Image Captioning

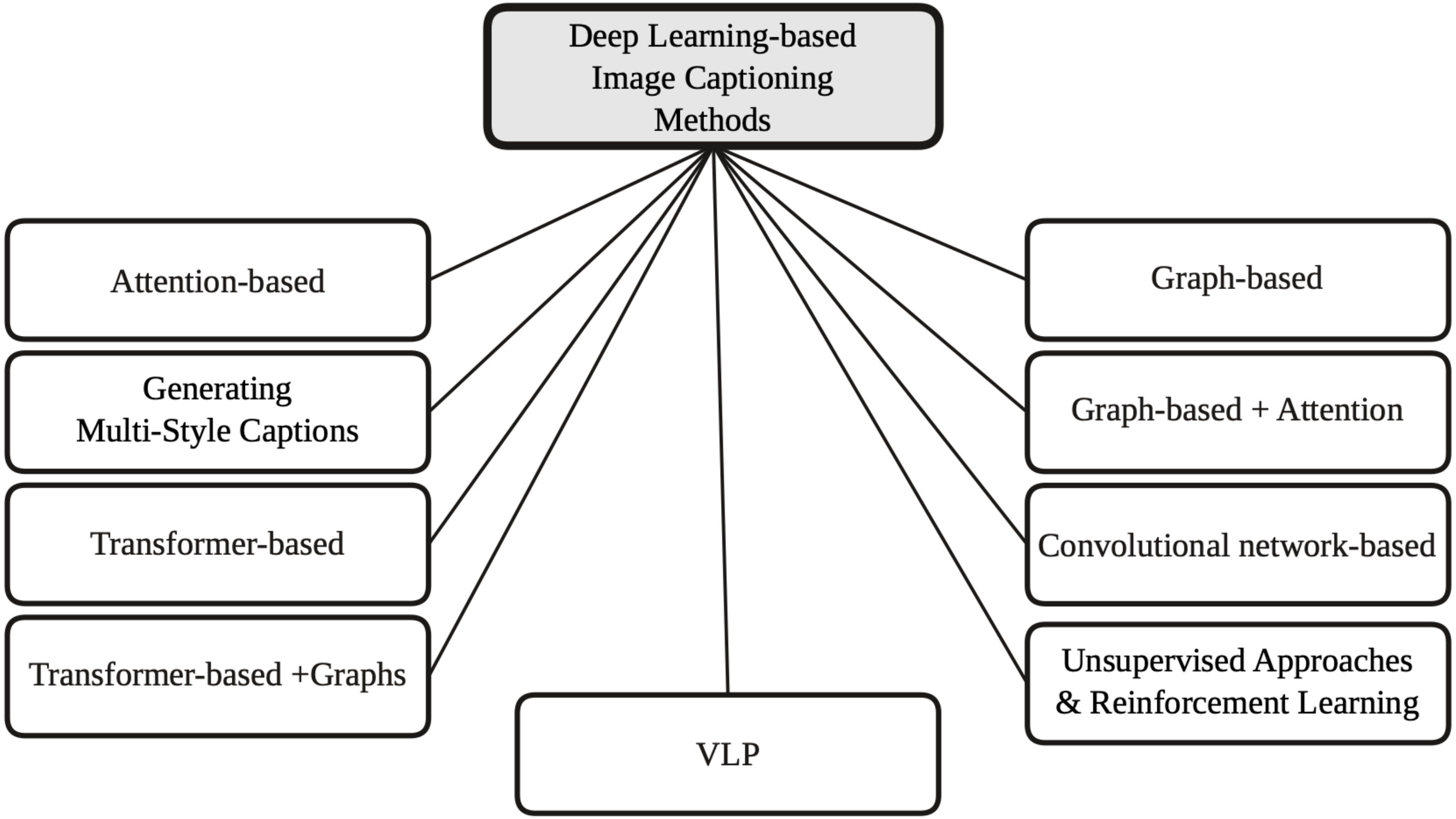
Olarik Surinta

Mahasarakham University

Image Captioning

- **An image caption** is a technology at the intersection of computer vision and natural language processing (NLP).
- It is used for generating a sentence that describes the content information in an image.
- This technology is very challenging due to the complexity involved in **learning spatial** and **semantic features** from images and then ***creating a descriptive text*** distribution.

Source: https://arxiv.org/abs/2201.12944



Deep Learning-based Image Captioning Methods

Attention-based

Generating Multi-Style Captions

Transformer-based

Transformer-based +Graphs

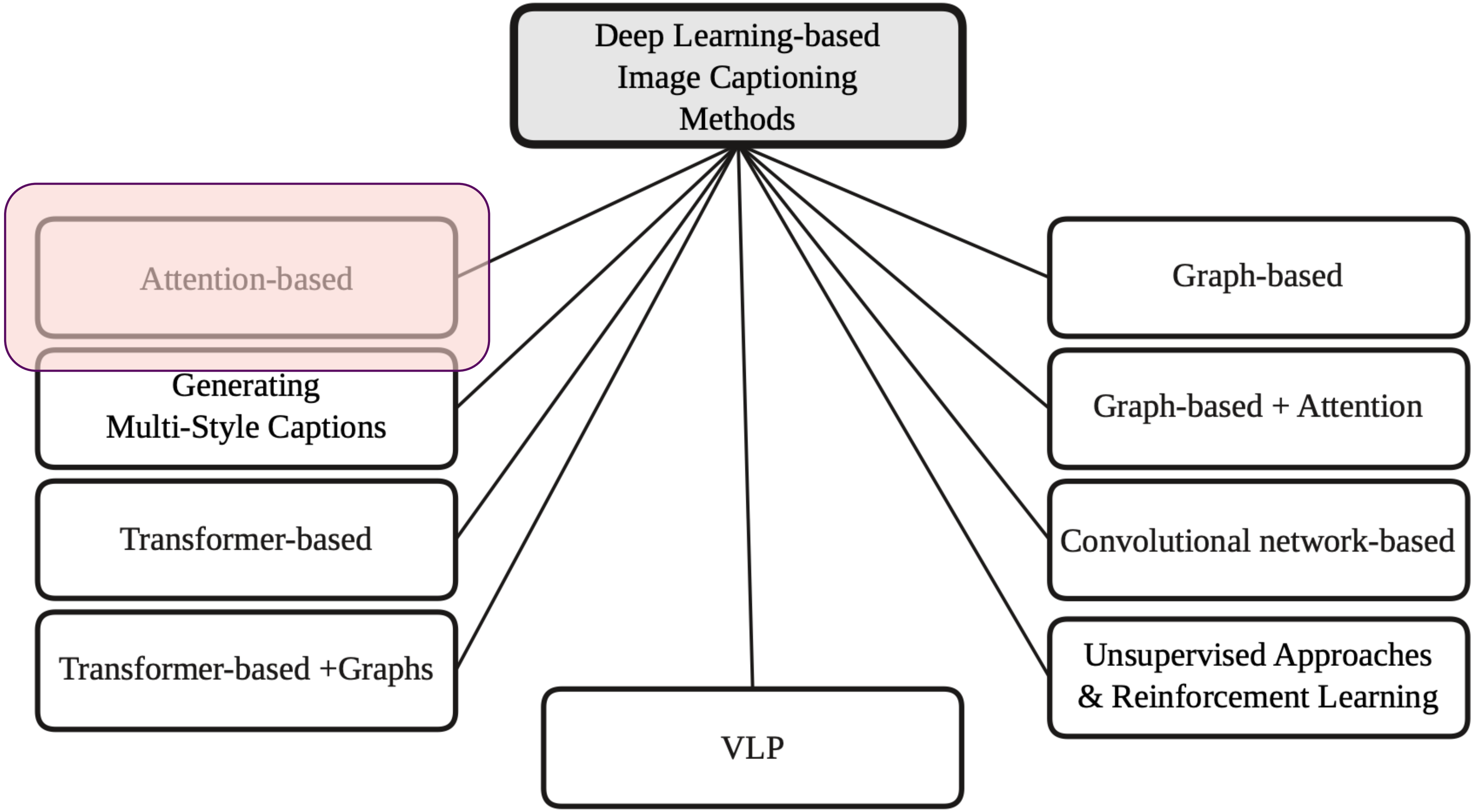
VLP

Graph-based

Graph-based + Attention

Convolutional network-based

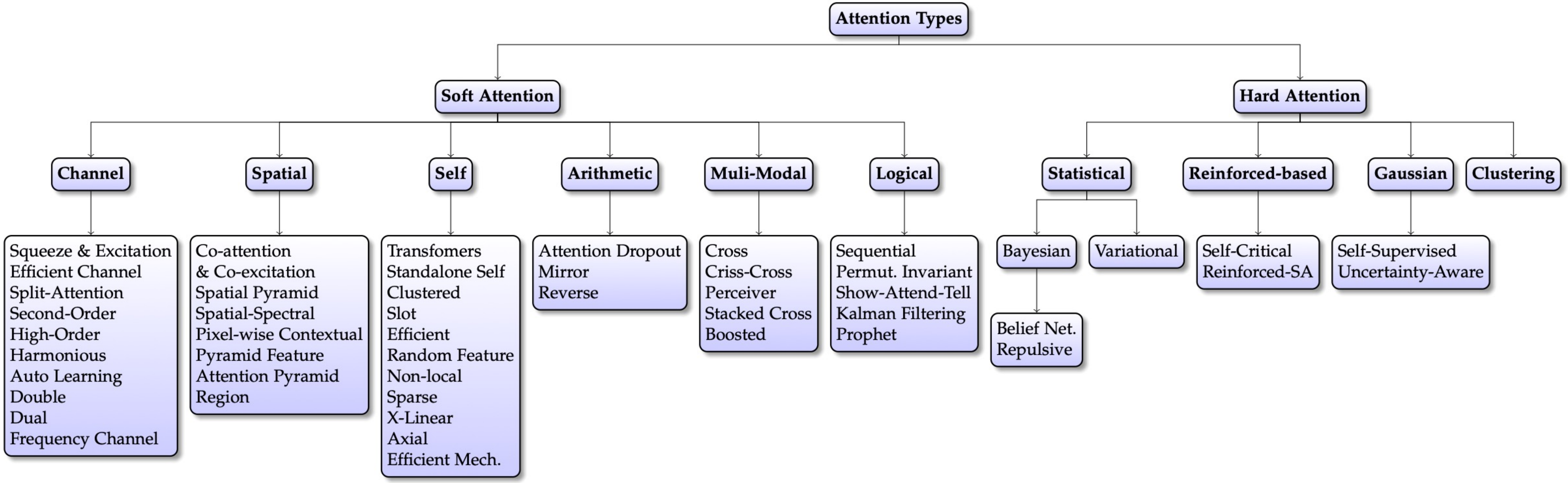
Unsupervised Approaches & Reinforcement Learning


Deep Learning-based
Image Captioning
Methods
Attention-based
Generating
Multi-Style Captions
Transformer-based
Transformer-based +Graphs
VLP
Graph-based
Graph-based + Attention
Convolutional network-based
Unsupervised Approaches
& Reinforcement Learning

Attention-based Methods

- The methods that fall under the attention-based category utilize attention mechanisms to emphasize the most relevant parts of the input image when generating captions.
- The attention mechanism in an **encoder-decoder** framework is typically used in machine translation.

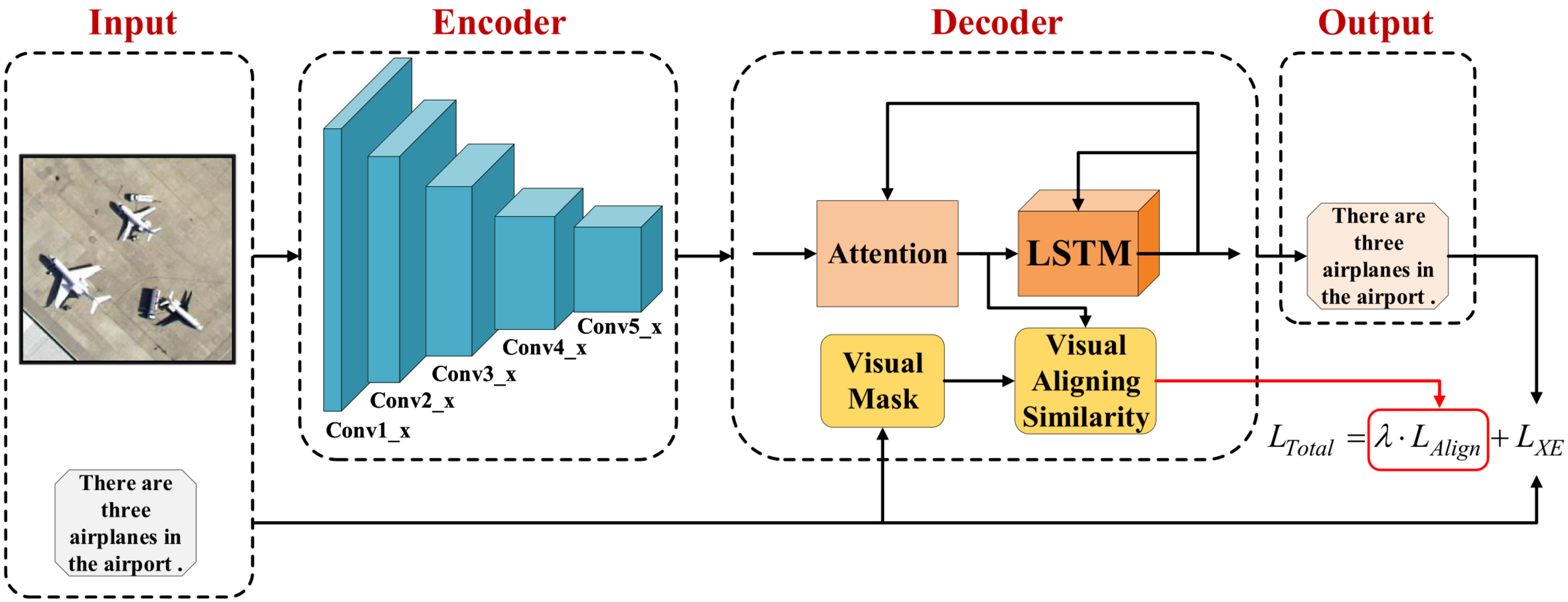
A Taxomony of Attention Types



- Attention Types
  - Soft Attention
    - Channel
      - Squeeze & Excitation
      - Efficient Channel
      - Split-Attention
      - Second-Order
      - High-Order
      - Harmonious
      - Auto Learning
      - Double
      - Dual
      - Frequency Channel
    - Spatial
      - Co-attention
      - & Co-excitation
      - Spatial Pyramid
      - Spatial-Spectral
      - Pixel-wise Contextual
      - Pyramid Feature
      - Attention Pyramid
      - Region
    - Self
      - Transfomers
      - Standalone Self
      - Clustered
      - Slot
      - Efficient
      - Random Feature
      - Non-local
      - Sparse
      - X-Linear
      - Axial
      - Efficient Mech.
    - Arithmetic
      - Attention Dropout
      - Mirror
      - Reverse
    - Muli-Modal
      - Cross
      - Criss-Cross
      - Perceiver
      - Stacked Cross
      - Boosted
    - Logical
      - Sequential
      - Permut. Invariant
      - Show-Attend-Tell
      - Kalman Filtering
      - Prophet
  - Hard Attention
    - Statistical
      - Bayesian
        - Belief Net.
        - Repulsive
      - Variational
    - Reinforced-based
      - Self-Critical
      - Reinforced-SA
    - Gaussian
      - Self-Supervised
      - Uncertainty-Aware
    - Clustering

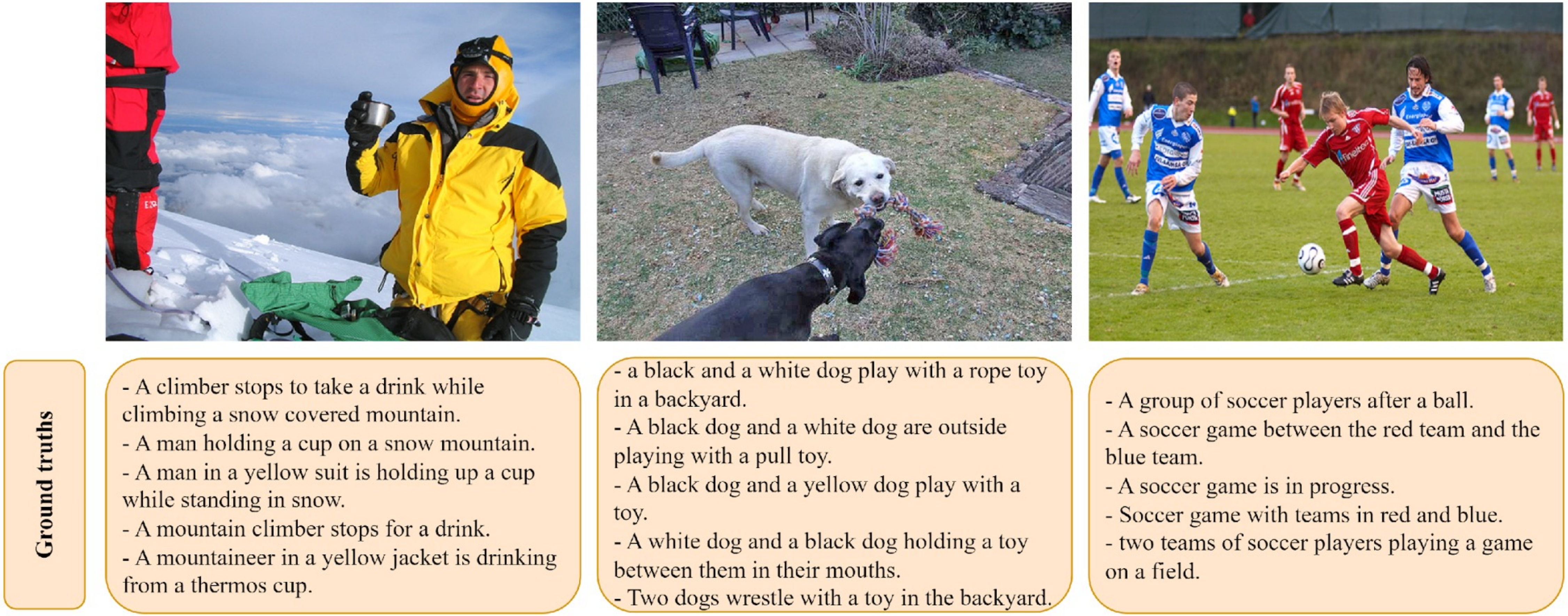
Source: https://arxiv.org/abs/2204.07756

Visual Attention Mechanism


Input
Encoder
Decoder
Output
There are three airplanes in the airport .
Conv1_x
Conv2_x
Conv3_x
Conv4_x
Conv5_x
Attention
LSTM
Visual Mask
Visual Aligning Similarity
There are three airplanes in the airport .
$L_{Total} = \lambda \cdot L_{Align} + L_{XE}$

Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8843891

Flickr8K Dataset



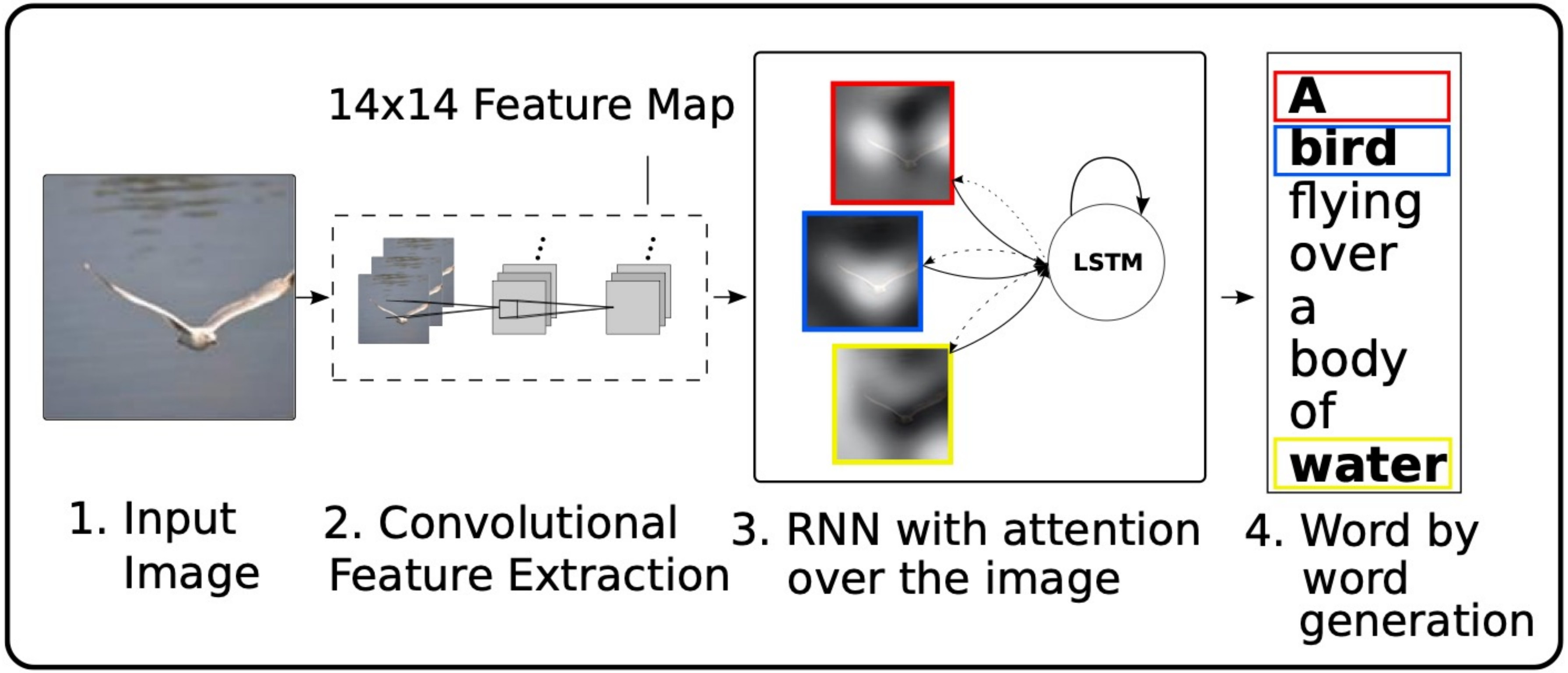
Ground truths

- A climber stops to take a drink while climbing a snow covered mountain.
- A man holding a cup on a snow mountain.
- A man in a yellow suit is holding up a cup while standing in snow.
- A mountain climber stops for a drink.
- A mountaineer in a yellow jacket is drinking from a thermos cup.

- a black and a white dog play with a rope toy in a backyard.
- A black dog and a white dog are outside playing with a pull toy.
- A black dog and a yellow dog play with a toy.
- A white dog and a black dog holding a toy between them in their mouths.
- Two dogs wrestle with a toy in the backyard.

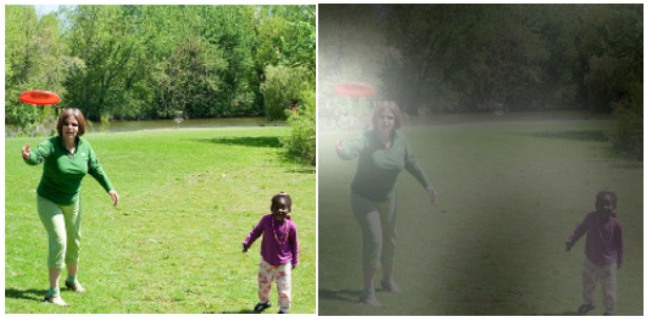
- A group of soccer players after a ball.
- A soccer game between the red team and the blue team.
- A soccer game is in progress.
- Soccer game with teams in red and blue.
- two teams of soccer players playing a game on a field.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

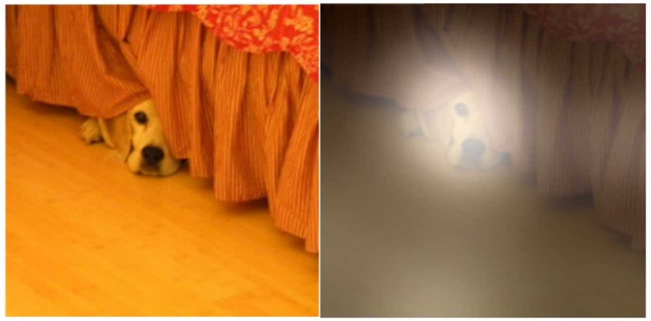

14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Source: https://arxiv.org/pdf/1502.03044.pdf

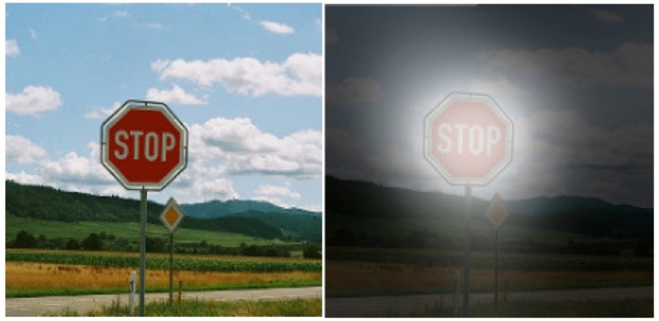
Examples of attending to the correct object



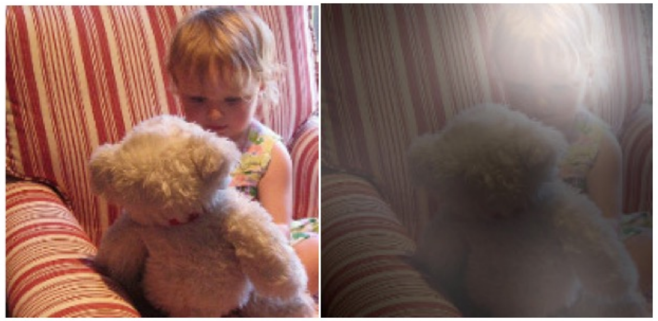
A woman is throwing a frisbee in a park.



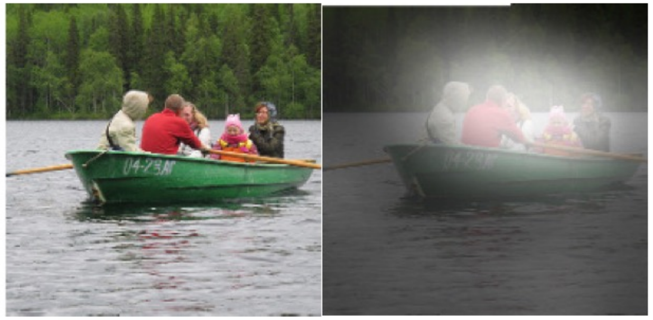
A dog is standing on a hardwood floor.



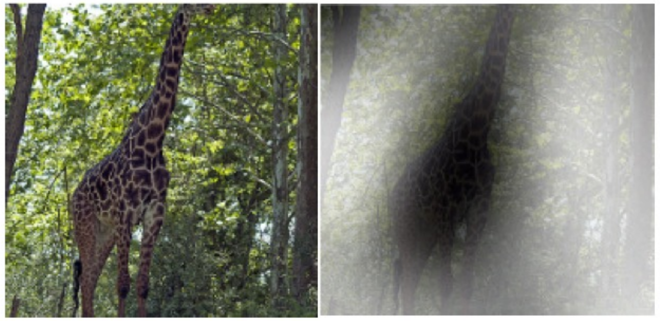
A stop sign is on a road with a mountain in the background.



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

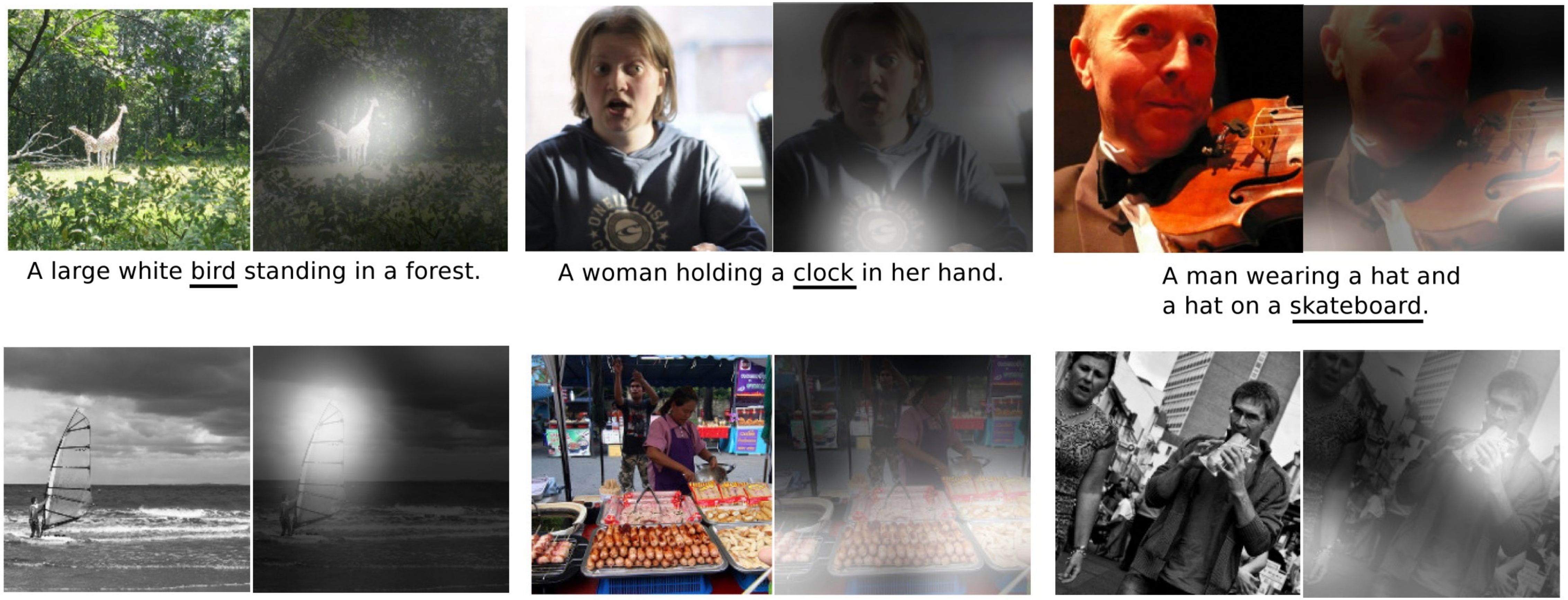


A group of people sitting on a boat in the water.



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

Examples of mistakes where we can use attention to gain intuition into what the model saw.



A large white bird standing in a forest.

A woman holding a clock in her hand.

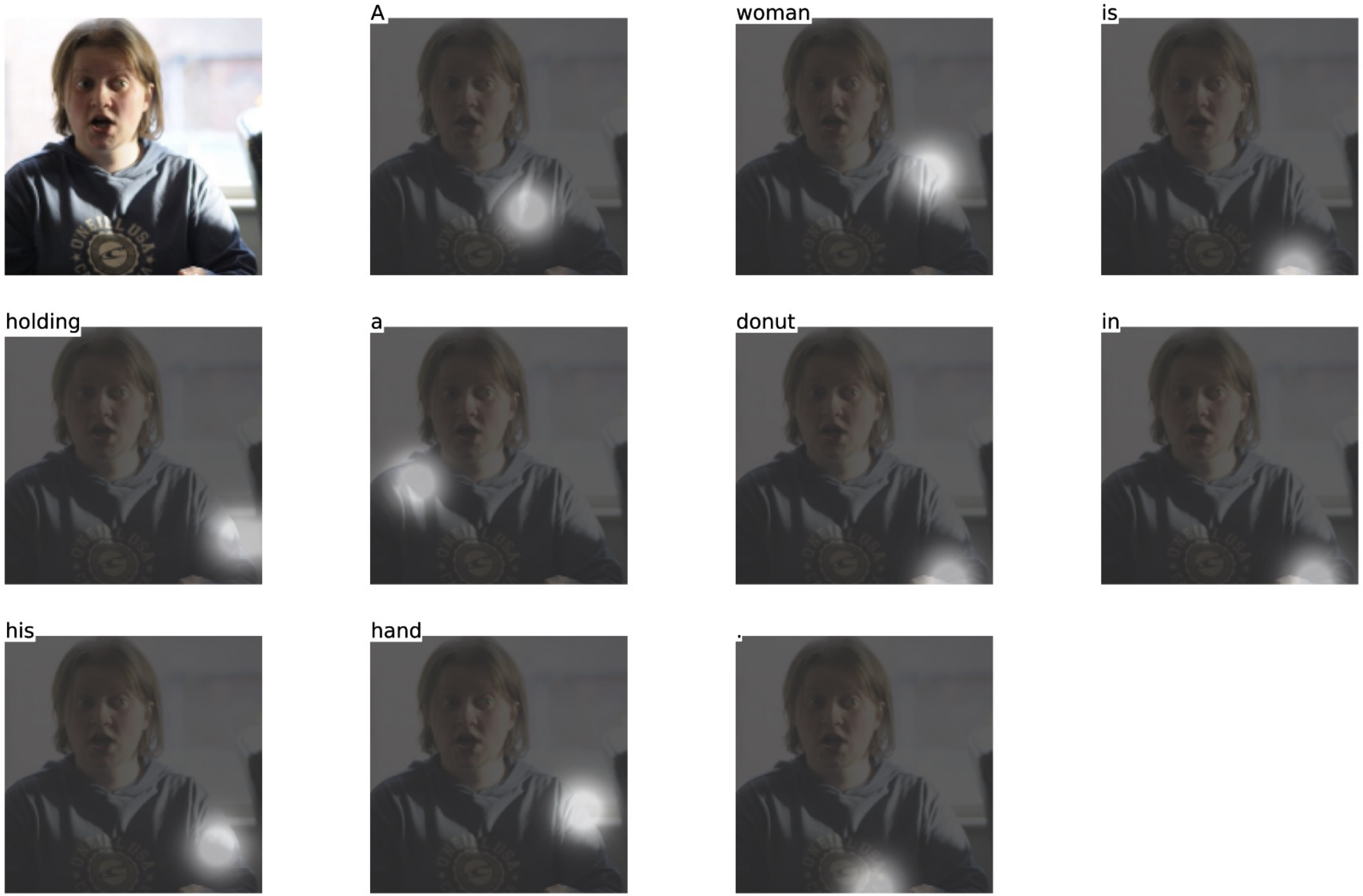
A man wearing a hat and a hat on a skateboard.

A person is standing on a beach with a surfboard.

A woman is sitting at a table with a large pizza.

A man is talking on his cell phone while another man watches.

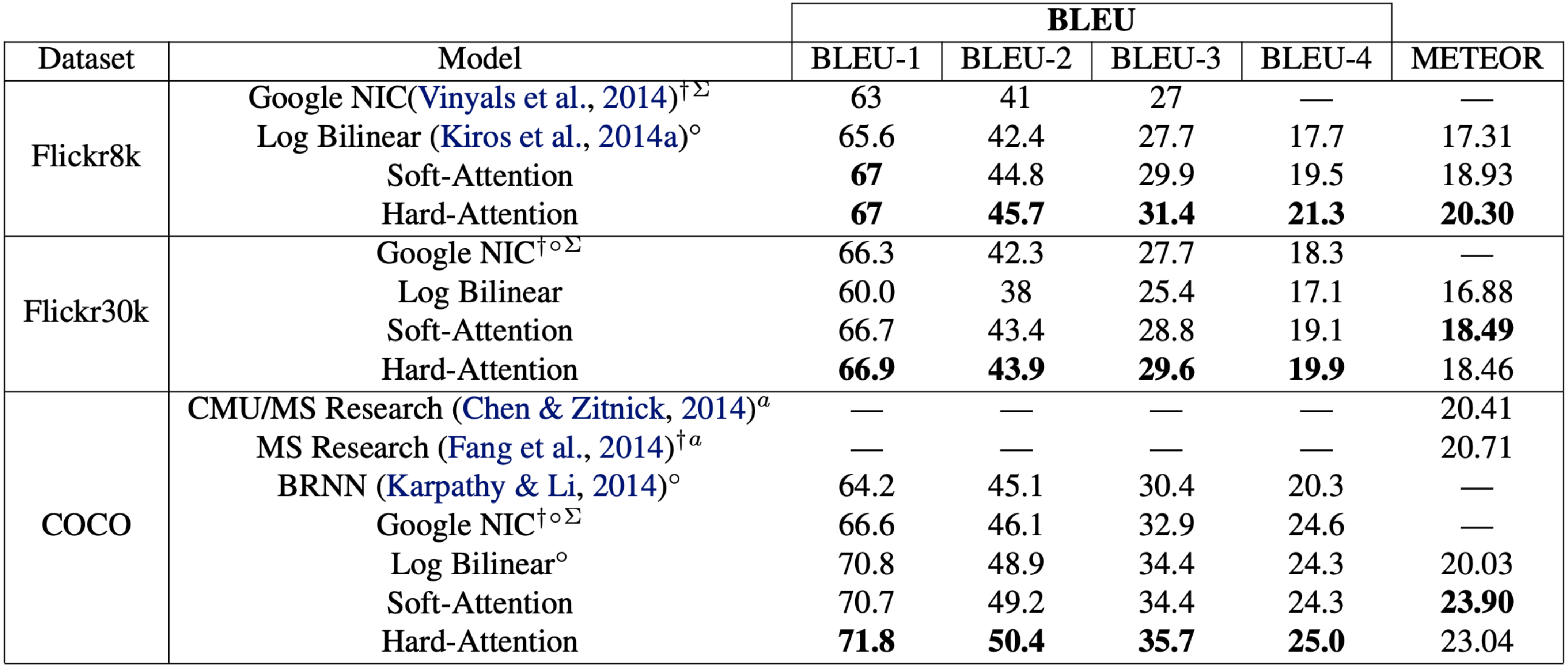
Source: https://arxiv.org/pdf/1502.03044.pdf



- A woman is holding a donut in his hand.
- A woman holding a clock in her hand.

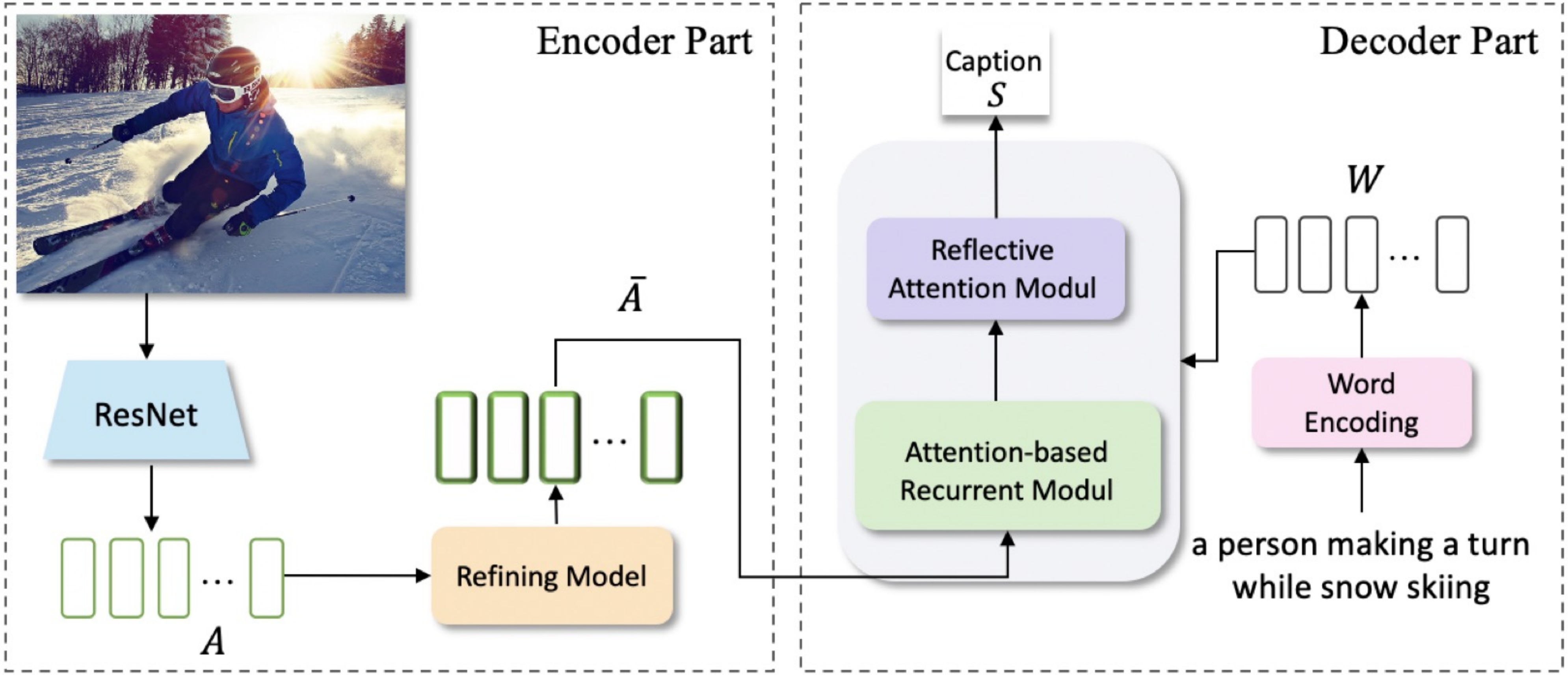
Source: https://arxiv.org/pdf/1502.03044.pdf

BLEU-1,2,3,4/METEOR metrics compared to other methods



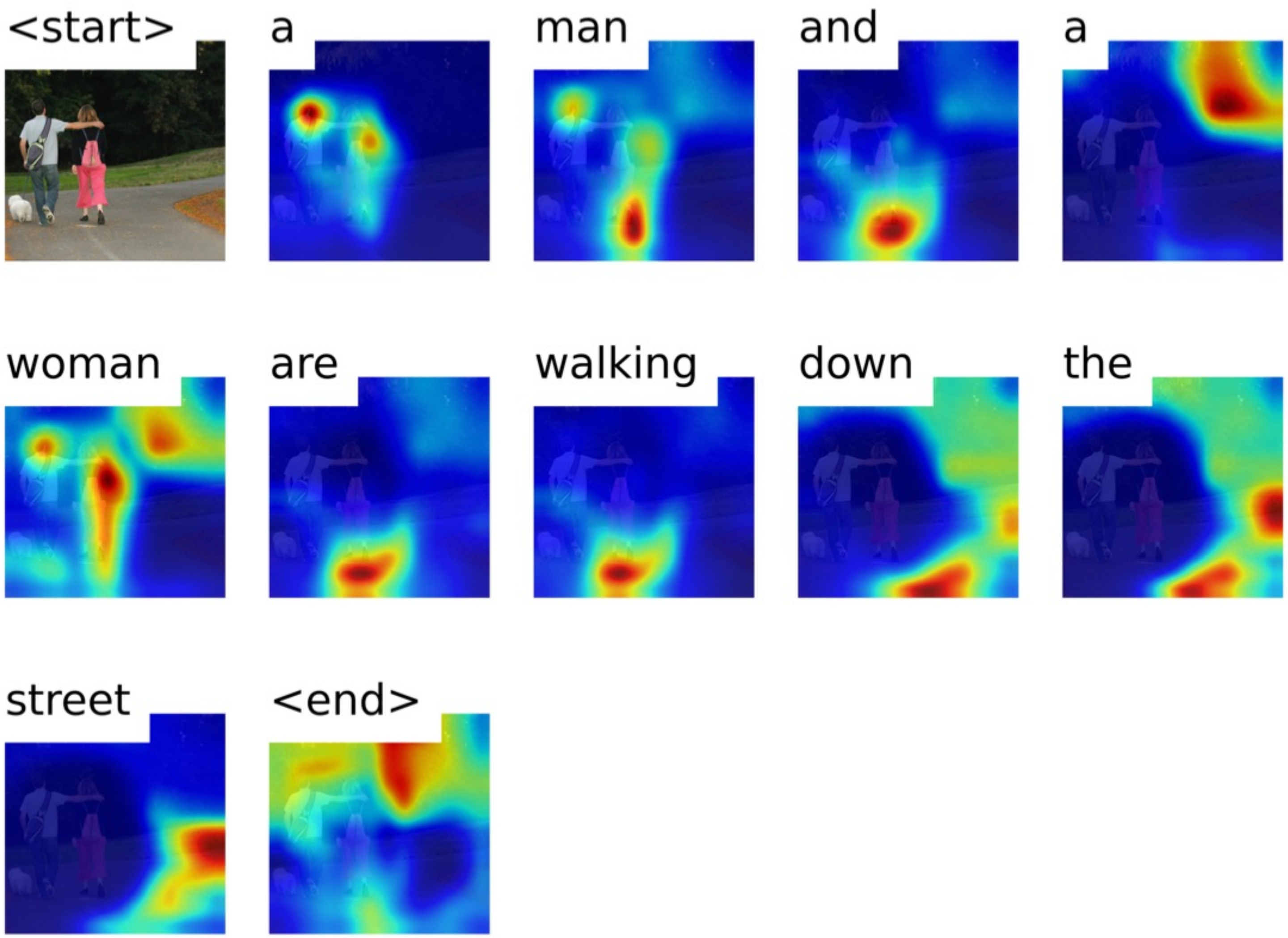
		BLEU				
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC$^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)^{a}	—	—	—	—	20.41
	MS Research (Fang et al., 2014)$^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)$^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC$^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear$^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

Image Captioning based on Feature Refinement and Reflective Decoding


Encoder Part
ResNet
A
Refining Model
$\bar{A}$
Decoder Part
Caption
S
Reflective Attention Modul
Attention-based Recurrent Modul
W
Word Encoding
a person making a turn while snow skiing

Source: https://arxiv.org/pdf/2206.07986.pdf

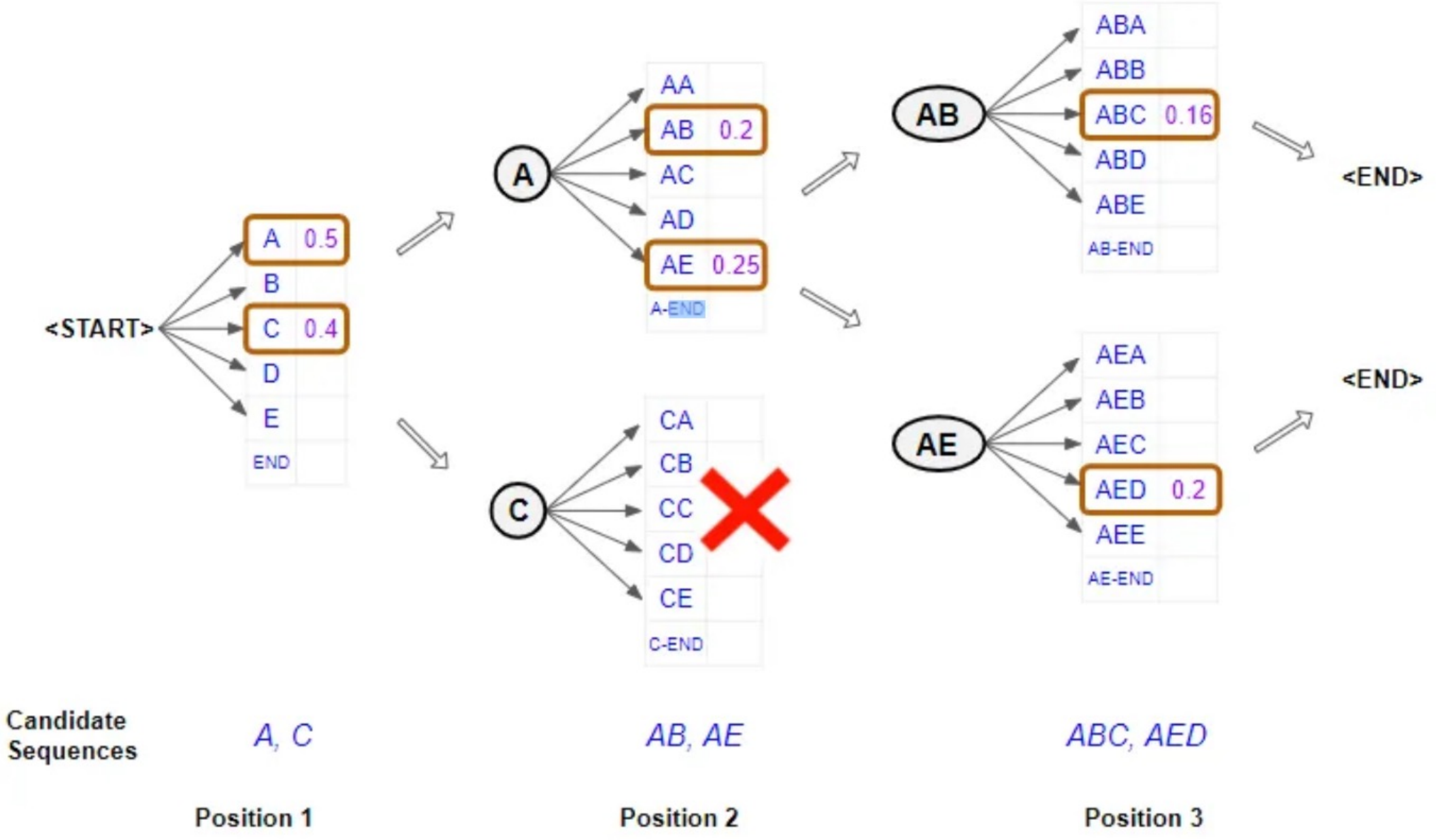
Visualization of attention weights learned by RefiningVisAttRefAtt model



Beam Search

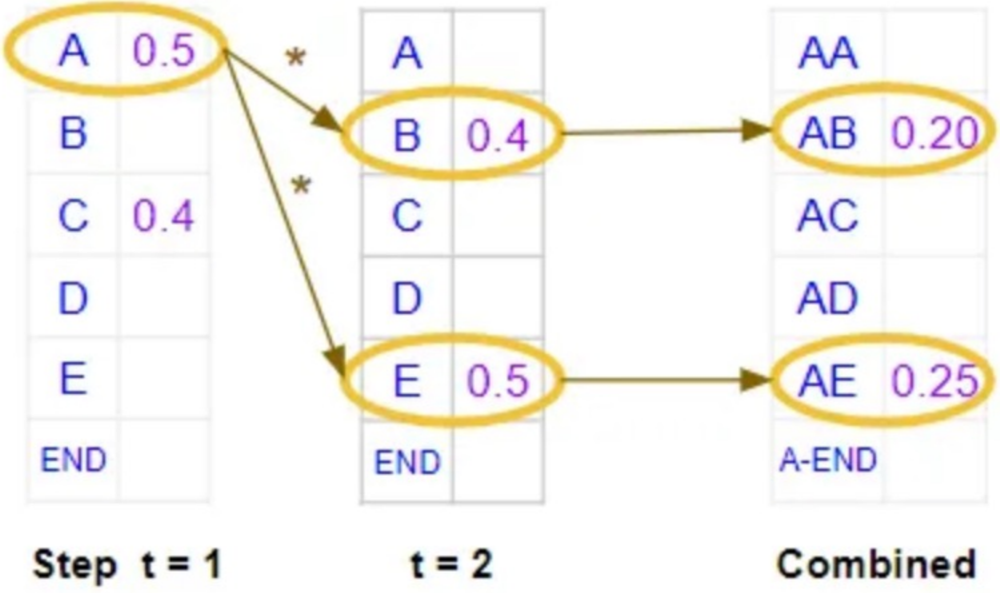
- **Beam search** is a heuristic search algorithm that explores a graph by expanding the most promising node in a limited set.
- Beginning from the start state in some search space, the possible successor states are generated and keep only the "**best**" *k* candidates.
- Then generate all the successors for those *k* states, again keep just the **top *k*** among these options, and so on. When the search is over, the best solution found so far.

Source: https://builtin.com/software-engineering-perspectives/beam-search#



Beam Search example, with width = 2 (Image by Author)

https://towardsdatascience.com/foundations-of-nlp-explained-visually-beam-search-how-it-works-1586b9849a24



A	0.5
B	
C	0.4
D	
E	
END	

Step t = 1

A	
B	0.4
C	
D	
E	0.5
END	

t = 2

AA	
AB	0.20
AC	
AD	
AE	0.25
A-END	

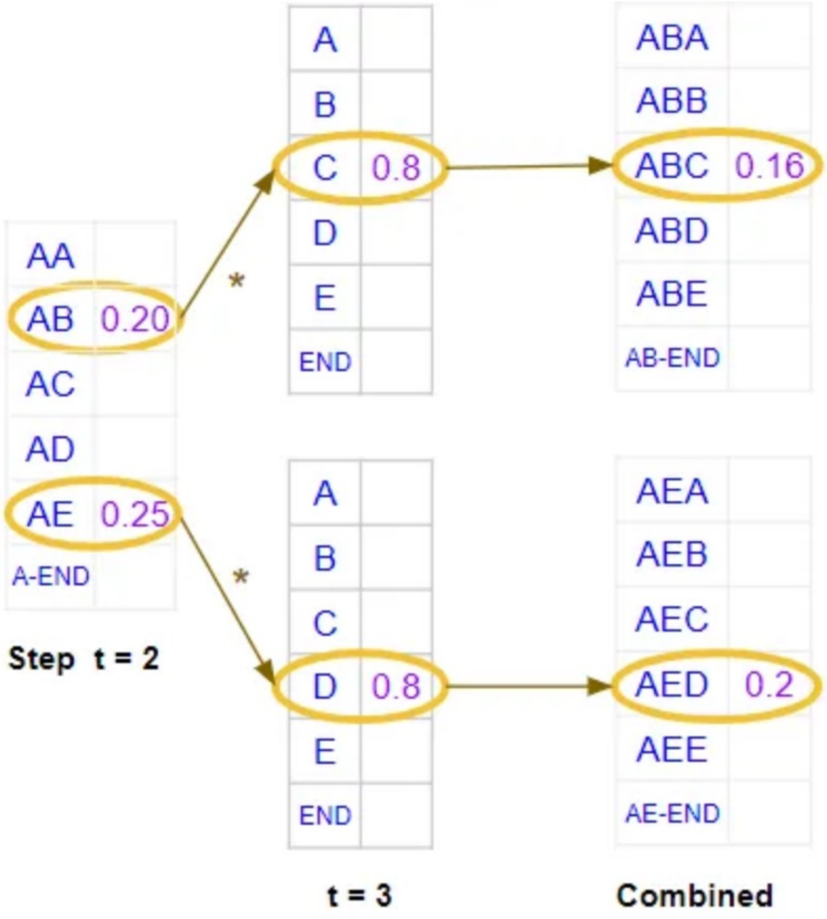
Combined

Prob (AB | input) = Prob (A | input) * Prob (B | A, input)

Prob (AB) = Prob (A) * Prob (B | A)

= 0.5 * 0.4

= 0.20



Prob (ABC) = Prob (AB) * Prob (C | AB)

= 0.2 * 0.8 = 0.16

Prob (AED) = Prob (AE) * Prob (D | AE)

= 0.25 * 0.8 = 0.2

Then chooses the sequence that has the **highest combined probability** to make its final prediction.

Towards Explanatory Interactive Image Captioning using Top-Down and Bottom-Up Features, Beam Search and Re-ranking

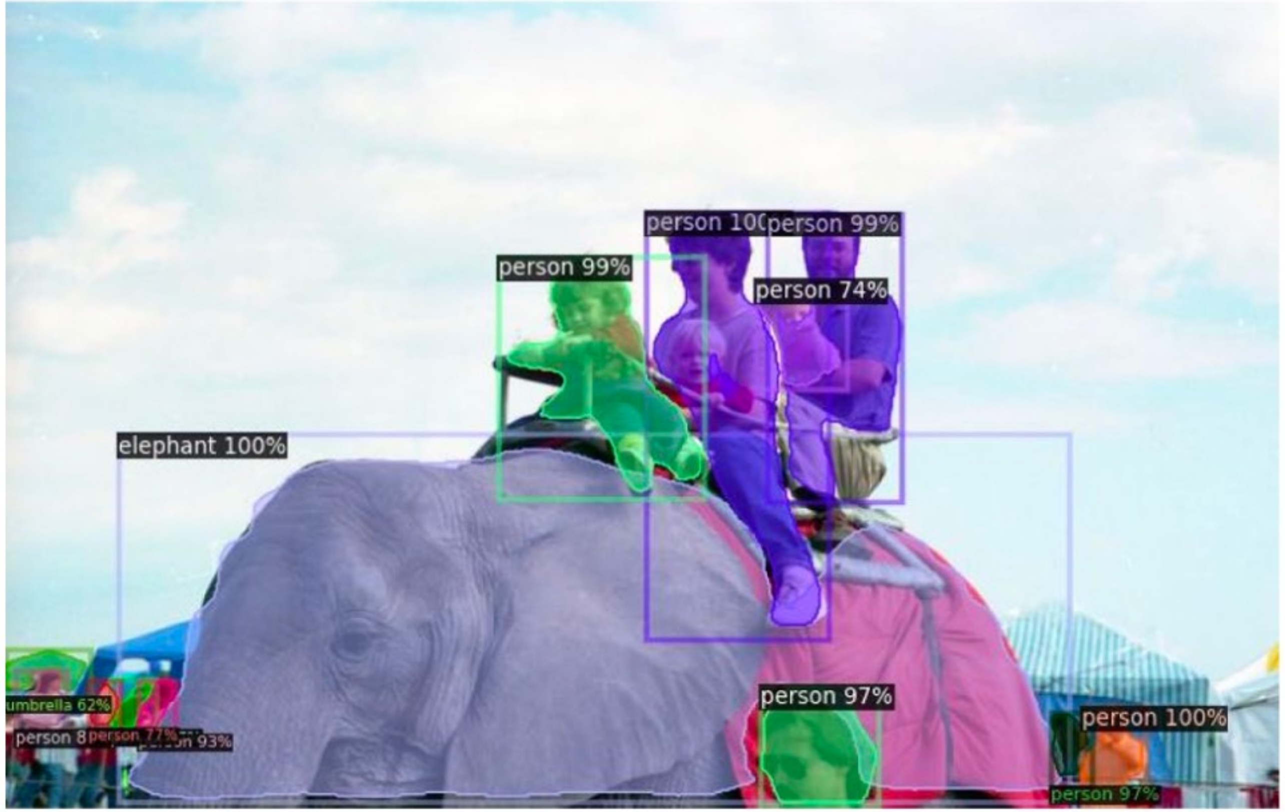
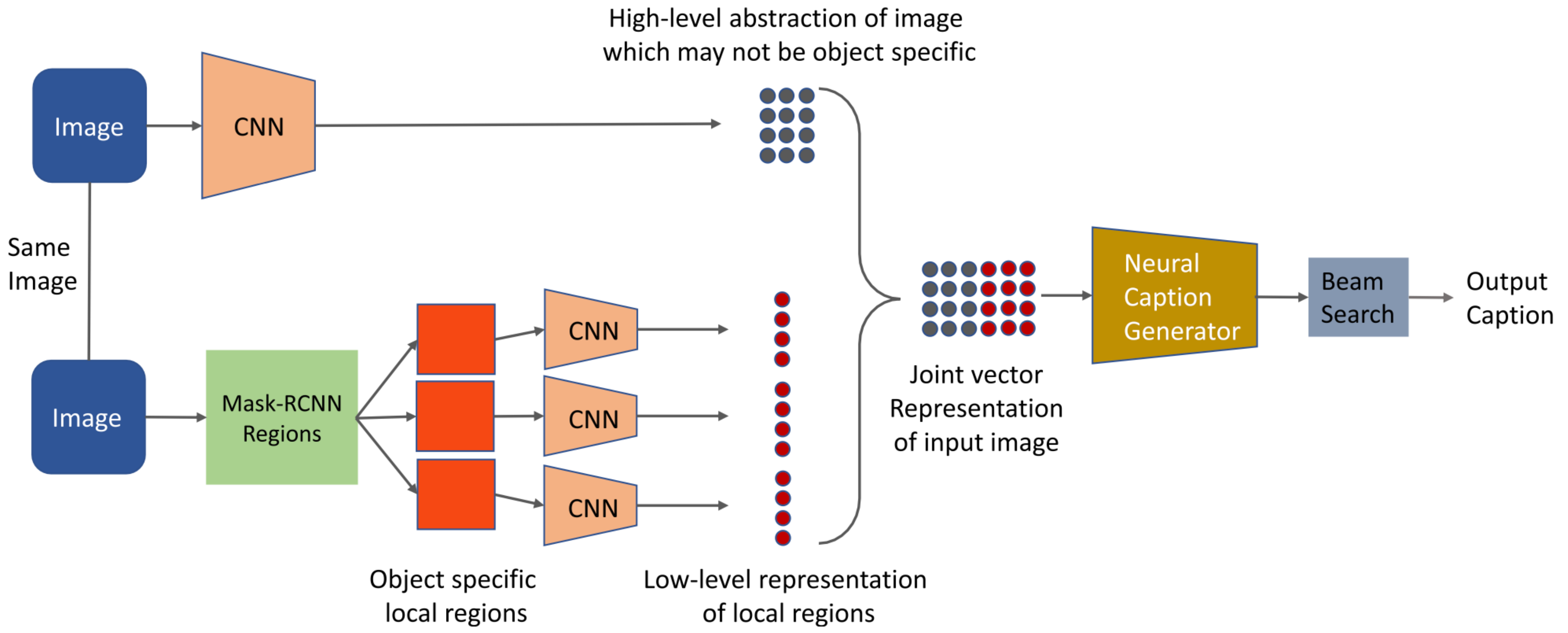


Image with caption generated using our approach: “**a group of people riding on top of an elephant**”

Source: https://link.springer.com/article/10.1007/s13218-020-00679-2

Caption generation with augmented visual attention


High-level abstraction of image which may not be object specific
Image
CNN
Same Image
Image
Mask-RCNN Regions
CNN
CNN
CNN
Joint vector Representation of input image
Neural Caption Generator
Beam Search
Output Caption
Object specific local regions
Low-level representation of local regions

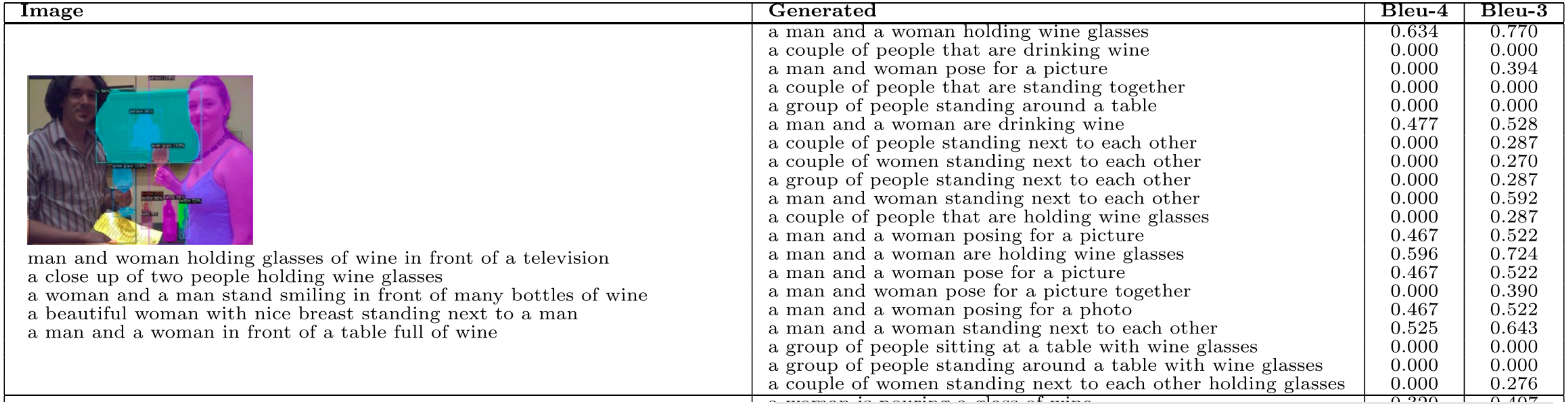


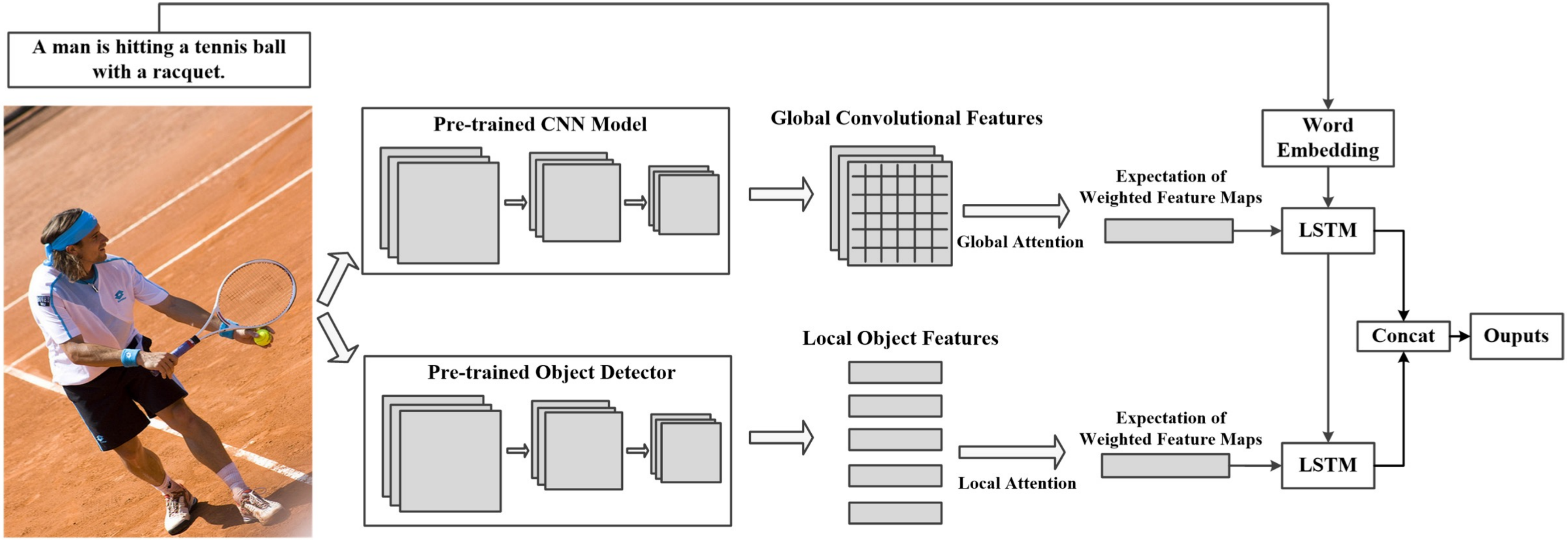
Image	Generated	Bleu-4	Bleu-3
man and woman holding glasses of wine in front of a television a close up of two people holding wine glasses a woman and a man stand smiling in front of many bottles of wine a beautiful woman with nice breast standing next to a man a man and a woman in front of a table full of wine	a man and a woman holding wine glasses	0.634	0.770
	a couple of people that are drinking wine	0.000	0.000
	a man and woman pose for a picture	0.000	0.394
	a couple of people that are standing together	0.000	0.000
	a group of people standing around a table	0.000	0.000
	a man and a woman are drinking wine	0.477	0.528
	a couple of people standing next to each other	0.000	0.287
	a couple of women standing next to each other	0.000	0.270
	a group of people standing next to each other	0.000	0.287
	a man and woman standing next to each other	0.000	0.592
	a couple of people that are holding wine glasses	0.000	0.287
	a man and a woman posing for a picture	0.467	0.522
	a man and a woman are holding wine glasses	0.596	0.724
	a man and a woman pose for a picture	0.467	0.522
	a man and woman pose for a picture together	0.000	0.390
	a man and a woman posing for a photo	0.467	0.522
	a man and a woman standing next to each other	0.525	0.643
	a group of people sitting at a table with wine glasses	0.000	0.000
	a group of people standing around a table with wine glasses	0.000	0.000
	a couple of women standing next to each other holding glasses	0.000	0.276
	a woman is pouring a glass of wine	0.320	0.407

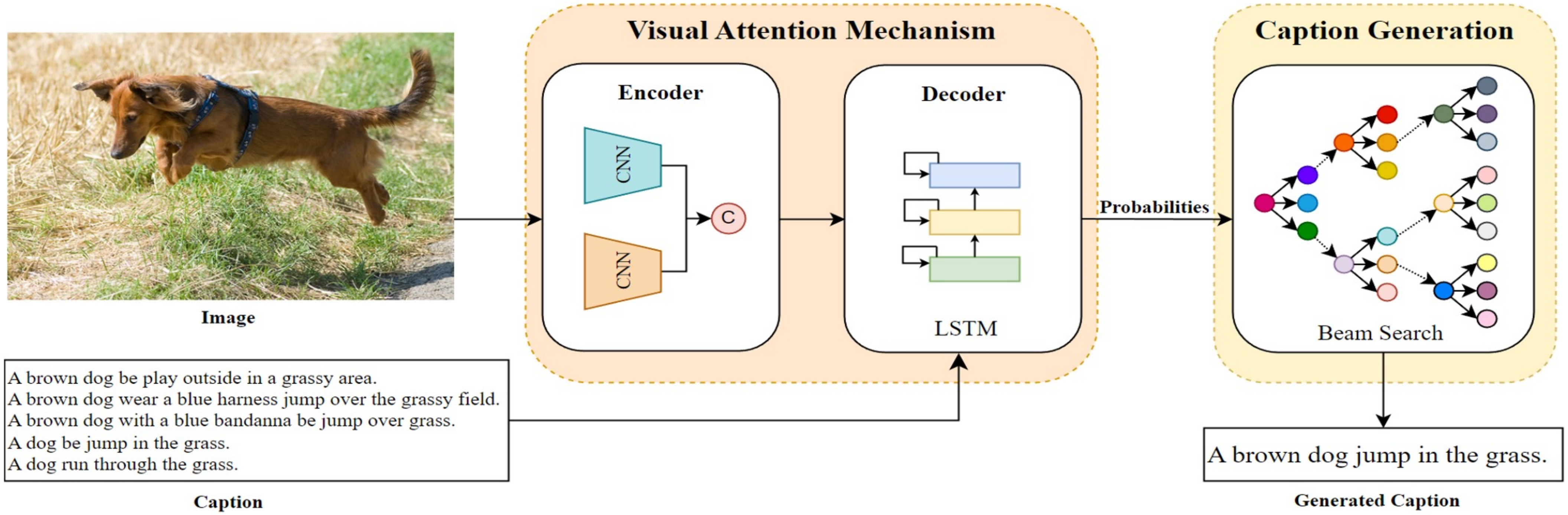
Image Captioning via a Hierarchical Attention Mechanism and Policy Gradient Optimization

- The hierarchical attention mechanism consists of two parts:
  - **Spatial attention mechanism** corresponds to global CNN features
  - **Local attention mechanism** which corresponds to object features.

Source: https://arxiv.org/pdf/1811.05253.pdf

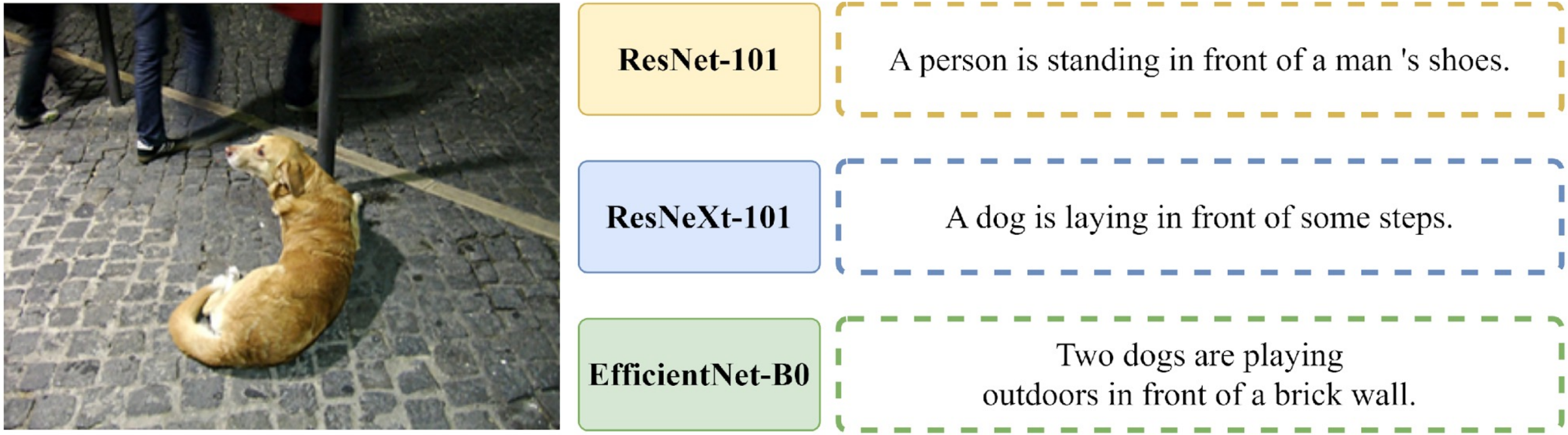
Hierarchical Attention Mechanism


A man is hitting a tennis ball with a racquet.
Pre-trained CNN Model
Global Convolutional Features
Word Embedding
Expectation of Weighted Feature Maps
Global Attention
LSTM
Concat
Ouputs
Local Object Features
Pre-trained Object Detector
Expectation of Weighted Feature Maps
Local Attention
LSTM

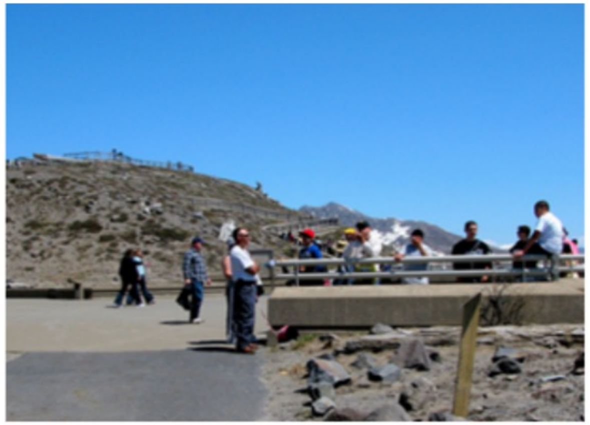
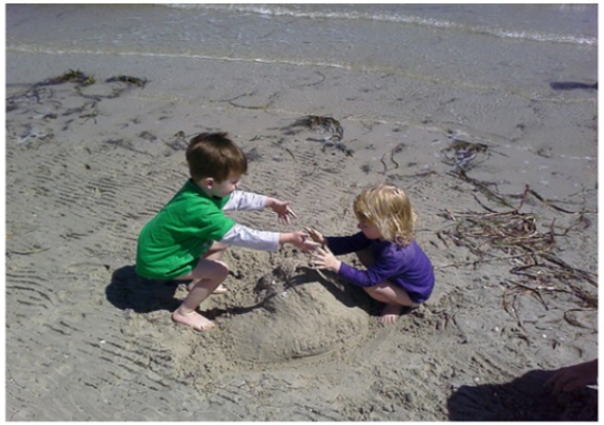
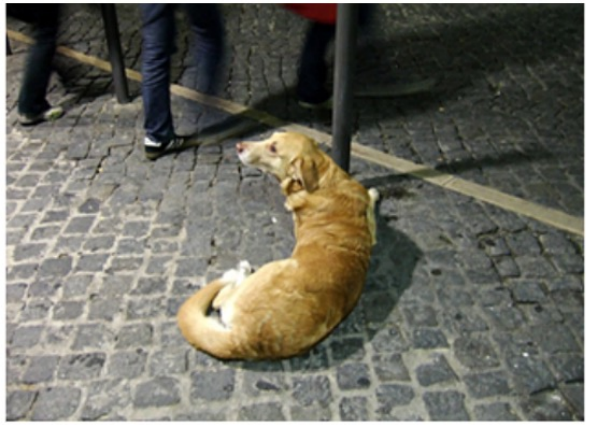


A brown dog be play outside in a grassy area.
A brown dog wear a blue harness jump over the grassy field.
A brown dog with a blue bandanna be jump over grass.
A dog be jump in the grass.
A dog run through the grass.

A brown dog jump in the grass.



ResNet-101	A person is standing in front of a man 's shoes.
ResNeXt-101	A dog is laying in front of some steps.
EfficientNet-B0	Two dogs are playing outdoors in front of a brick wall.

			
Ground truths	- A group of people on a walkway. - A group of people on a zig zagging path through the mountains - People are standing around a scenic lookout on a sunny day. - People stand at the bottom of a meandering walkway that goes uphill. - tourists are standing a mountain viewpoint beneath a clear blue sky.	- A little girl and boy play in the sand on the beach. - A young boy and girl play together in the sand. - boy and girl make a sand castle - Two children build a sand castle on the beach. - two kids building a sand castle	- A brown dog is sitting on a cobbled pavement. - A brown god relaxes on a brick sidewalk. - A dog lies down on a cobblestone street. - A yellow dog is lying near where people are walking. - The dog is lying on the cobblestone street.
Dual-CNN	A group of people are walking along a rocky path.	Three children playing in the sand.	A brown dog is laying on the sidewalk.
BLEU Scores	**BLEU-1:** 66.41 **BLEU-2:** 42.65 **BLEU-3:** 33.63 **BLEU-4:** 25.48	**BLEU-1:** 71.43 **BLEU-2:** 59.76 **BLEU-3:** 52.62 **BLEU-4:** 43.47	**BLEU-1:** 77.78 **BLEU-2:** 62.36 **BLEU-3:** 48.43 **BLEU-4:** 36.89

Code

- https://shorturl.at/cBVX9 (Using the model)
- https://shorturl.at/ahmxz (Training the model)
- https://shorturl.at/akqFL (Fused-CNN)

Thank you for your attention