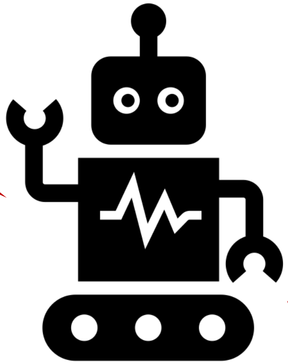
Learning from vision and language

Zero-shot learning

- Question: can we teach a machine to recognize classes based on just textual descriptions?

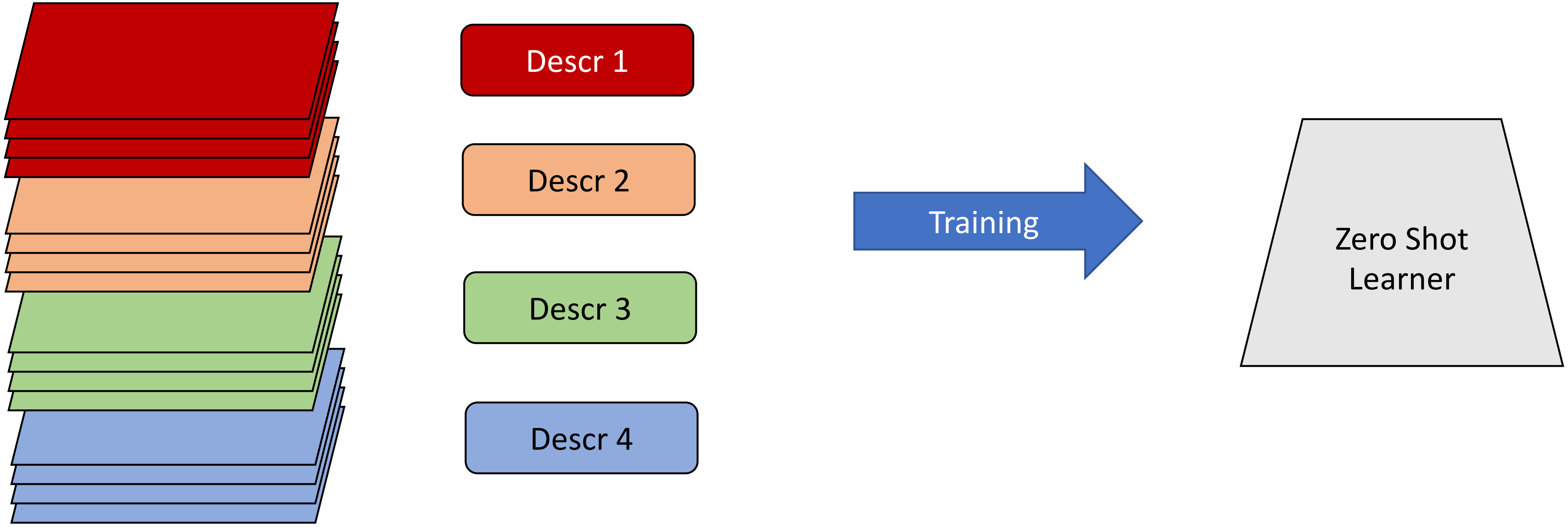
An okapi is like a horse but with stripes on its legs

Gotcha



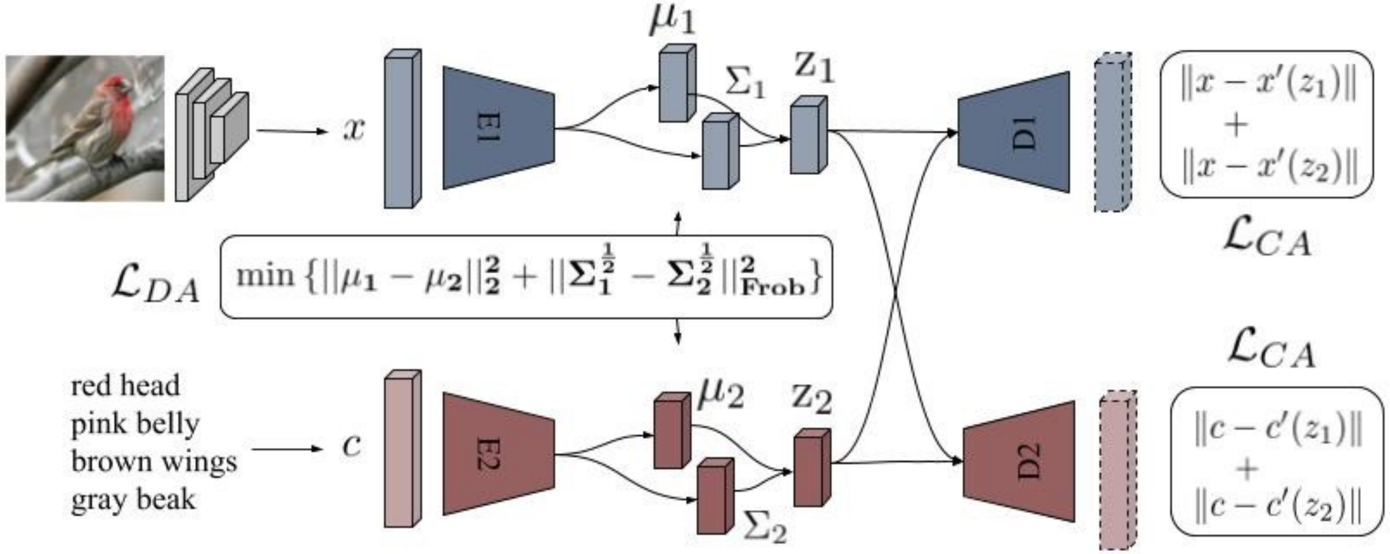
Okapi!

Zero-shot learning – Training the learner



Zero-shot learning – training the learner

- General approach: align embeddings of images and words



Zero-shot learning

- Typically zero-shot learning techniques use attribute descriptions of classes
- But can use textual descriptions as well
- Need a text encoder

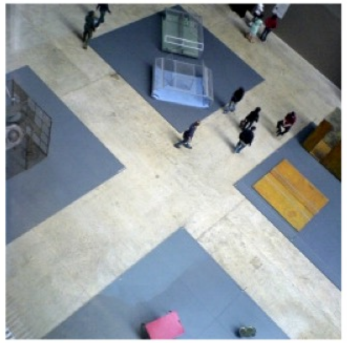
Weak supervision

- Zero-shot learning is typically performed on a particular domain
- Can we do "generic" zero-shot learning?

The world of internet images and captions

the veranda hotel portixol palma

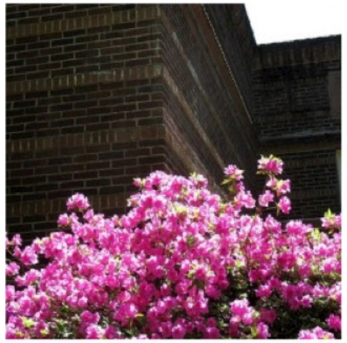
plane approaching zrh avro regional jet rj



not as impressive as embankment that s for sure

student housing by lungaard tranberg architects in copenhagen click here to see where this photo was taken

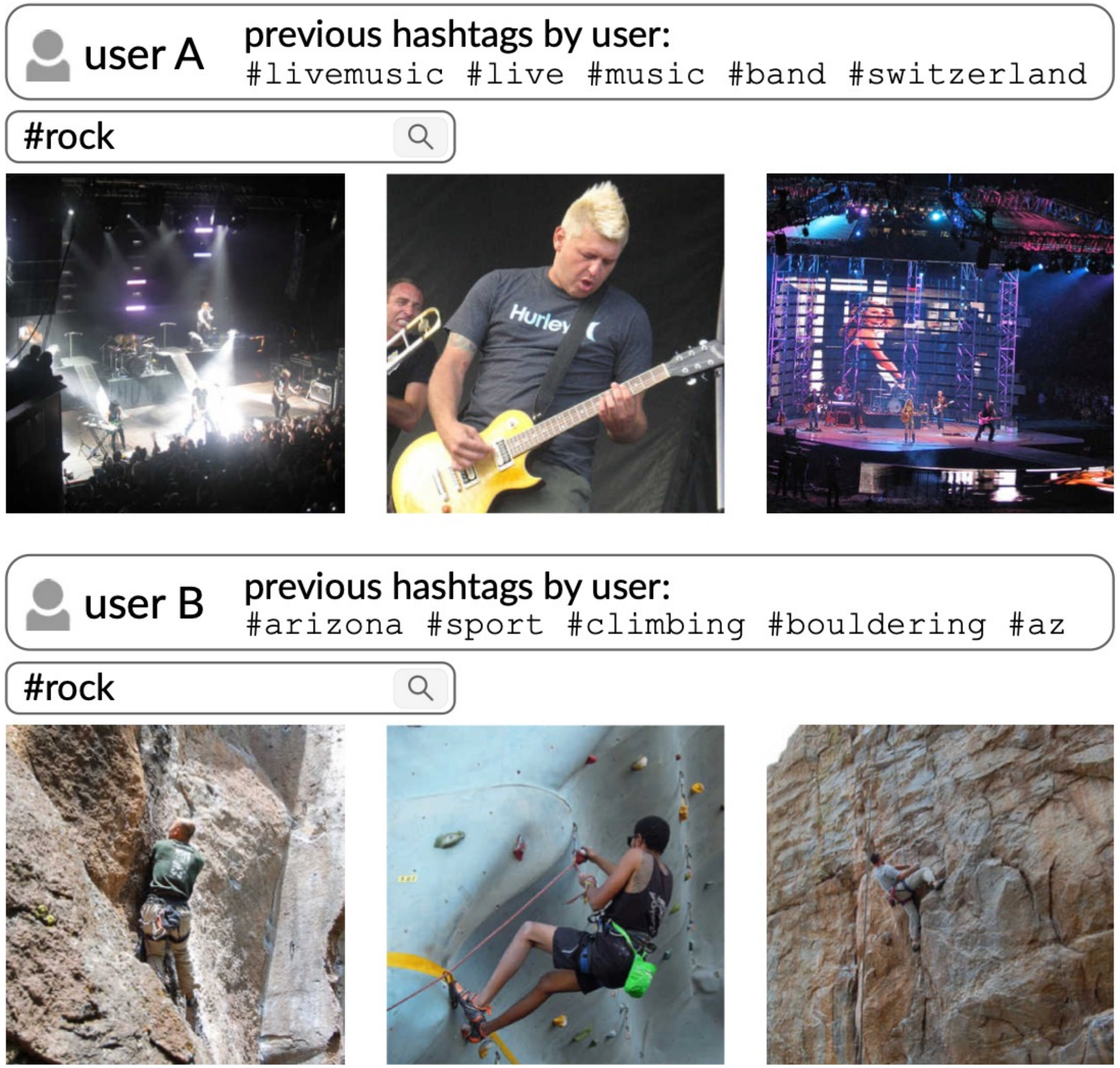
article in the local paper about all the unusual things found at otto s home



this was another one with my old digital camera i like the way it looks for some things though slow and lower resolution than new cameras another problem is that it s a bit of a brick to carry and is a pain unless you re carrying a bag with some room it s nearly x x and weighs ounces new one is x x and weighs ounces i underexposed this one a bit did exposure bracketing script underexposure on that camera looks melty yummy gold kodak film like

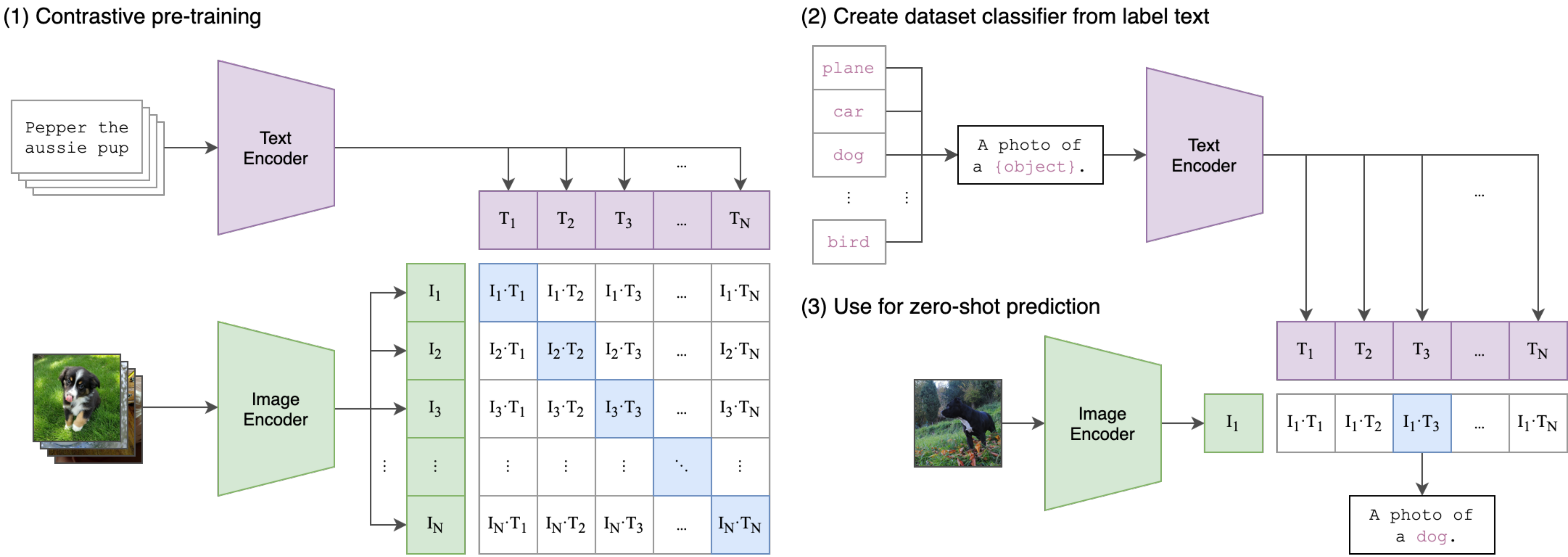
A. Joulin*, L.J.P. van der Maaten*, A. Jabri, and N. Vasilache (*both authors contributed equally). **Learning Visual Features from Large Weakly Supervised Data**. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67-84, 2016.

The world of internet images and captions



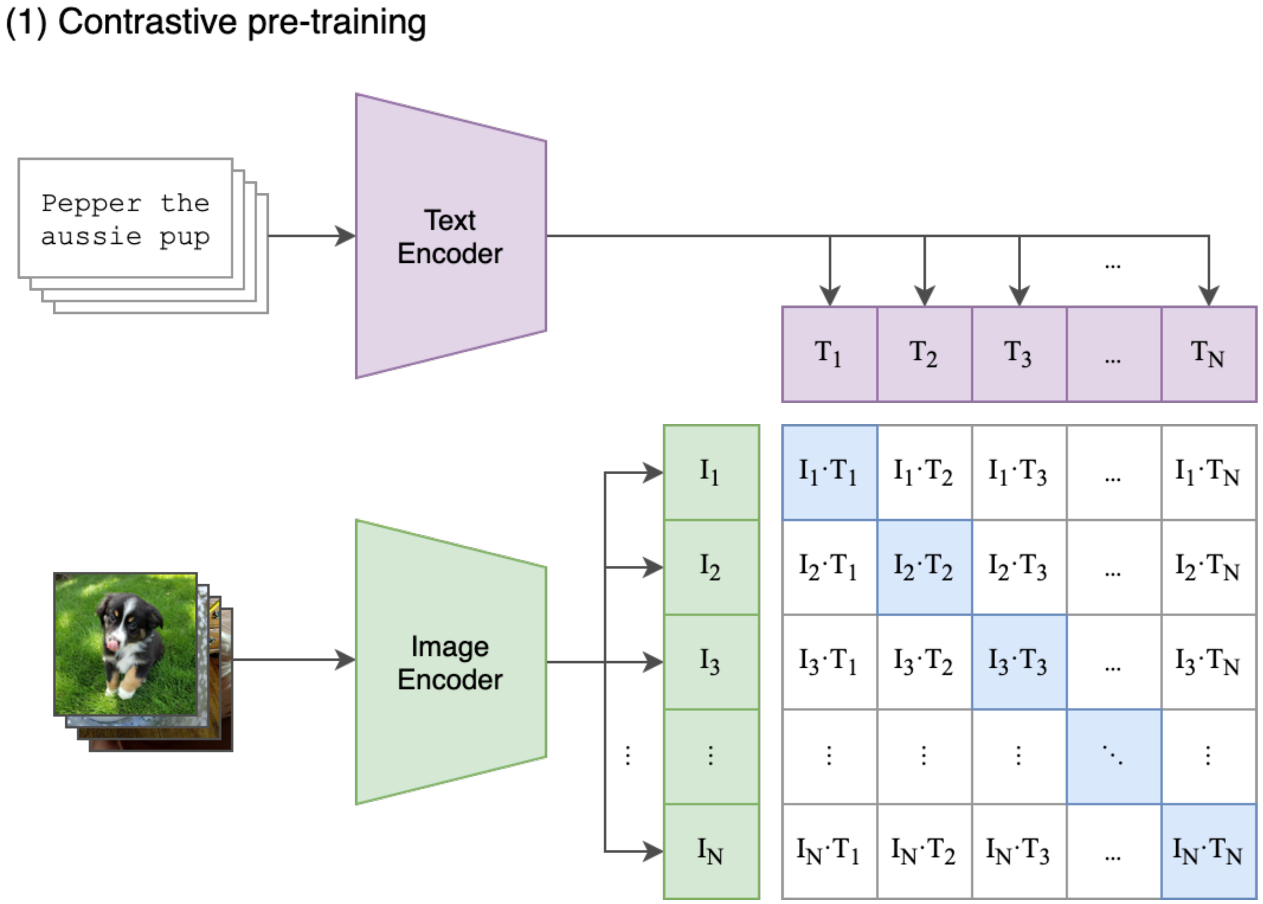
A. Veit, M. Nickel, S. Belongie, and L.J.P. van der Maaten. **Separating Self-Expression and Visual Content in Hashtag Supervision**. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5919-5927, 2018

Vision-language pre-training



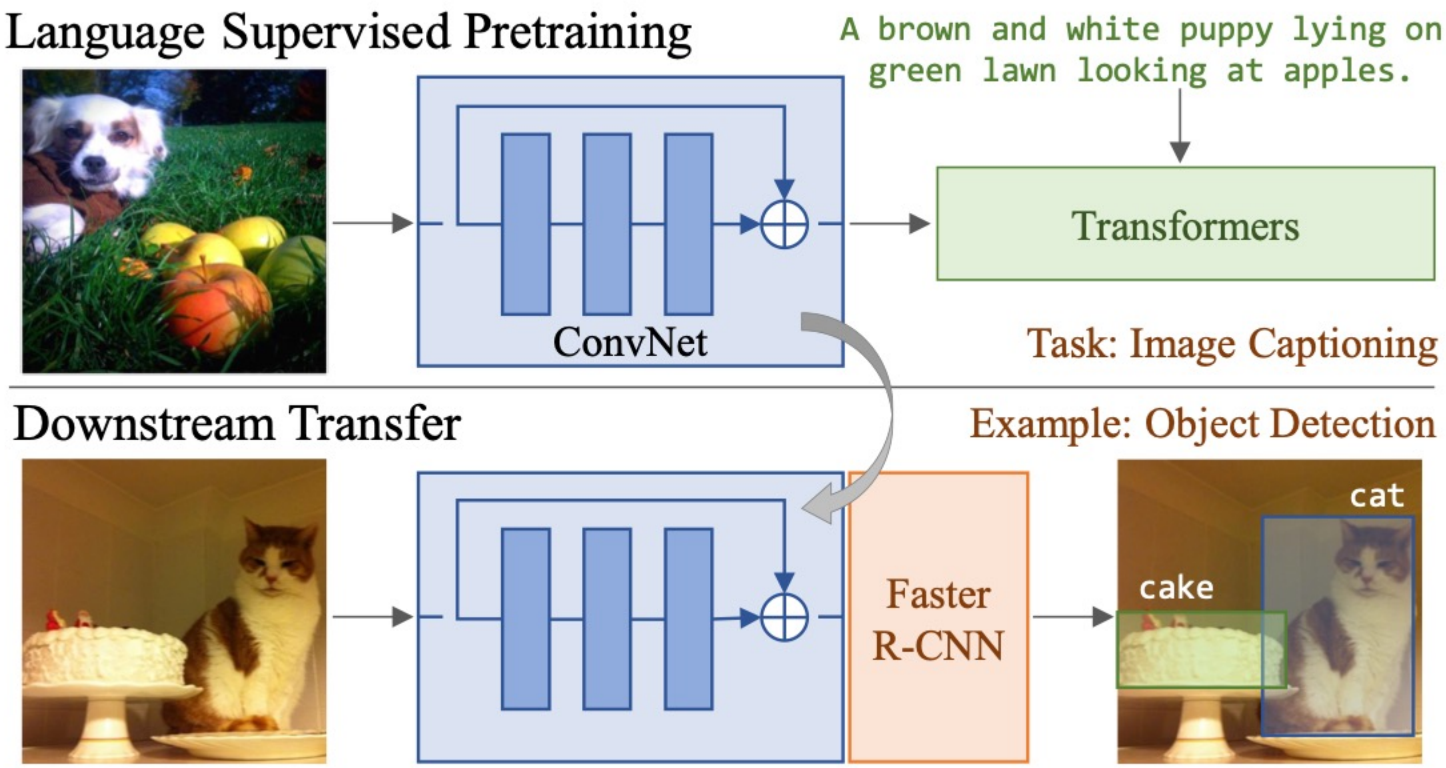
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748-8763). PMLR.

Vision –language pre-training - CLIP




```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

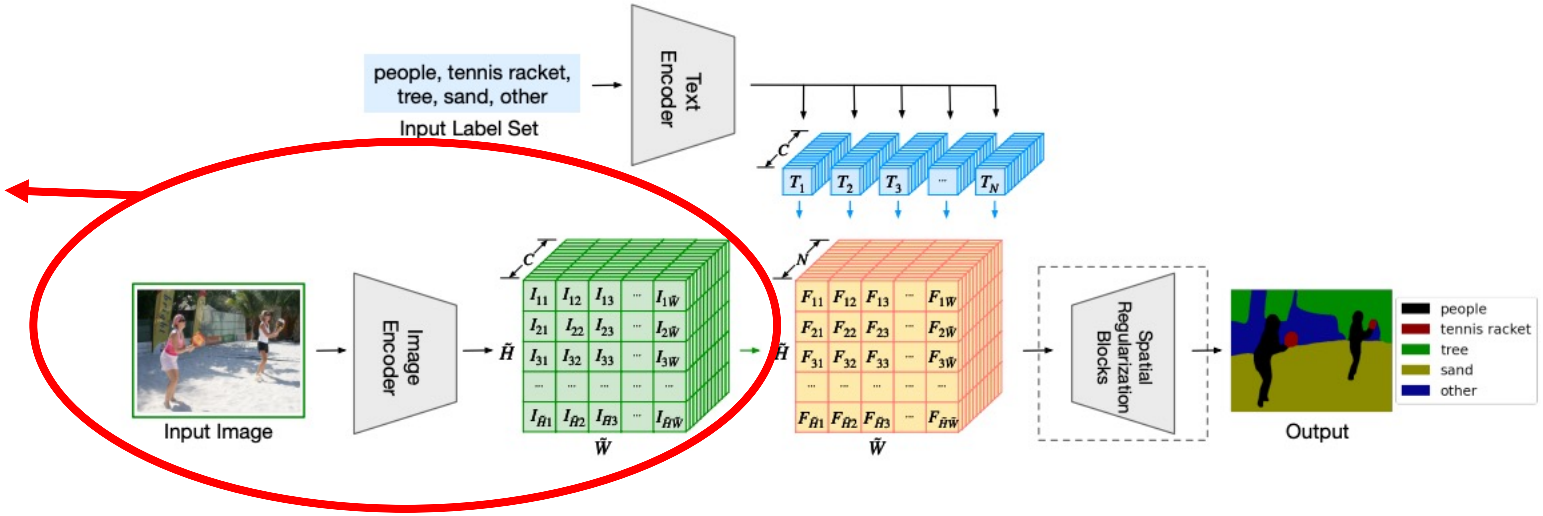
Vision-language pre-training

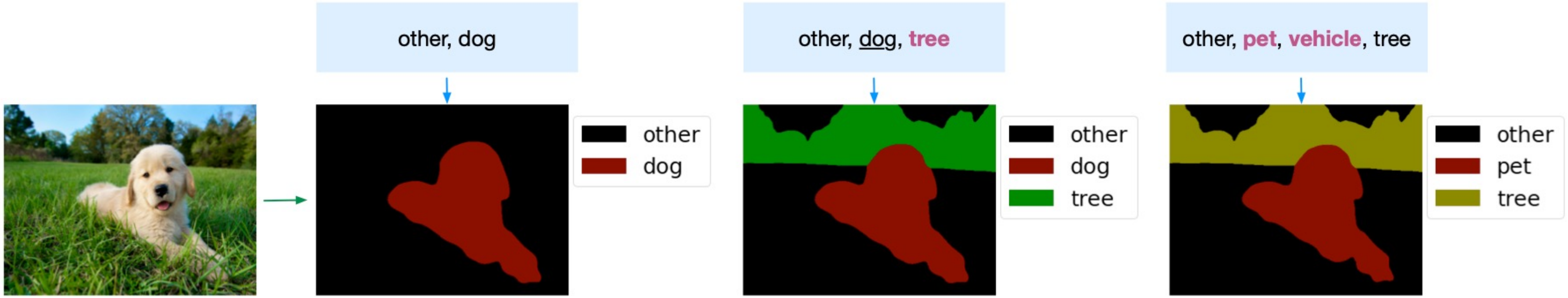


Desai, Karan, and Justin Johnson. "Virtex: Learning visual representations from textual annotations." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.

Using vision-language embeddings

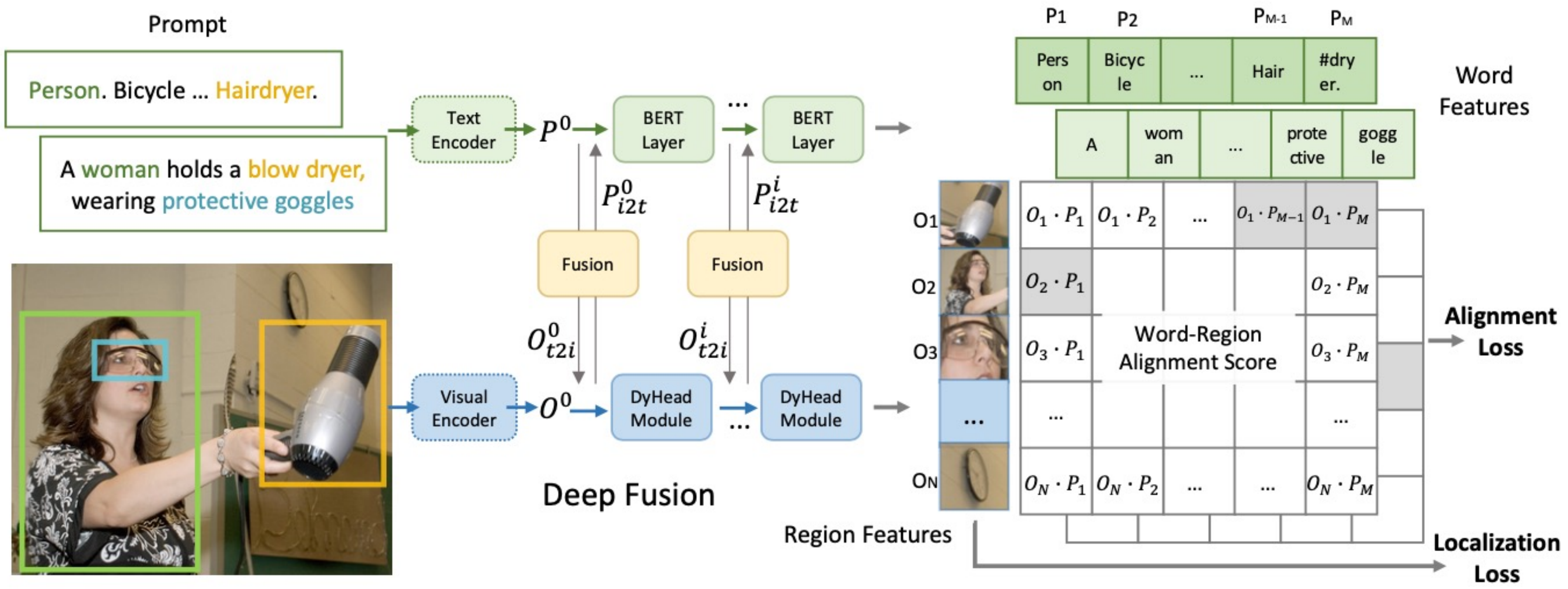
Trained on
COCO
Segmentation





Li, B., Weinberger, K. Q., Belongie, S., Koltun, V., & Ranftl, R. (2021, September). Language-driven Semantic Segmentation. In *International Conference on Learning Representations*.

GLIP : Vision-language for object detection



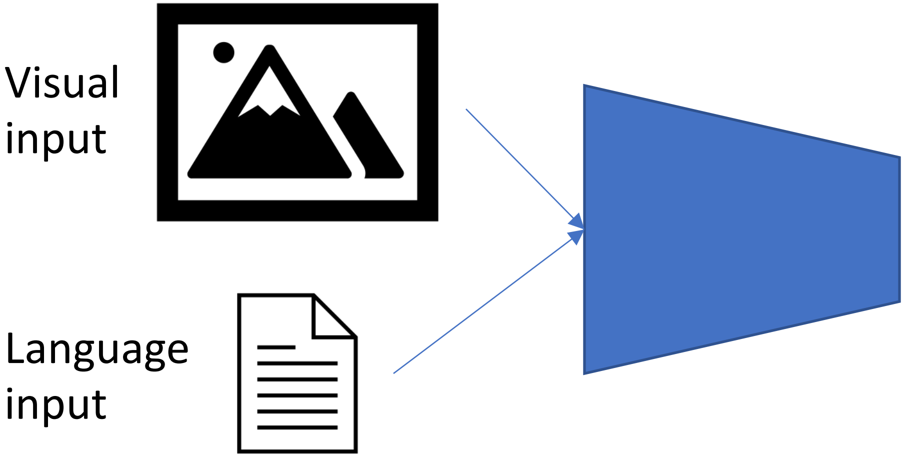
Li, Liunian Harold, et al. "Grounded language-image pre-training." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Vision and Communication with Language

Vision, Language and Embodiment

- Embodied agents have to communicate
- Humans must be able to communicate with machines
- Multimodal input provides strong supervisory signal

The many flavors of vision and language tasks



Grounding
Zero-shot learning
Visual question answering

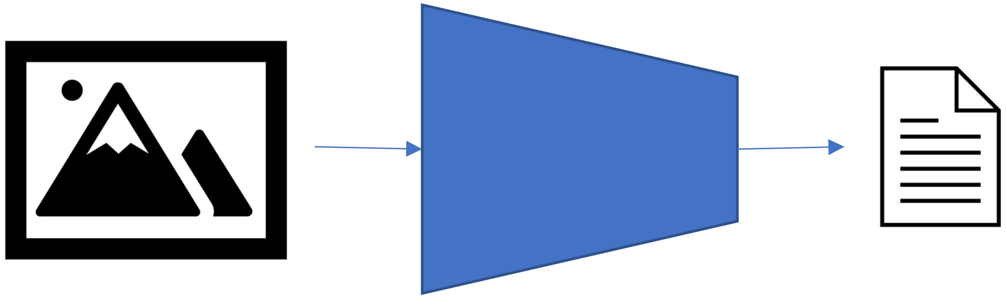


Image captioning

Image captioning - The task

A group of young men playing soccer.

Image captioning - why?

- Alt-text for visually impaired
- Test for true understanding?

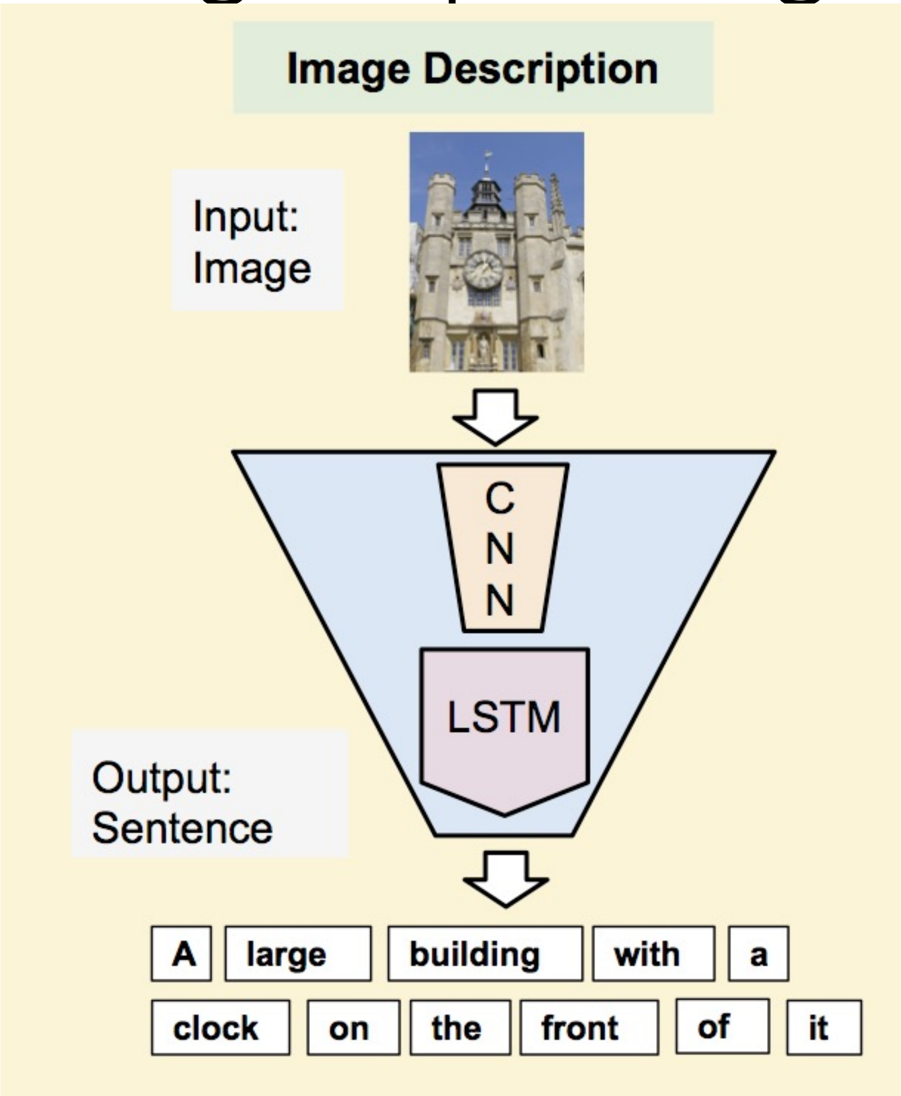
Image captioning - evaluation

- Given computer-generated caption and human caption, compute match
- BLEU from machine translation community
- Computes (modified) n-gram precision

Reference: A group of people playing soccer

Candidate: People playing baseball.
BLEU: 1/3

Image captioning



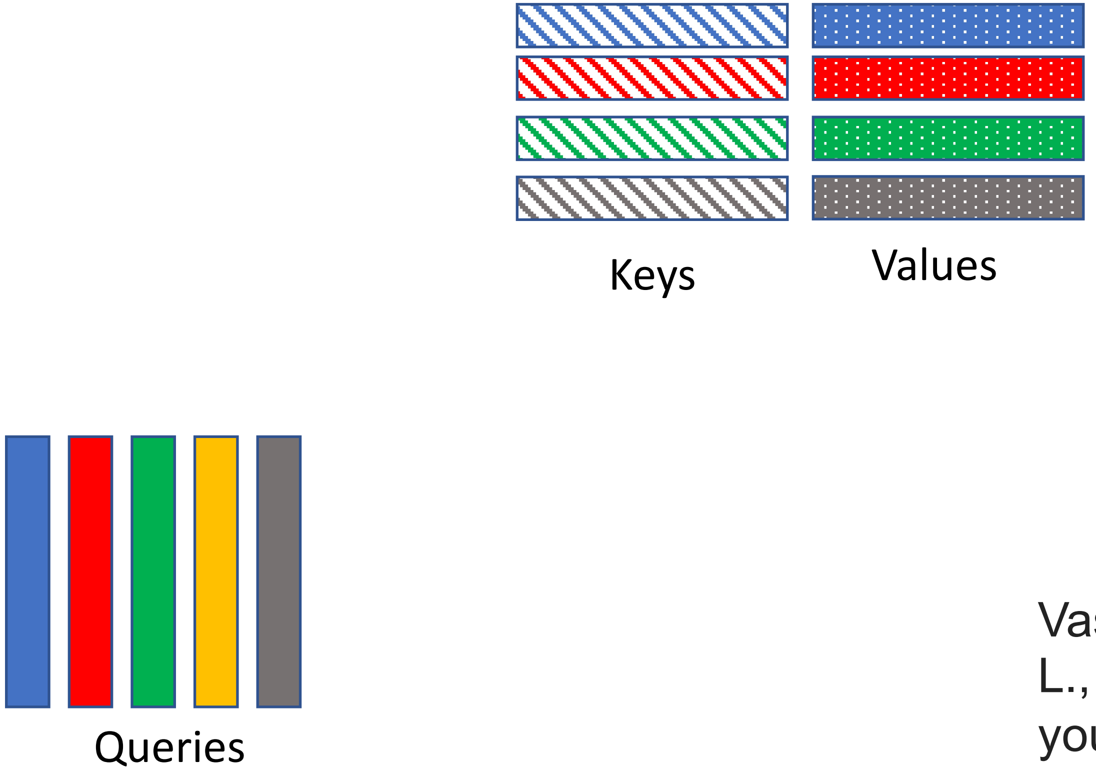
Long-term Recurrent Convolutional Networks. J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell. In *CVPR,* 2015.

Deep Visual-Semantic Alignments for Generating Image Descriptions. Andrej Karpathy and Li Fei-Fei. In *CVPR,* 2015

Show and tell: A neural image caption generator
Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan.
In *CVPR,* 2015.

Attention (Transformers)

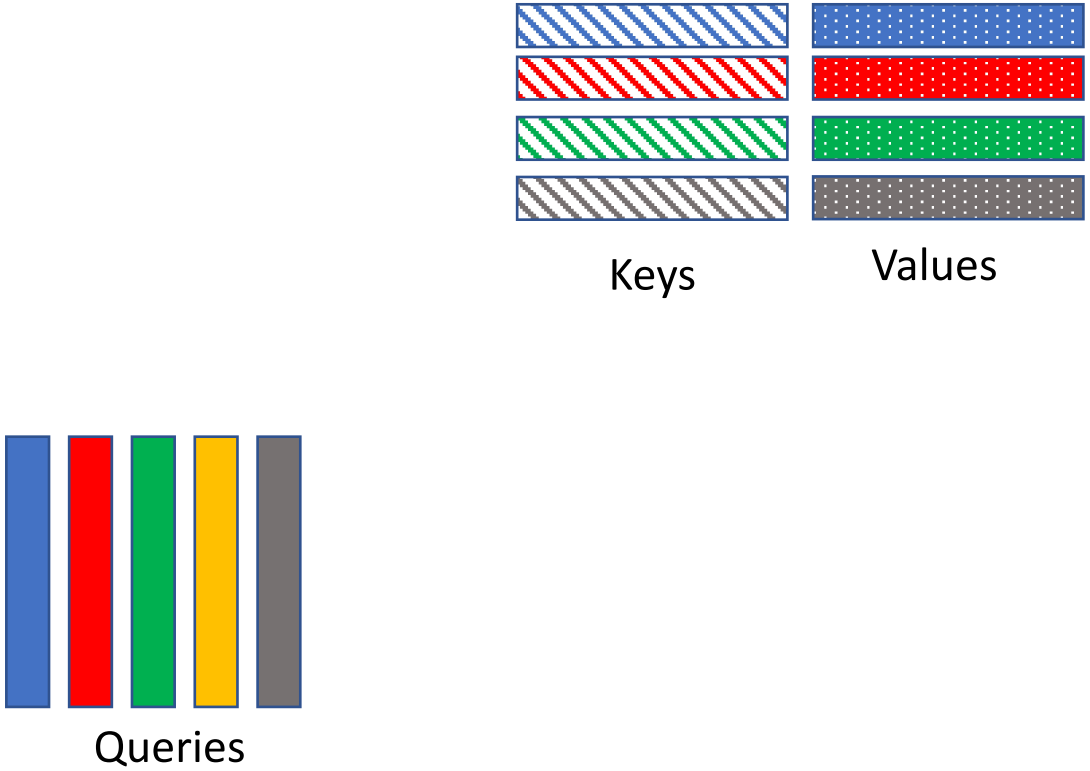
- Comes from the NLP community
- Is an approach for processing sets



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*(pp. 5998-6008).

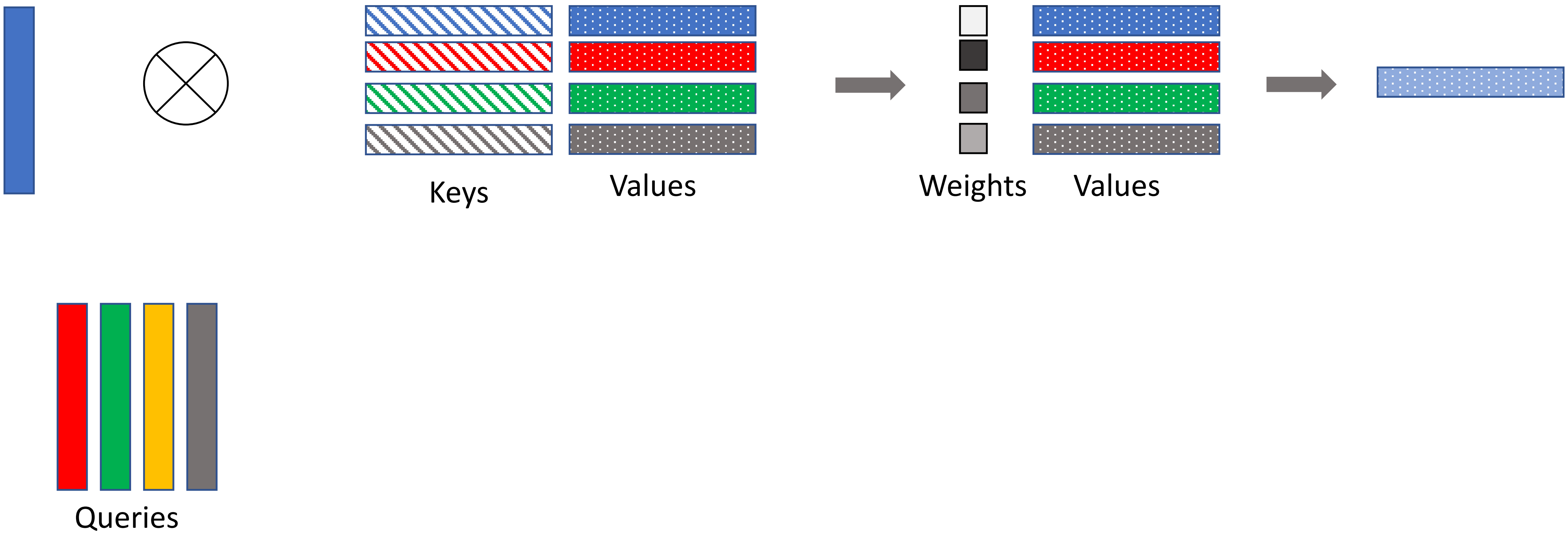
Attention (Transformers)

- Comes from the NLP community
- Is an approach for processing sets

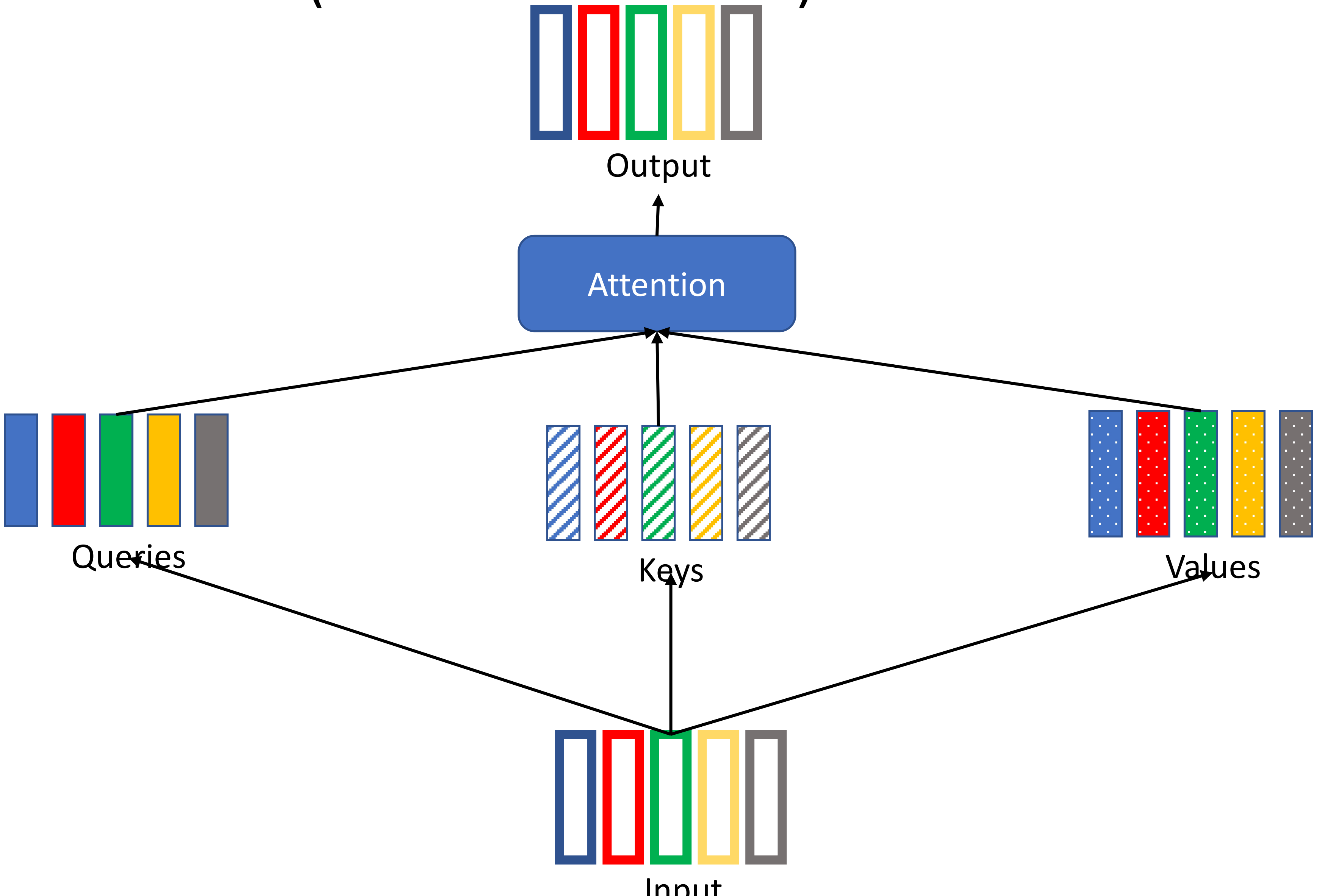


Attention (Transformers)

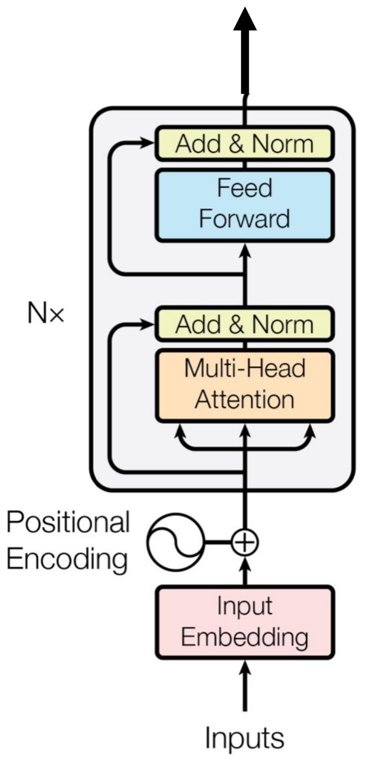
- Comes from the NLP community
- Is an approach for processing sets



Attention (Transformers)

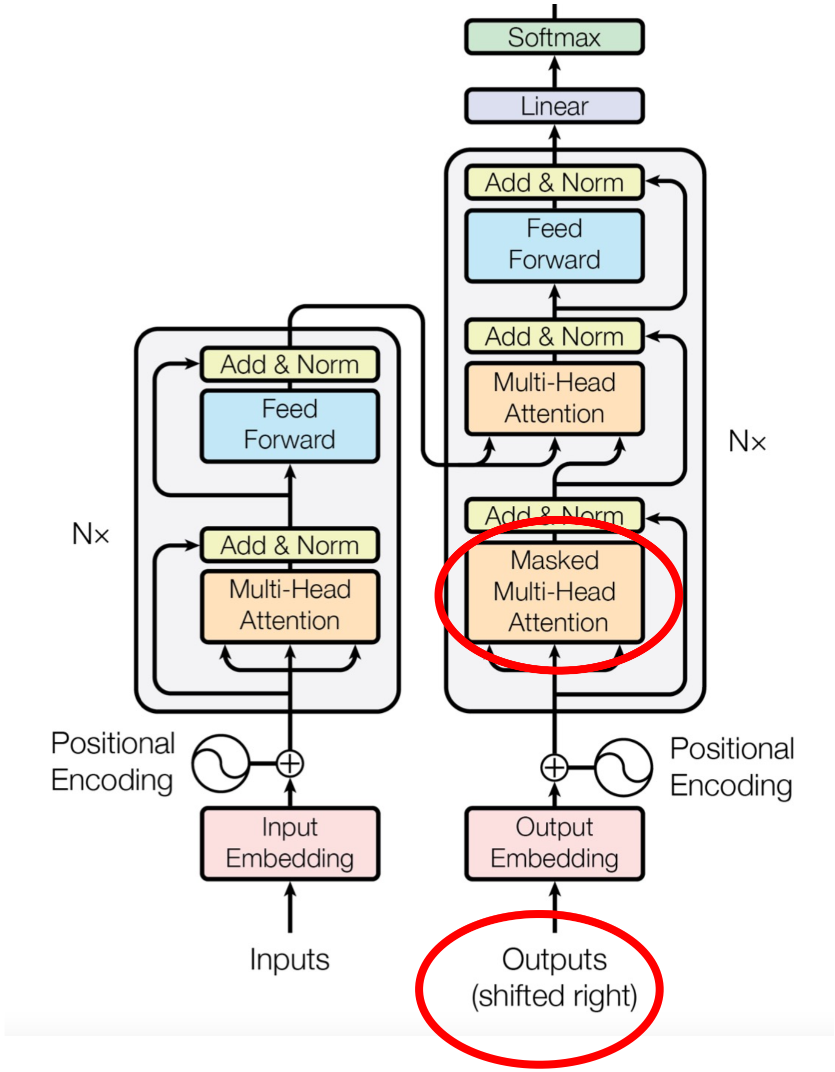


Attention (Transformers) for Encoding sequences



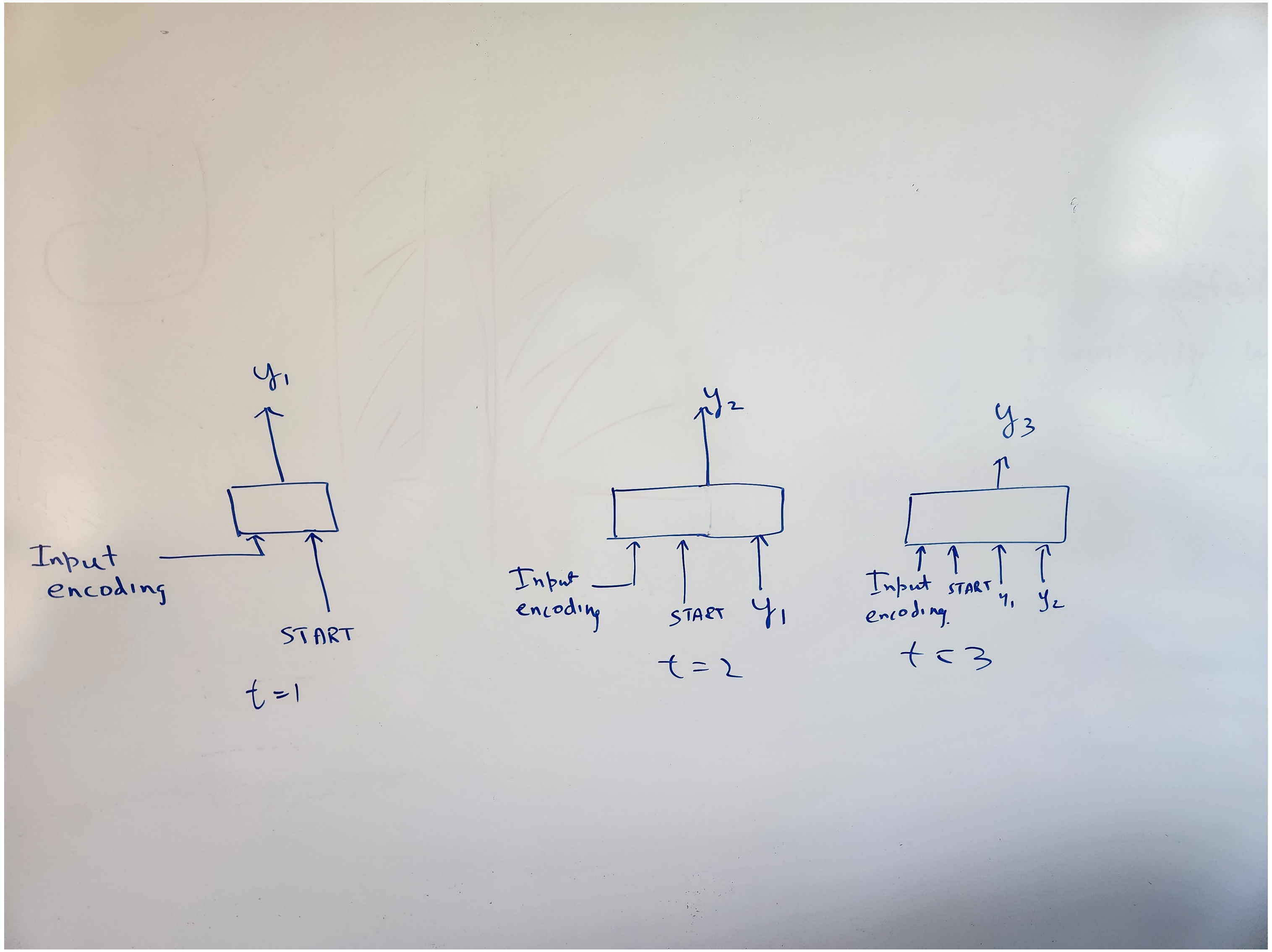
Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

Attention for Outputing Sequences



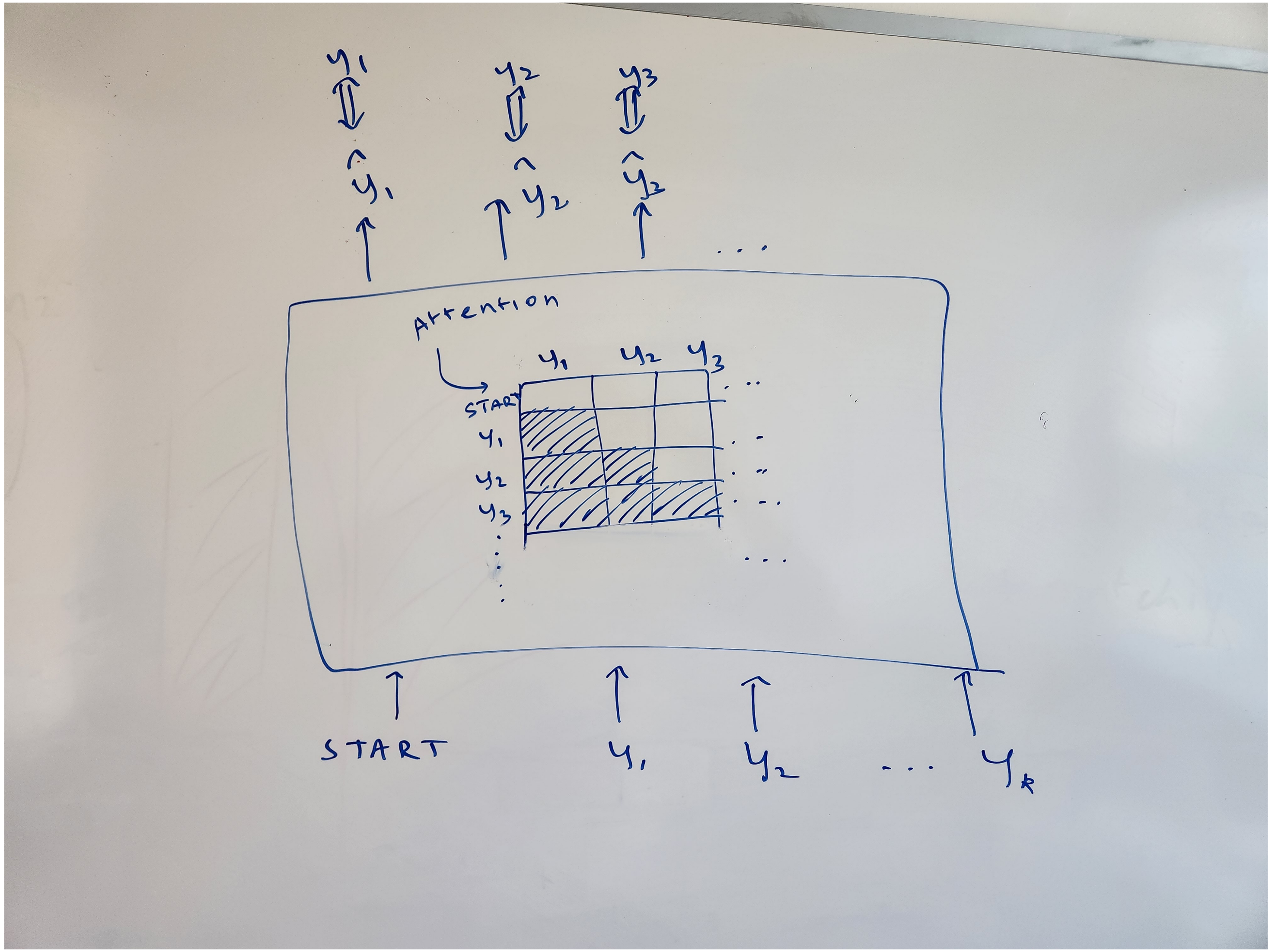
Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

Attention for Outputing Sequences



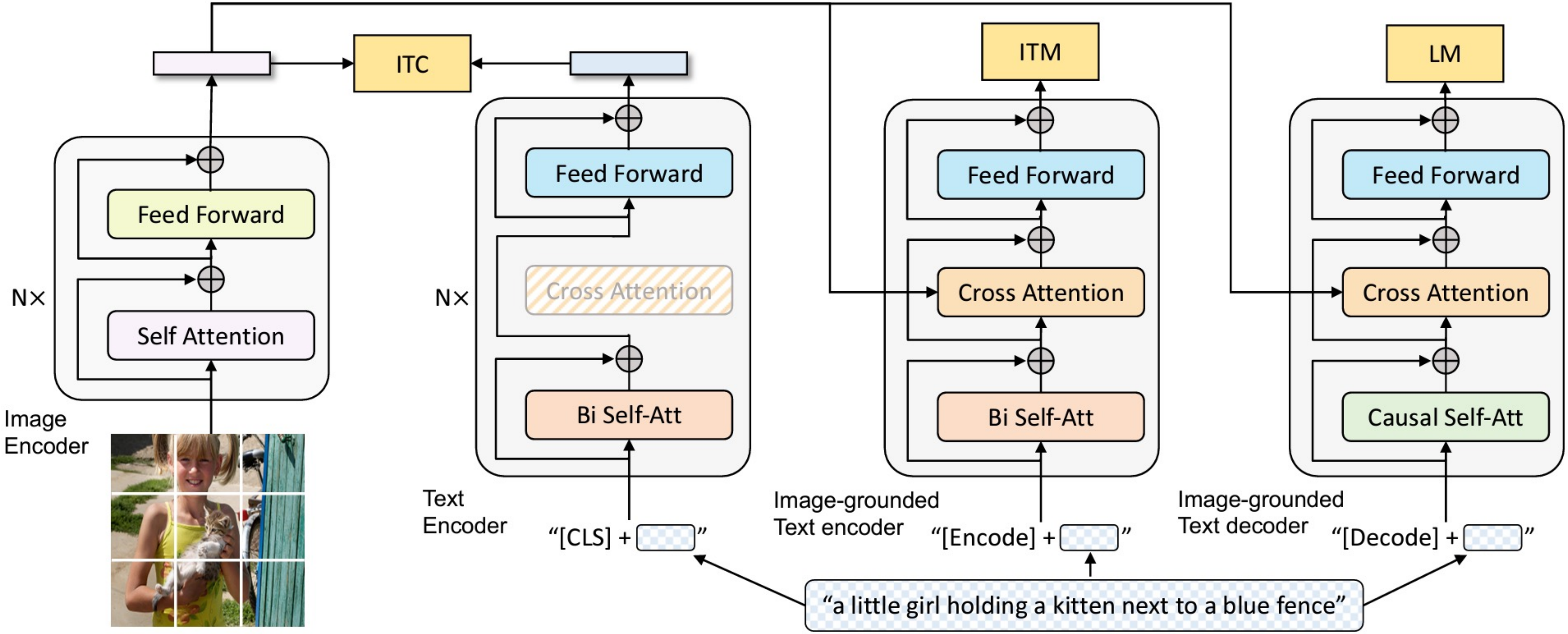
Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

Attention for Outputing Sequences



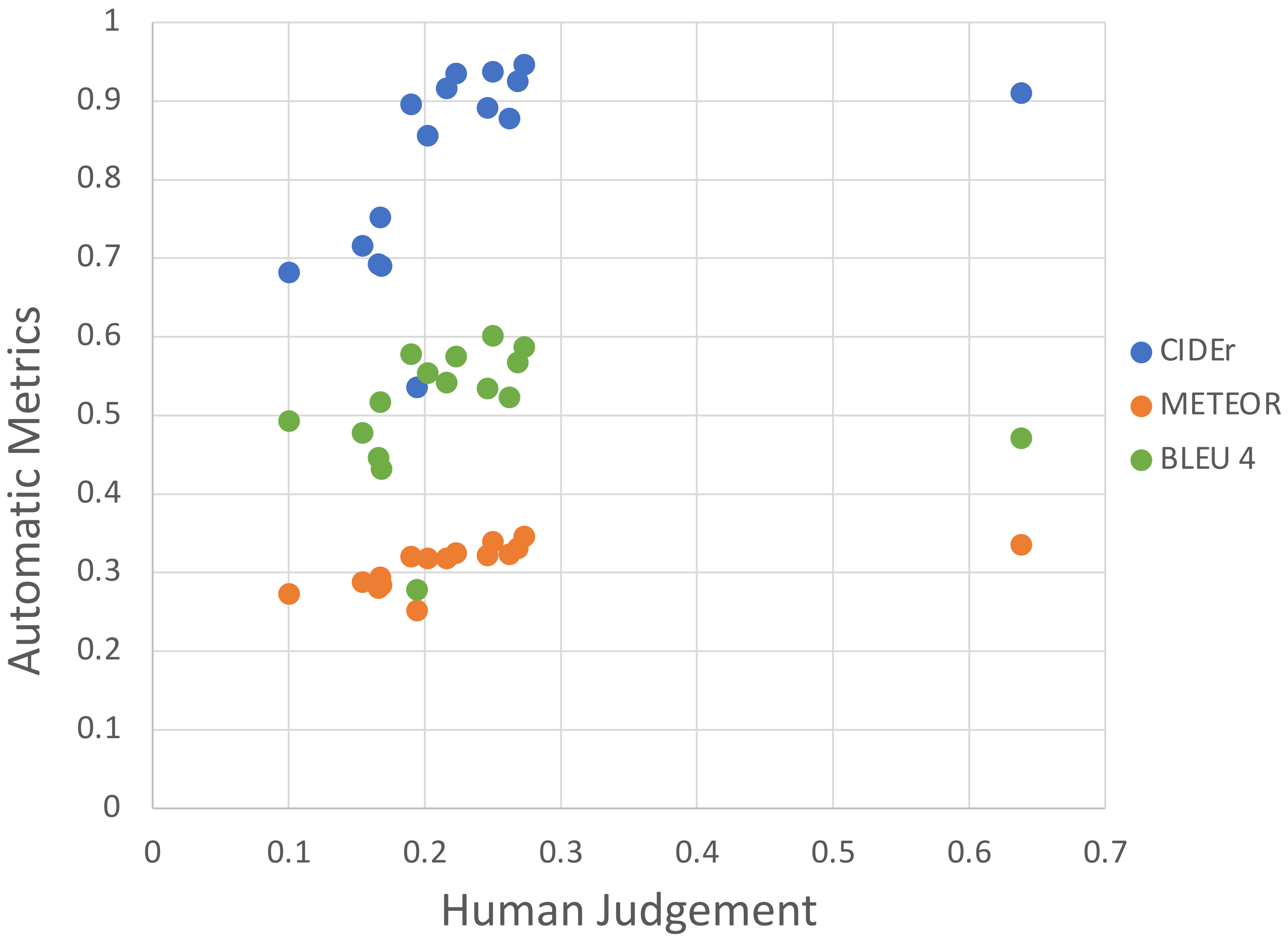
Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

Modern image captioning - BLIP



Li, Junnan, et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." *International Conference on Machine Learning*. PMLR, 2022.

Evaluation Metrics



Slide credit: Larry Zitnick

Evaluation Metrics

Human captions

Slide credit: Larry Zitnick

A man riding a wave on a surfboard in the water.

Slide credit: Larry Zitnick

A man riding a wave on a surfboard in the water.

“surfboard”

Slide credit: Larry Zitnick

The post-captioning world

- Captioning is hard to evaluate!
  - Frame task so that it is easy to evaluate objectively
- Datasets are biased!
  - Control dataset bias

Stephanie Melnick
@unicornsteph96

Follow

I'm going to crush the rebellion... but first, let me take a selfie. #captionbot

I am not really confident, but I think it's a man taking a selfie in front of a building.

Reasoning

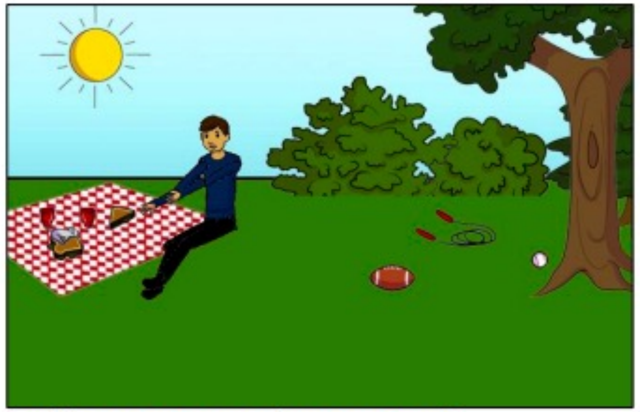
- Want vision systems to reason about what is going on
  - Identify objects and scenes
  - Identify relationships between objects
  - Understand physics of the world
  - Understand social interactions, intent etc.
  - Incorporate knowledge: common sense, pop culture, …

Visual Question Answering

- Direct motivation: assistive technology
- Indirect motivation: sandbox for reasoning

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

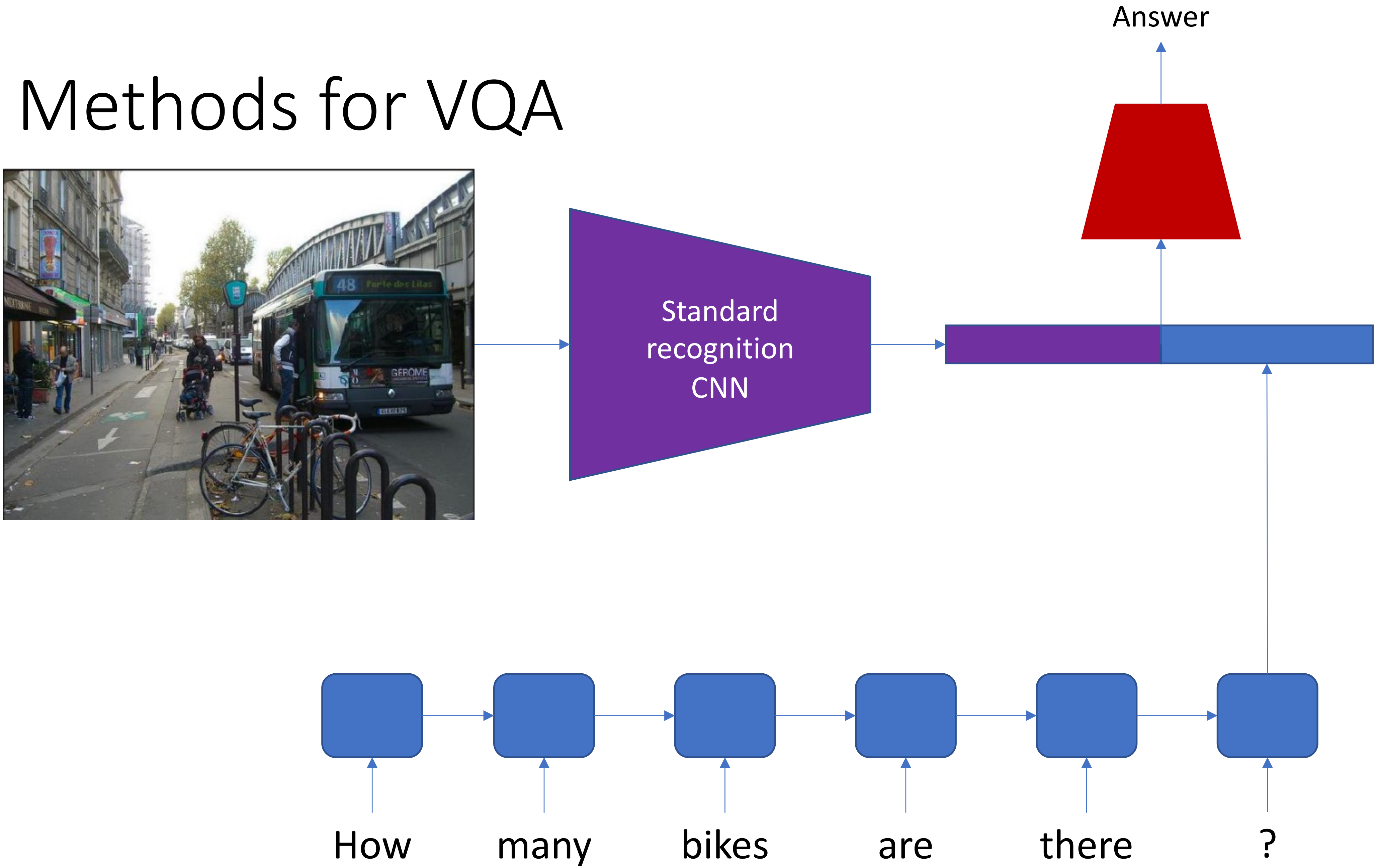
Does it appear to be rainy?
Does this person have 20/20 vision?

VQA: Visual Question Answering. Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh. In *ICCV,* 2015.

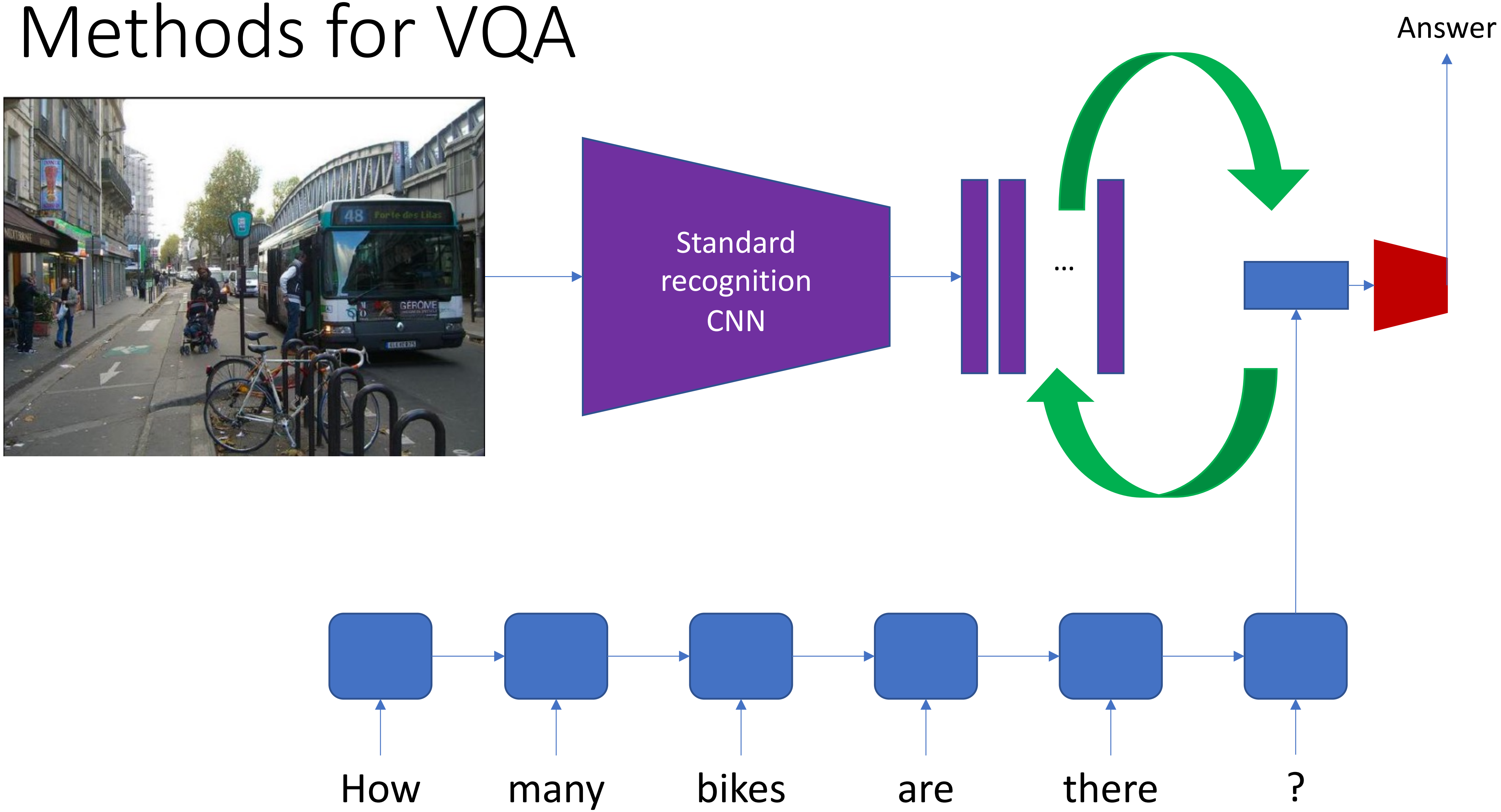
Visual Question Answering

"We have built a smart robot. It understands a lot about images. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene (e.g., kitchen, beach), people's expressions and poses, and properties of objects (e.g., color of objects, their texture). Your task is to stump this smart robot! Ask a question about this scene that this smart robot probably can not answer, but any human can easily answer while looking at the scene in the image."

Methods for VQA

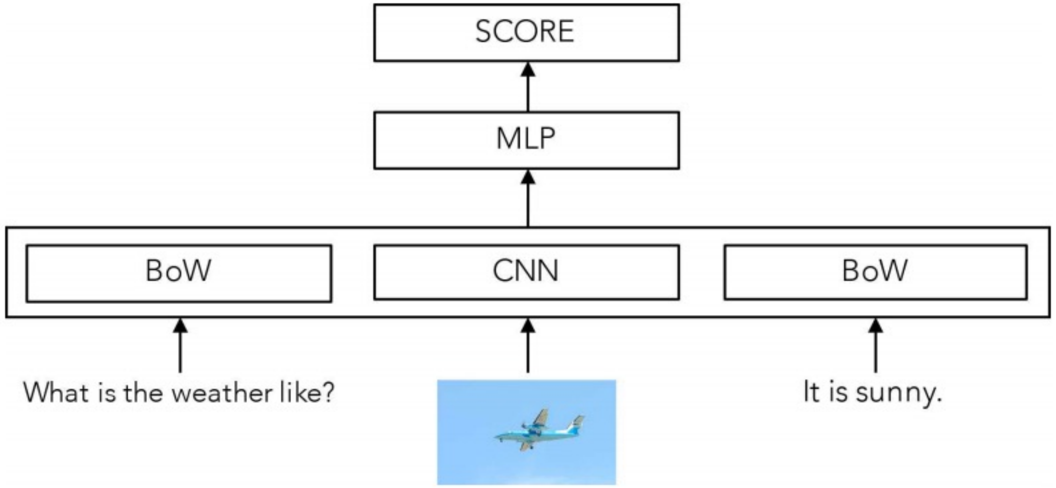


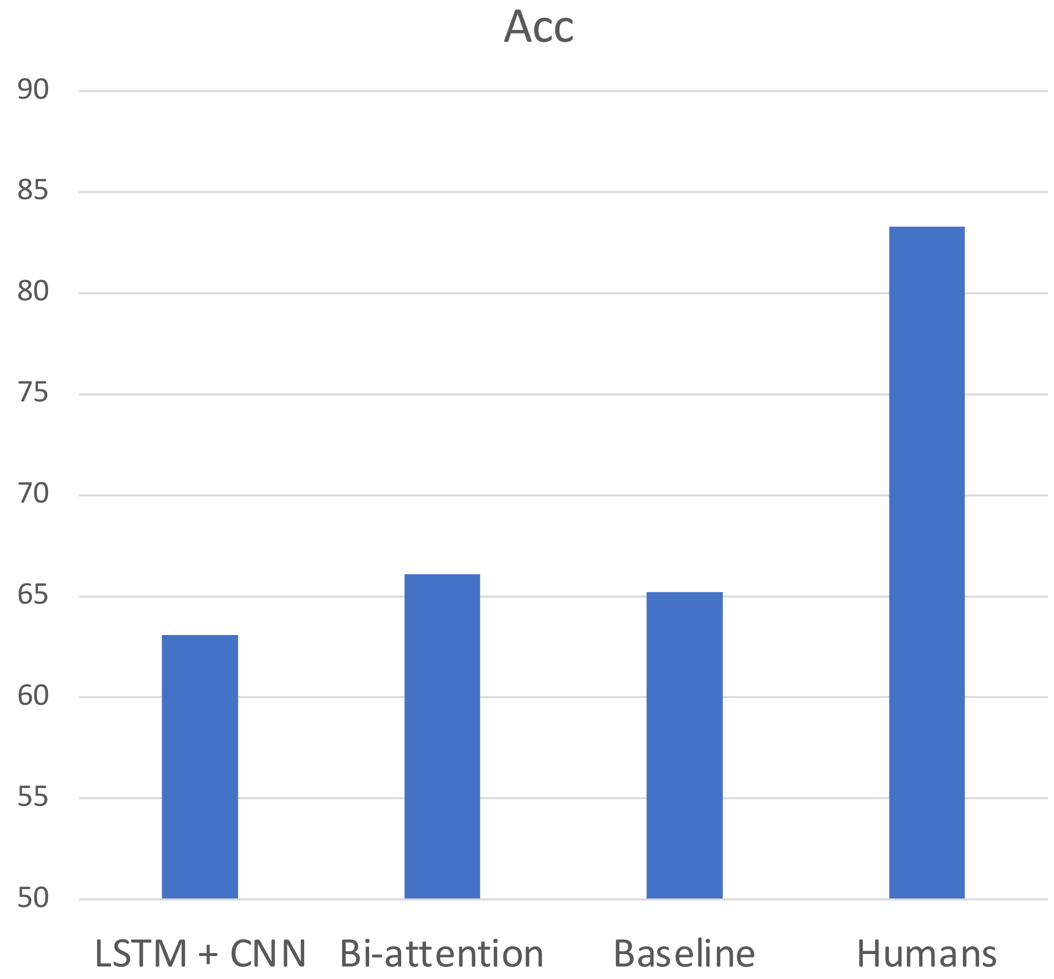
Methods for VQA



Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In CVPR, 2016

The Unreasonable Effectiveness of Baselines





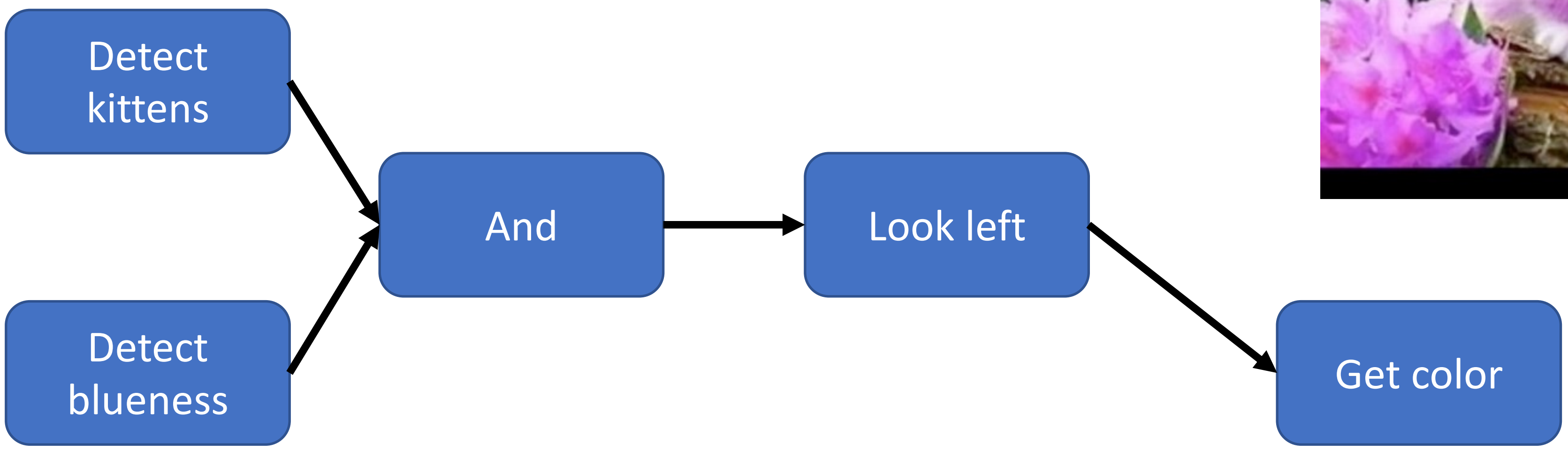
A. Jabri, A. Joulin, L. van der Maaten. Revisiting visual question answering baselines. In *ECCV,* 2016.

Compositional reasoning

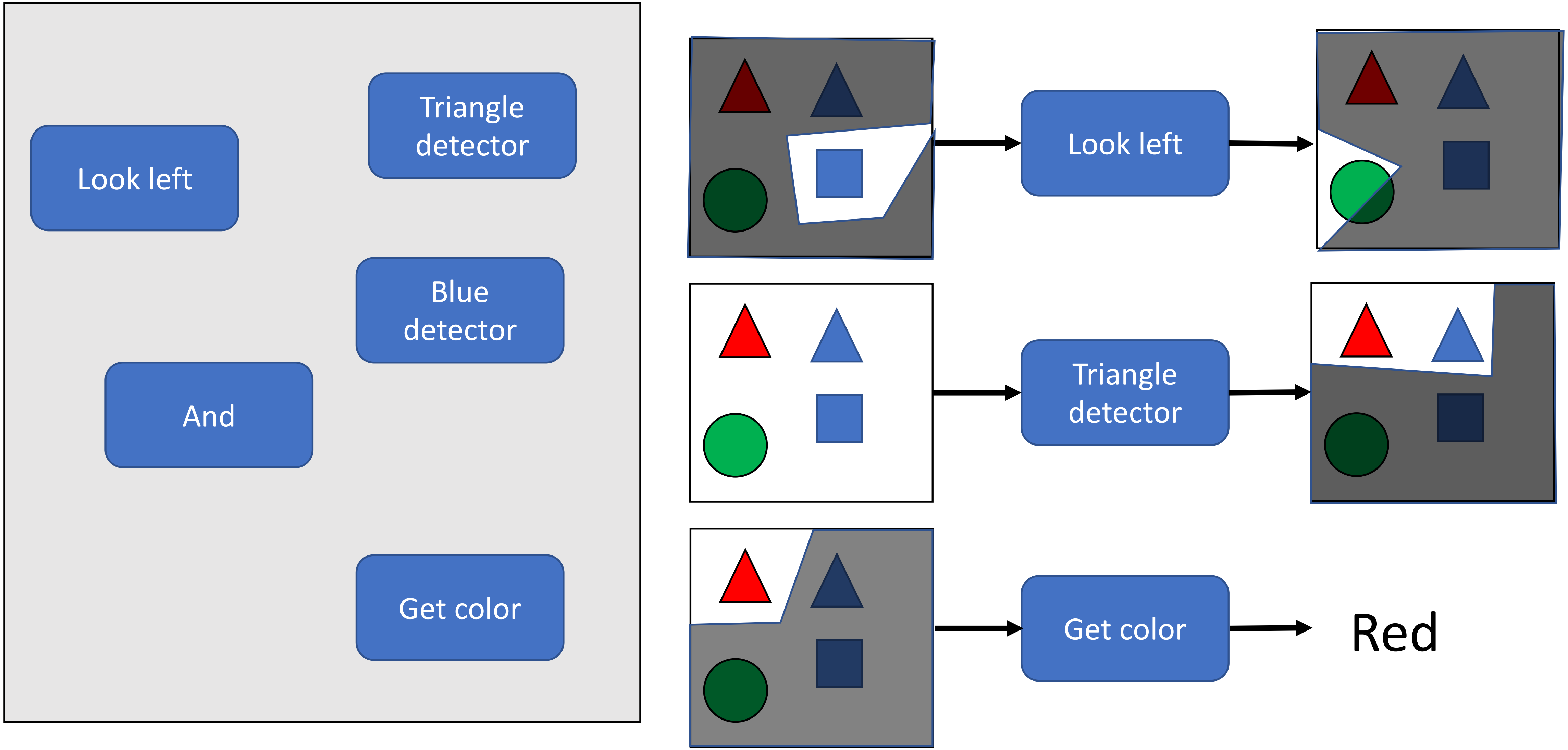
What is the color of the kitten to the left of the blue kitten?

Compositional reasoning

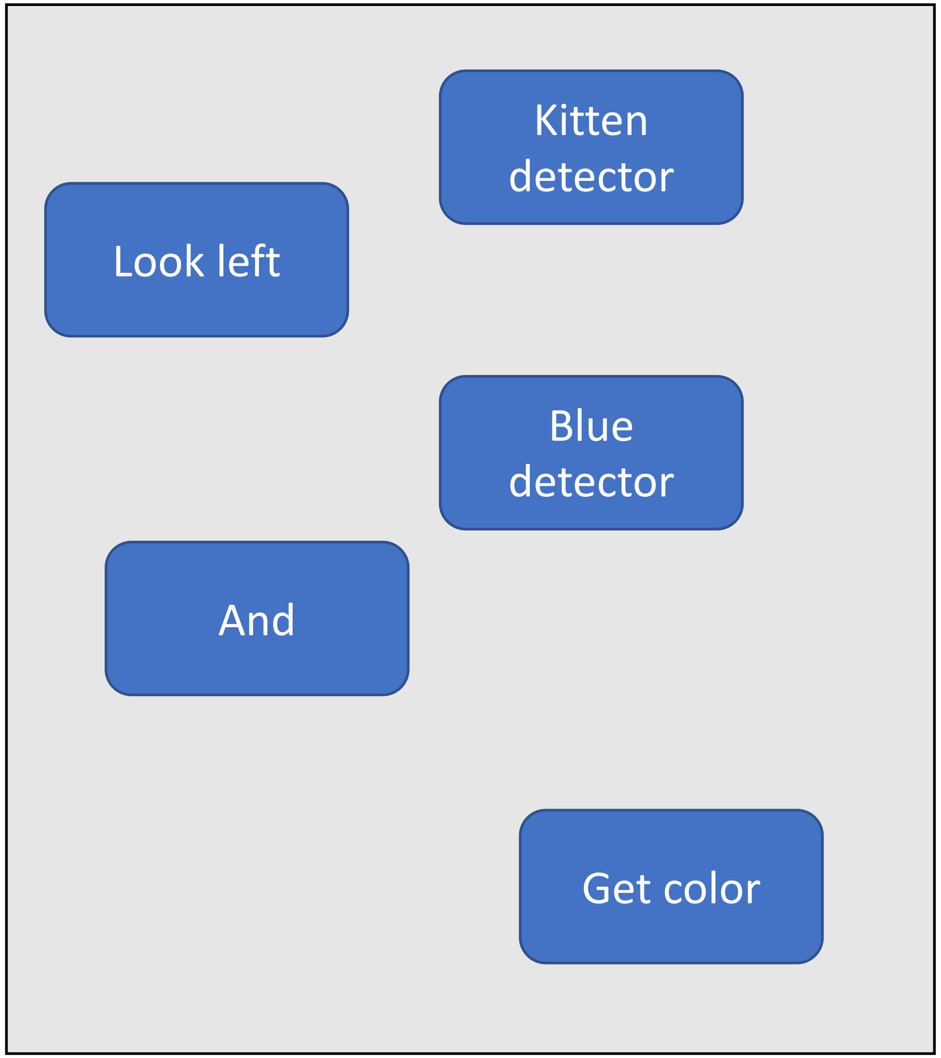
What is the color of the kitten to the left of the blue kitten?



Compositional reasoning

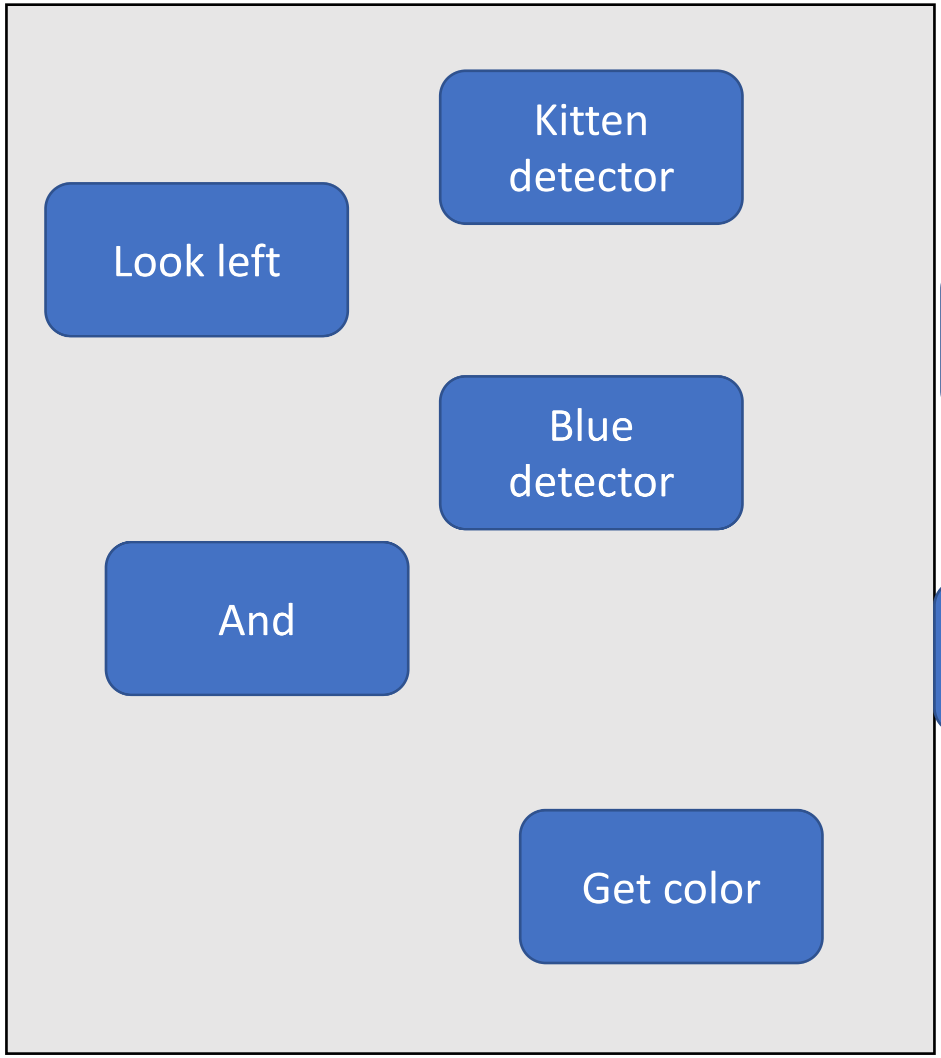


Compositional reasoning

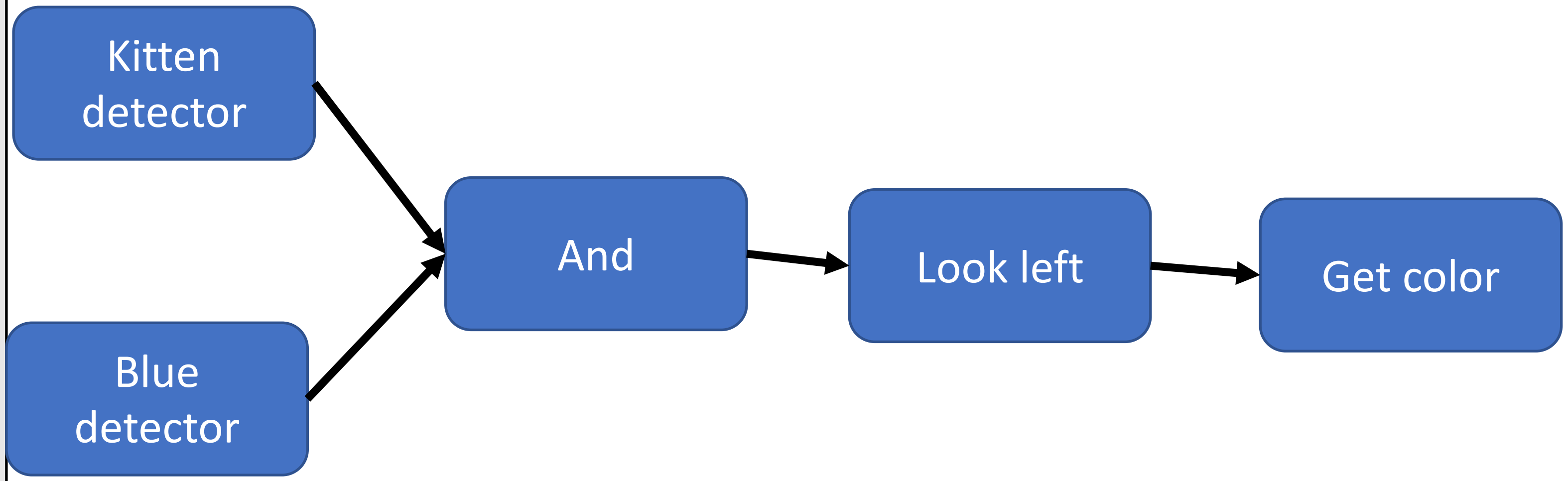


What is the color of the kitten to the left of the blue kitten?

Compositional reasoning



What is the color of the kitten to the left of the blue kitten?



Compositional reasoning

- How do we learn a mapping from language to trees?
  - Problem: semantic parsing
  - Option 1: Syntactic parsing
  - Option 2: Use supervision

*Neural module networks.*Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Dan Klein.CVPR 2016

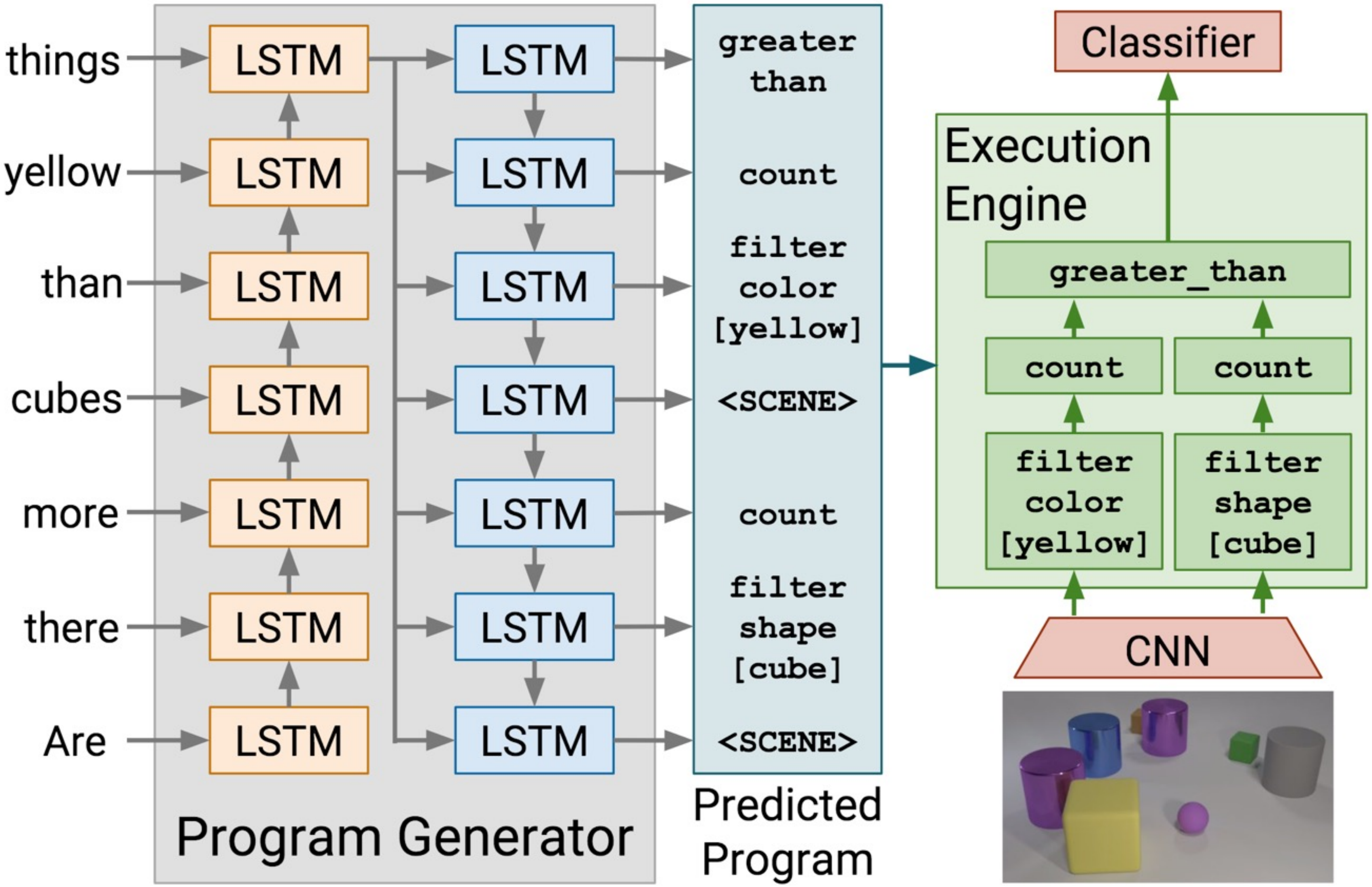
*Learning to compose neural networks for question answering.*Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Dan Klein.NAACL 2016

*Learning to reason: End-to-end module networks for visual question answering.*Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Kate Saenko.ICCV 2017

Inferring and Executing Programs for Visual Reasoning
Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, Ross Girshick. *ICCV*, 2017

Compositional reasoning

Question: *Are there more cubes than yellow things?* **Answer**: *Yes*



The problem with VQA

- Dataset biases allow cheating
  - Only-question Bag-of-Words: 53.7% (vs ~65% for state-of-the-art)
- Require common sense to answer
  - “What is the moustache made of?”
- Hard to diagnose error
  - Is the problem understanding the question?
  - Or understanding the image?

What color are her eyes?
What is the mustache made of?

Clever Hans