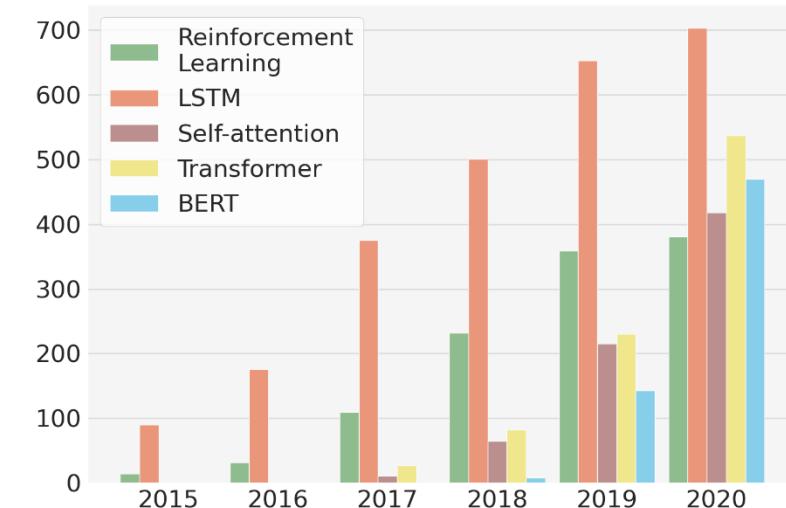
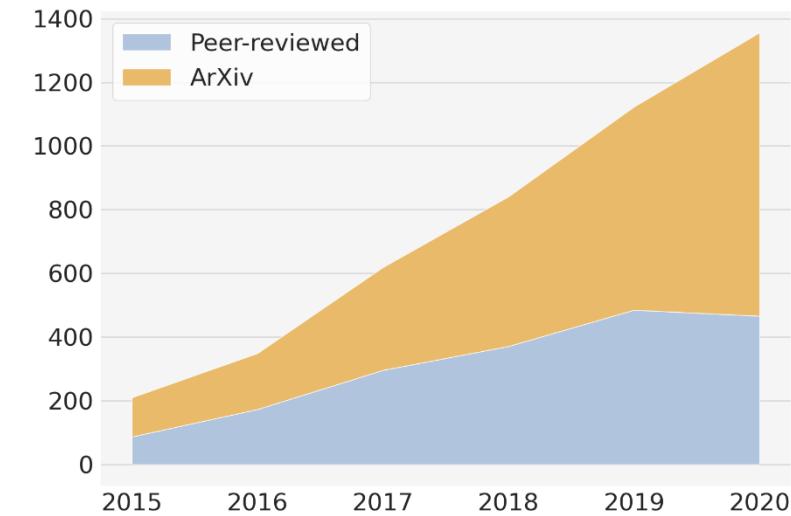
**Goal:** describe a visual input in natural language.

**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*.

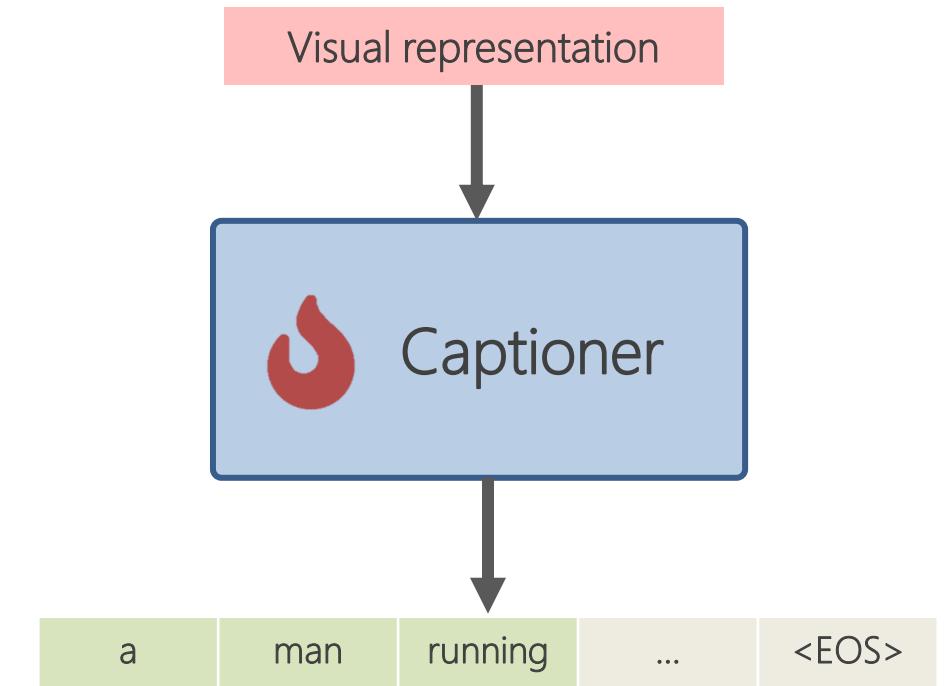


"Image Captioning" and related keywords in the text of recent papers:



# Anatomy of a Captioning System

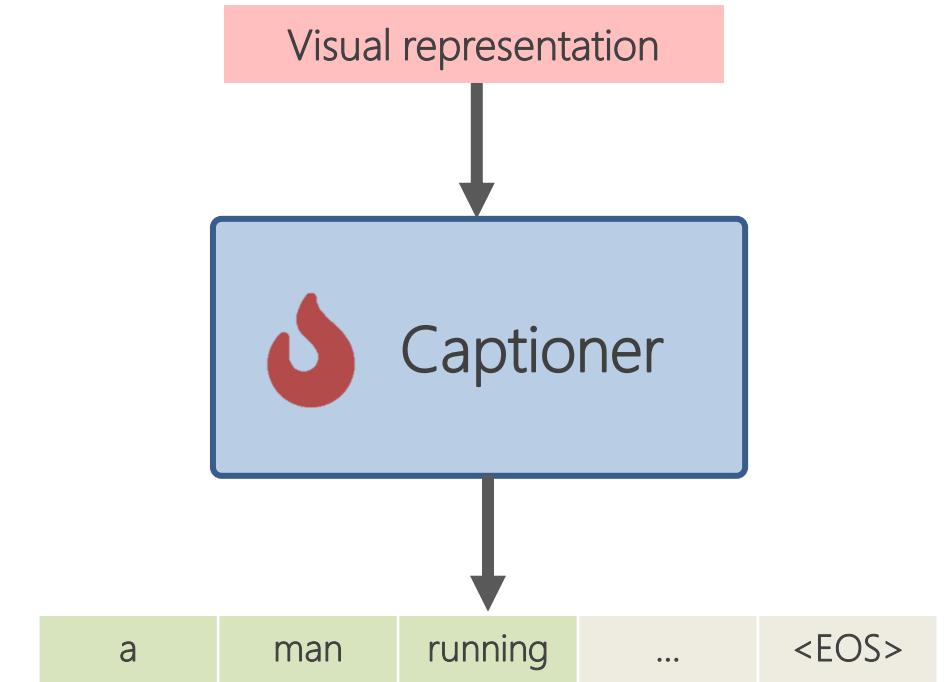
Base technical idea:  
combine visual feature extractors with language models.



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

*Many possibilities*

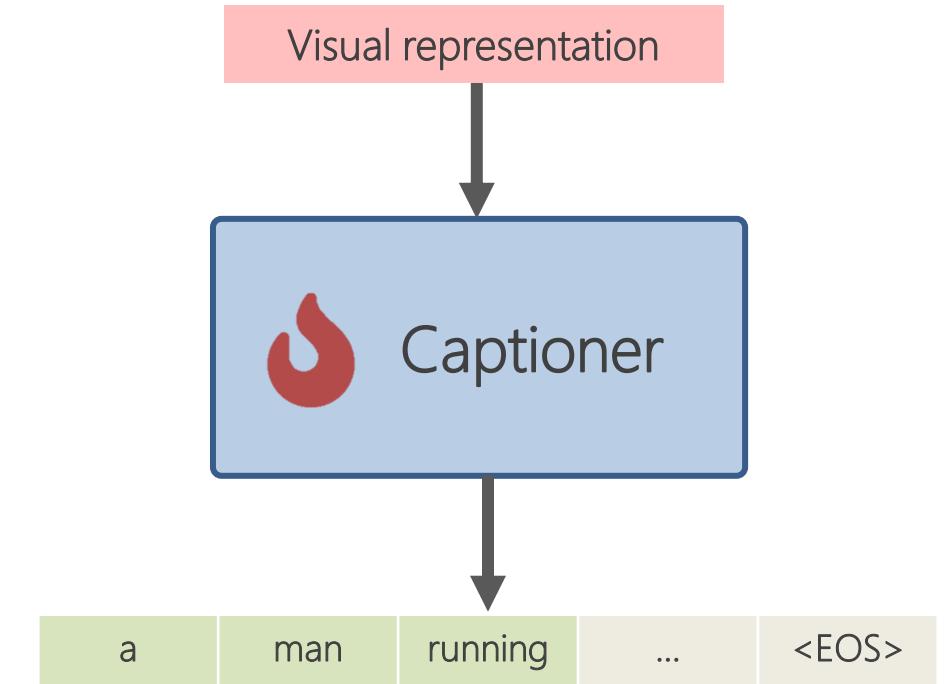
- Language model:
  - RNN and variants (LSTM)
  - 1-d CNN
  - Transformer-based (recently)



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

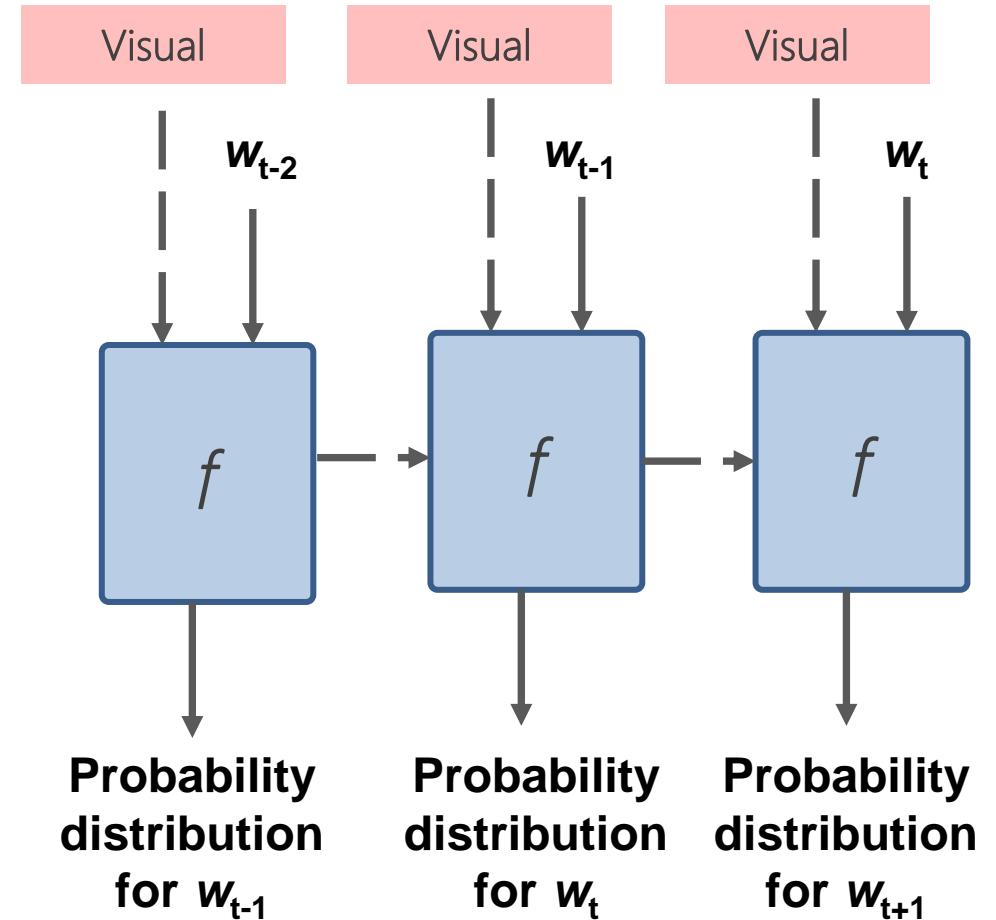
*Many possibilities*

- 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)



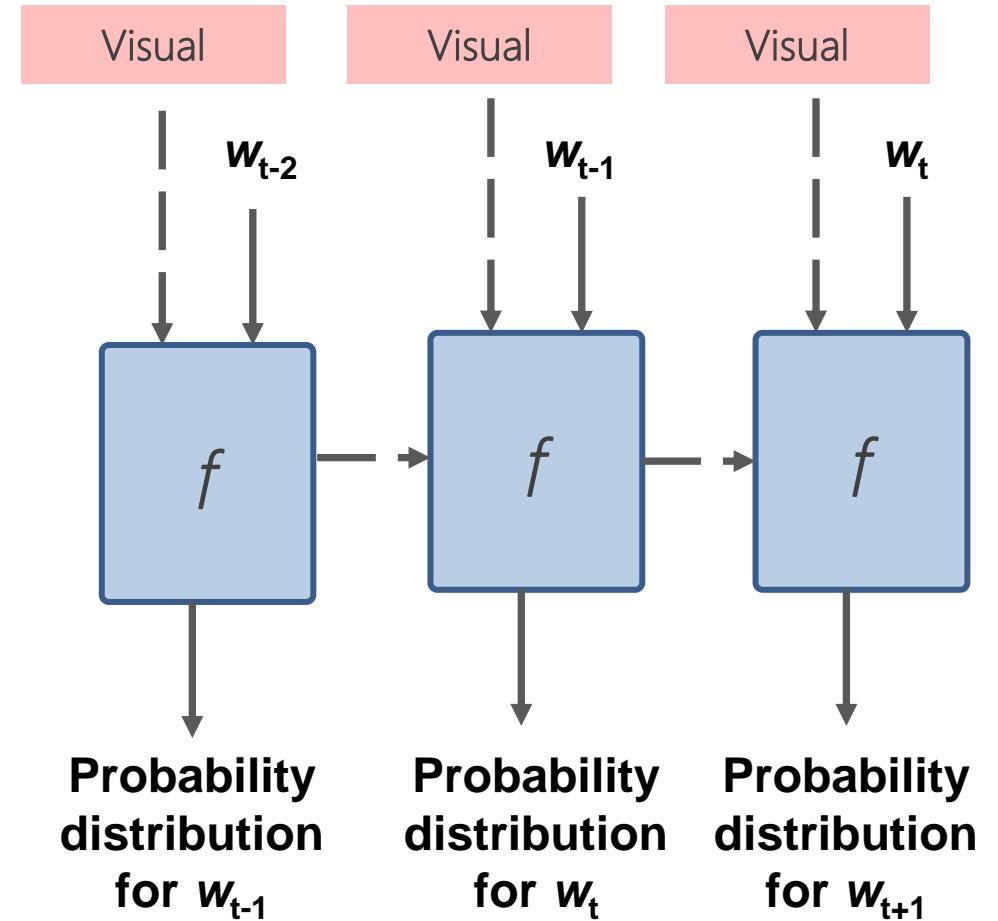
## Language model

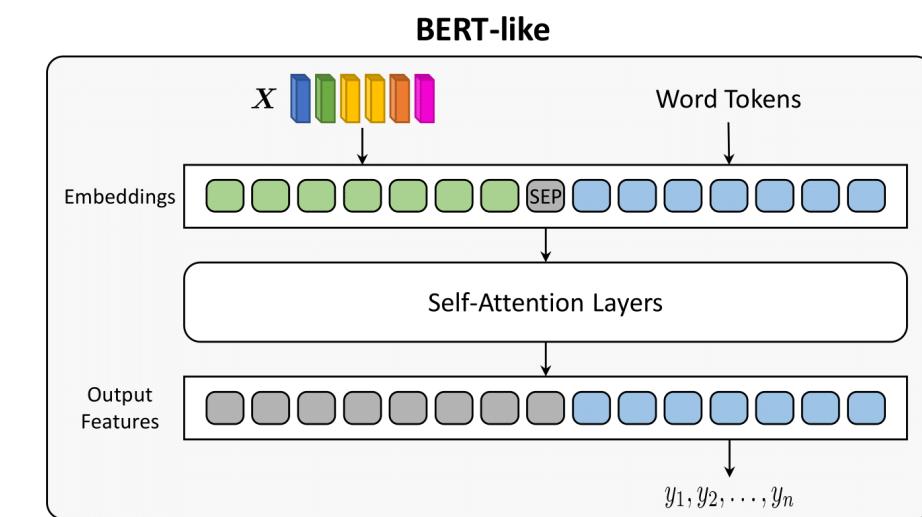
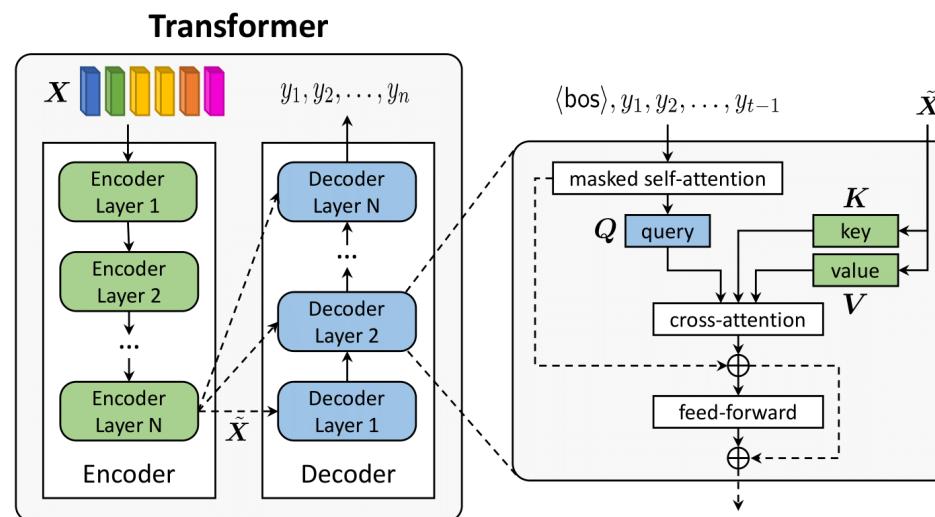
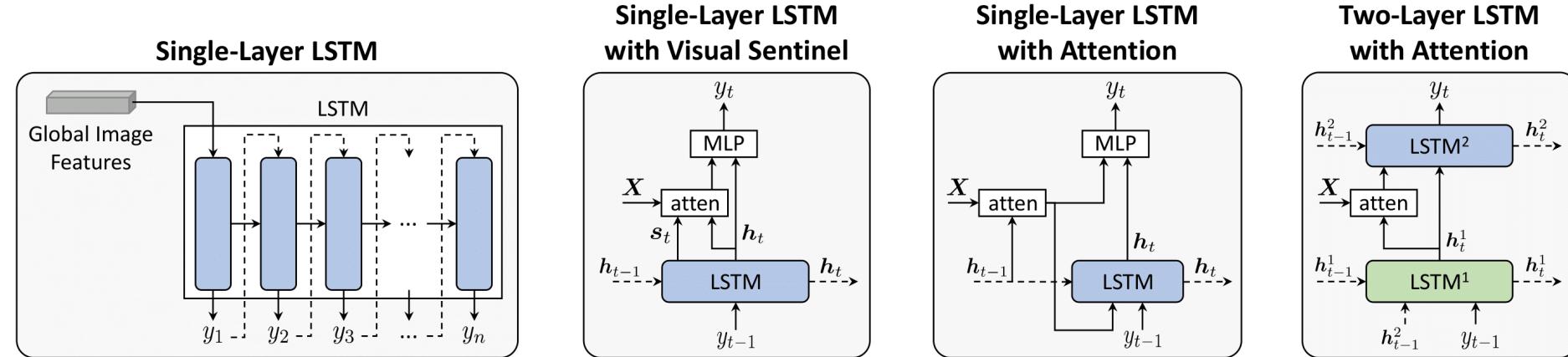
- Prediction process is always **sequential**, i.e. we model the probability of outputting a word given previous words in the sentence.
- The probability distribution for  $w_t$  is conditioned on  $w_{t-1}, w_{t-2}, \dots, w_0$



## Language model

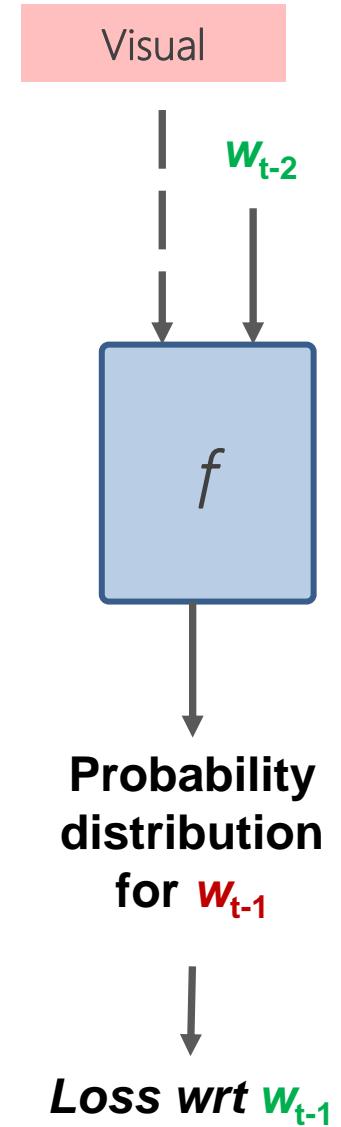
- Prediction process is always **sequential**, i.e. we model the probability of outputting a word given previous words in the sentence.
- The probability distribution for  $w_t$  is conditioned on  $w_{t-1}, w_{t-2}, \dots, w_0$
- A function  $f$  models the computational graph for predicting the word at each step (the “**step function**”).
  - *Any of {RNN, CNN, Transformer, ...}*





### At training time

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

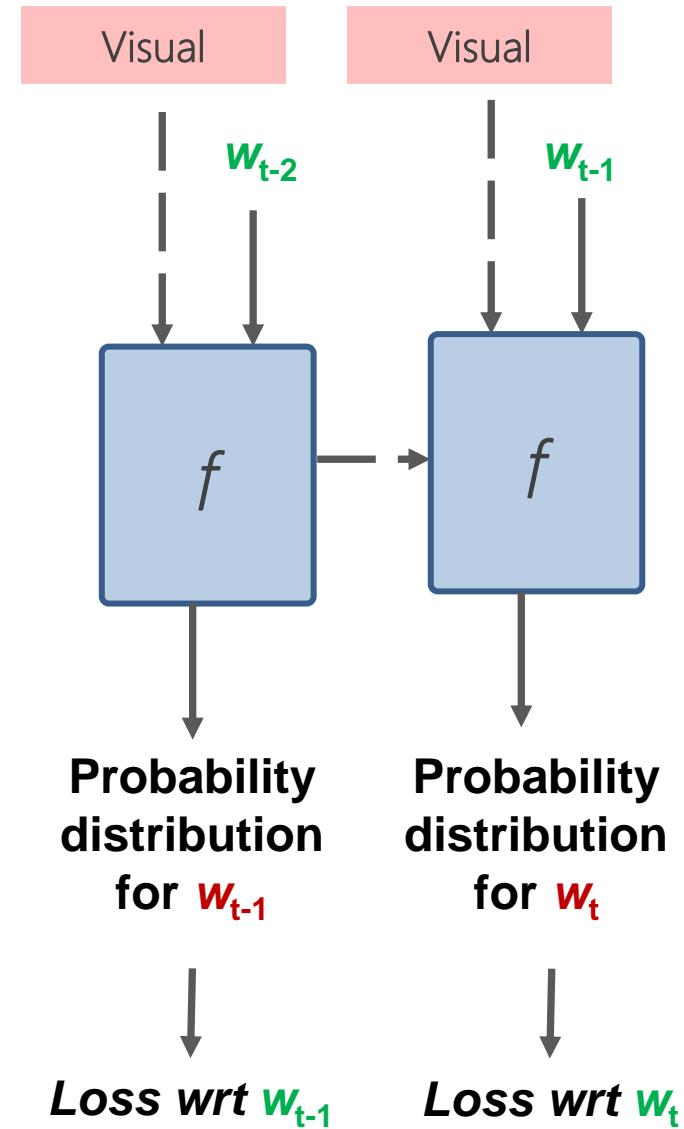


$w_t$ :ground-truth words

# Anatomy of a Captioning System

## At training time

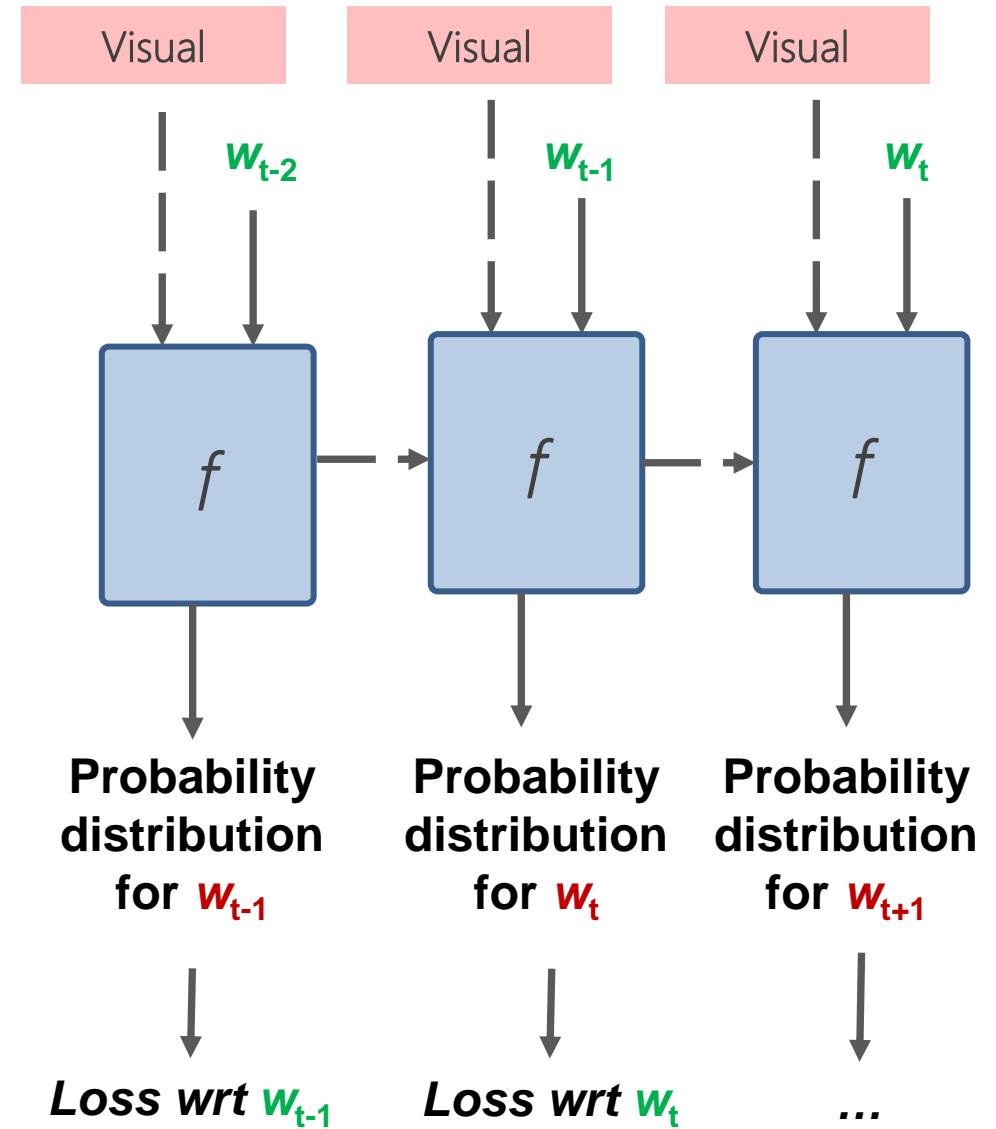
- Train the model to predict a word given the previous ground-truth words.



$w_t$ :ground-truth words

### At training time

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



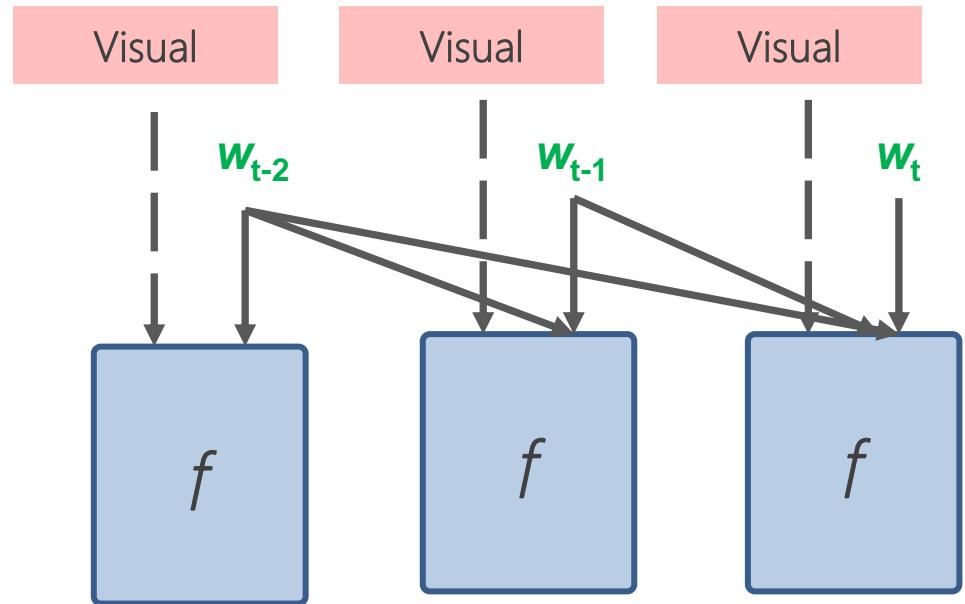
$w_t$ :ground-truth words

### 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

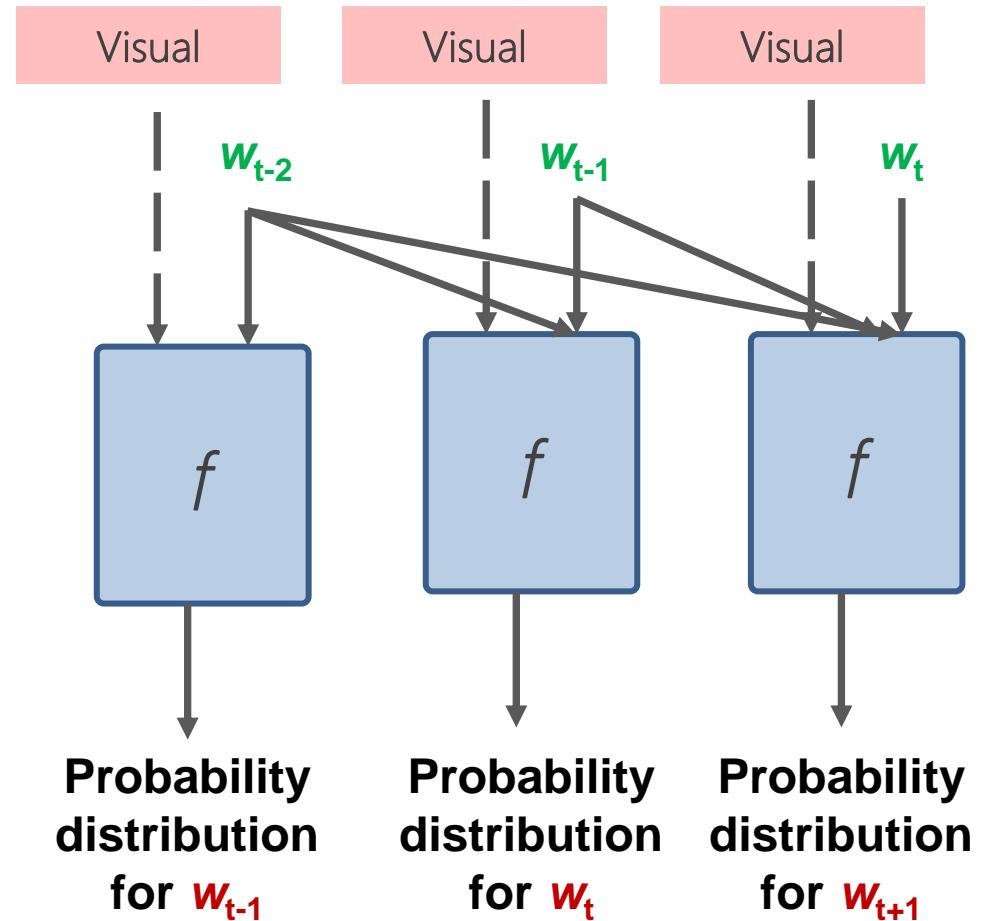
### At training time

- 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



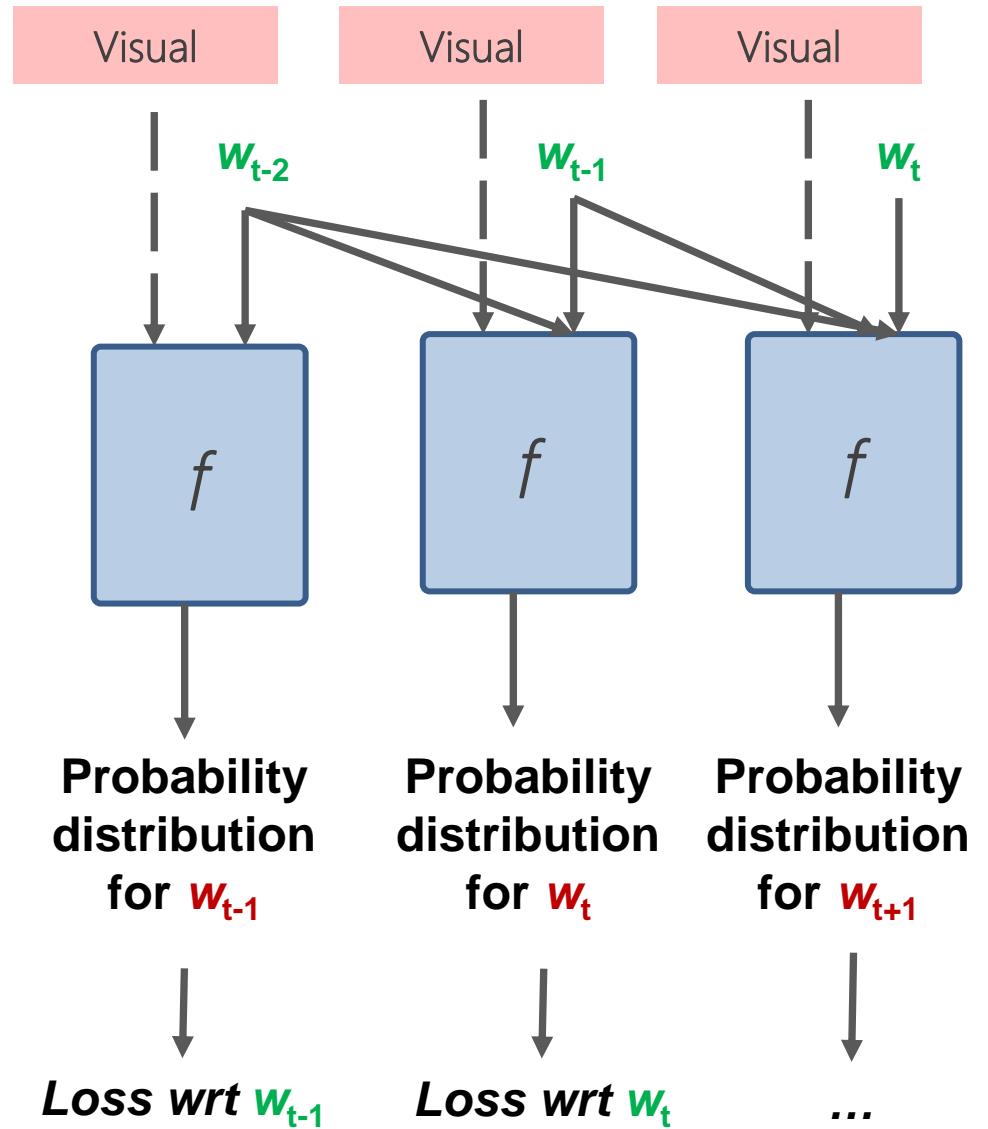
### At training time

- 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:

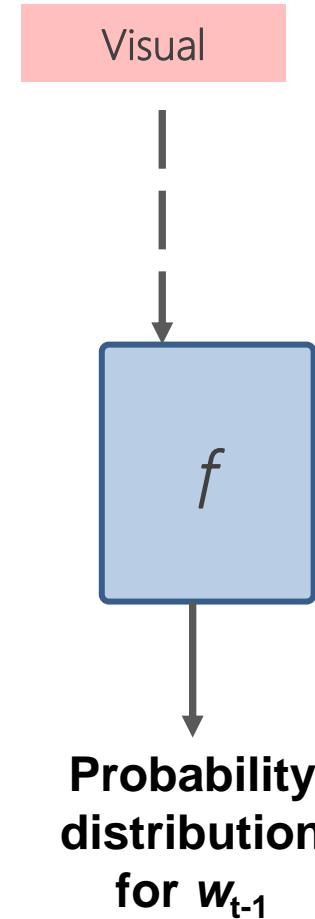
- We can parallelize over the  $t$  axis.
  - → Training time reduction
  - E.g. Conv1D, Transformer

$w_t$ :ground-truth words



### At prediction time (sampling)

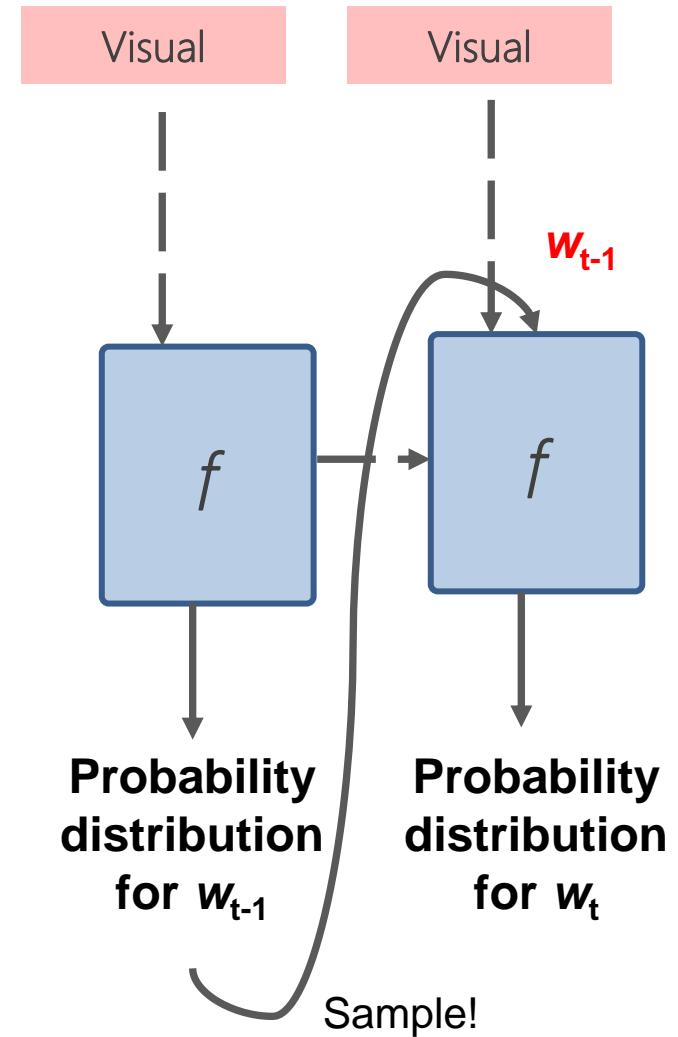
- We sample one word from the previous output, and use this as an input.



$w_t$ : sampled words

### At prediction time (sampling)

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

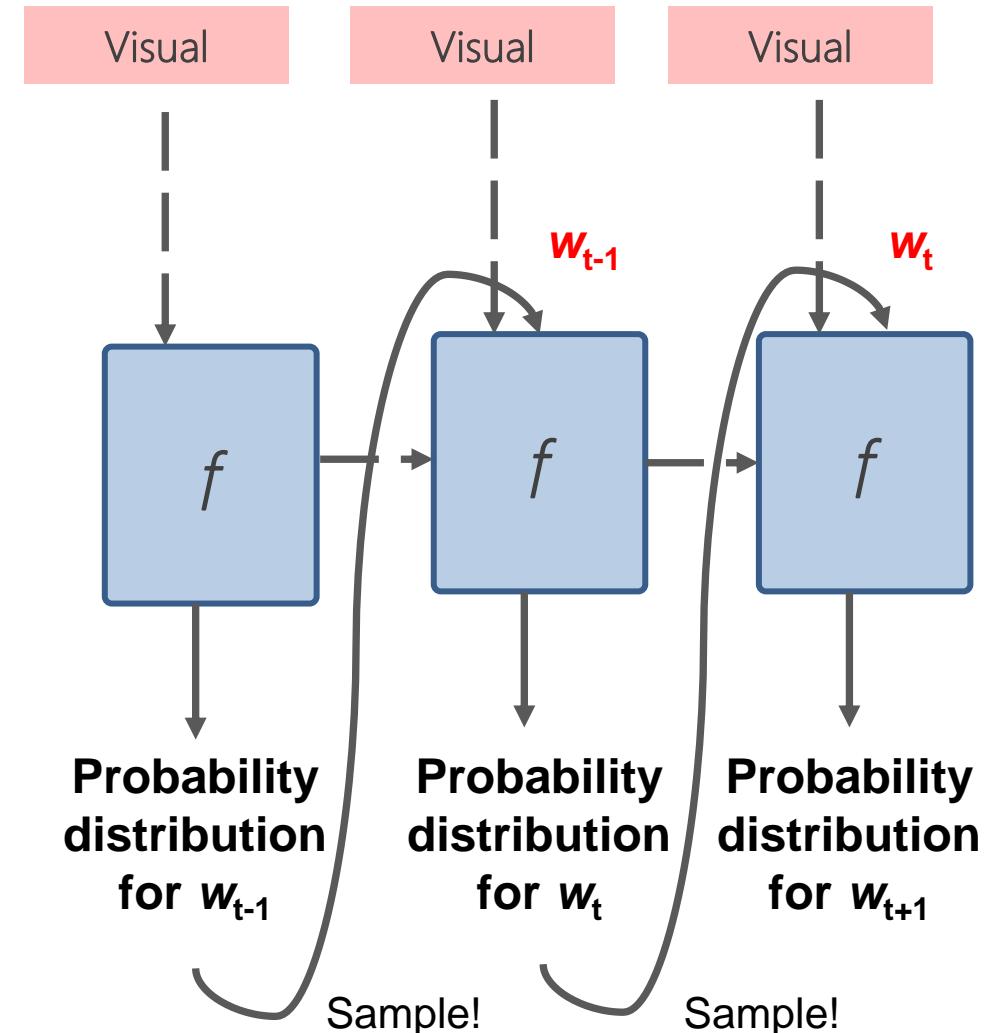


$w_t$ : sampled words

### At prediction time (sampling)

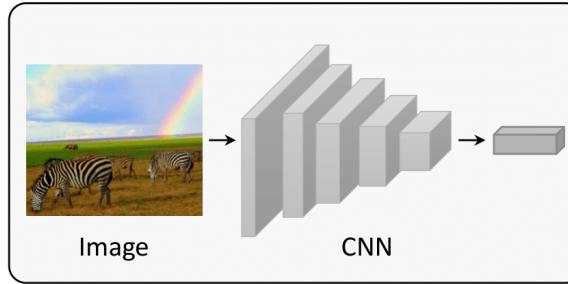
- We sample one word from the previous output, and use this as an input.
- Possible strategies:
  - Always sample the most probable word
  - Build a tree of possible choices, then select the chain of predictions with maximum probability (*beam search*)

$w_t$ : sampled words

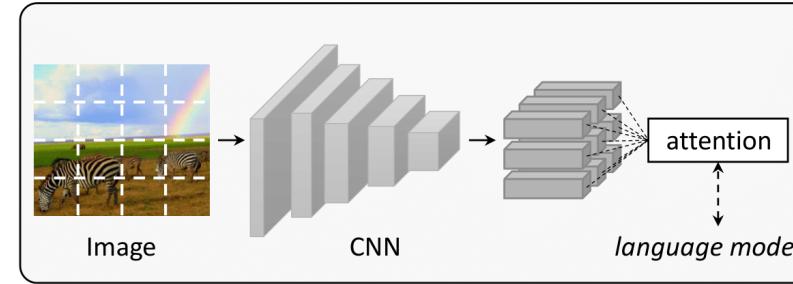


# What About Image Encoding

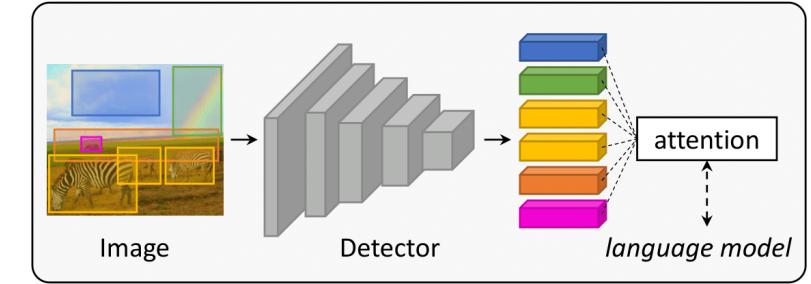
**Global CNN Features**



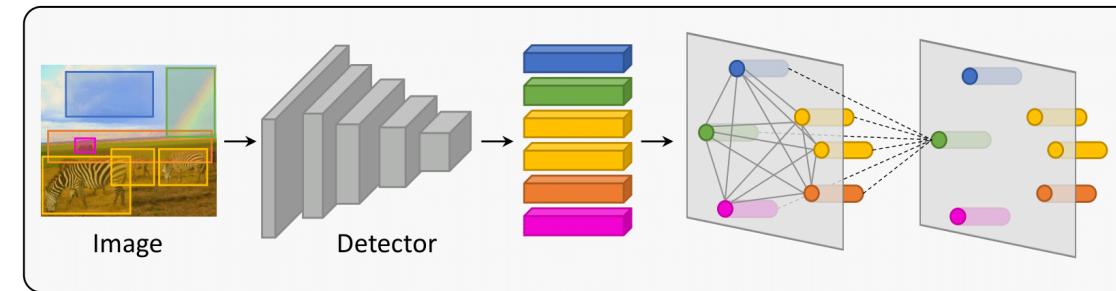
**Attention Over Grid of CNN Features**



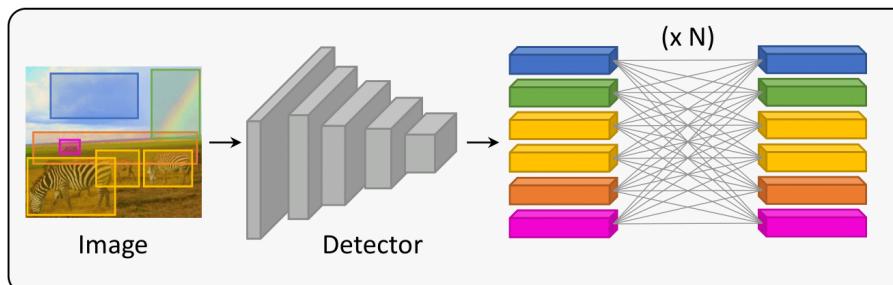
**Attention Over Visual Regions**



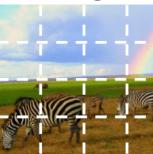
**Graph-based Encoding**



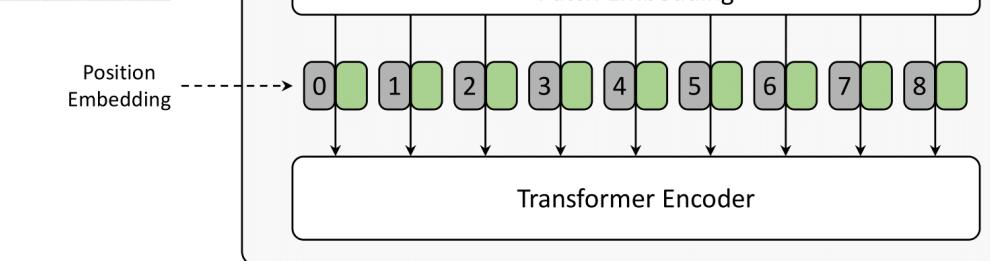
**Self-Attention Encoding**



Image



Position  
Embedding

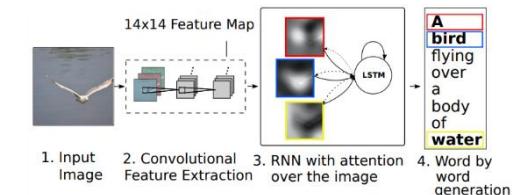


# Describing Images in Natural Language

## • RNN-based Captioning

Vinyals, *et al.* Show and Tell - *CVPR*, 2015

Karpathy, *et al.* Deep Visual-Semantic Alignments - *CVPR*, 2015



attention on spatial features

Xu, *et al.* Show, Attend and Tell - *ICML*, 2015

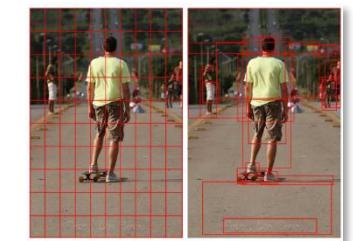
- Attentive models

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



cross-entropy loss

- Training strategies

reinforcement learning

Rennie, *et al.* Self-Critical Sequence Training - *CVPR*, 2017

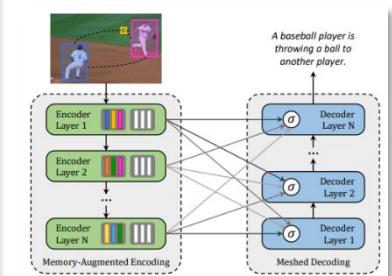
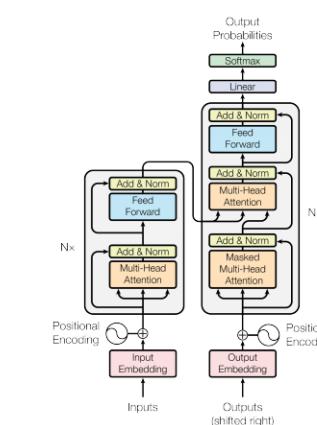
## • Transformer-based Captioning

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<sup>2</sup> Transformer - *CVPR*, 2020



• ...

- **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

A photograph of a large billboard in Yakima, Washington. The billboard features a red rose at the top right and the text "Welcome to YAKIMA The Palm Springs of Washington" in a serif font. The word "YAKIMA" is in large capital letters, and "The Palm Springs of Washington" is in smaller capital letters below it.

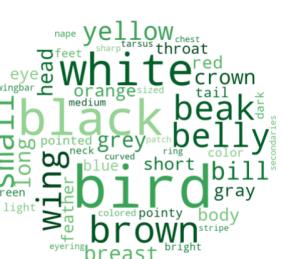
The image is a composite of two photographs. On the left, a person wearing a black helmet and riding a brown horse with a white blaze on its forehead is jumping over a blue and white striped pole. The horse is in mid-air, and the rider is leaning forward. On the right, a clear glass bowl is filled with sliced strawberries and peeled tangerine segments.

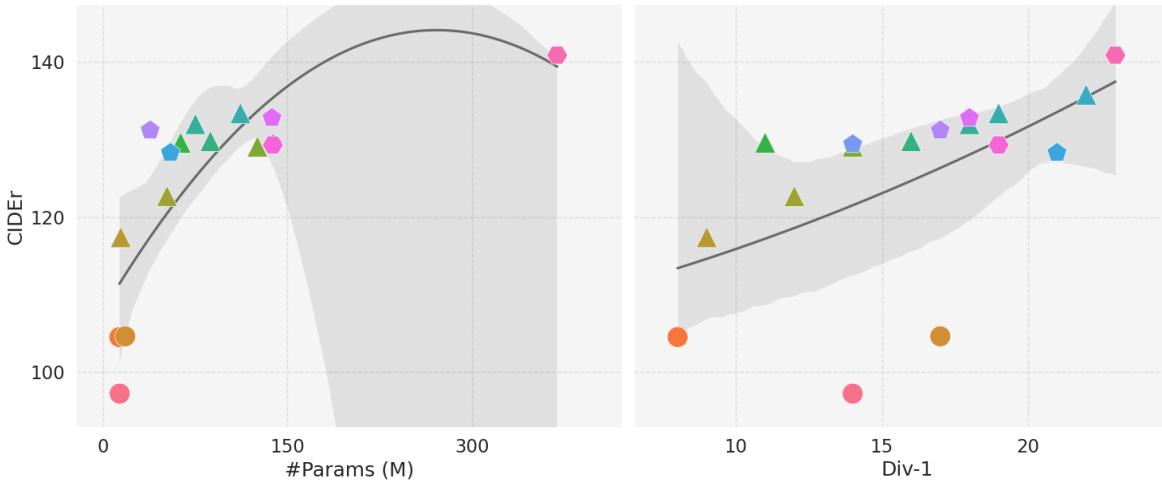
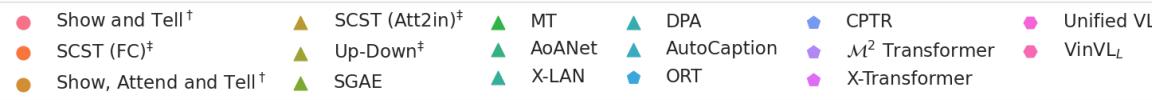
**Fashion Captioning**

en  
d  
glass  
book  
page phone  
title  
sit  
read  
display  
be  
sign  
white  
designer  
called  
Wearring  
setting

A photograph of a blue jay standing on a dark wooden surface. The bird has a bright blue back and wings, a white belly, and a distinctive white patch on its wing. It is facing right, showing its profile.

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





		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	✓	(✓)

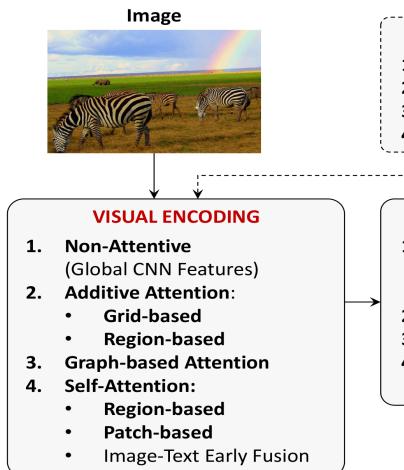
But looking at the captions is key...

# From Show to Tell: A survey on Image Captioning

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

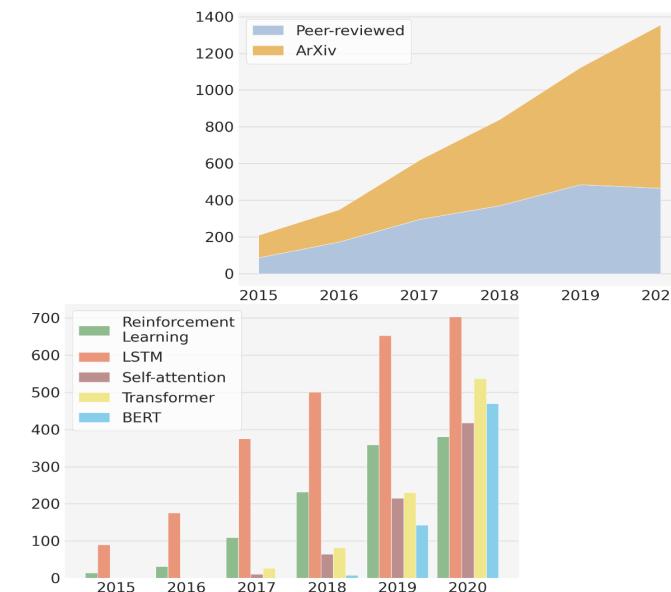
TPAMI 2023 - <https://arxiv.org/pdf/2107.06912.pdf>

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



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

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



**From Show to Tell:  
A Survey on Image Captioning**

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli,  
Giuseppe Fiameni, and Rita Cucchiara

**Abstract**—Connecting Vision and Language plays an essential role in Generative Intelligence. For this reason, in the last few years, a large research effort has been devoted to image captioning, i.e. the task of describing images with syntactically and semantically meaningful sentences. Starting from 2015 the task has generally been addressed with pipelines composed of a visual encoding step and a language model for text generation. During these years, both components have evolved considerably through the exploitation of object regions, attributes, and relationships and the introduction of multi-modal connections, fully-attentive approaches, and BERT-like early-fusion strategies. However, regardless of the impressive results obtained, research in image captioning has not reached a conclusive answer yet. This work aims at providing a comprehensive overview and categorization of image captioning approaches, from visual encoding and text generation to training strategies, used datasets, and evaluation metrics. In this respect, we quantitatively compare many relevant state-of-the-art approaches to identify the most impactful technical innovations in image captioning architectures and training strategies. Moreover, many variants of the problem and its open challenges are analyzed and discussed. The final goal of this work is to serve as a tool for understanding the existing state-of-the-art and highlighting the future directions for an area of research where Computer Vision and Natural Language Processing can find an optimal synergy.

**Index Terms**—Image Captioning, Vision-and-Language, Deep Learning, Survey.

---

**1 INTRODUCTION**

IMAGE captioning is the task of describing the visual content of an image in natural language, employing a visual understanding system and a language model capable of generating meaningful and syntactically correct sentences. Neuroscience research has clarified the link between human vision and language generation only in the last few years [1]. Similarly, in Artificial Intelligence, the design of architectures capable of processing images and generating language is a very recent matter. The goal of these research efforts is to find the most effective pipeline to process an input image, represent its content, and transform that into a sequence of words by generating connections between visual and textual elements while maintaining the fluency of language. In its standard configuration, image captioning is an image-to-sequence problem whose inputs are pixels. These are encoded as one or multiple feature vectors in the visual encoding step, which prepares the input for a second generative step, called the language model. This produces a sequence of words or sub-words decoded according to a given vocabulary. In these few years, the research community improved the models considerably: from the first deep learning-based proposals adopting Recurrent Neural Networks (RNNs) fed with global image descriptor extracted by a Convolutional Neural Network (CNN), methods have been enriched with attentive approaches and reinforcement learning, up to the breakthroughs of Transformers and self-attention to single-stream BERT-like approaches. At the same time, the Computer Vision and Natural Language Processing (NLP) communities have addressed the challenge of building proper evaluation protocols and evaluation metrics to compare results with human-generated ground-truths. Moreover, several domain-specific scenarios and variants of the task have been investigated. However, the achieved results are still far from setting an ultimate solution. With the aim of providing a testament to the journey that captioning has taken so far, and with that of encouraging new ideas, in this paper, we trace a holistic overview of the models developed in the last years.

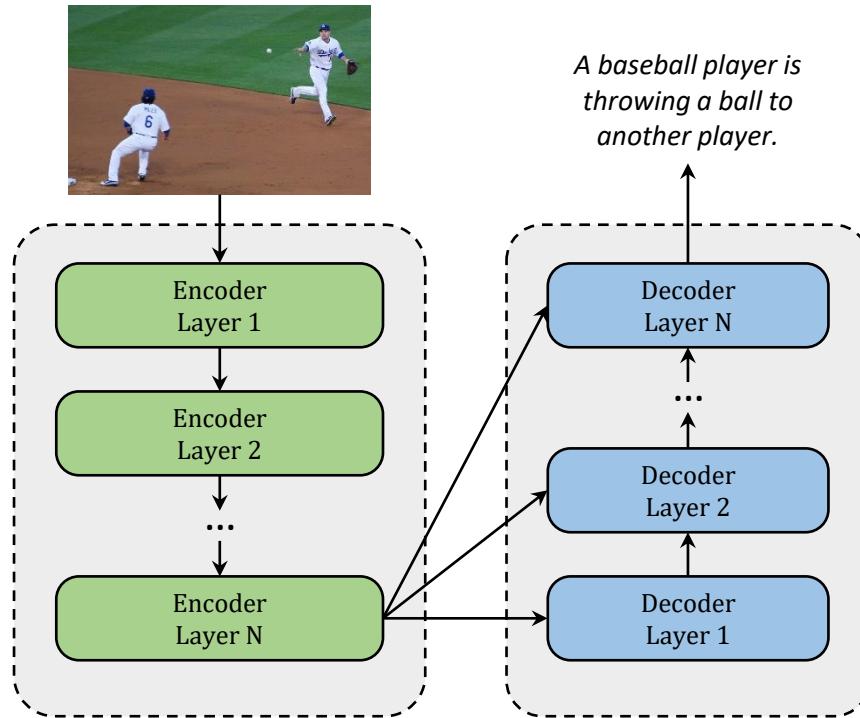
Following the inherent dual nature of captioning models, we develop a taxonomy of both the visual encoding and the language modeling approaches, focusing on their key aspects and limitations. We also focus on the training strategies followed in the literature over the past years, from cross-entropy loss to reinforcement learning and the recent advancement obtained by the pre-training paradigm. Furthermore, we review the main datasets used to explore image captioning from domain-generic benchmarks to domain-specific datasets collected to investigate specific aspects of the problem and analyze standard and non-standard metrics adopted for performance evaluation, which capture different aspects of the produced caption quality. An additional contribution of this work is a quantitative comparison of main image captioning methods, which considers both standard and non-standard metrics, and a discussion on their relationships, which sheds light on performance, differences, and characteristics of the most important models. Finally, we give an overview of many variants of the problem and discuss some open challenges and future directions.

• M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, and R. Cucchiara are with the Department of Engineering “Enzo Ferrari”, University of Modena and Reggio Emilia, Modena, Italy.  
 E-mail: {matteo.stefanini, marcella.cornia, lorenzo.baraldi, silvia.cascianelli, rita.cucchiara}@unimore.it.  
 • G. Fiameni is with NVIDIA AI Technology Center, Italy.  
 E-mail: gfiameni@nvidia.com

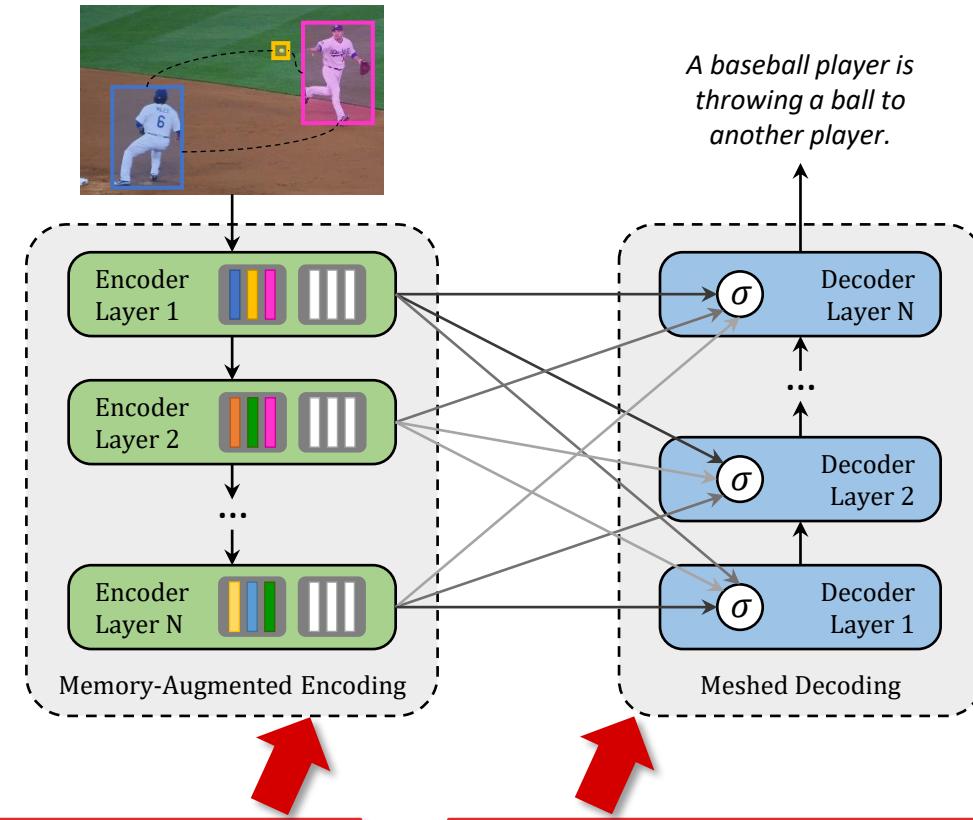
# Meshed-Memory Transformer

# Meshed-Memory Transformer

## Original Transformer



## M<sup>2</sup> Transformer



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

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

- 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.

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

### 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.

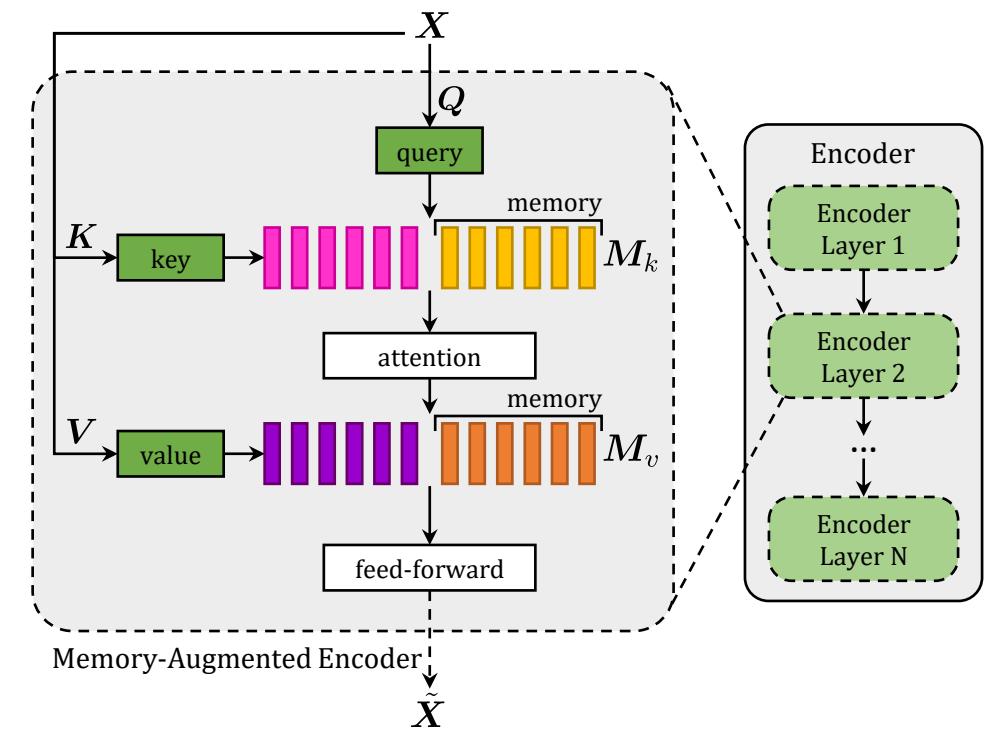
### Cross-Attention (decoder only)

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

- 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:

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

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



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

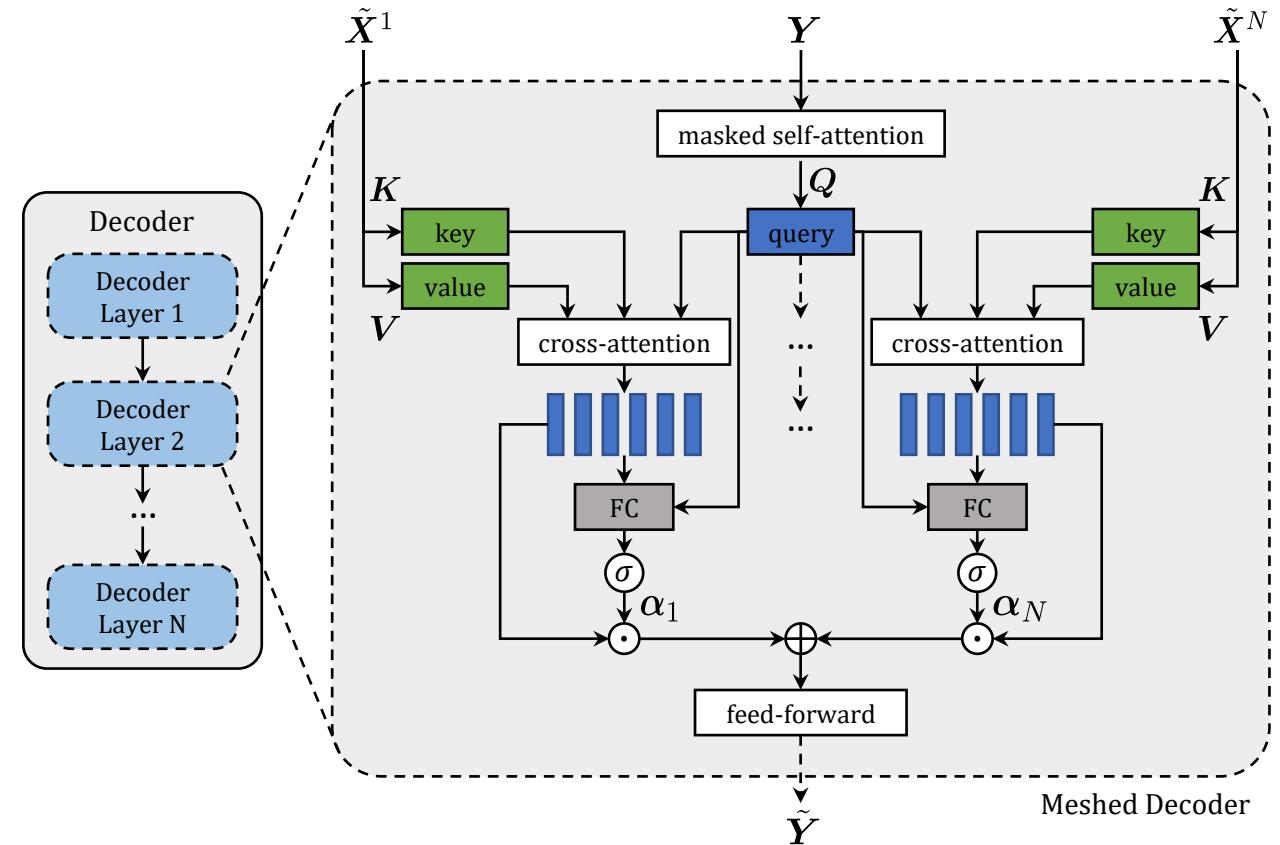
$$\mathcal{M}_{\text{mesh}}(\tilde{\mathcal{X}}, \mathbf{Y}) = \sum_{i=1}^N \alpha_i \odot \mathcal{C}(\tilde{\mathcal{X}}^i, \mathbf{Y})$$

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

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

- Weights in  $\alpha_i$  modulate the contribution of each encoding layer and the relative importance between different layers.

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



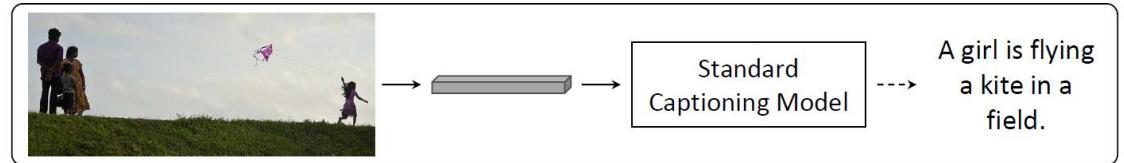
	CVPR 2017	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE		CIDEr	
		c5	c40	c5	c40										
CVPR 2017	SCST [33]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
CVPR 2018	Up-Down [4]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
ICCV 2019	RDN [18]	80.2	95.3	-	-	-	-	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]	<b>81.6</b>	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
<b><math>M^2</math> Transformer</b>		<b>81.6</b>	<b>96.0</b>	<b>66.4</b>	<b>90.8</b>	<b>51.8</b>	<b>82.7</b>	<b>39.7</b>	<b>72.8</b>	<b>29.4</b>	<b>39.0</b>	<b>59.2</b>	<b>74.8</b>	<b>129.3</b>	<b>132.1</b>

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

# Controllable Captioning

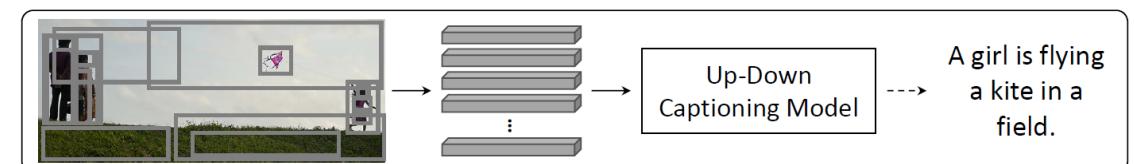
## Early captioning approaches:

- Global image feature vector



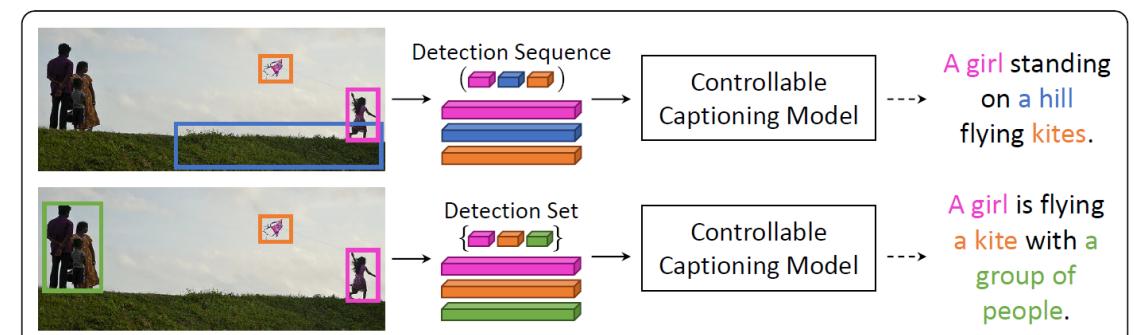
## Attention-based approaches:

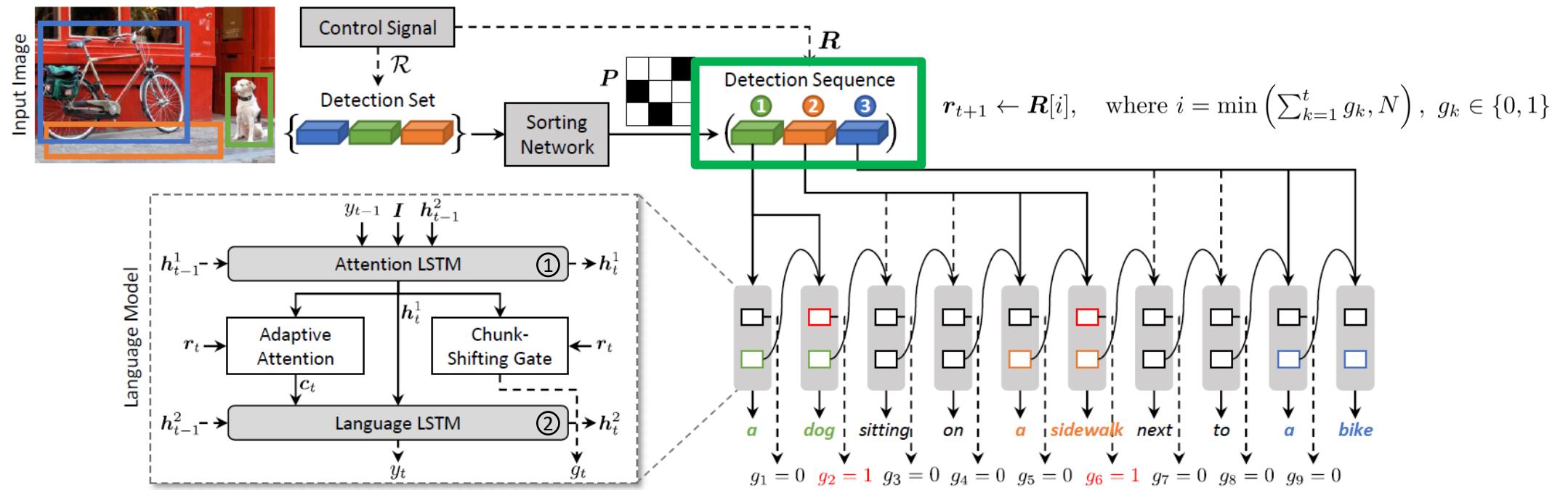
- Weakly interpretable (through attention)
- Not controllable.
  - We can't decide which regions get processed
  - No control over the generation process.



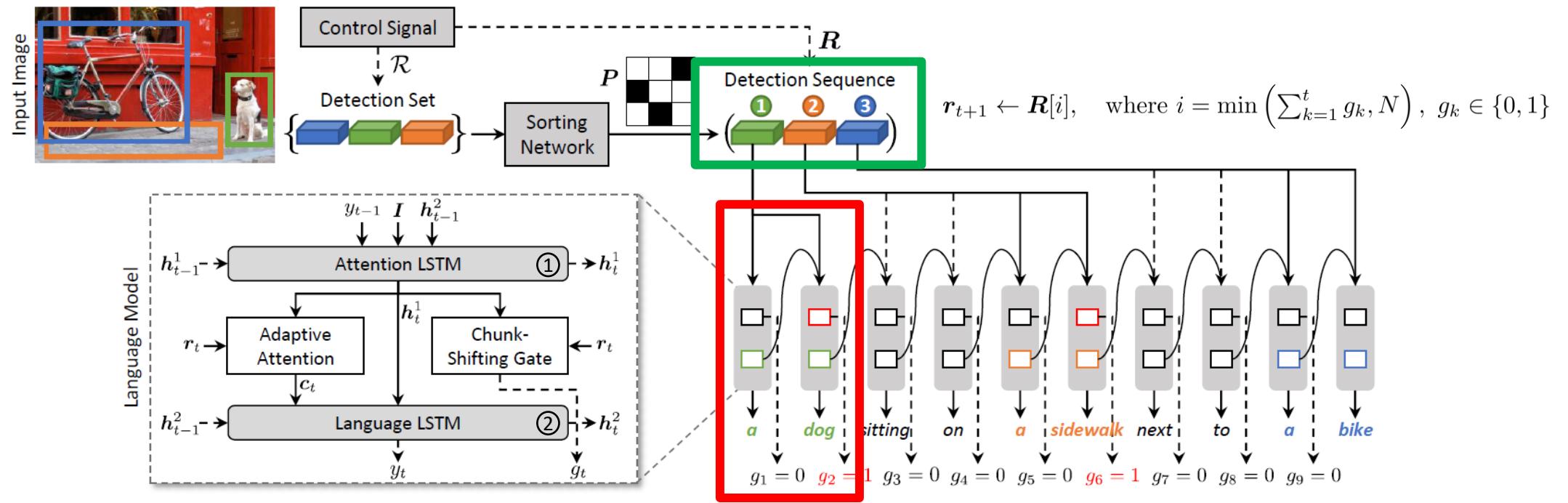
## Show, control and tell

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

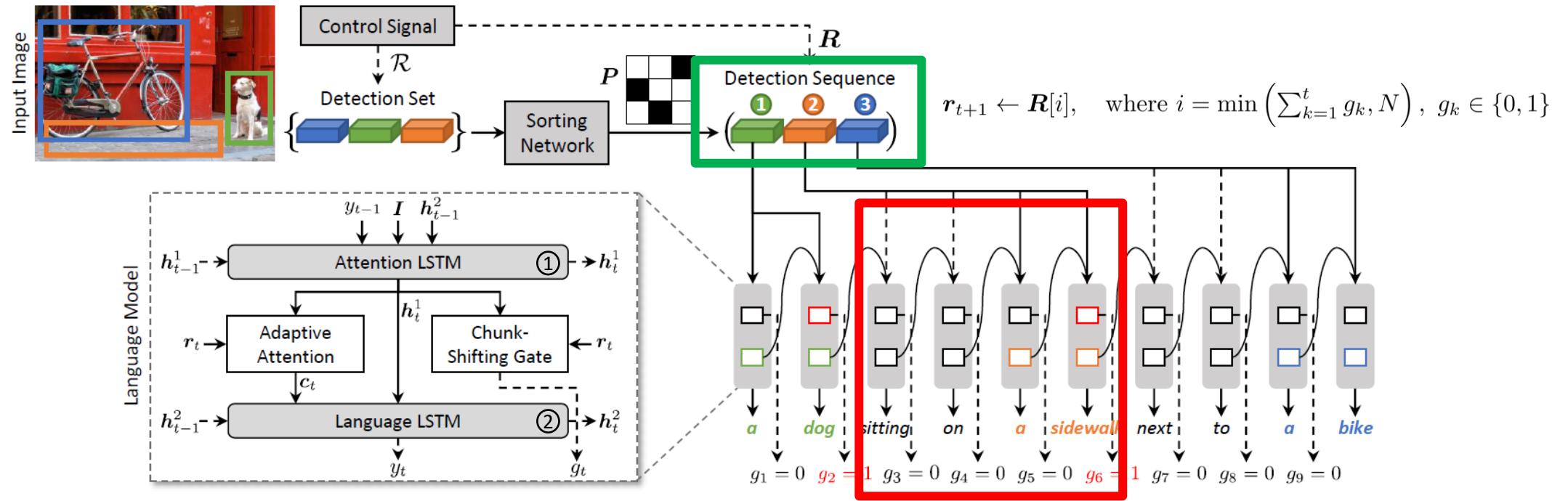




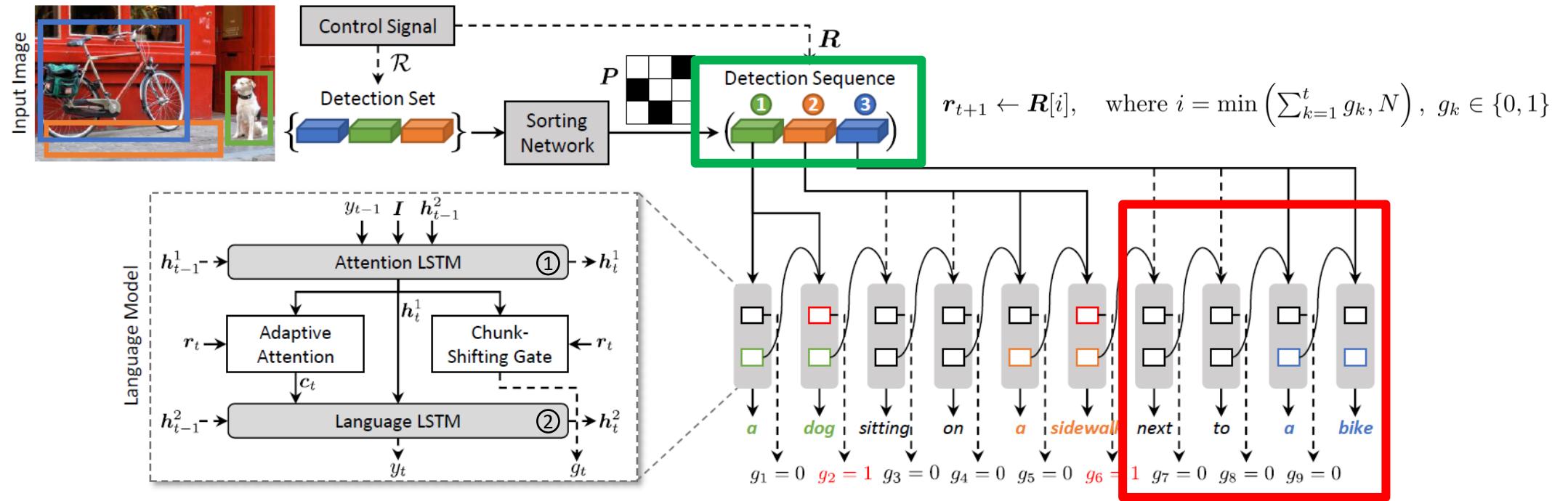
- Language model takes as input **a sequence of regions**
- Switches between one region and the next one via a learned chunk-shifting gate
  - When it's done with the generation of chunk, it moves to the next region in the sequence



- 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

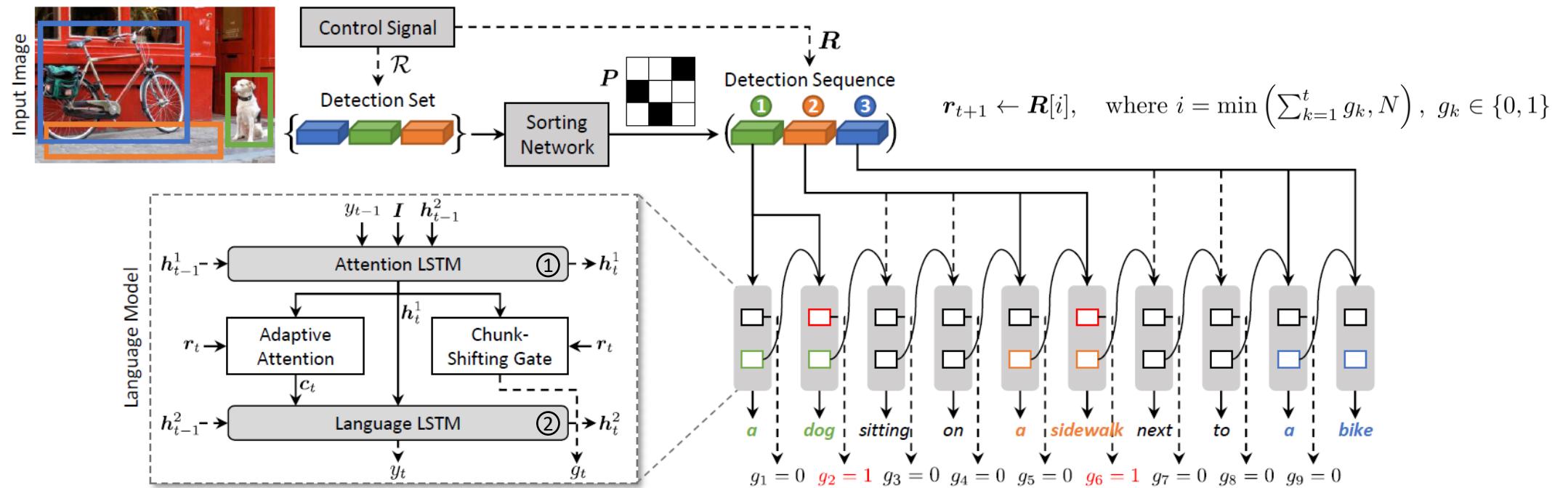


- Language model takes as input a sequence of regions
- **Switches between one region and the next one via a learned chunk-shifting gate**
  - When it's done with the generation of chunk, it moves to the next region in the sequence



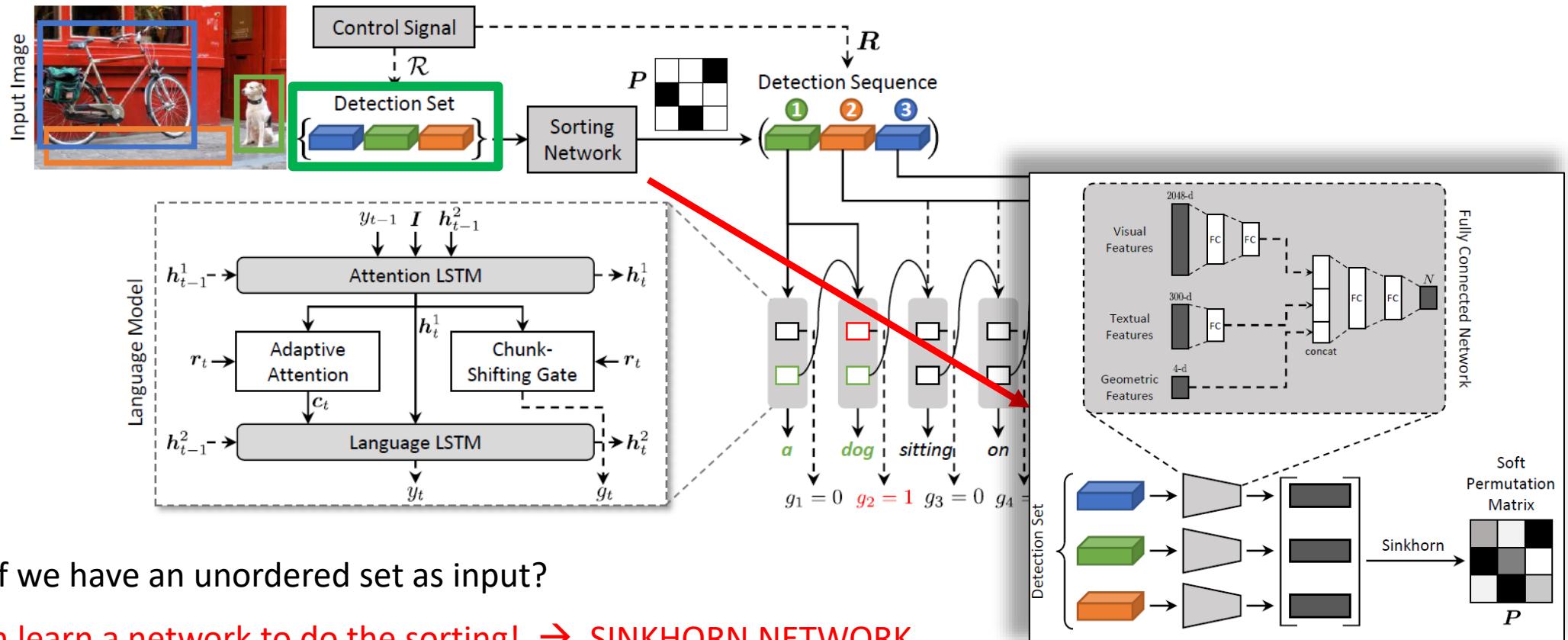
- Language model takes as input a sequence of regions
- **Switches between one region and the next one via a learned chunk-shifting gate**
  - When it's done with the generation of chunk, it moves to the next region in the sequence

# Generating Controllable Captions



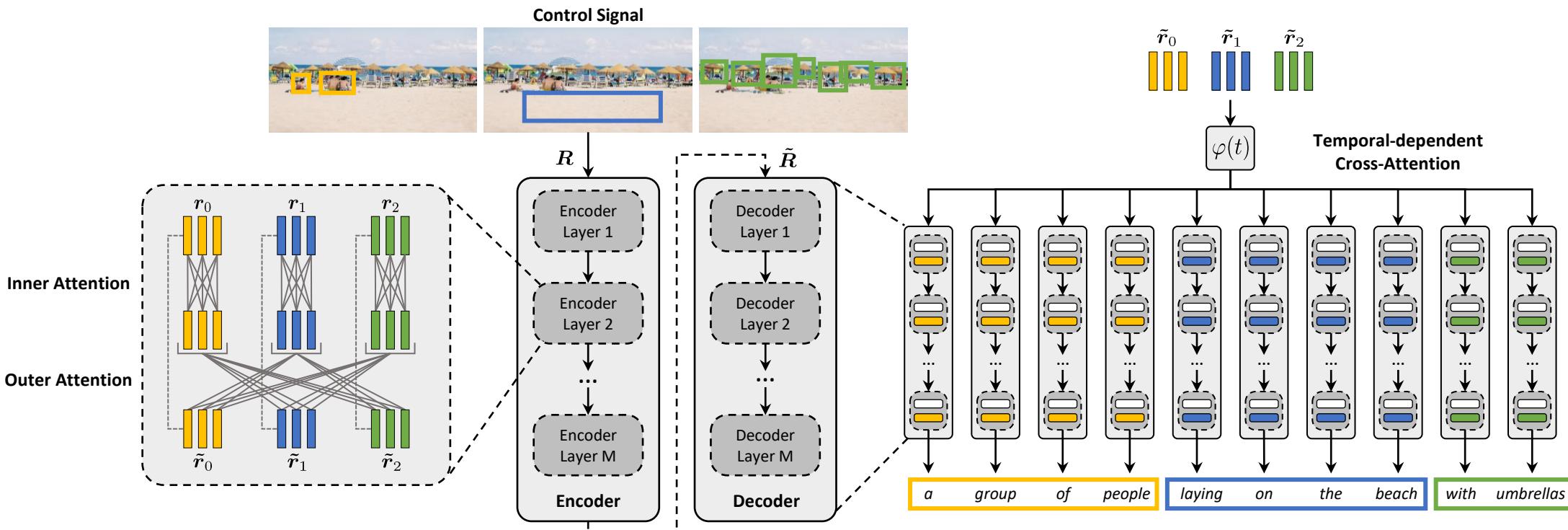
- Train on GT words and shifting gate values (obtained via NLP)

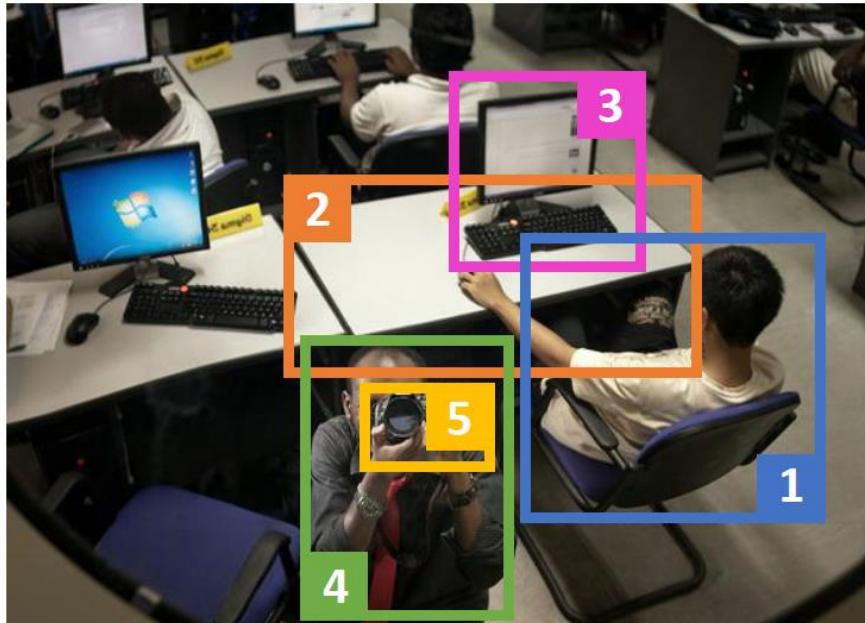
$$\begin{aligned}
 L(\theta) = & - \sum_{t=1}^T \left( \underbrace{\log p(y_t^* | r_{1:t}^*, y_{1:t-1}^*)}_{\text{Word-level probability}} + \right. \\
 & + g_t^* \log p(g_t = 1 | r_{1:t}^*, y_{1:t-1}^*) + \\
 & \left. + (1 - g_t^*) \log(1 - \underbrace{p(g_t = 1 | r_{1:t}^*, y_{1:t-1}^*)}_{\text{Chunk-level probability}}) \right)
 \end{aligned}$$



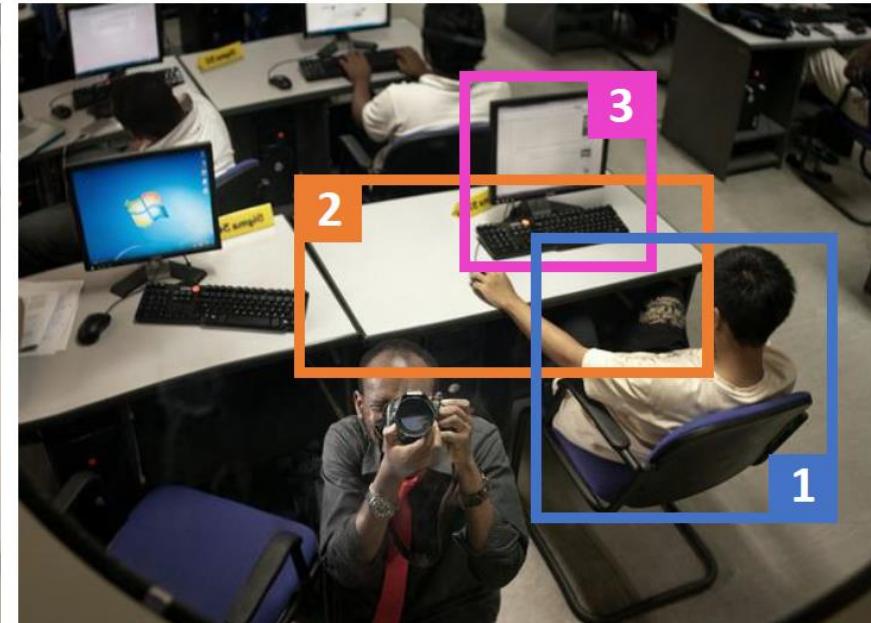
- What if we have an unordered set as input?
- We can learn a network to do the sorting! → **SINKHORN NETWORK**
  - Approximates a derivable permutation matrix
  - Train on real data, then use the Hungarian to get the true permutation matrix.

# Transformer-based Show, Control and Tell



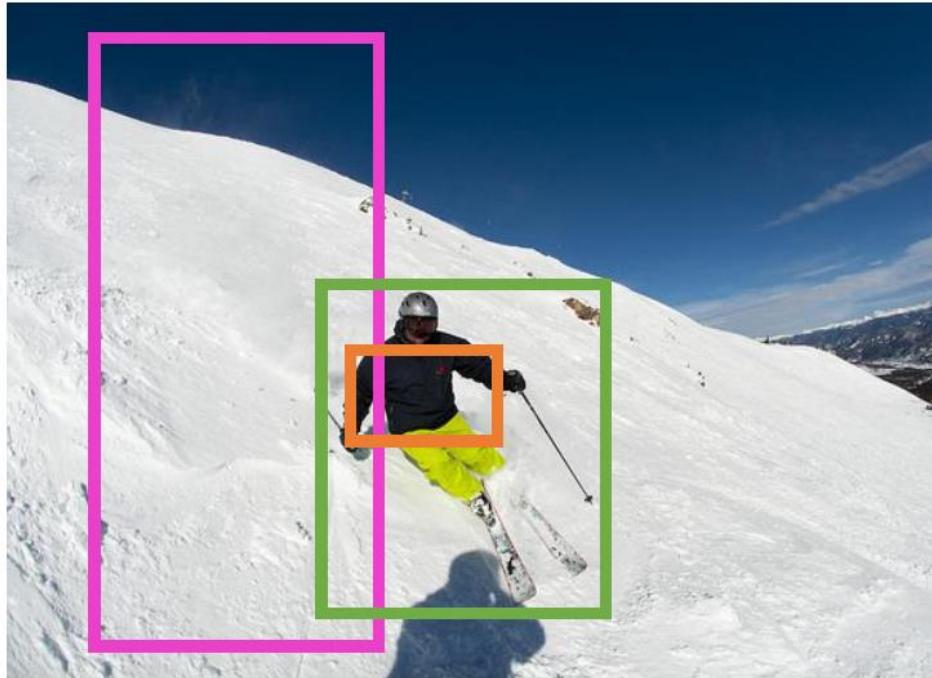
**Results when controlling with a sequence of regions**

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

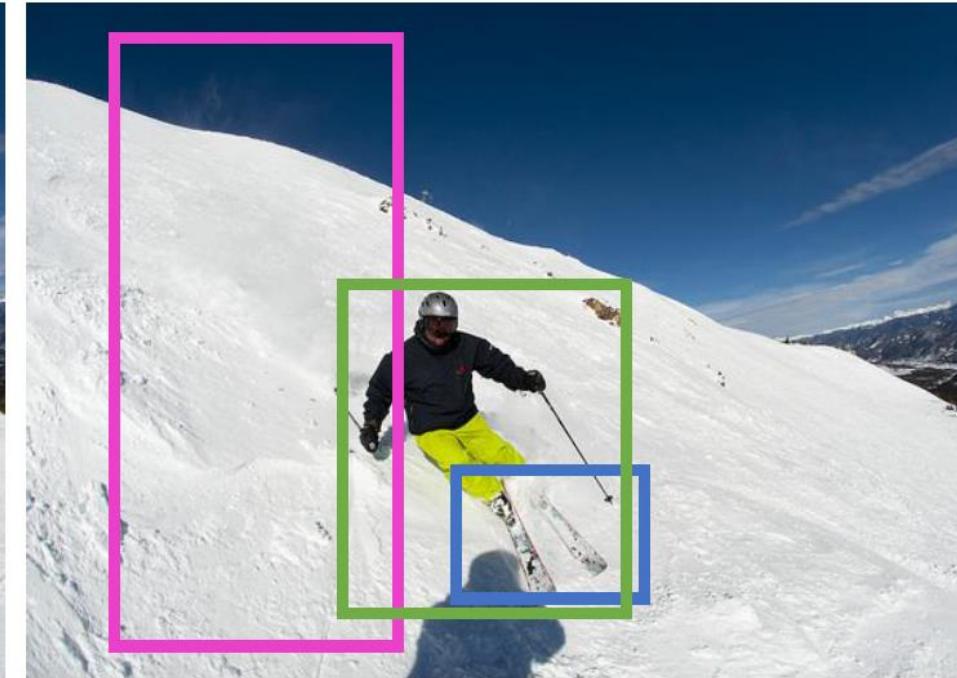


A man sitting at a desk with a computer.

## 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.

# 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

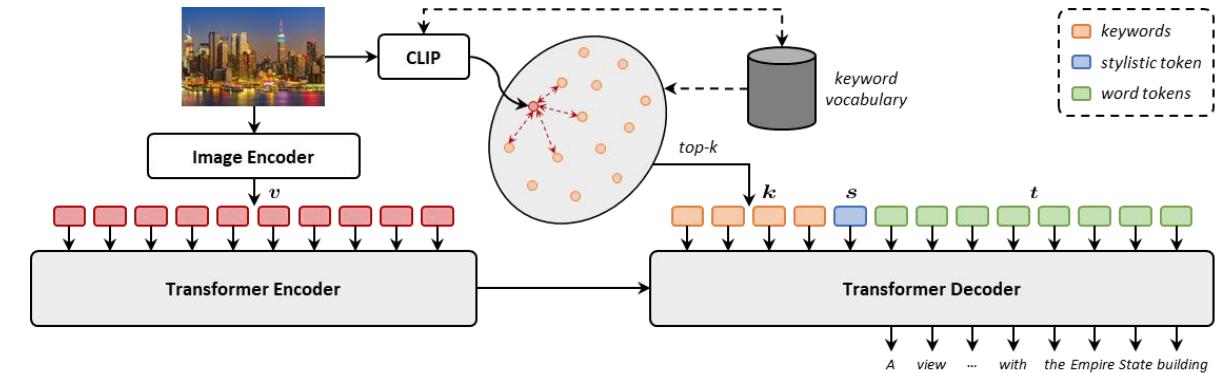
- **CNN feature extractors** which can directly take raw pixels as input and avoid the need of using object detectors;
- **Textual keywords** extracted with large-scale cross-modal models;
- **Stylistic tokens** to separate hand-collected and web-based image-caption pairs.

## 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**.



**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 <sup>base</sup>	40.5	29.7	-	137.6	22.8
VinVL <sup>base</sup>	40.9	<b>30.9</b>	-	140.6	<b>25.1</b>
<b>UniversalCap<sup>tiny</sup></b>	40.8	29.9	59.9	140.4	23.4
<b>UniversalCap<sup>small</sup></b>	<b>41.2</b>	30.4	<b>60.2</b>	143.0	24.1
<b>UniversalCap<sup>base</sup></b>	40.8	30.4	<b>60.2</b>	<b>143.4</b>	24.2

**Results on Open Images**

	CLIP-S	# Long-tail Words	# Proper Nouns
VinVL <sup>base</sup>	0.708	149	55
VinVL <sup>large</sup>	0.715	186	60
<b>UniversalCap<sup>tiny</sup></b>	0.728	821	410
<b>UniversalCap<sup>small</sup></b>	0.732	866	432
<b>UniversalCap<sup>base</sup></b>	<b>0.739</b>	<b>1,071</b>	<b>469</b>



**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.*

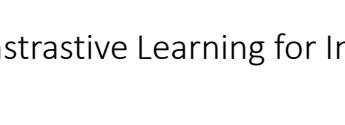
**Universal Captioner:**  
*A [Harry Potter](#) and the [Prisoner of Azkaban](#) concert poster.*

# Evaluation Metrics

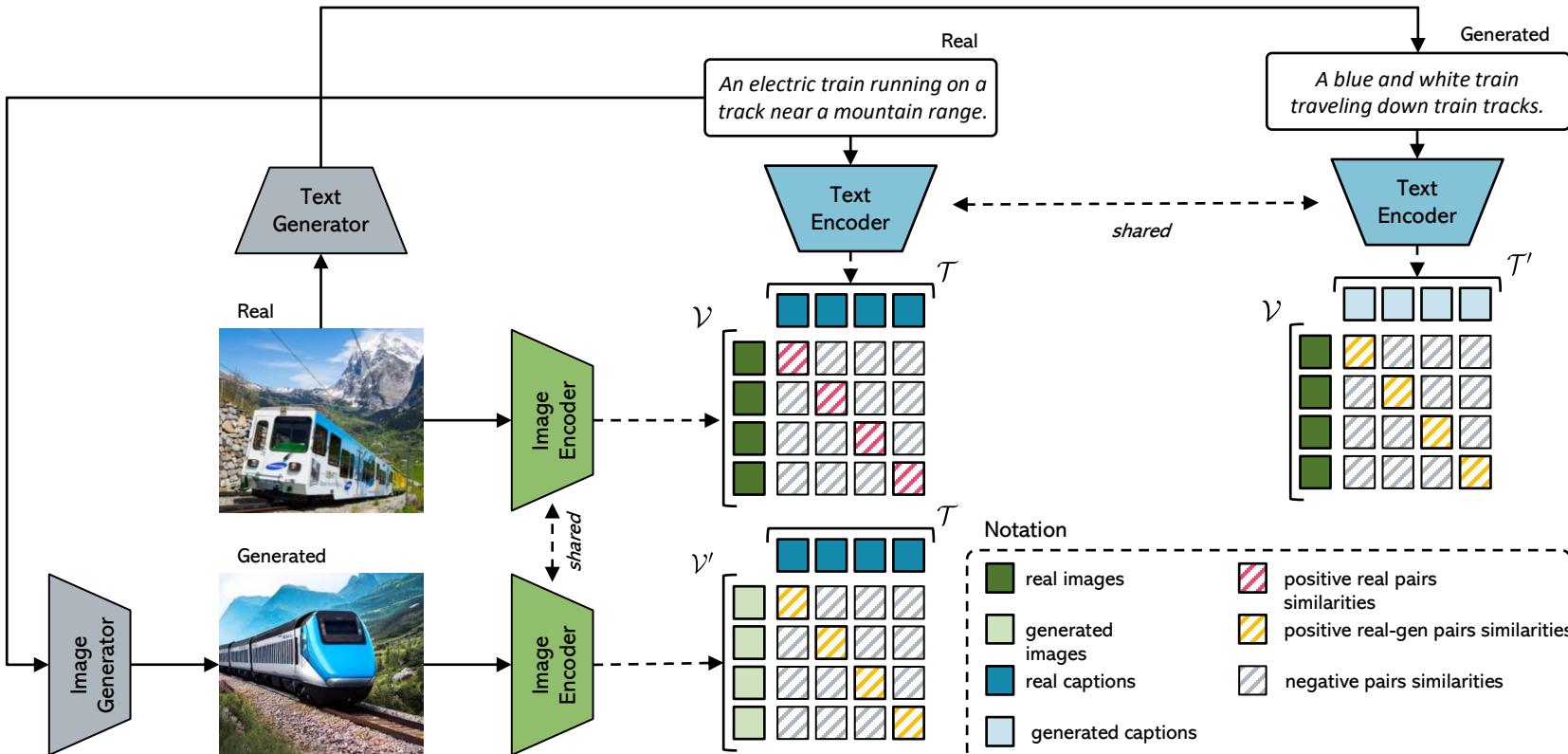
# A New Evaluation metric

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

# 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<sup>1</sup> and BLIP<sup>2</sup>, 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 $\tau_b$	Kendall $\tau_c$	Kendall $\tau_b$	Kendall $\tau_c$
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
<b>PAC-S</b>	<b>53.9</b>	<b>54.3</b>	<b>36.0</b>	<b>18.6</b>
	(+2.8)	(+3.1)	(+1.6)	(+0.9)
RefCLIP-S	52.6	53.0	36.4	18.8
<b>RefPAC-S</b>	<b>55.4</b>	<b>55.8</b>	<b>37.6</b>	<b>19.5</b>
	(+2.8)	(+2.8)	(+1.2)	(+0.7)

	Composite	
	Kendall $\tau_b$	Kendall $\tau_c$
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
<b>PAC-S</b>	<b>51.5</b>	<b>55.7</b>
	(+1.7)	(+1.9)
RefCLIP-S	51.2	55.4
<b>RefPAC-S</b>	<b>52.8</b>	<b>57.1</b>
	(+1.6)	(+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	<b>99.3</b>	96.5	72.0	80.9
<b>PAC-S</b>	<b>60.6</b>	<b>99.3</b>	<b>96.9</b>	<b>72.9</b>	<b>82.4</b>
	(+4.7)	(+0.0)	(+0.4)	(+0.9)	(+1.5)
RefCLIP-S	64.9	<b>99.5</b>	95.5	73.3	83.3
<b>RefPAC-S</b>	<b>68.2</b>	<b>99.5</b>	<b>95.6</b>	<b>75.9</b>	<b>84.8</b>
	(+3.3)	(+0.0)	(+0.1)	(+2.6)	(+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

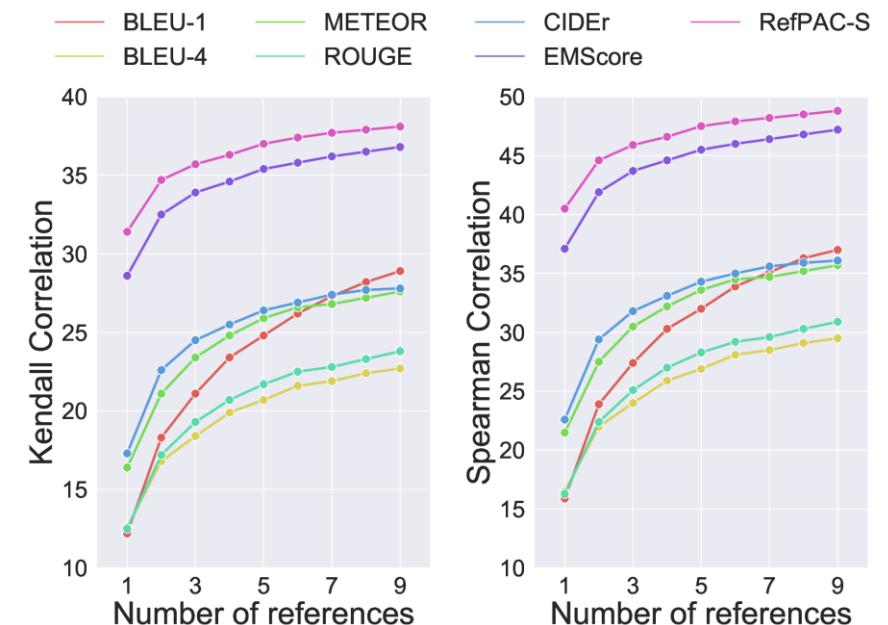
# Video Captioning Correlation with Human Judgment

It works well on videos too!

	No Ref		1 Ref		9 Refs	
	Kendall $\tau_b$	Spearman $\rho$	Kendall $\tau_b$	Spearman $\rho$	Kendall $\tau_b$	Spearman $\rho$
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
<b>PAC-S / RefPAC-S</b>	<b>25.1</b> (+1.9)	<b>32.6</b> (+2.3)	<b>31.4</b> (+2.8)	<b>40.5</b> (+3.4)	<b>38.1</b> (+1.3)	<b>48.8</b> (+1.6)

Human judgment correlation scores on the VATEX-EVAL<sup>1</sup> dataset.

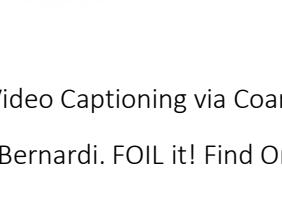
We show Kendall  $\tau_B$  correlation score at varying of the number of reference captions.



And it hallucinates less than previous metrics 😊

	<b>FOIL</b>		<b>ActivityNet-FOIL</b>
	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
<b>PAC-S</b>	<b>89.9</b> (+2.7)	<b>89.9</b> (+2.7)	<b>90.1</b> (+0.6)
RefCLIP-S	91.0	92.6	-
EMScoreRef	-	-	92.4
<b>RefPAC-S</b>	<b>93.8</b> (+2.8)	<b>95.2</b> (+2.6)	<b>93.5</b> (+1.1)

We extend our analysis to two datasets for detecting hallucinations in textual sentences, namely FOIL<sup>2</sup> and ActivityNet<sup>1</sup>.

Image	Candidate Captions	Evaluation Scores
	A <b>silver knife</b> containing many carrots with long, green stems.	CLIP-S <b>0.942</b> PAC-S 0.854
	A <b>silver bowl</b> containing many carrots with long, green stems.	CLIP-S 0.912 PAC-S <b>0.893</b>
	A person tries to catch a <b>ball</b> on a beach.	CLIP-S <b>0.781</b> PAC-S 0.798
	A person tries to catch a <b>frisbee</b> on a beach.	CLIP-S 0.759 PAC-S <b>0.828</b>
	A <b>baby horse</b> is seen standing in between another elephant's legs.	CLIP-S <b>0.782</b> PAC-S 0.793
	A <b>baby elephant</b> is seen standing in between another elephant's legs.	CLIP-S 0.769 PAC-S <b>0.820</b>
	Different kinds of food on a plate with a <b>cup</b> .	CLIP-S <b>0.682</b> PAC-S 0.758
	Different kinds of food on a plate with a <b>fork</b> .	CLIP-S 0.676 PAC-S <b>0.789</b>

# 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 $\tau_b$	Kendall $\tau_c$	Kendall $\tau_b$	Kendall $\tau_c$	Kendall $\tau_b$	Spearman $\rho$	Accuracy	Accuracy	Accuracy
<b>CLIP ViT-B/16</b>	CLIP-S	51.7	52.1	34.9	18.0	-	81.1	90.6	-
	EMScore	-	-	-	-	24.1	31.4	-	90.0
	<b>PAC-S</b>	<b>54.5</b> (+2.8)	<b>54.9</b> (+2.8)	<b>35.9</b> (+1.0)	<b>18.5</b> (+0.5)	<b>26.8</b> (+2.7)	<b>34.7</b> (+3.3)	<b>82.9</b> (+1.8)	<b>91.1</b> (+0.5)
<b>CLIP ViT-L/14</b>	CLIP-S	52.6	53.0	35.2	18.2	-	81.7	90.9	-
	EMScore	-	-	-	-	26.7	34.7	-	89.0
	<b>PAC-S</b>	<b>55.4</b> (+2.8)	<b>55.8</b> (+2.8)	<b>36.8</b> (+1.6)	<b>19.0</b> (+0.8)	<b>28.9</b> (+2.2)	<b>37.4</b> (+2.7)	<b>82.0</b> (+0.3)	<b>91.9</b> (+1.0)
<b>OpenCLIP ViT-B/32</b>	CLIP-S	52.3	52.6	35.4	18.3	-	81.2	88.9	-
	EMScore	-	-	-	-	24.8	32.2	-	88.2
	<b>PAC-S</b>	<b>53.6</b> (+1.3)	<b>53.9</b> (+1.3)	<b>36.1</b> (+0.7)	<b>18.6</b> (+0.3)	<b>25.4</b> (+0.6)	<b>33.1</b> (+0.9)	<b>82.4</b> (+1.2)	<b>90.1</b> (+1.2)
<b>OpenCLIP ViT-L/14</b>	CLIP-S	54.4	54.5	36.6	18.9	-	82.5	92.2	-
	EMScore	-	-	-	-	27.0	35.0	-	90.7
	<b>PAC-S</b>	<b>55.3</b> (+0.9)	<b>55.7</b> (+1.2)	<b>37.0</b> (+0.4)	<b>19.1</b> (+0.2)	<b>27.8</b> (+0.8)	<b>36.1</b> (+1.1)	<b>82.7</b> (+0.2)	<b>93.1</b> (+0.9)

# Qualitative Results

Image	Candidate Captions	Evaluation Scores		Image	Candidate Captions	Evaluation Scores	
		CLIP-S	PAC-S			CLIP-S	PAC-S
	Two white dogs running.	CLIP-S <b>0.530</b>	PAC-S 0.500		A man and young girl eat a meal on a city street .	CLIP-S <b>0.769</b>	PAC-S 0.764
	A man riding a motorbike kicks up dirt.	CLIP-S 0.486	PAC-S <b>0.542</b>		A small brown and white dog running through tall grass.	CLIP-S 0.752	PAC-S <b>0.820</b>
	Little girl in bare feet sitting in a circle.	CLIP-S <b>0.524</b>	PAC-S 0.431		A man jumps while snow skiing.	CLIP-S <b>0.512</b>	PAC-S 0.503
	A white dog runs in the grass.	CLIP-S 0.426	PAC-S <b>0.456</b>		A man is hiking on a snow-covered trail.	CLIP-S 0.464	PAC-S <b>0.567</b>
	Four woman wearing formal gowns pose together and smile.	CLIP-S <b>0.700</b>	PAC-S 0.730		Two girls walking down the street.	CLIP-S <b>0.583</b>	PAC-S 0.556
	A man in a wetsuit surfs.	CLIP-S 0.613	PAC-S <b>0.762</b>		A dog lies down on a cobblestone street.	CLIP-S 0.550	PAC-S <b>0.562</b>
	Boy with a red crown in a shopping cart.	CLIP-S <b>0.385</b>	PAC-S 0.467		A woman is signaling is to traffic , as seen from behind.	CLIP-S <b>0.753</b>	PAC-S 0.767
	People stand outside near a concrete wall and a window.	CLIP-S 0.359	PAC-S <b>0.509</b>		A man rides a bike through a course.	CLIP-S 0.714	PAC-S <b>0.800</b>

# Qualitative Results

Image	Candidate Captions	Evaluation Scores				Image	Candidate Captions	Evaluation Scores			
		METEOR	CIDEr	CLIP-S	PAC-S			METEOR	CIDEr	CLIP-S	PAC-S
	A blue bird being held by a handler.	35.2	96.3	80.1	80.0		A passenger train in the snow.	26.8	89.7	83.5	83.1
	A blue bird perched on a gloved hand.	18.6	39.0	76.1	82.1		A red train driving through a snow covered city.	27.2	72.6	81.4	85.7
	A black boxer dog with a white underbelly and brown collar looks at the camera.	35.1	26.6	77.5	82.3		A dog pokes its head out from under a pile of stuff.	25.8	60.5	67.5	75.6
	A close up of a black pug.	11.6	21.1	71.0	83.5		A dog underneath a wooden beam.	22.0	38.9	63.9	81.6
	Trains amble by the rail yard.	26.2	68.8	81.9	75.4		A large green coach with a bridge in the background	28.3	32.0	87.1	76.7
	The red train and the yellow train on the tracks.	14.7	28.3	79.8	81.6		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

# Want to know more?

arXiv:2303.12112v1 [cs.CV] 21 Mar 2023

## Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation

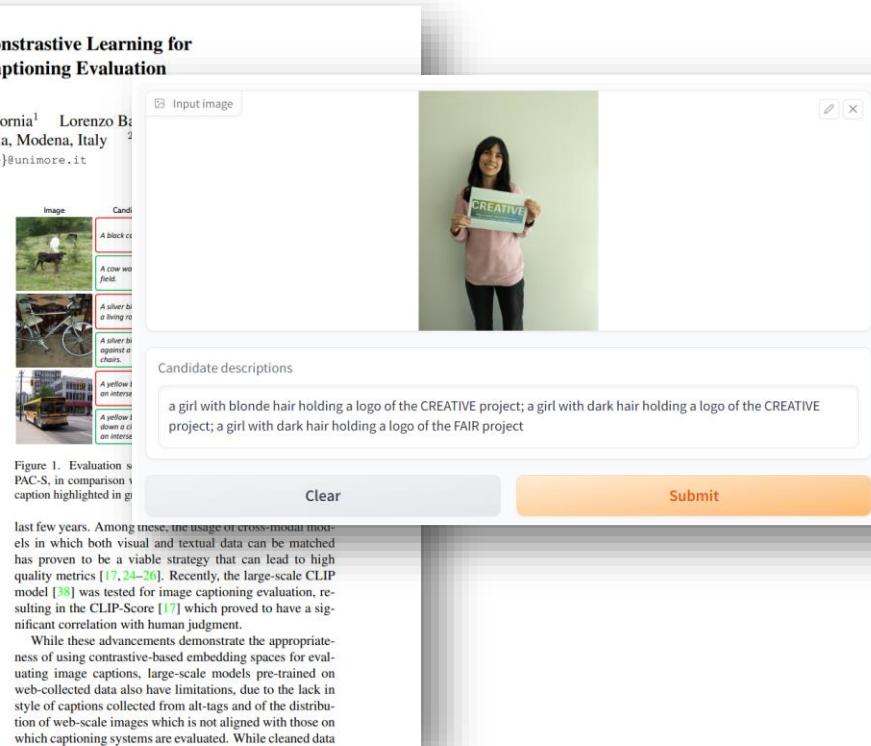
Sara Sarto<sup>1</sup> Manuele Barraco<sup>1</sup> Marcella Cornia<sup>1</sup> Lorenzo Baraldi<sup>2</sup>  
<sup>1</sup>University of Modena and Reggio Emilia, Modena, Italy  
<sup>2</sup>{name.surname}@unimore.it

### Abstract

The CLIP model has been recently proven to be very effective for a variety of cross-modal tasks, including the evaluation of captions generated from vision-and-language architectures. In this paper, we propose a new recipe for a contrastive-based evaluation metric for image captioning, namely Positive-Augmented Contrastive learning Score (PAC-S), that in a novel way unifies the learning of a contrastive visual-semantic space with the addition of generated images and text on curated data. Experiments spanning several datasets demonstrate that our new metric achieves the highest correlation with human judgments on both images and videos, outperforming existing reference-based metrics like CIDEr and SPICE and reference-free metrics like CLIP-Score. Finally, we test the system-level correlation of the proposed metric when considering popular image captioning approaches, and assess the impact of employing different cross-modal features. Our source code and trained models are publicly available at: <https://github.com/aimagelab/pacscore>.

### 1. Introduction

The task of image captioning, which requires an algorithm to describe visual contents with natural language sentences, has been gaining considerable attention from the research community in the past few years [22, 53, 61]. As such, the task has witnessed methodological and architectural innovations, ranging from the usage of self-attentive models [10, 16, 19, 36] to the development of better connections between visual and textual modalities with the ad-

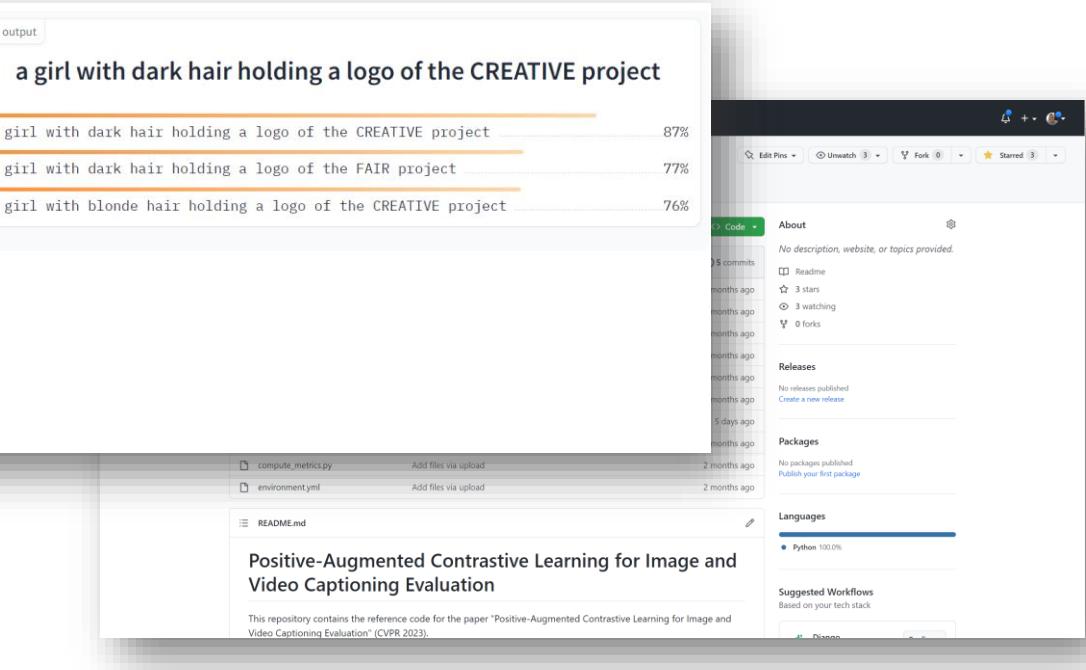


Read the paper

<https://arxiv.org/abs/2303.12112>

<https://github.com/aimagelab/pacscore>

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



Use it in your projects ☺

JUNE 18-22, 2023  
CVPR  
VANCOUVER, CANADA

# Thank you!



Manuele Barraco



Sara Sarto



Nicholas Moratelli



Marcella Cornia



Lorenzo Baraldi



Rita Cucchiara