DeepMind

Interpretability in Vision

Mateusz Malinowski

Plan

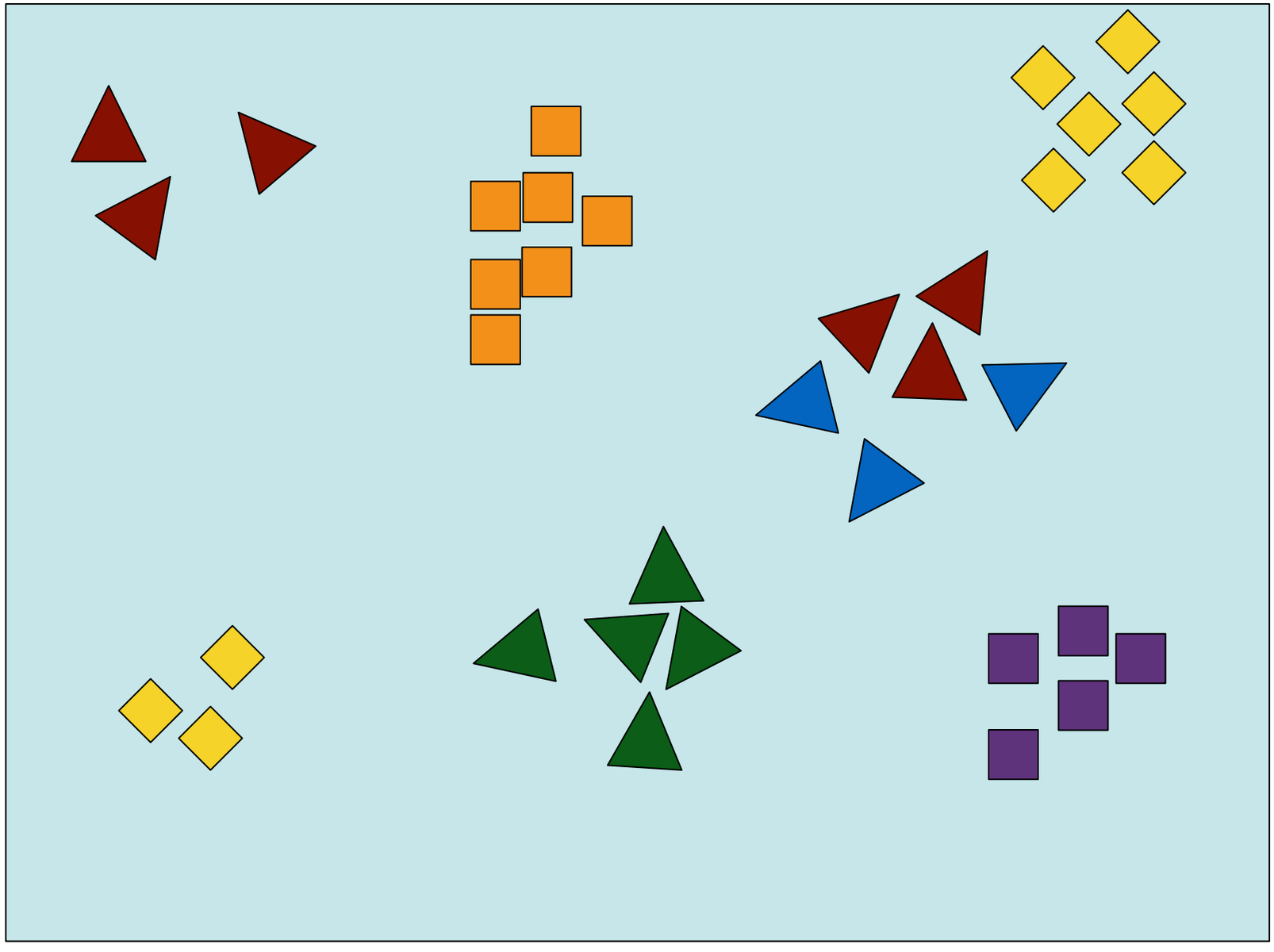
- Attention
- Behavioural Tests
- Using art for explanation

DeepMind

Attention

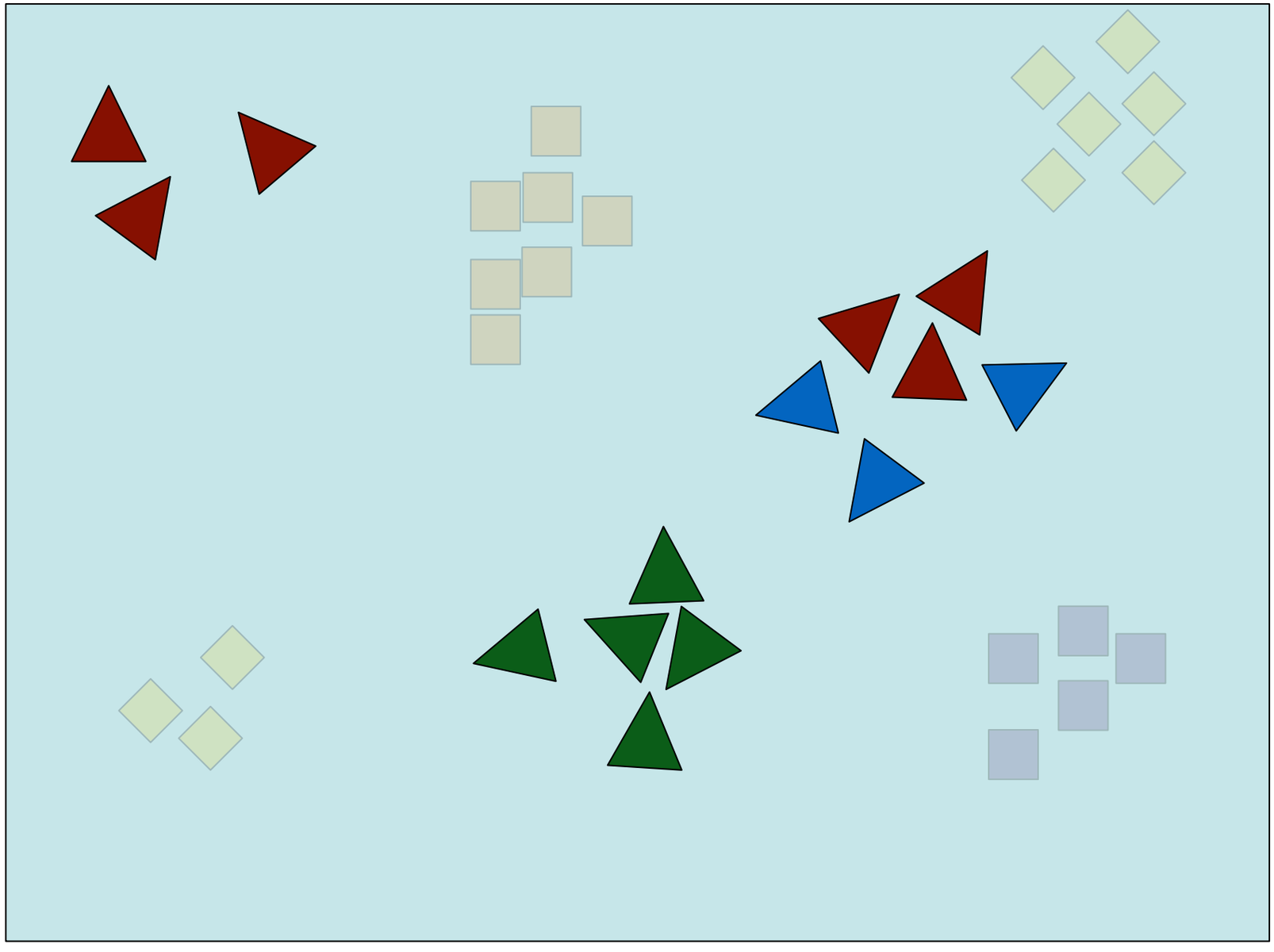
Mateusz Malinowski

What is attention?



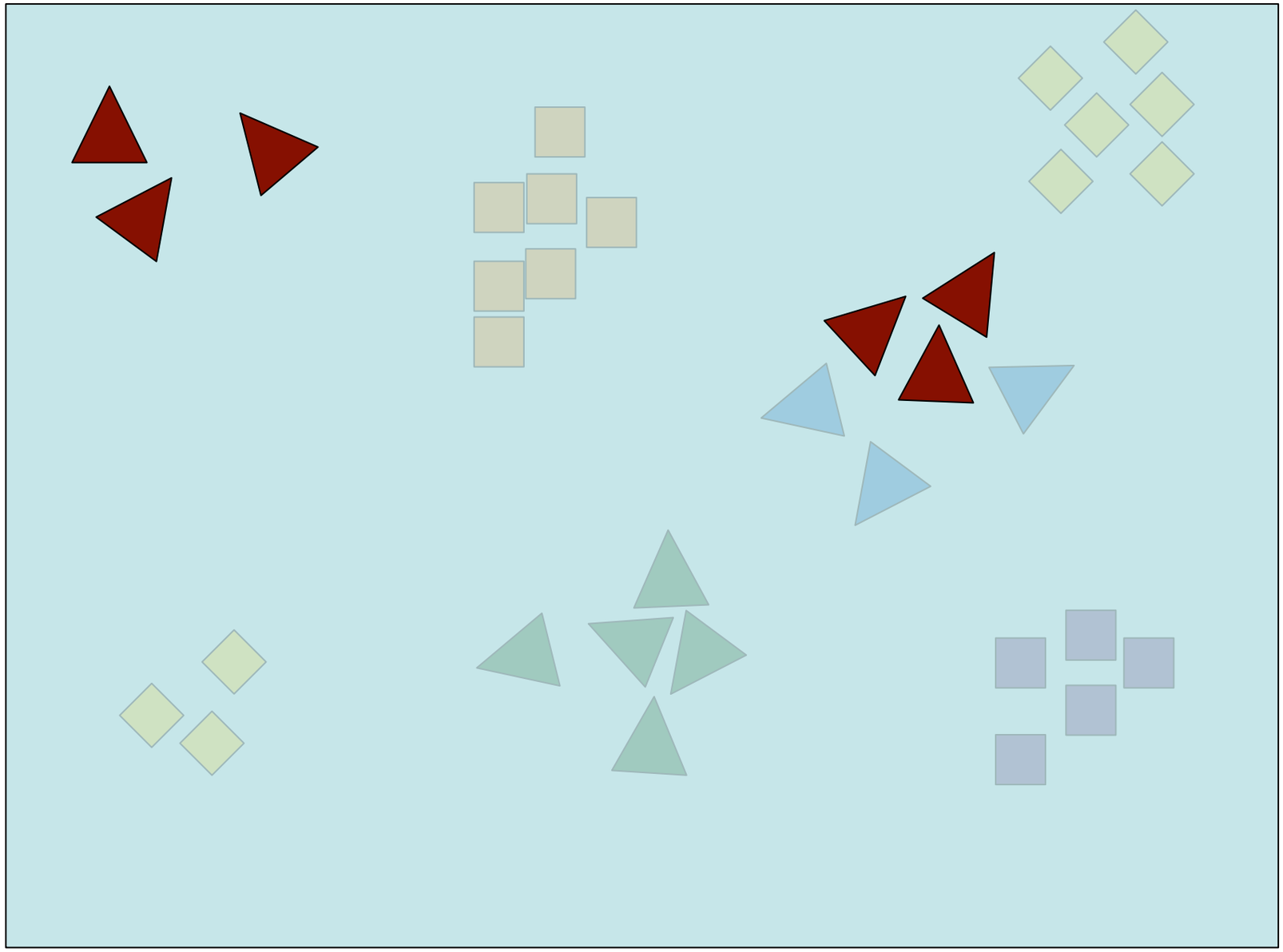
How many red triangles are there?

What is attention?



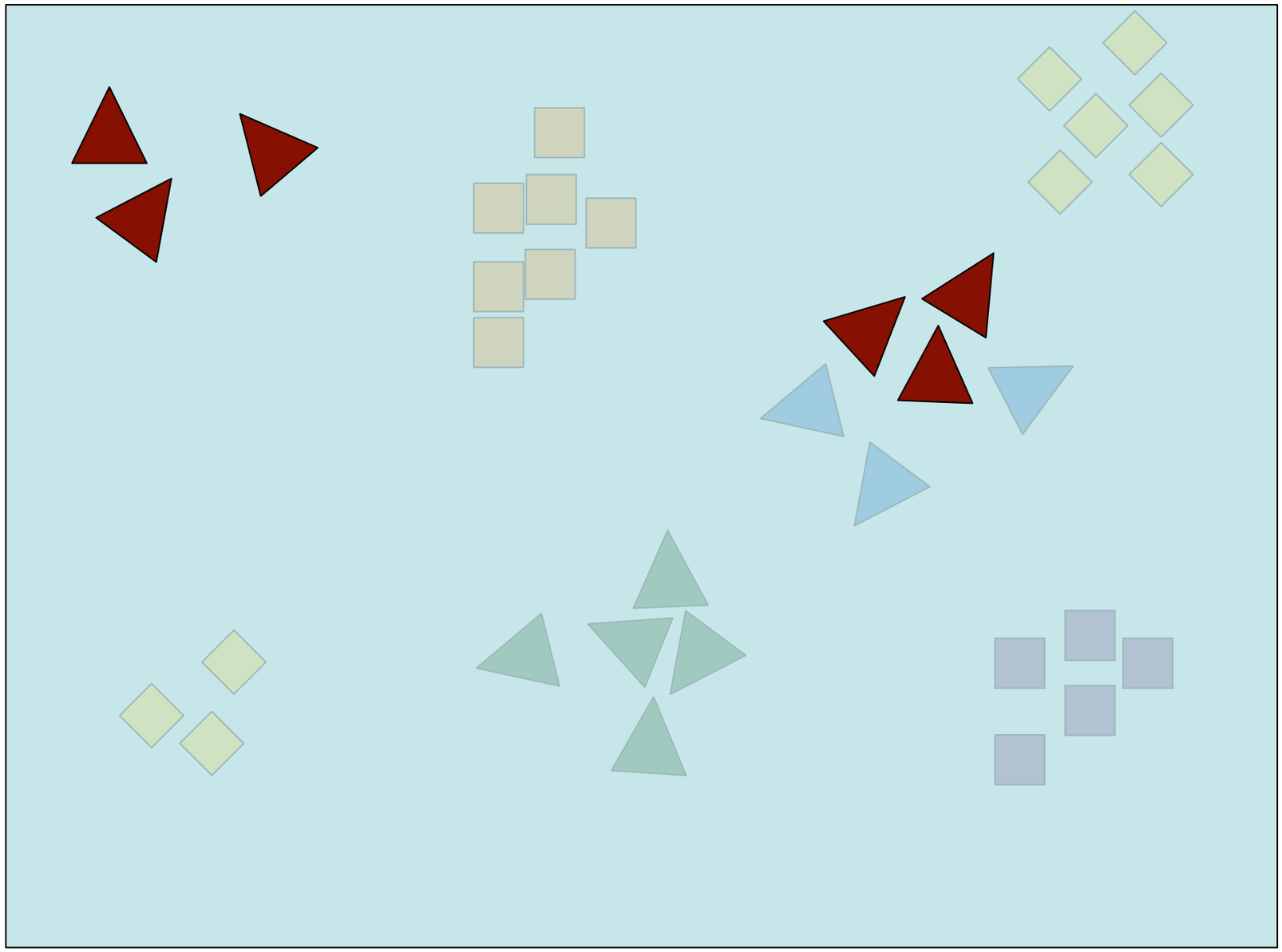
How many red **triangles** are there?

What is attention?



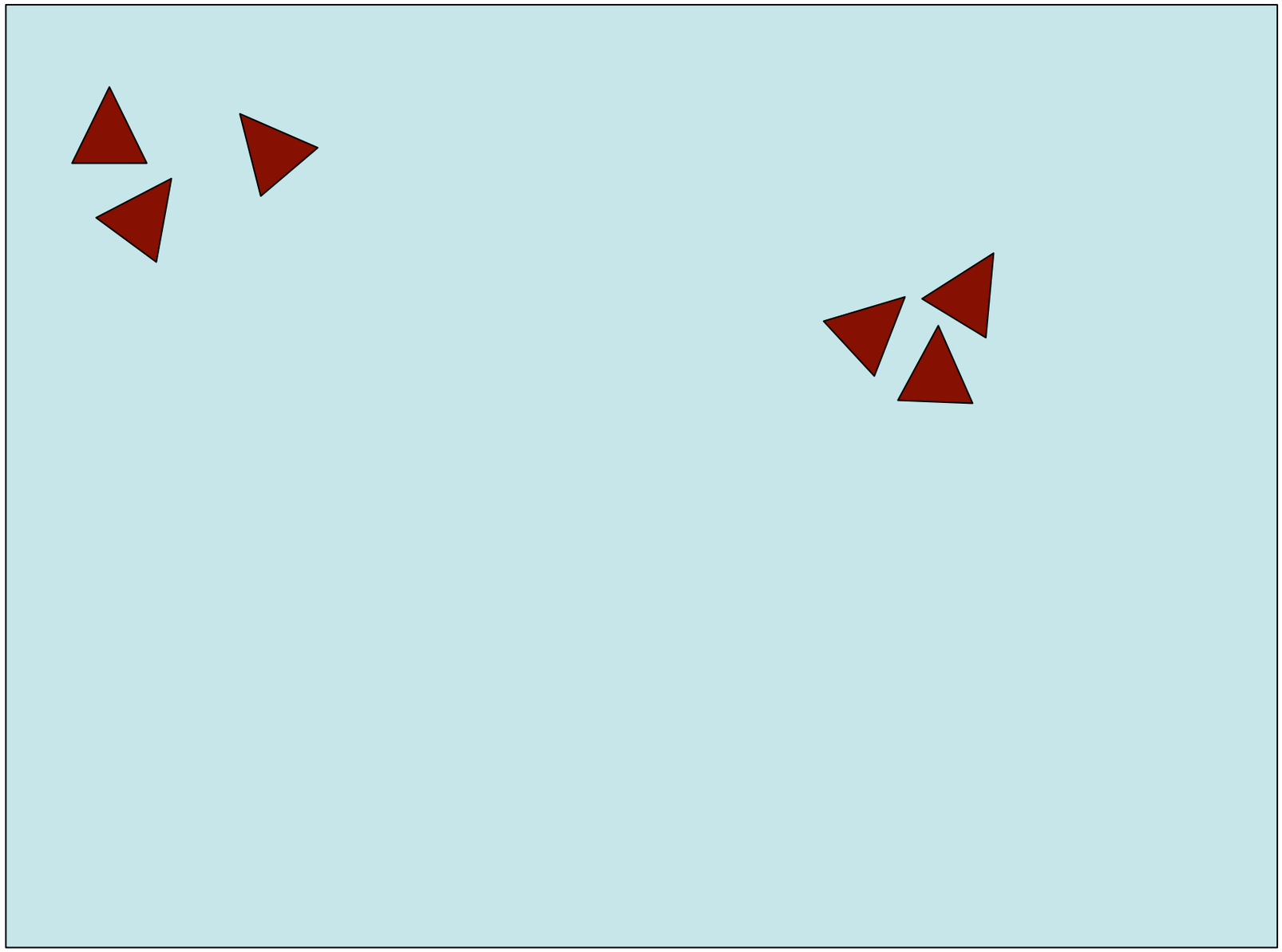
How many **red triangles** are there?

Soft-attention: “everything is visible but with different weighting”



How many **red triangles** are there?

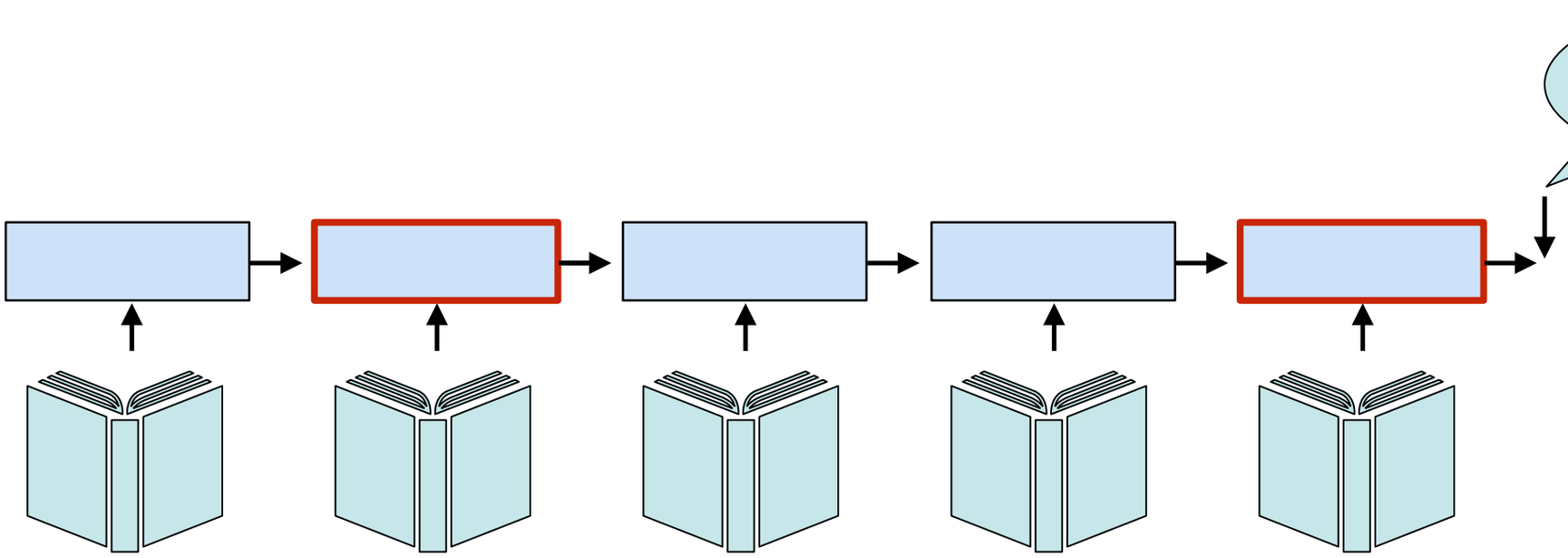
Hard-attention: “Only things of interest are visible”



How many **red triangles** are there?

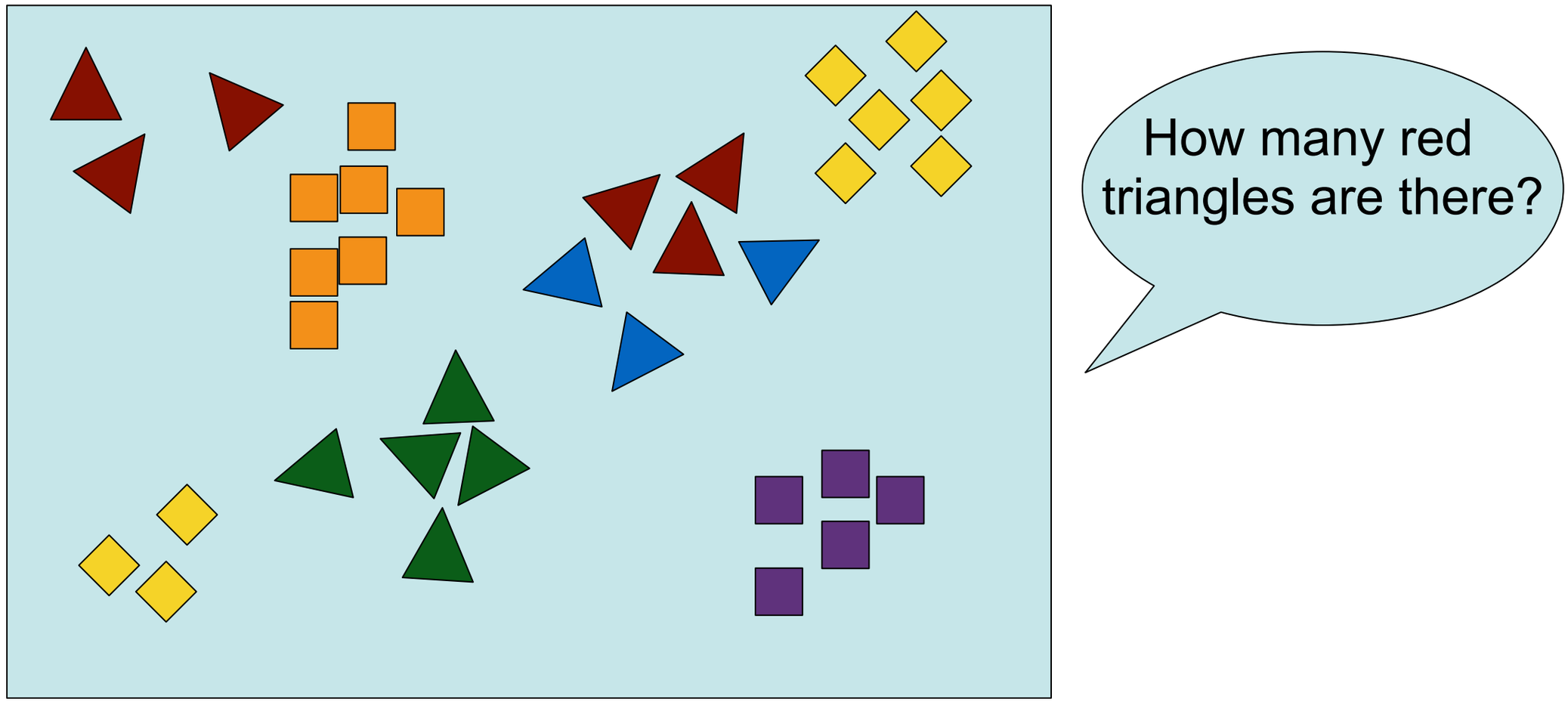
Why attention?

- **Long term memories** - attending to memories
  - Dealing with gradient vanishing problem



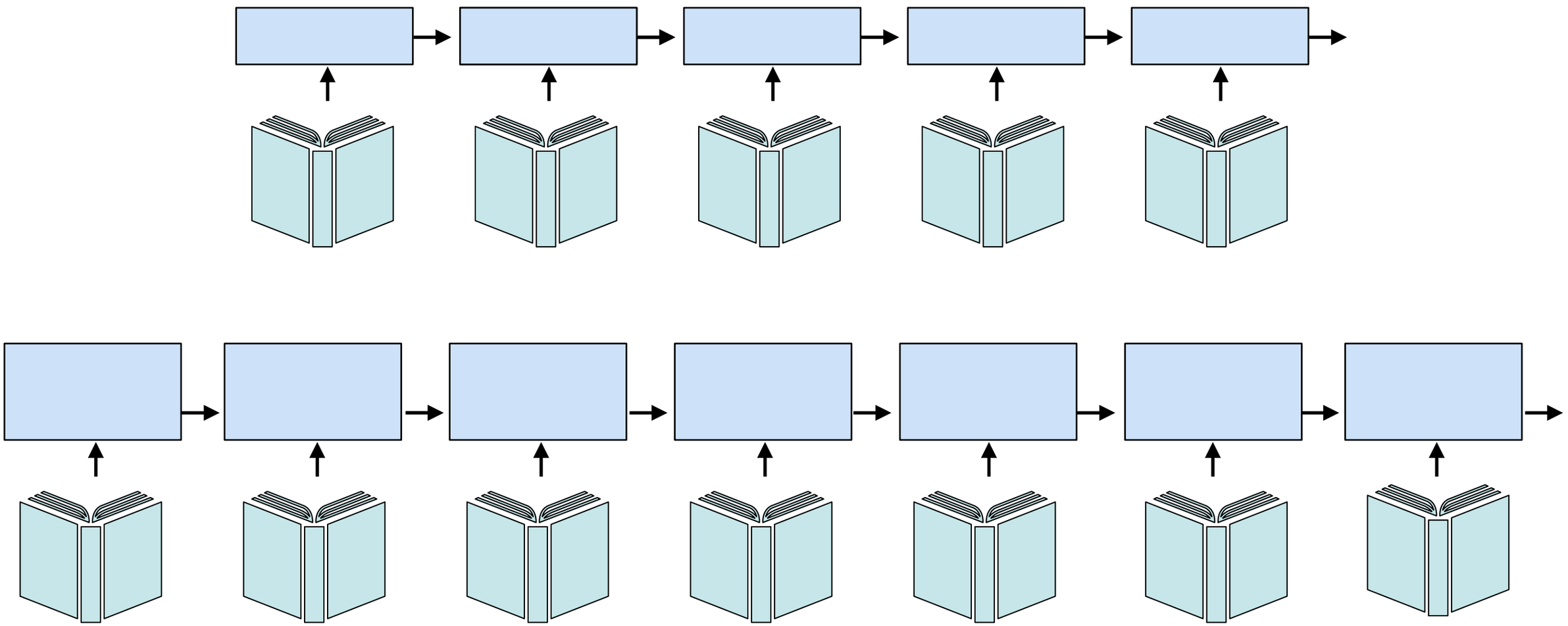
Why attention?

- Long term memories - attending to memories
- **Exceeding limitations of a global representation**
  - Attending/focusing to smaller parts of data
    - patches in images
    - words or phrases in sentences



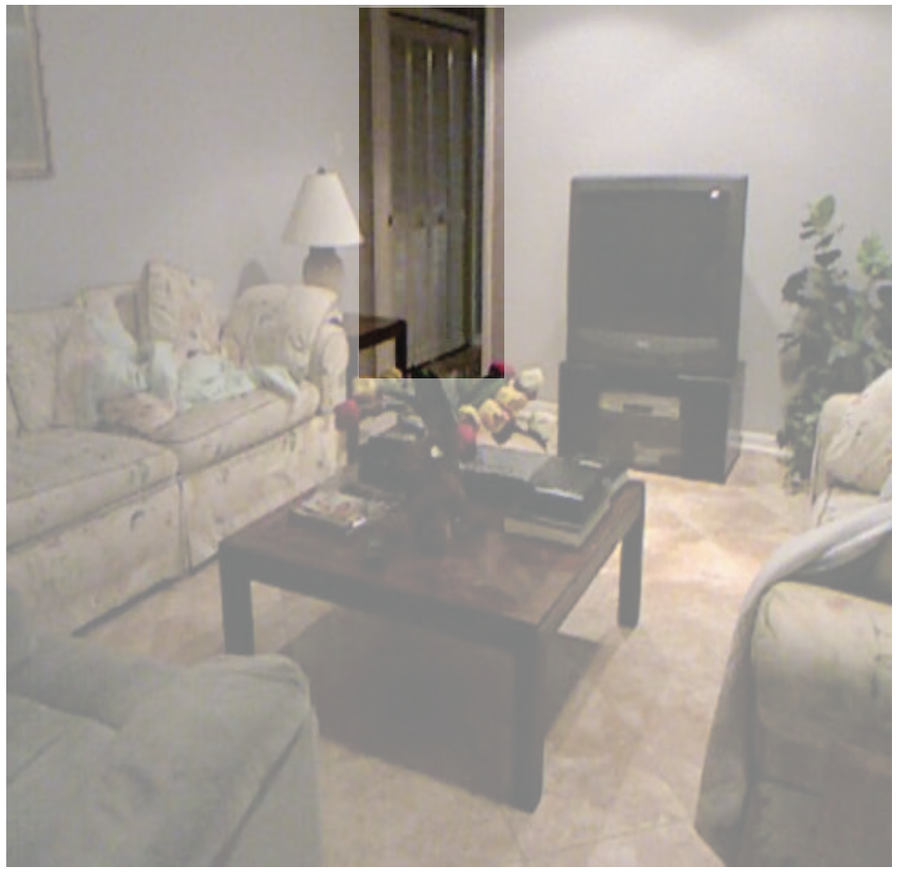
Why attention?

- Long term memories - attending to memories
- Exceeding limitations of a global representation
- **Decoupling representation from a problem**
  - LSTM with longer sentences requires larger vectors



Why attention?

- Long term memories - attending to memories
- Exceeding limitations of a global representation
- Decoupling representation from a problem
  - LSTM with longer sentences requires larger vectors
- **Adds some interpretability to the models (error inspection)**



What is the colour of a lamp? White

Why attention?

- Long term memories - attending to memories
- Exceeding limitations of a global representation
- Decoupling representation from a problem
  - LSTM with longer sentences requires larger vectors
- Adds some interpretability to the models (error inspection)

Models might be **right** for **wrong reasons**!

What color is the book on the table? Black

Computational models of attention

- **Soft-** vs **Hard**-Attention

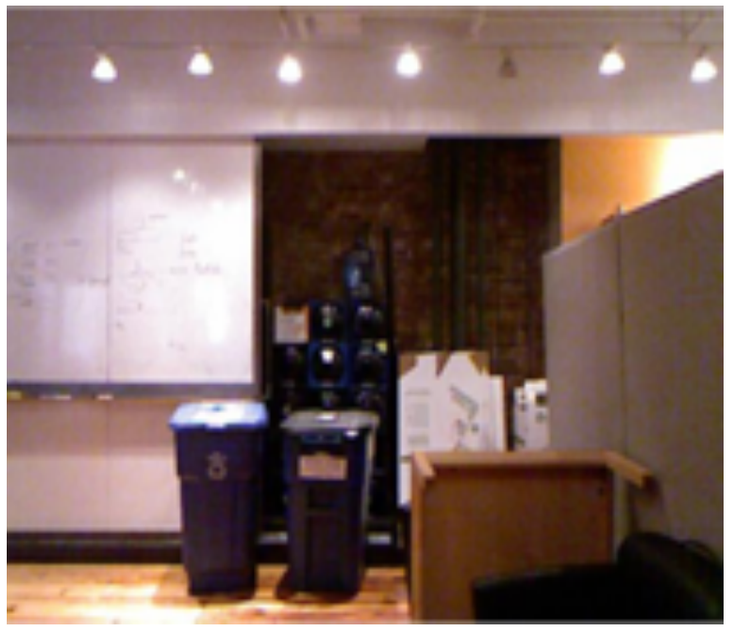
DeepMind

Soft-Attention

Mateusz Malinowski

Visual Question Answering

- Questions about Images
  - Images
  - Questions



What is behind the table?
sofa

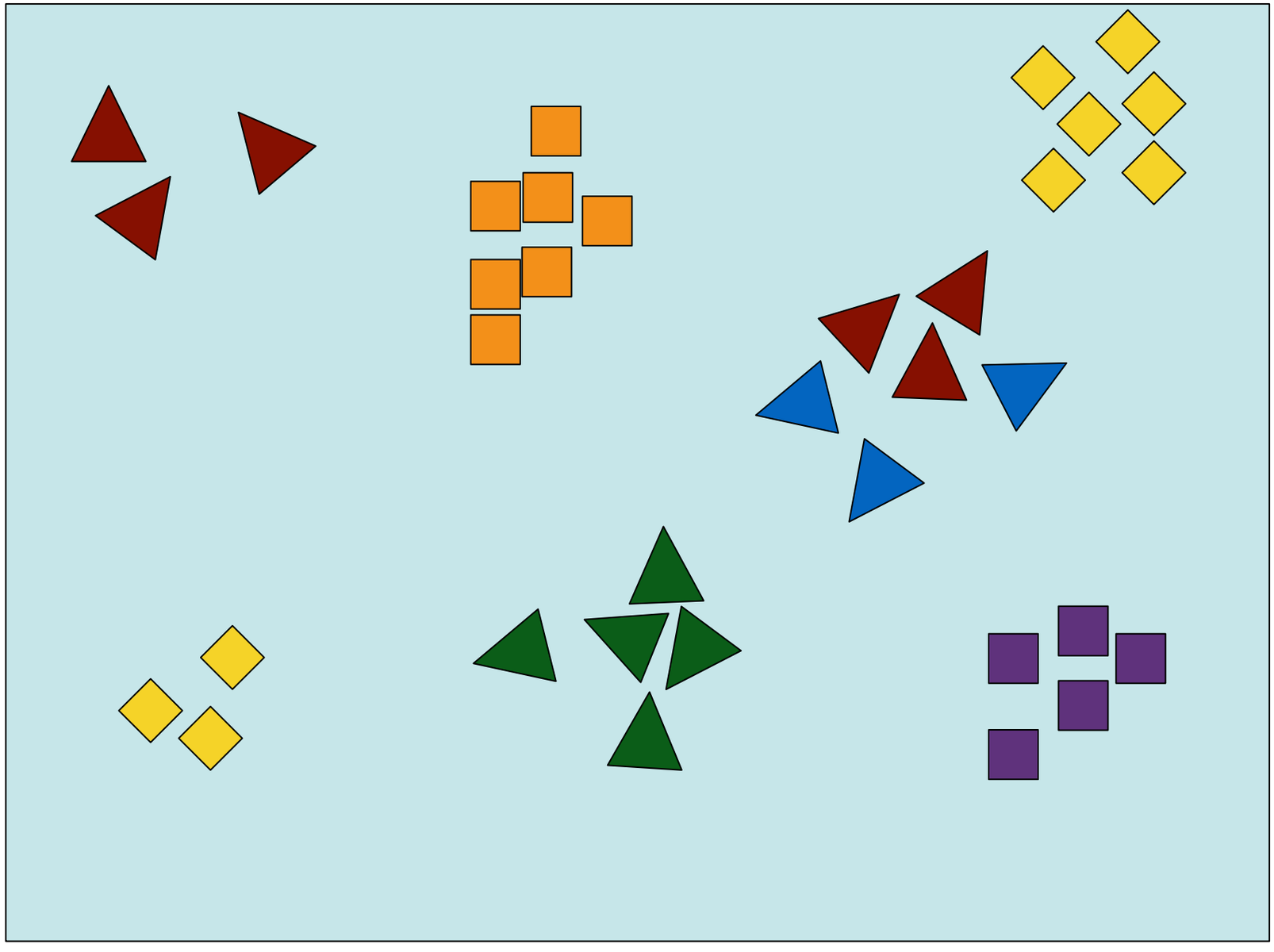
What is on the refrigerator?
magnet, paper

What color are the cabinets?
brown

How many lamps are there?
2

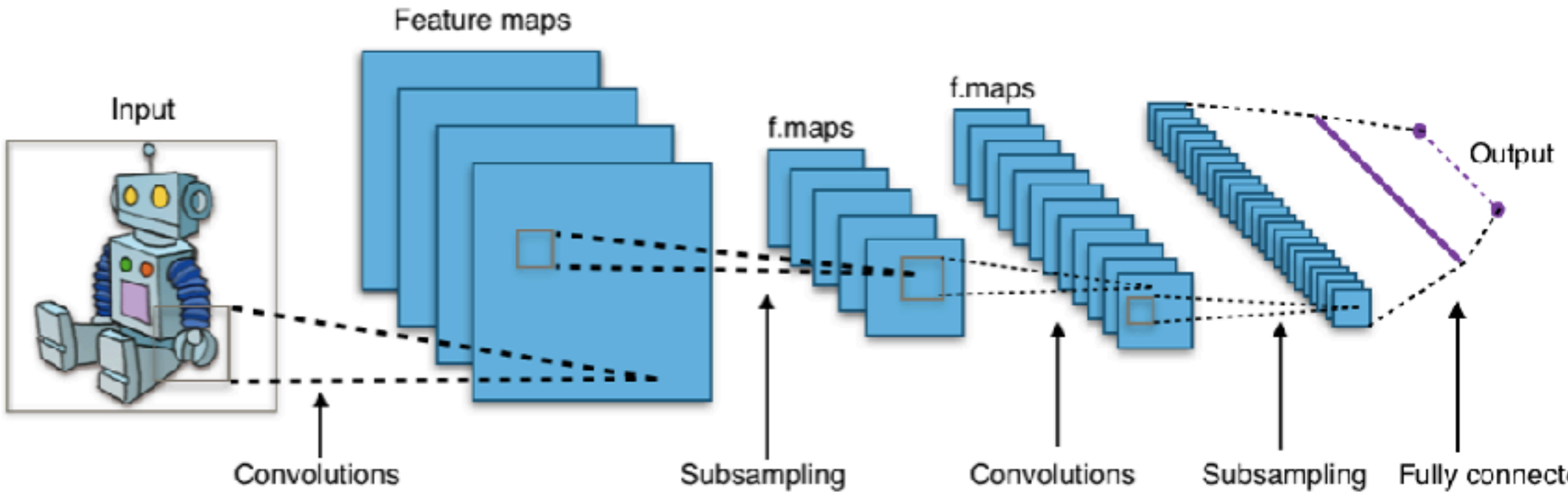
Malinowski and Fritz “A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input”

Visual Question Answering



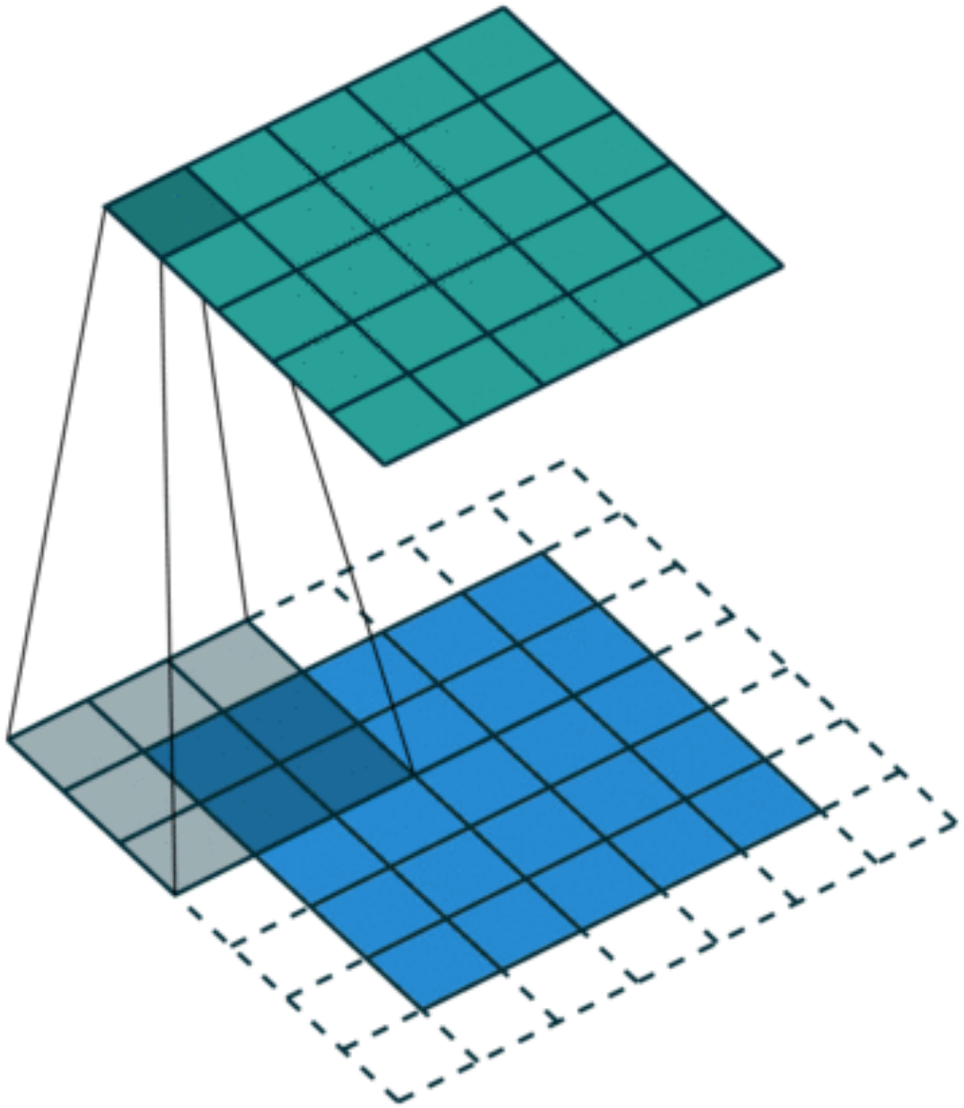
How many red triangles are there?

CNNs



Wikipedia

CNNs

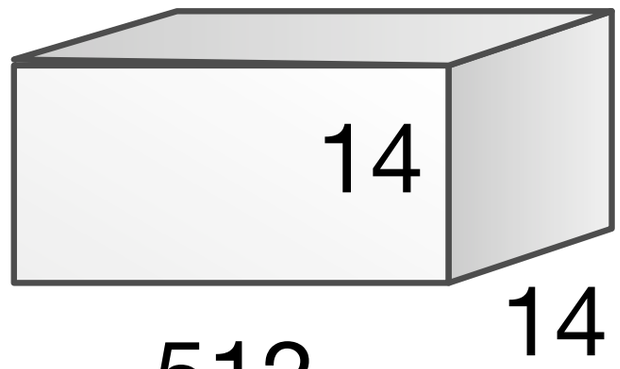


https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Soft-Attention in Visual Question Answering

- Based on “Stacked Attention Networks” (Yang et al.)
  - CNN encoder of the image

$$f_I = \mathbf{CNN}_{vgg}(I)$$

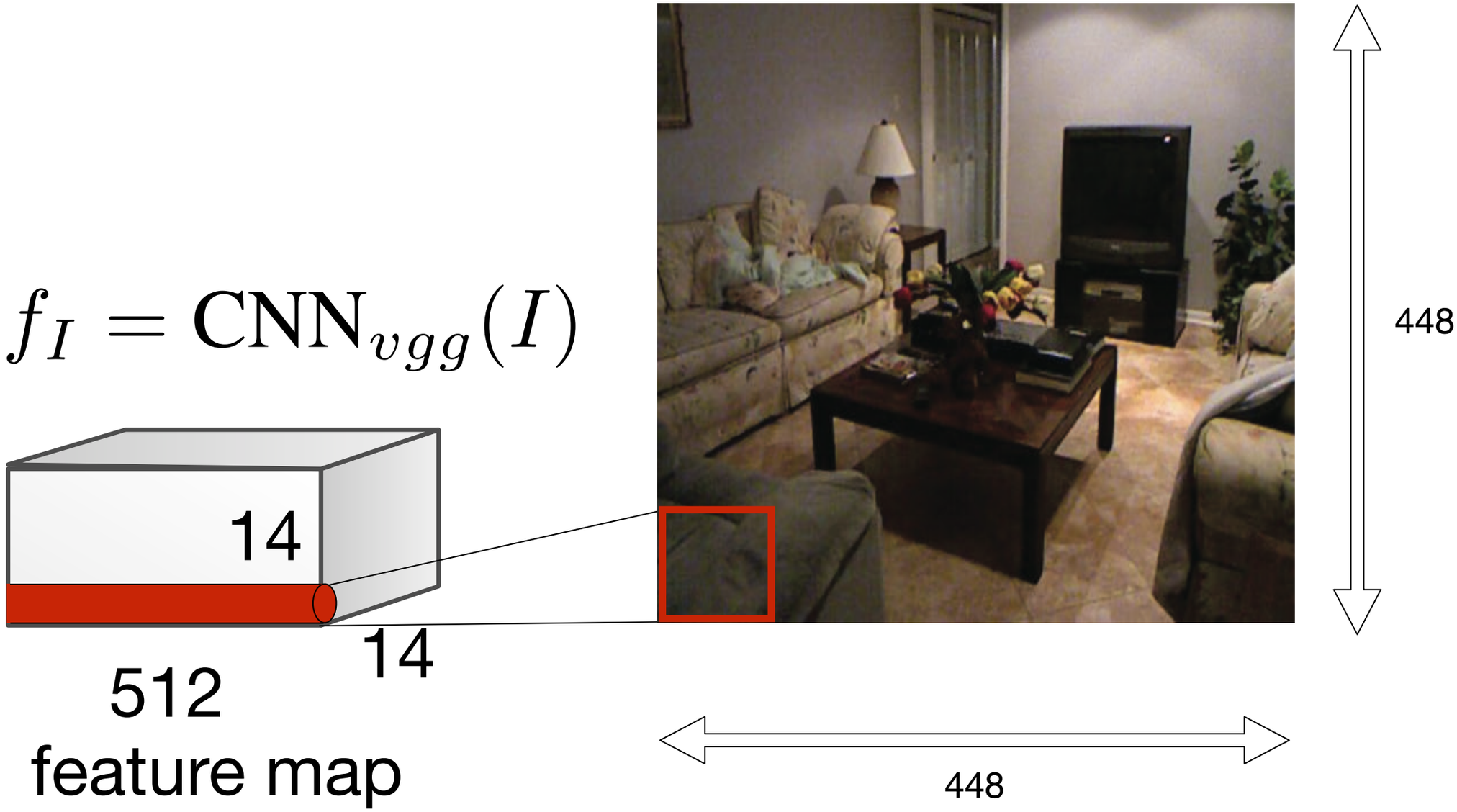


Yang et al. “Stacked Attention Networks”

Soft-Attention in Visual Question Answering

- Based on “Stacked Attention Networks” (Yang et al.)
  - CNN encoder of the image
  - Take a Convolutional Feature Map

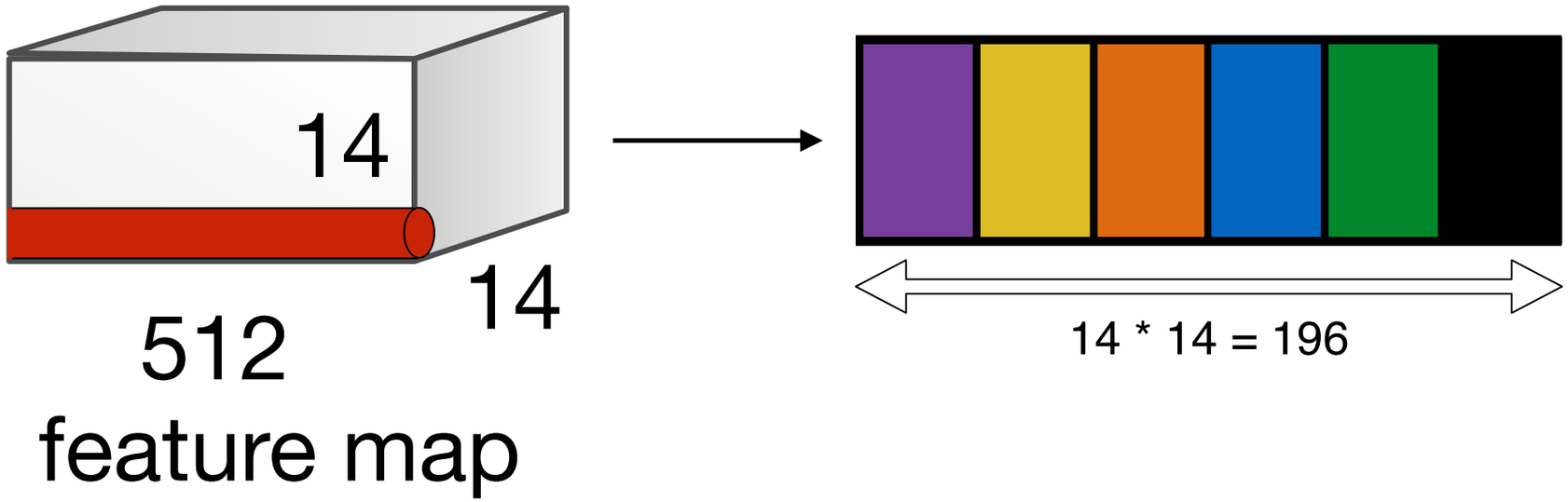
$$f_I = \mathrm{CNN}_{vgg}(I)$$



Yang et al. “Stacked Attention Networks”

Soft-Attention in Visual Question Answering

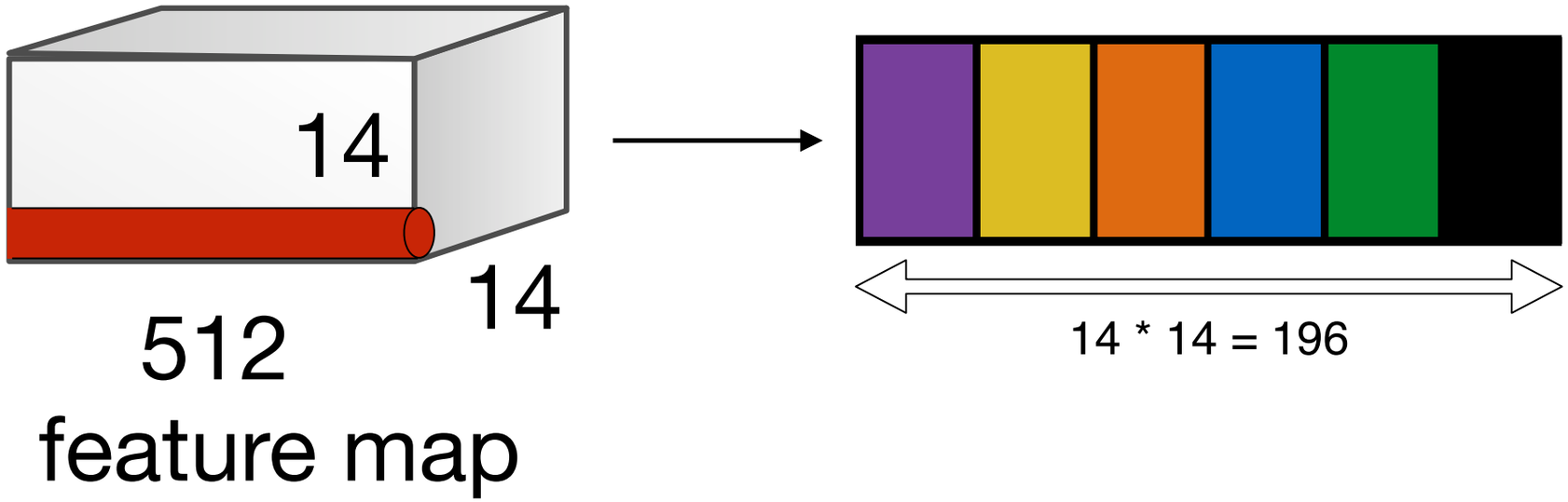
$$f_I = \mathbf{CNN}_{vgg}(I)$$



Yang et al. “Stacked Attention Networks”

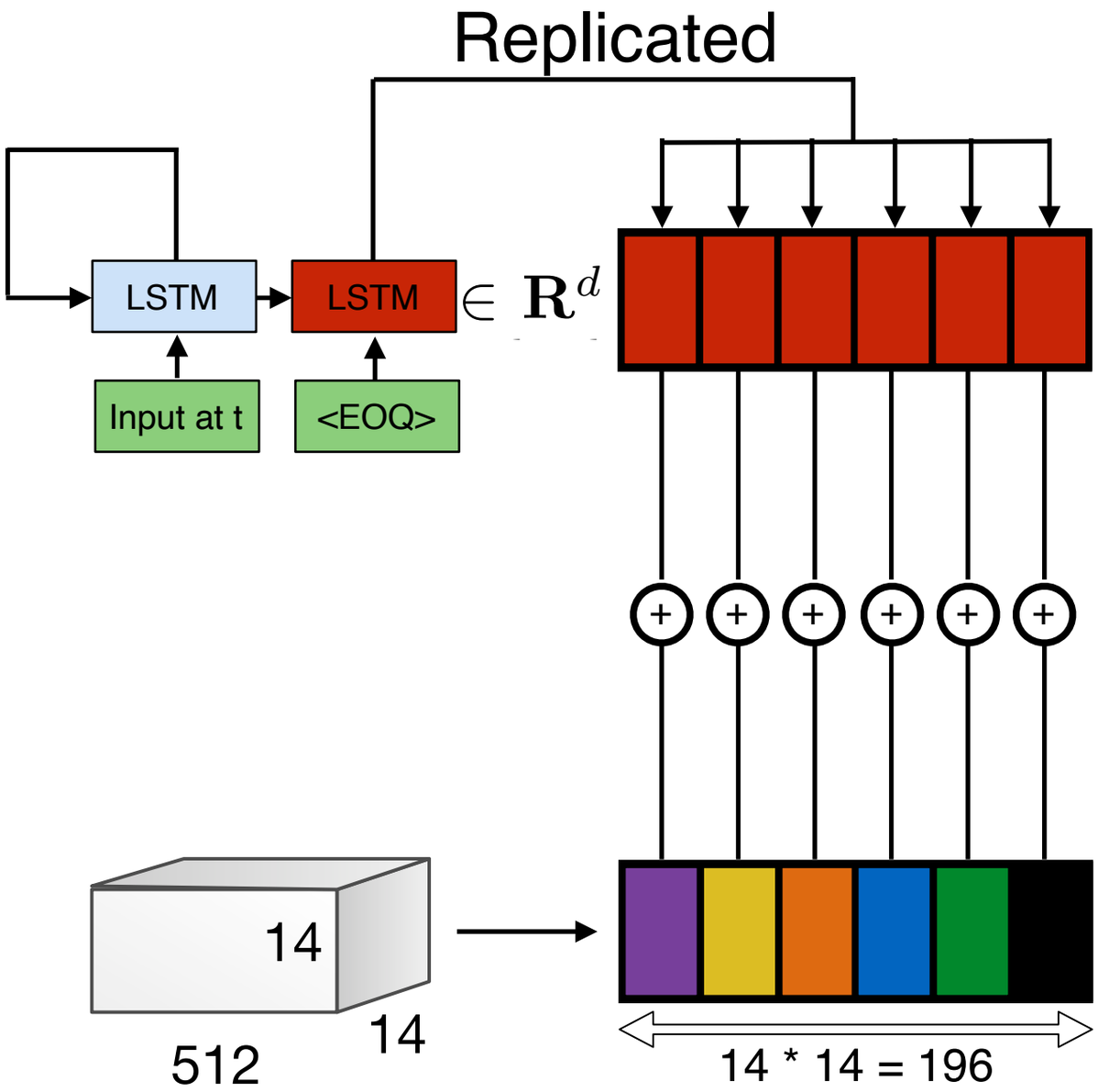
Soft-Attention in Visual Question Answering

$$f_I = \mathbf{CNN}_{vgg}(I)$$

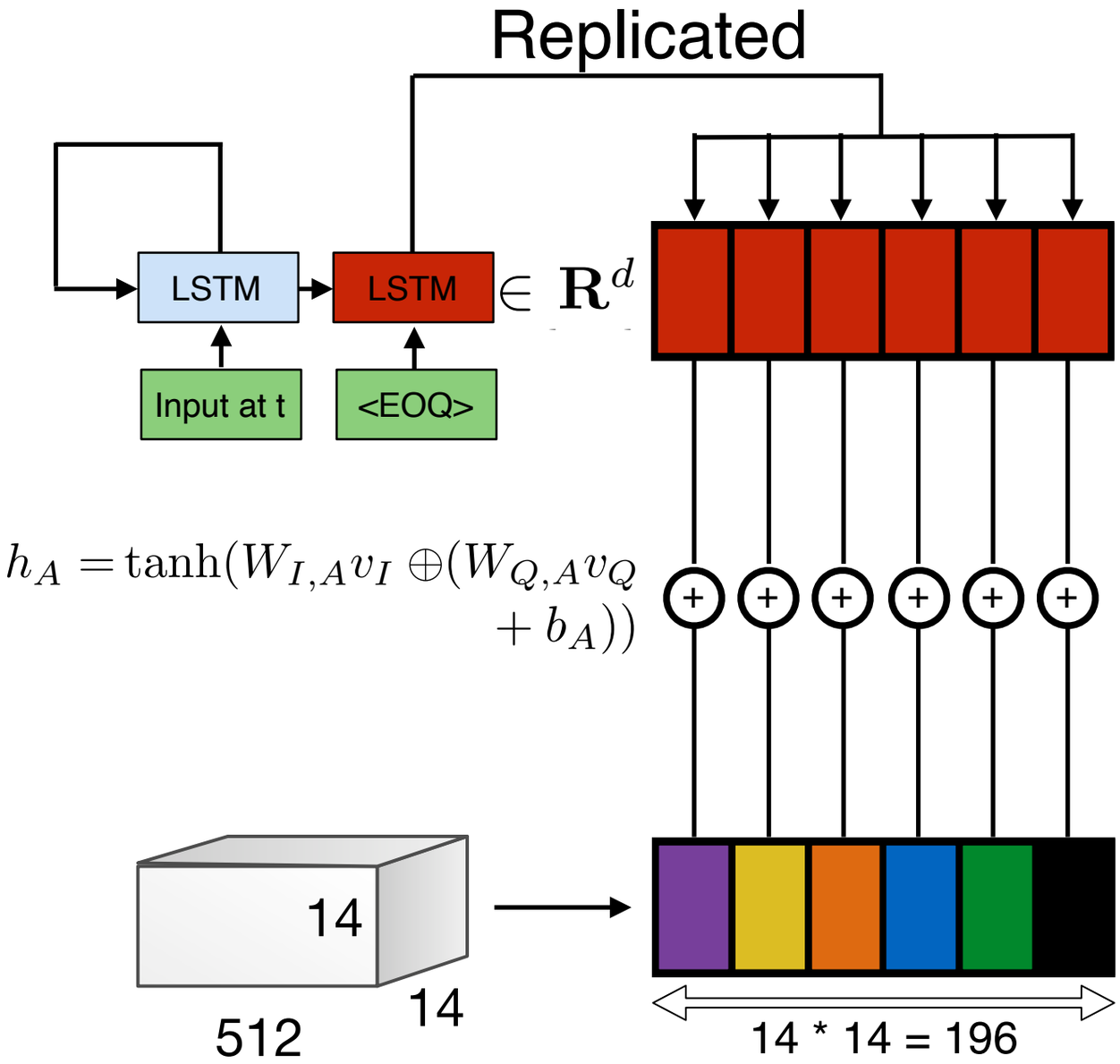


Yang et al. “Stacked Attention Networks”

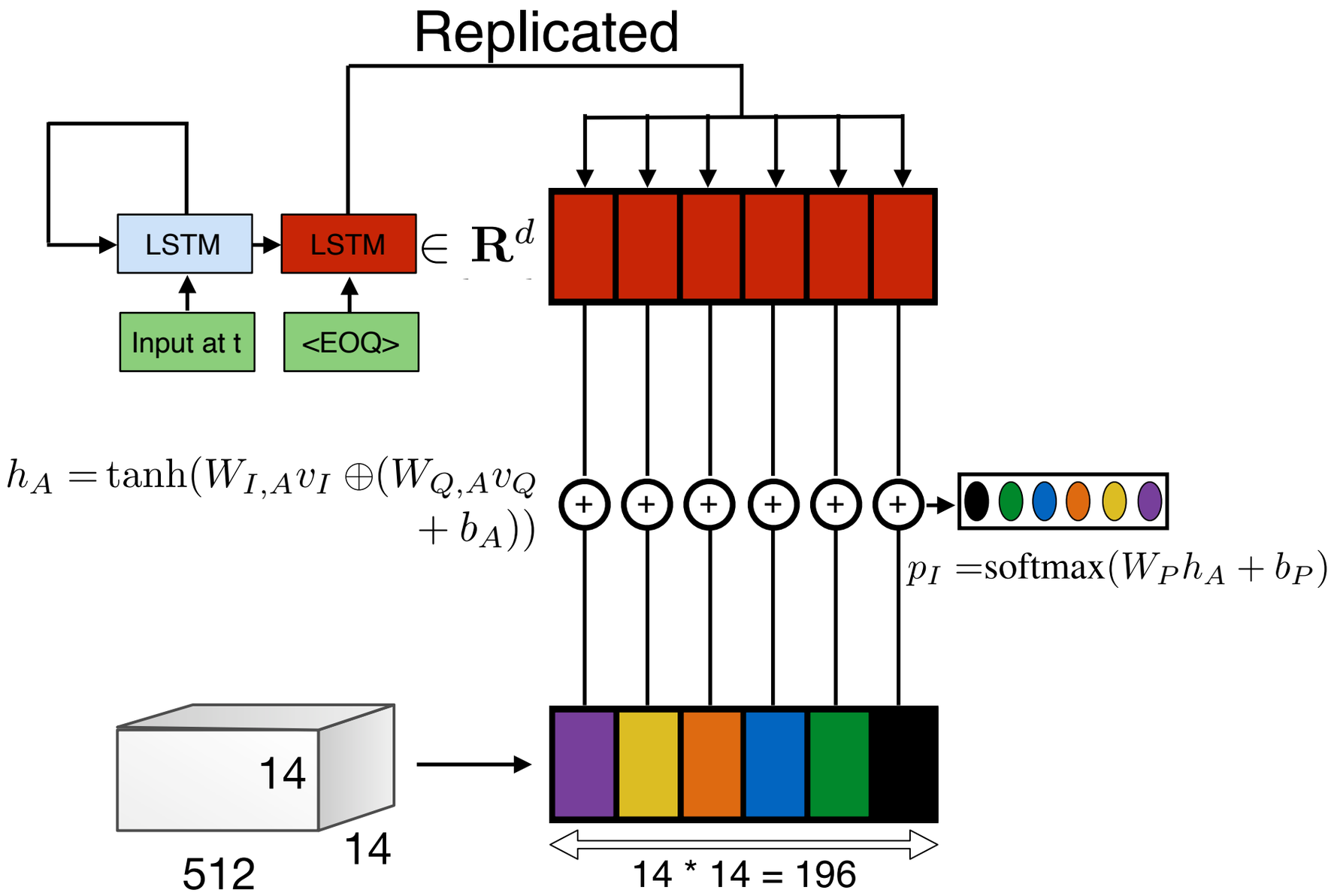
Soft-Attention in Visual Question Answering



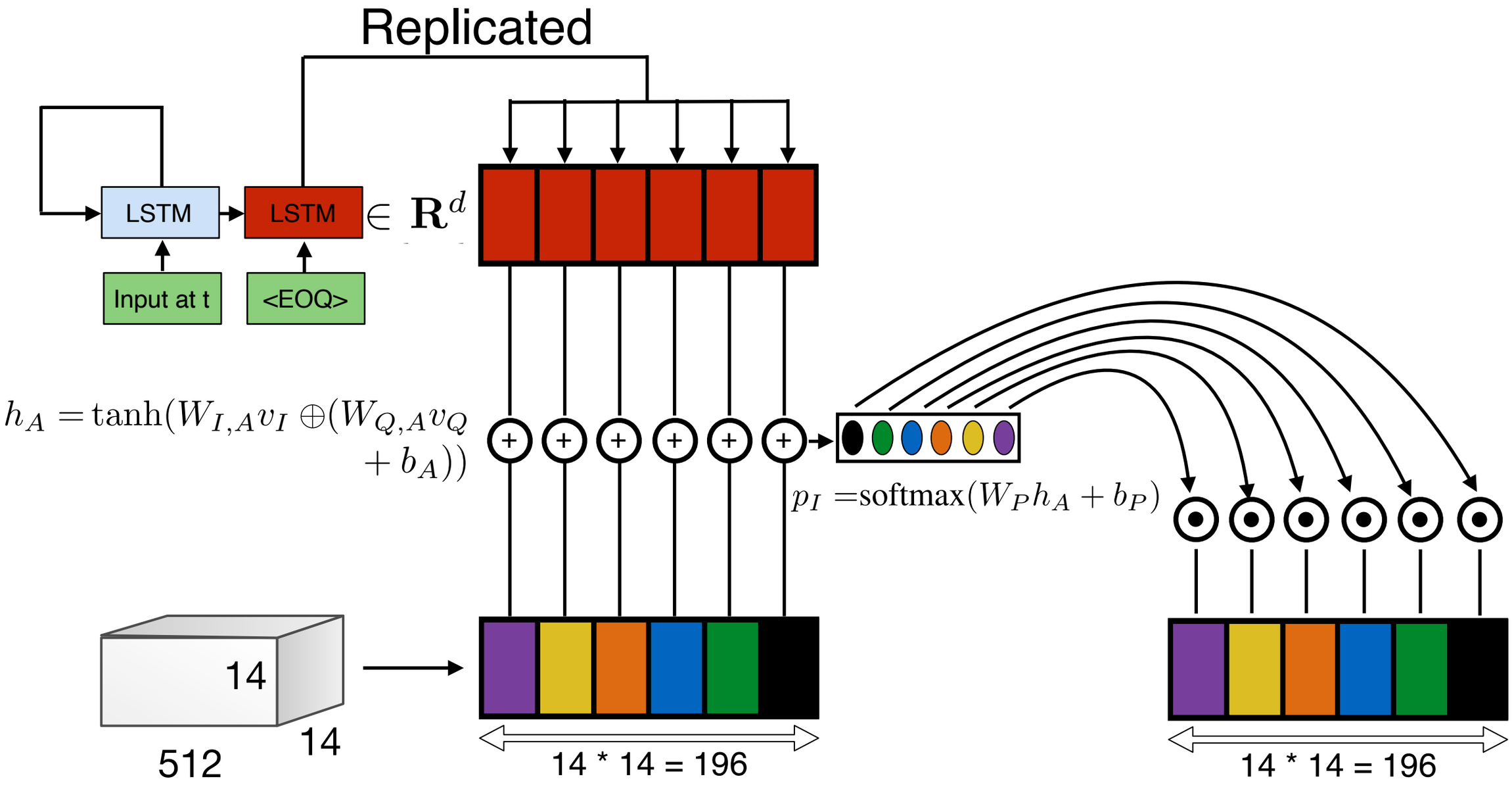
Soft-Attention in Visual Question Answering



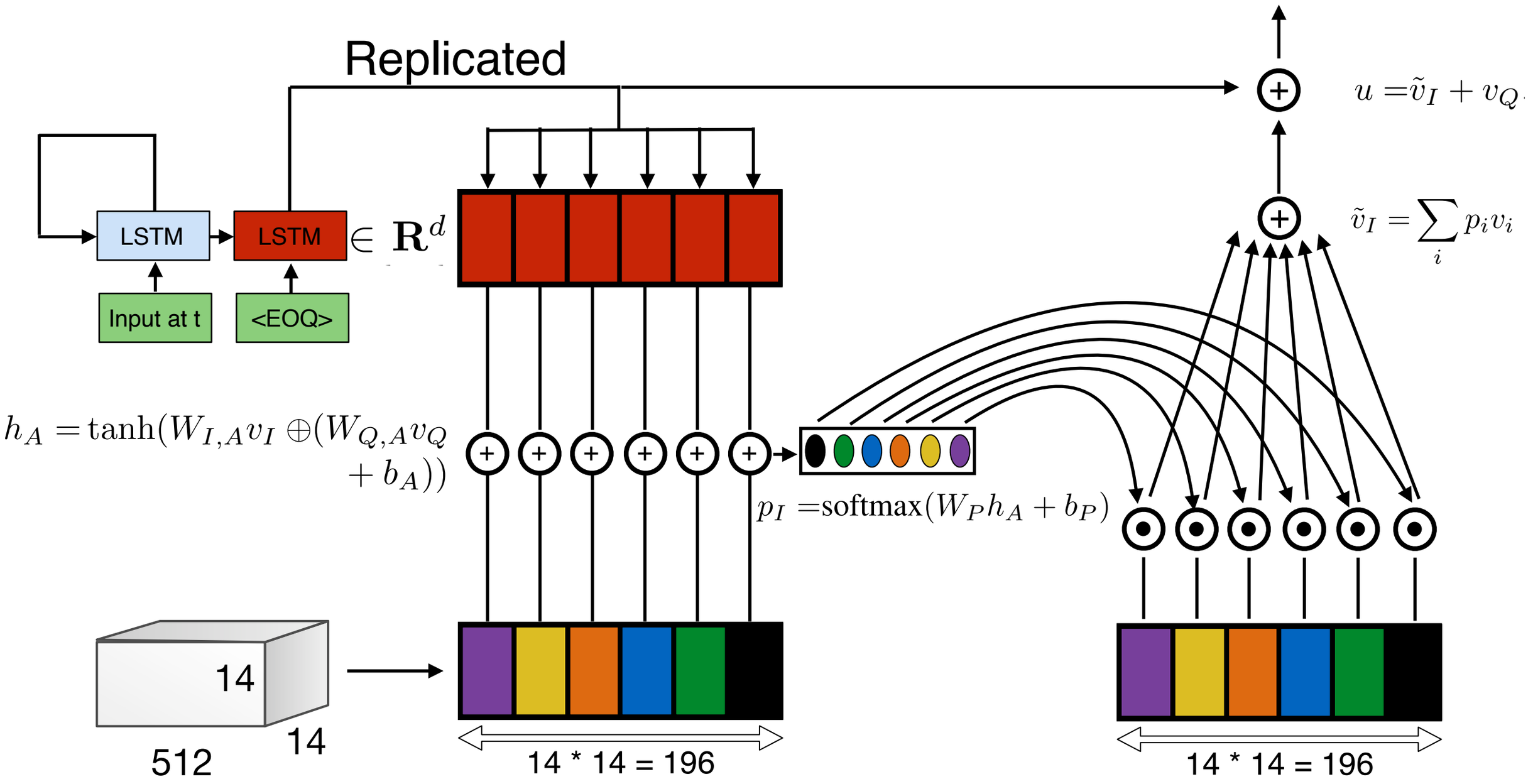
Soft-Attention in Visual Question Answering



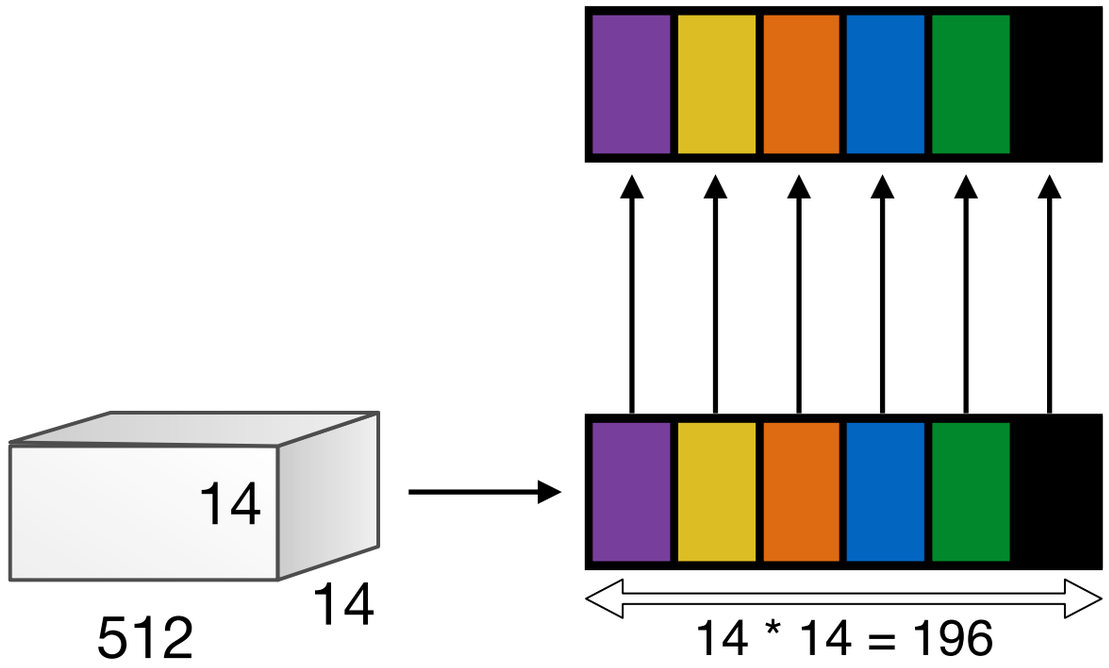
Soft-Attention in Visual Question Answering



Soft-Attention in Visual Question Answering

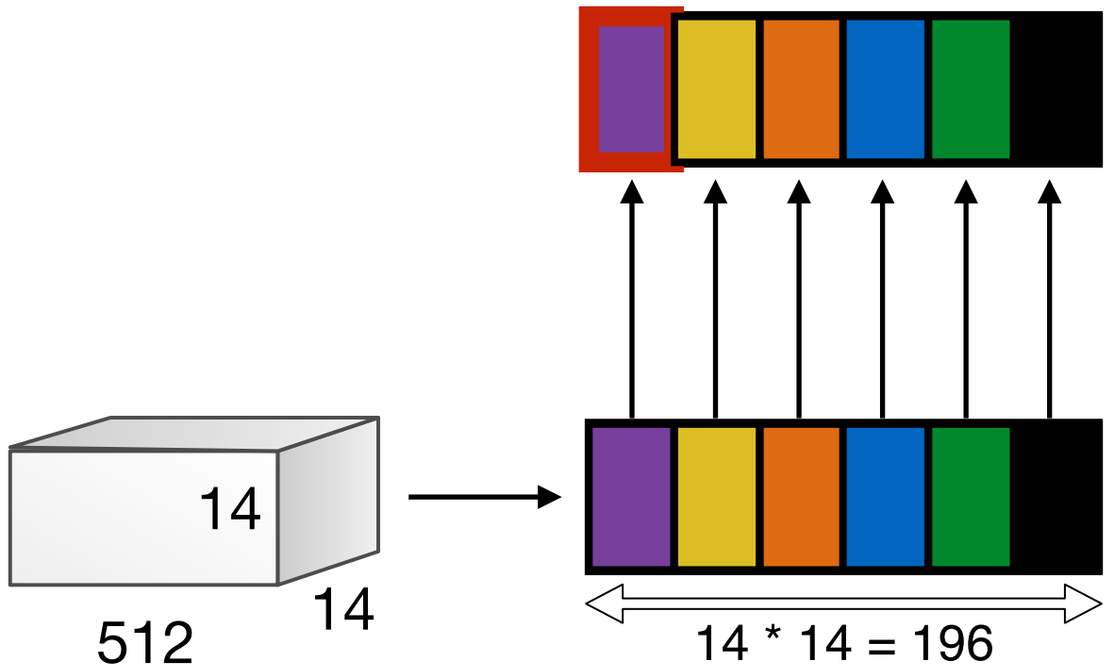


Self-(Soft-)Attention in Visual Question Answering



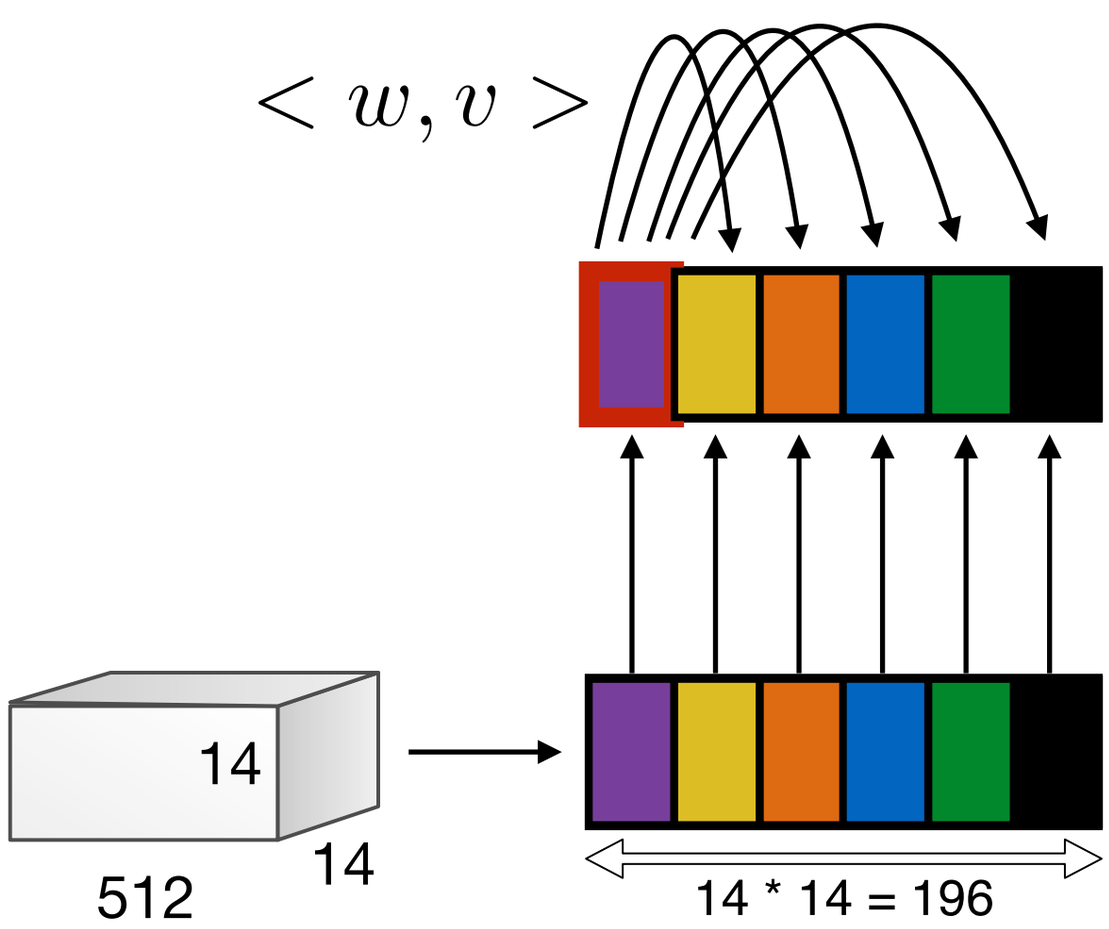
Vaswani et al. “Attention is All You Need”

Self-(Soft-)Attention in Visual Question Answering



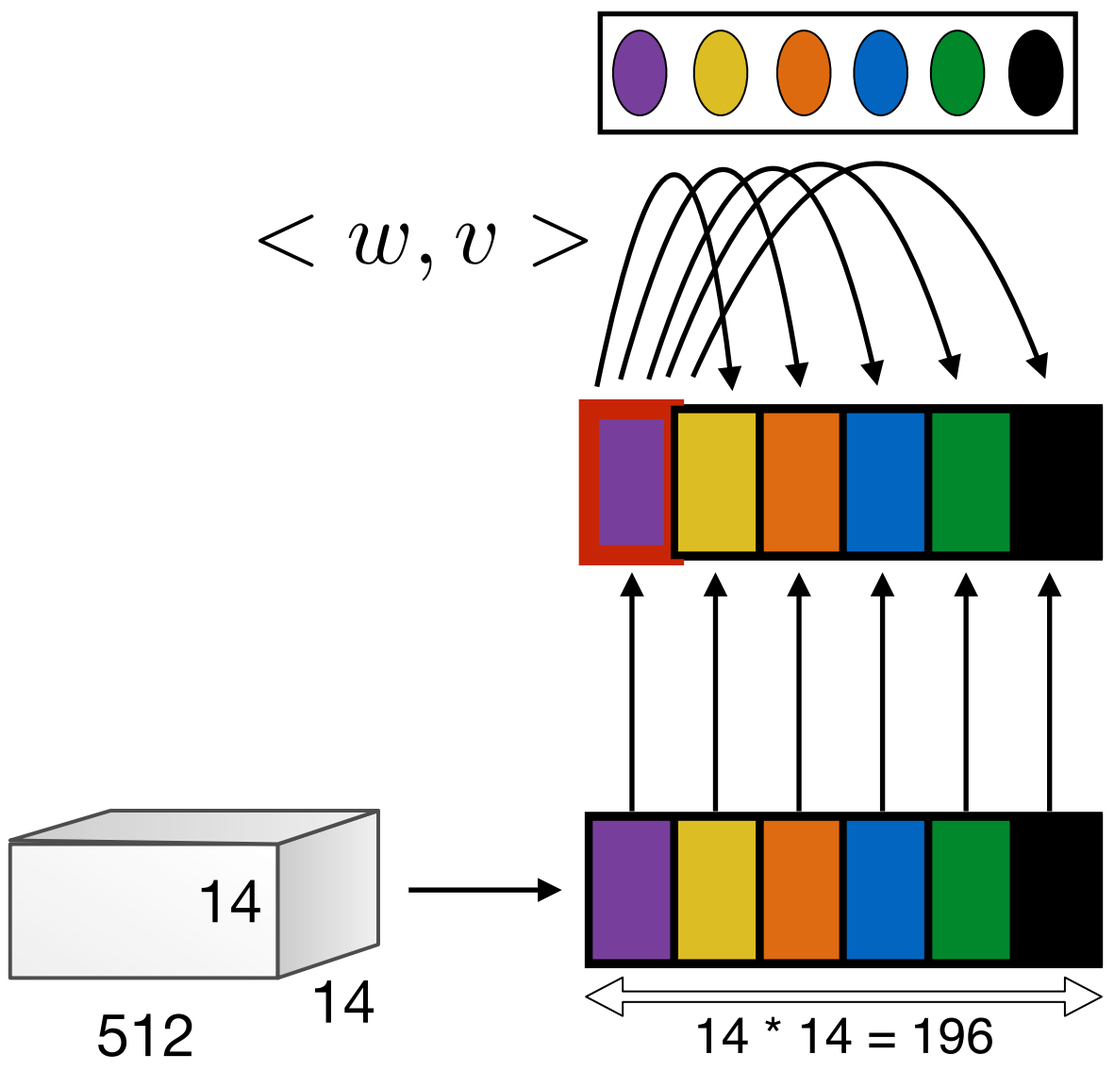
Vaswani et al. “Attention is All You Need”

Self-(Soft-)Attention in Visual Question Answering



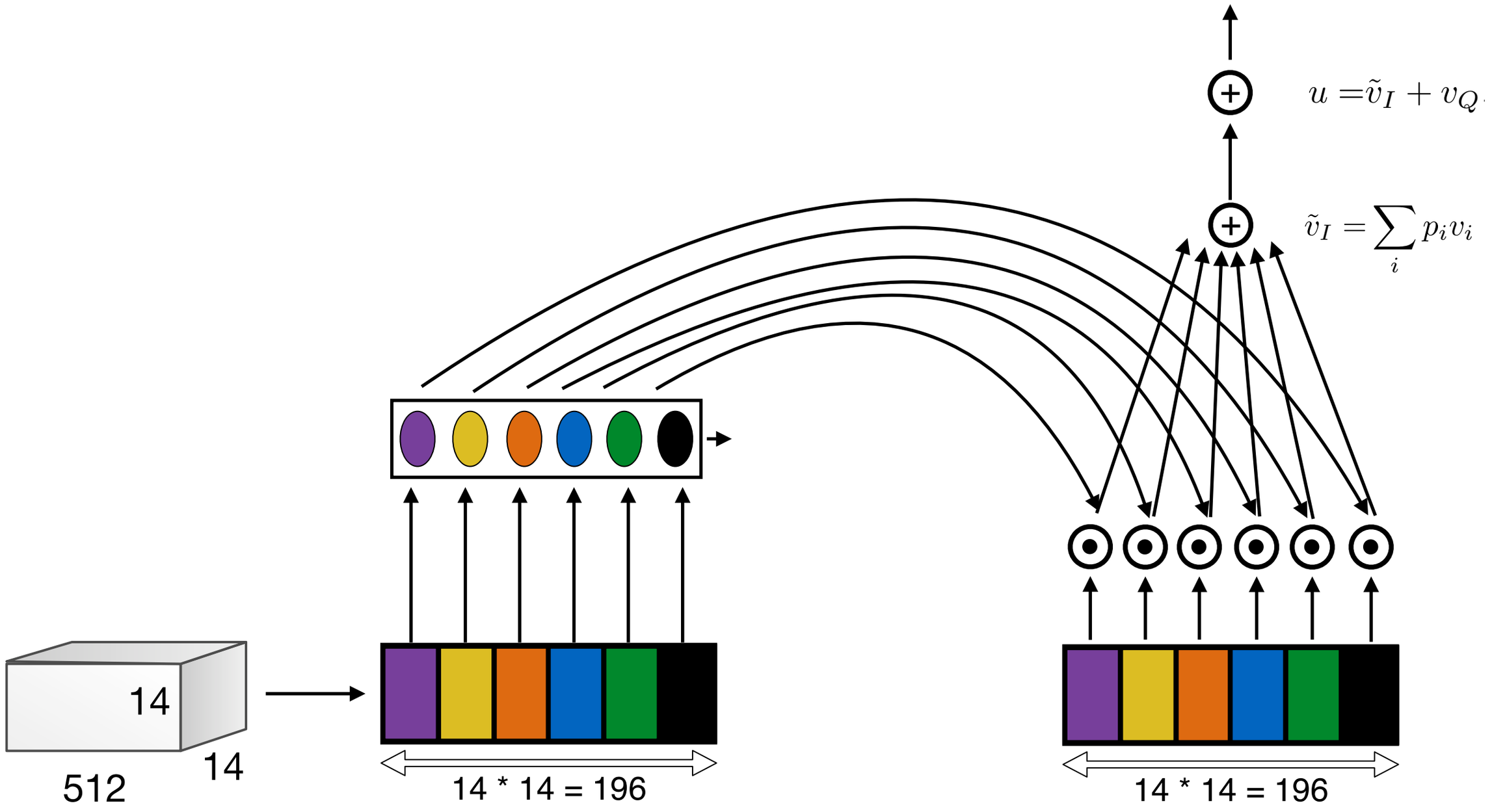
Vaswani et al. “Attention is All You Need”

Self-(Soft-)Attention in Visual Question Answering



Vaswani et al. “Attention is All You Need”

Self-(Soft-)Attention in Visual Question Answering

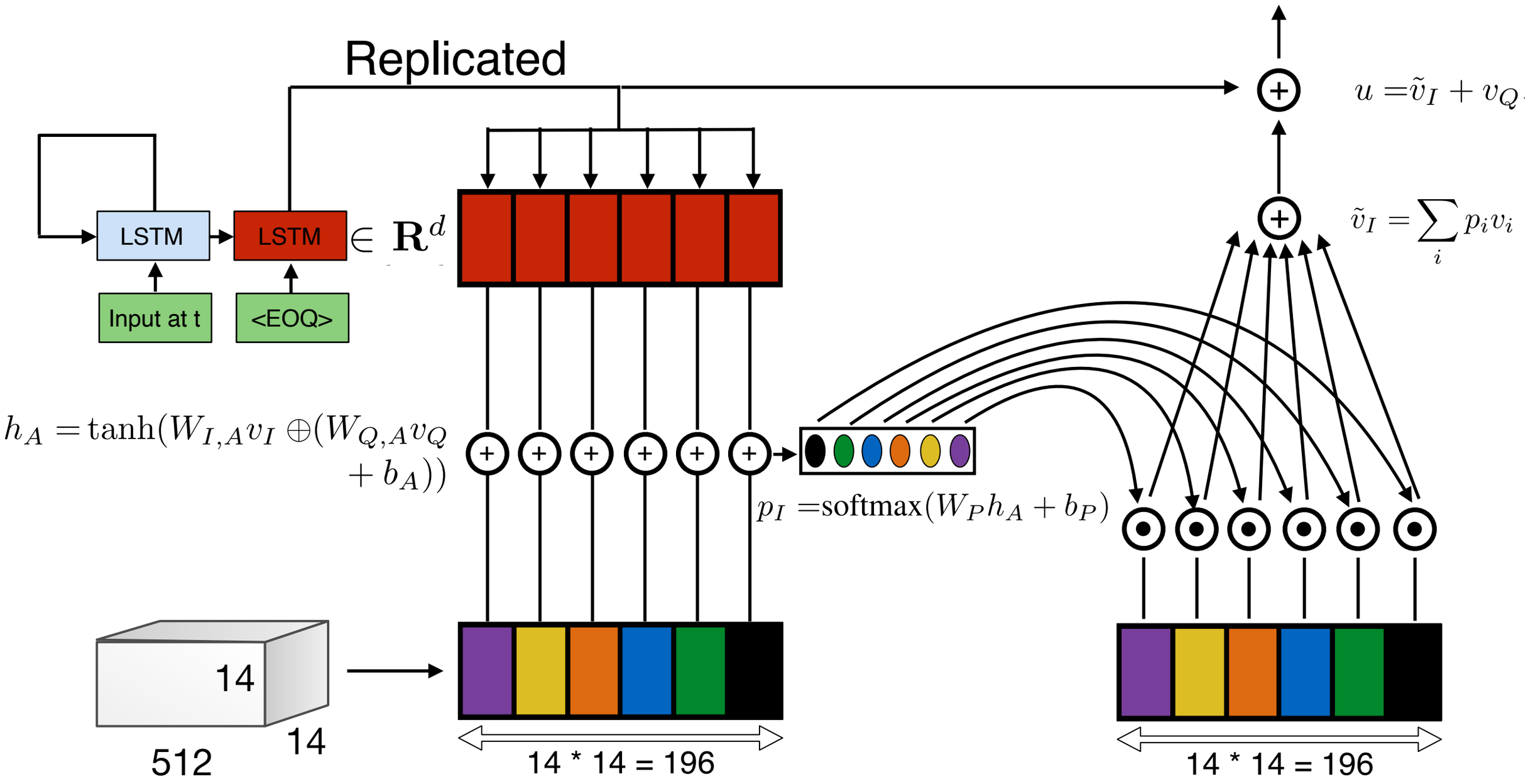


$$v_I = \tanh(W_I f_I + b_I) \in \mathbf{R}^{d \times m}$$

$$f_I = \mathbf{CNN}_{vgg}(I)$$

Vaswani et al. “Attention is All You Need”

Soft-Attention in Visual Question Answering



DeepMind

Hard-Attention

Mateusz Malinowski

Soft- vs Hard-Attention

- In soft-attention we compute the attention mask over the whole image

$$\alpha_i \in [0, 1]$$

Soft- vs Hard-Attention

- In soft-attention we compute the attention mask over the whole image

$$\alpha_i \in [0, 1]$$

- In hard-attention we compute a binary mask over the image, or equivalently we select a subset of interesting regions

$$\alpha_i \in \{0, 1\}$$

Hard-Attention in Computer Vision

- Challenges
  - Involves integers and this is a hard optimization problem
    - Even integer linear programming is already NP-hard
  - Integers mean non-differentiability, while our training frameworks are based on back propagation, and hence 'some form of differentiability' is required

Hard-Attention in Computer Vision

- Challenges
  - Involves integers and this is a hard optimization problem
    - Even integer linear programming is already NP-hard
  - Integers mean non-differentiability, while our training frameworks are based on back propagation, and hence 'some form of differentiability' is required
- On the other hand soft-attention can model to some extend hard-attention when used with 'softmax'

$$\alpha_i = \mathrm{softmax}(\phi(\hat{v}_i, \hat{q}; \theta_2))$$

  - Data=[1, 2, 4, 3], softmax=[.03, .08, .6, .2], argmax=[0, 0, 1, 0]
  - Data'=1000*Data, softmax=[0, 0, 1, 0], argmax=[0, 0, 1, 0]
  - Relationship between softmax and argmax

Hard-Attention in Computer Vision

- Challenges
  - Involves integers and this is a hard optimization problem
    - Even integer linear programming is already NP-hard
  - Integers mean non-differentiability, while our training frameworks are based on back propagation, and hence 'some form of differentiability' is required
- On the other hand soft-attention can model to some extend hard-attention when used with 'softmax'

$$\alpha_i = \text{softmax}(\phi(\hat{v}_i, \hat{q}; \theta_2))$$

  - Data=[1, 2, 4, 3], softmax=[.03, .08, .6, .2], argmax=[0, 0, 1, 0]
  - Data'=1000*Data, softmax=[0, 0, 1, 0], argmax=[0, 0, 1, 0]
  - Relationship between softmax and argmax

$$\lim_{\epsilon \to 0^+} \text{softmax}(\epsilon^{-1}[a_1, a_2, a_3, ...]) = [z_1, z_2, z_3, ...]/Z$$

$$\text{where } z_i = \begin{cases} 1 & \text{if } i = \arg\max_j([a_1, a_2, a_3, ...]) \\ 0 & \text{otherwise} \end{cases} \text{ and } Z = |\{i \mid i = \arg\max_j([a_1, a_2, a_3, ...])\}|$$

Hard-Attention in Computer Vision

- Cutting edge of Machine Learning and Computer Vision research
  - **Narrower question**:
    How to incorporate Hard-Attention into a backprop-based frameworks
  - **Broader question:**
    How to deal with discrete variables in backdrop-based frameworks?

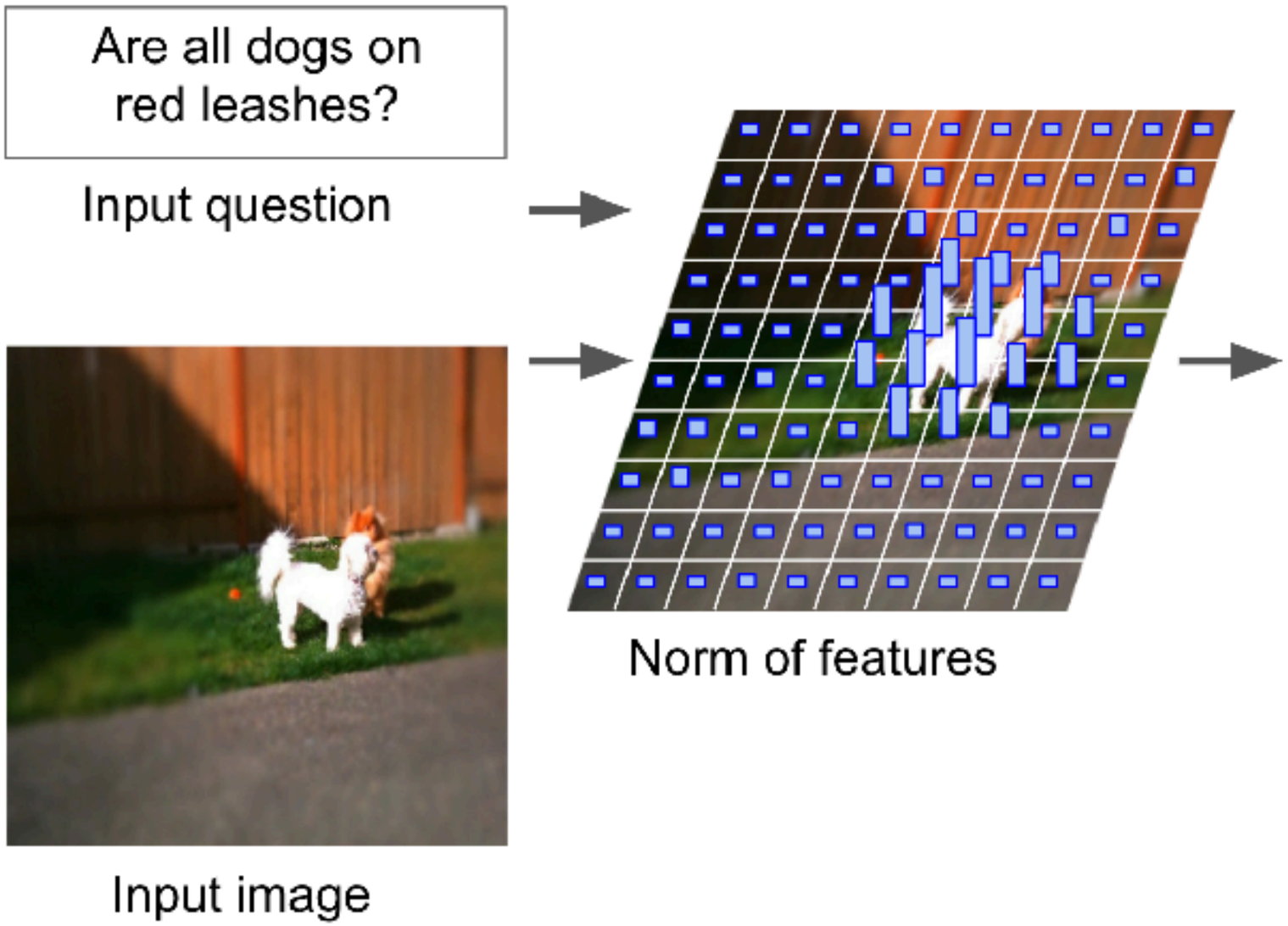
Hard-Attention in Computer Vision

- Many approaches
  - Gumbel samples
  - Straight-through estimators
  - Reinforcement learning / REINFORCE
  - A combination of max-pooling and L2-norm

Why Hard-Attention?

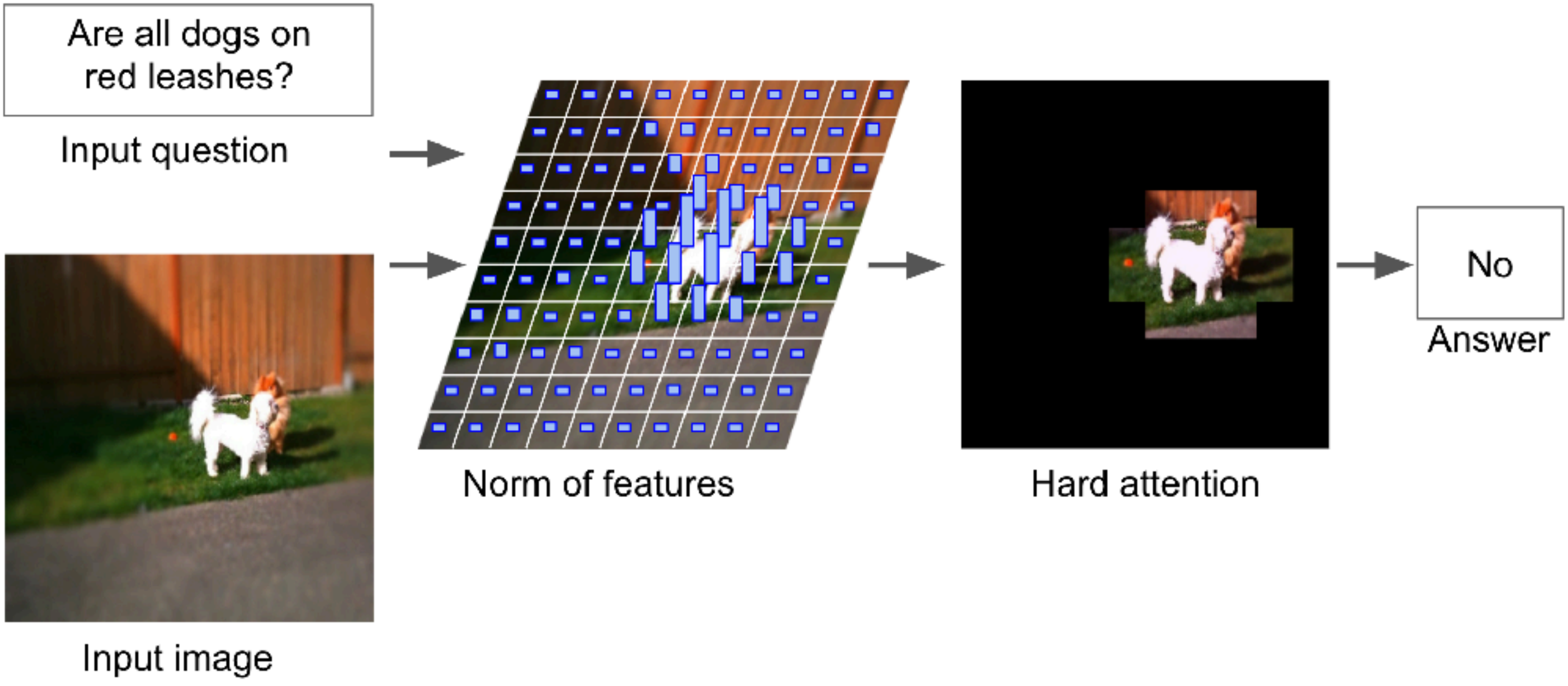
- Computationally more efficient
  - Working with subset of input data
- Statistically more efficient
  - We ‘do not mix vectors together’ but instead we choose the relevant subset
- Biologically more plausible
  - Animals arguably have both attention mechanisms
- Somehow more elegant

Hard-Attention [optional]



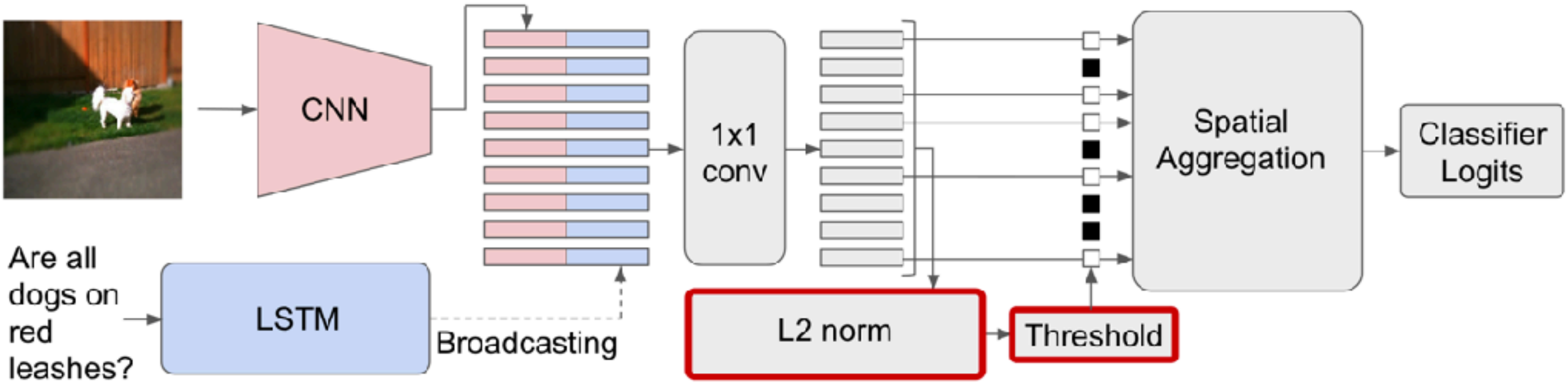
Malinowski et al. “Learning Visual Question Answering by Bootstrapping Hard-Attention”

Hard-Attention [optional]



Malinowski et al. “Learning Visual Question Answering by Bootstrapping Hard-Attention”

Hard-Attention [optional]



$$max(X, top = 1)$$

Generalization of Max-Pooling

$$max(X, \mathcal{M}(X), top = k)$$

Malinowski et al. “Learning Visual Question Answering by Bootstrapping Hard-Attention”

Results on CLEVR

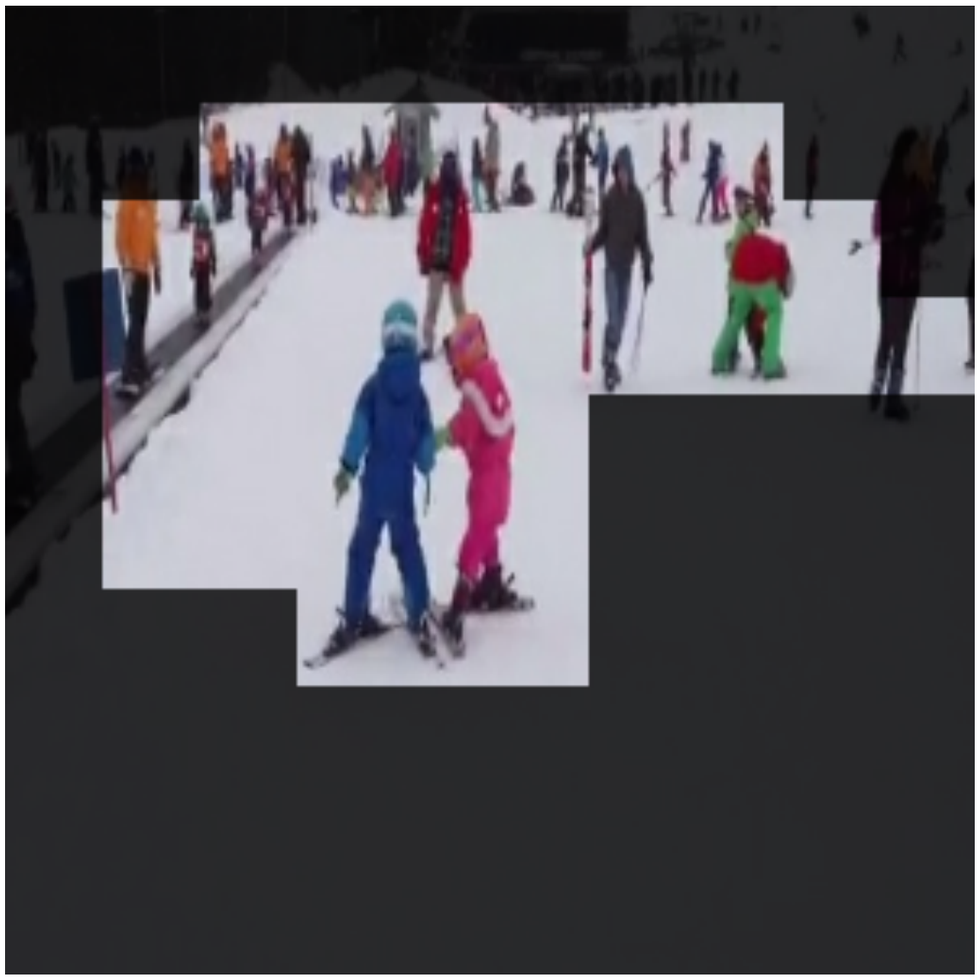
HAN + RN, 25%	95.2
HAN + RN (+), 25%	96.9
HAN + RN (++), 33% (64 cells [same as RN])	**98.8**
RN [1]	95.5
Human [2]	92.6

Hard-Attention [optional]

Are this lions? No

Malinowski et al. “Learning Visual Question Answering by Bootstrapping Hard-Attention”

Hard-Attention [optional]



Are all these people moving?

Yes

What color is this train?Yellow

Malinowski et al. “Learning Visual Question Answering by Bootstrapping Hard-Attention”

DeepMind

Behavioural Tests

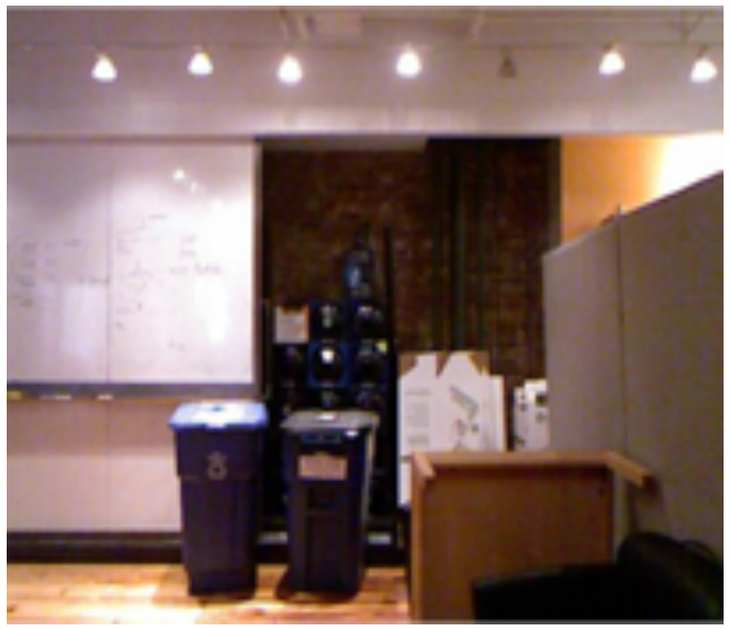
Mateusz Malinowski

Sufficiently complex system needs to be tested as a whole

- Car crash tests
- Wind tunnels
- Acceptance tests
- End-to-end tests

Visual Question Answering

- Questions about Images
  - Images
  - Questions
- **Not** is evaluated internal mechanisms, but the **whole behaviour**



What is behind the table?
sofa

What is on the refrigerator?
magnet, paper

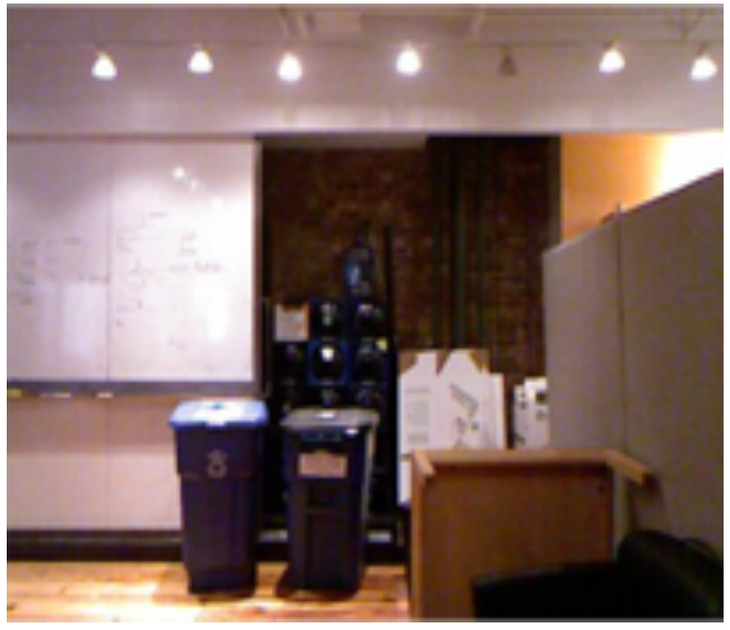
What color are the cabinets?
brown

How many lamps are there?
2

Malinowski and Fritz “A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input”

Visual Question Answering

- Questions about Images
  - Images
  - Questions
- **Not** is evaluated internal mechanisms, but the **whole behaviour**
  - **Scene understanding by asking questions**



What is behind the table?
sofa

What is on the refrigerator?
magnet, paper

What color are the cabinets?
brown

How many lamps are there?
2

Malinowski and Fritz “A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input”

But …

- It is easy to be fooled by such tests
  - Giving a right answer for wrong reasons

Malinowski and Fritz “A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input”

Let’s build a wonderful VQA system together

How many eyes are in the picture?

Let’s build a wonderful VQA system together

How many eyes are in the picture? -> 2

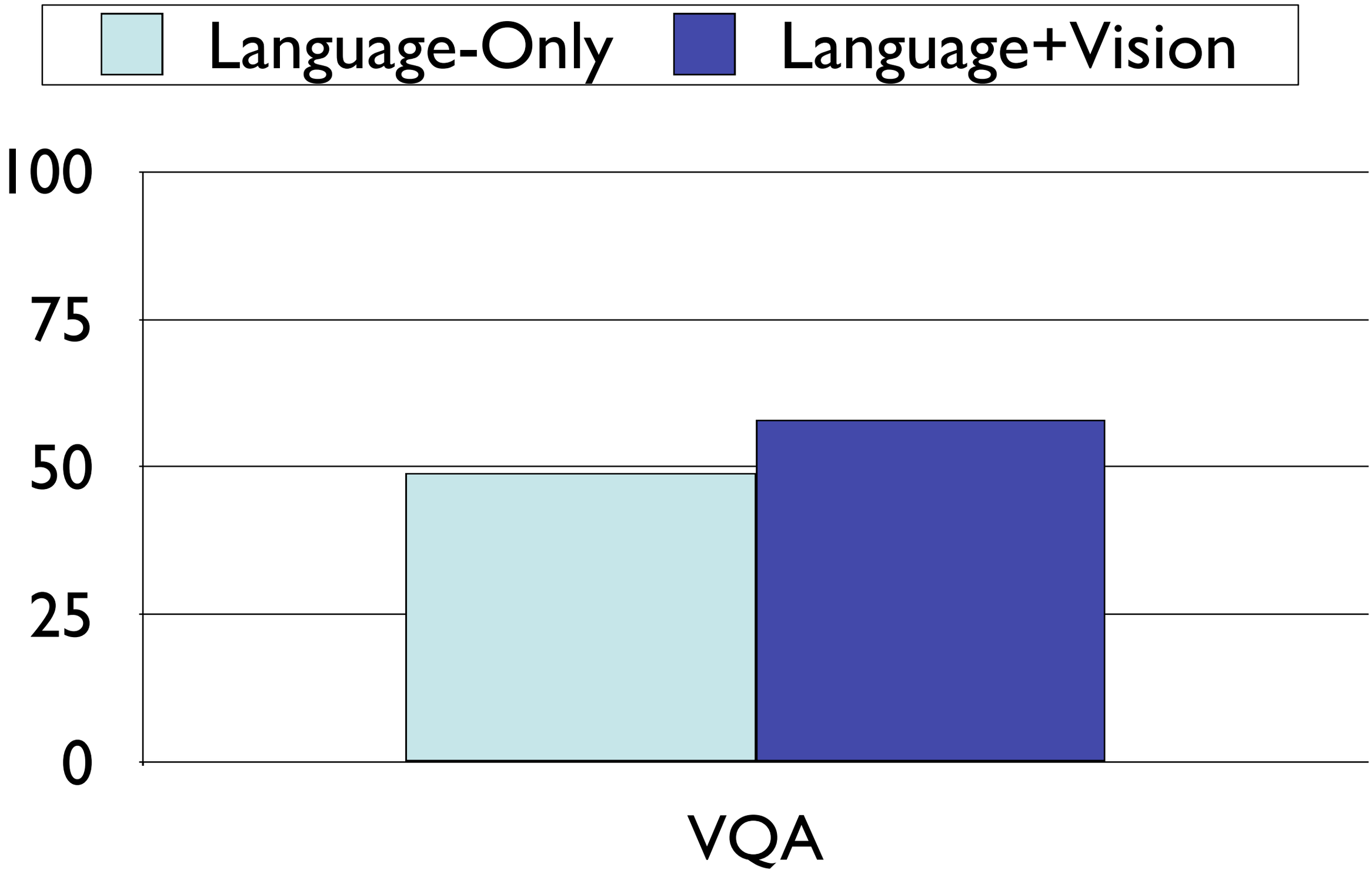
Let’s build a wonderful VQA system together

How many eyes are in the picture?

Let’s build a wonderful VQA system together

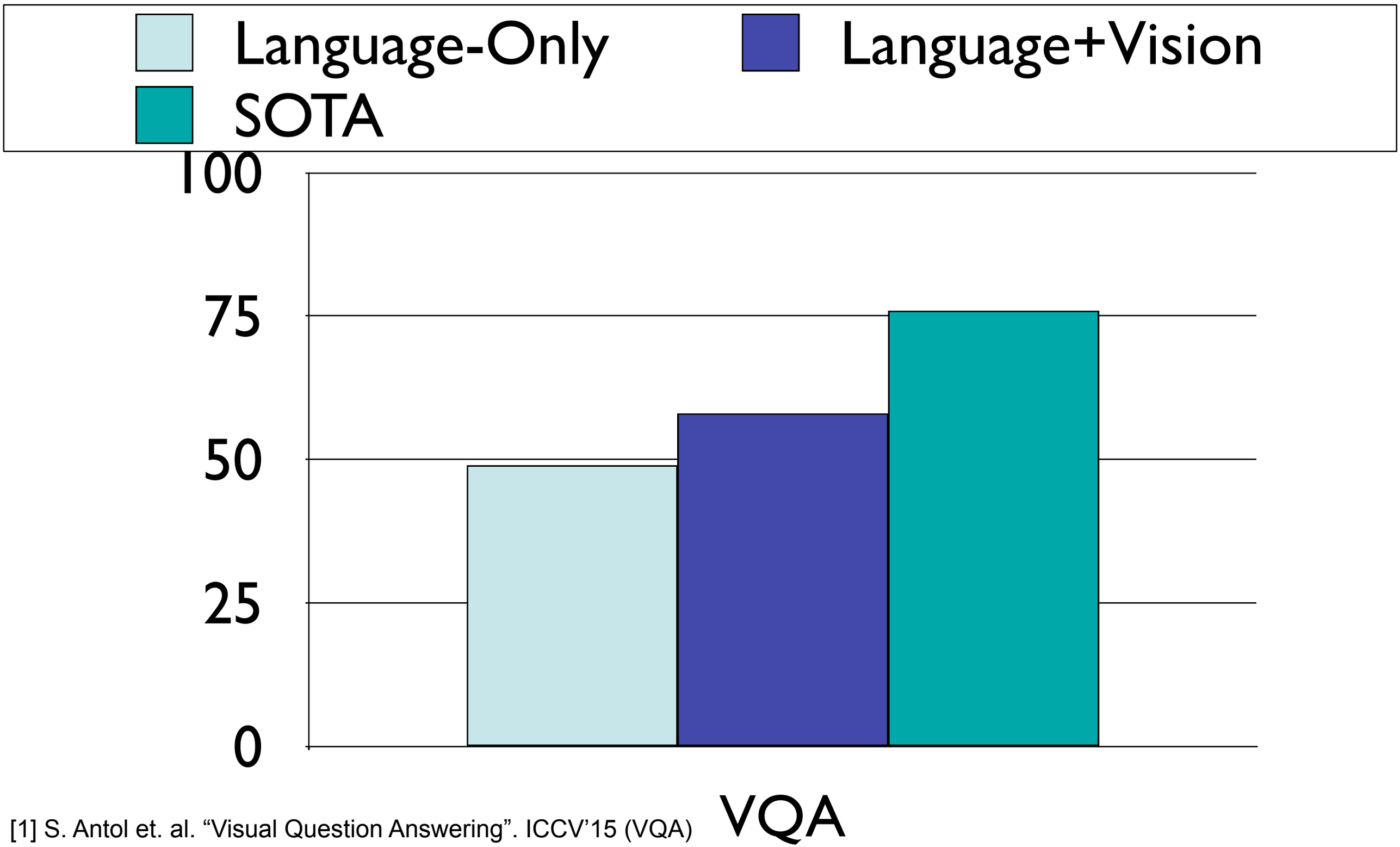
How many eyes are in the picture? -> 2

Performance on VQA



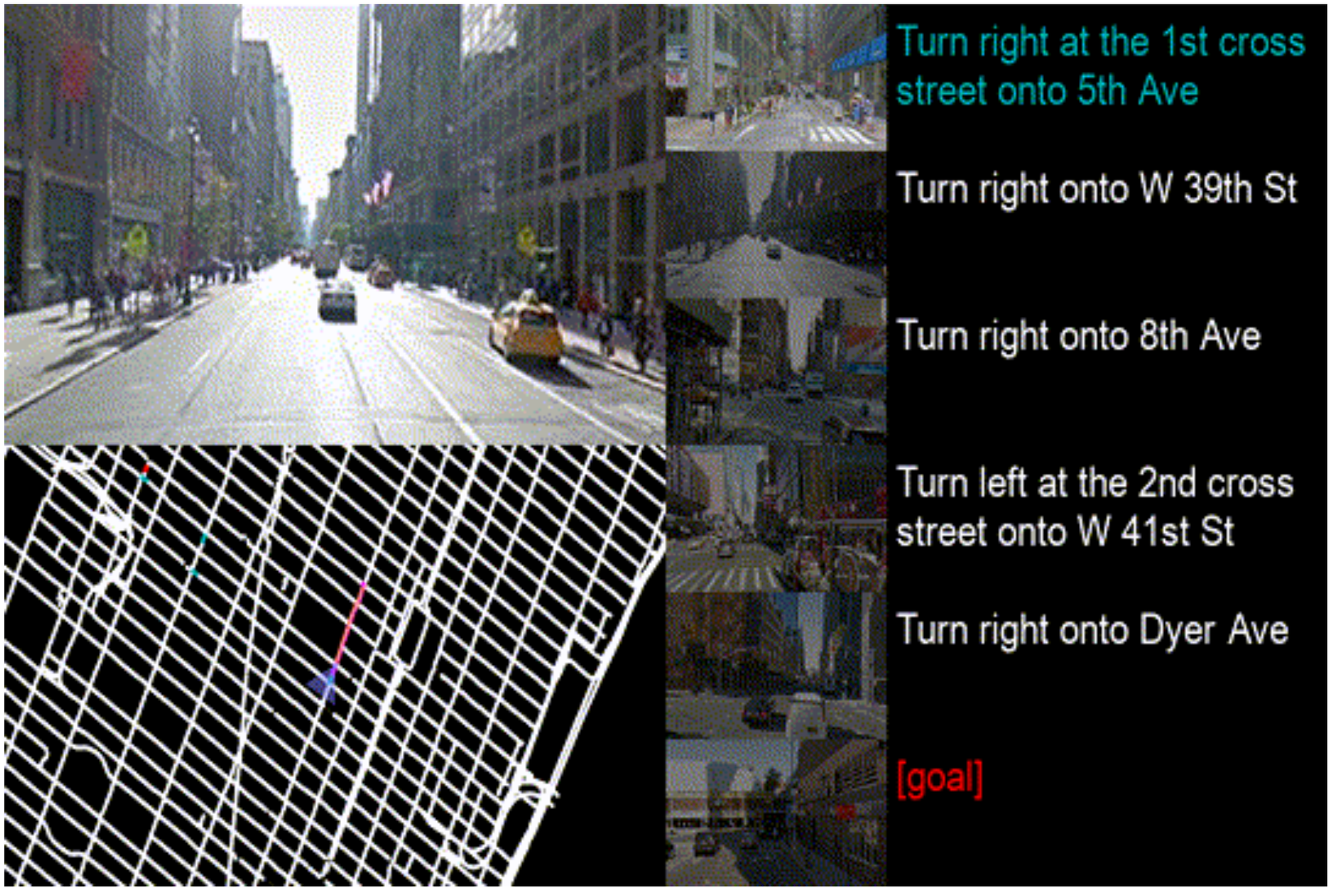
[1] S. Antol et. al. “Visual Question Answering”. ICCV’15 (VQA)
[2] M. Malinowski et. al. “Ask Your Neurons: A Deep Learning Approach to Visual Question Answering”. NeurIPS’14 (DAQUAR)

Performance on VQA



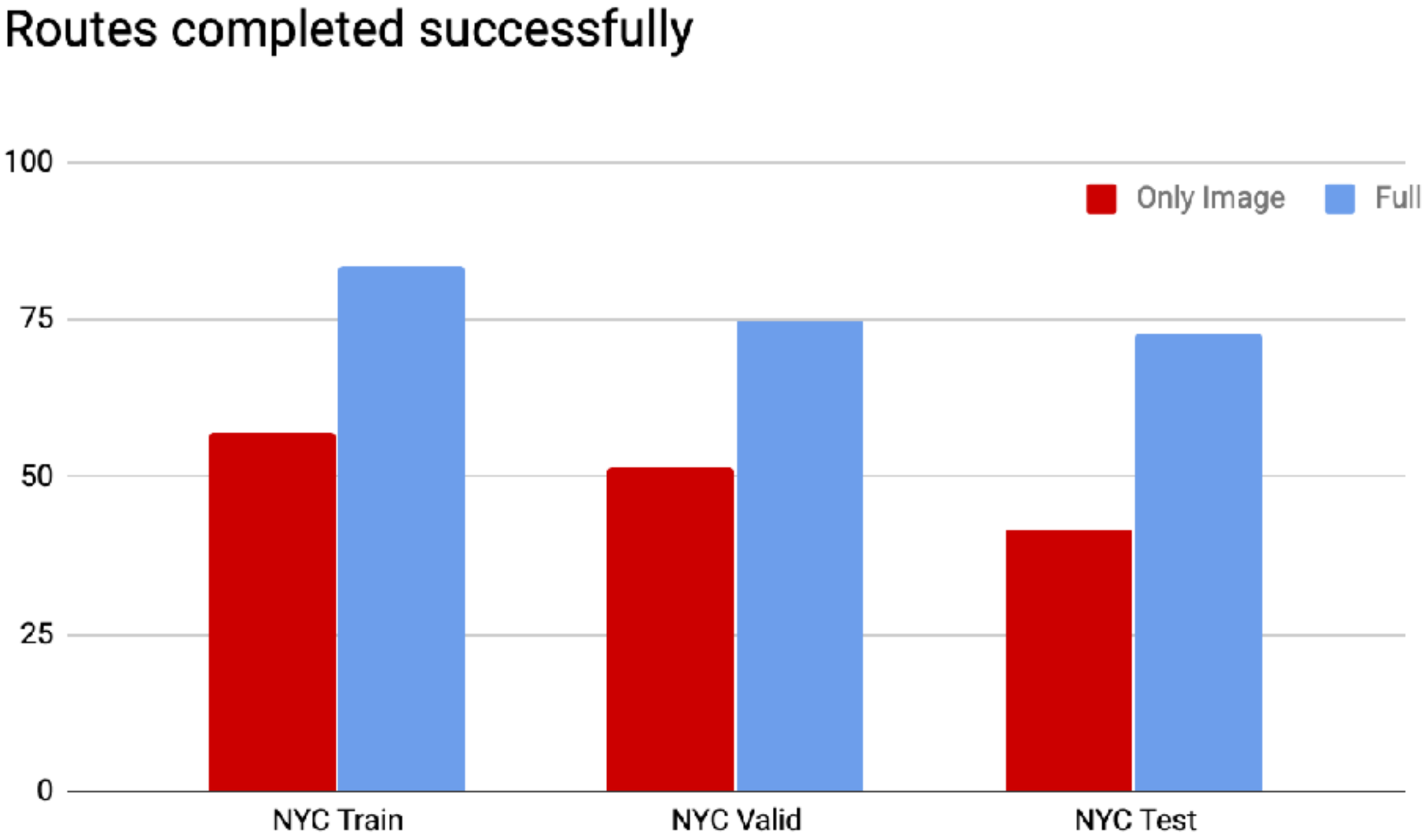
[1] S. Antol et. al. “Visual Question Answering”. ICCV’15 (VQA)
[2] M. Malinowski et. al. “Ask Your Neurons: A Deep Learning Approach to Visual Question Answering”. NeurIPS’14 (DAQUAR)
[3] D. Nguyen et al. Grid Features + MoVie (SOTA)

Bias in StreetLearn



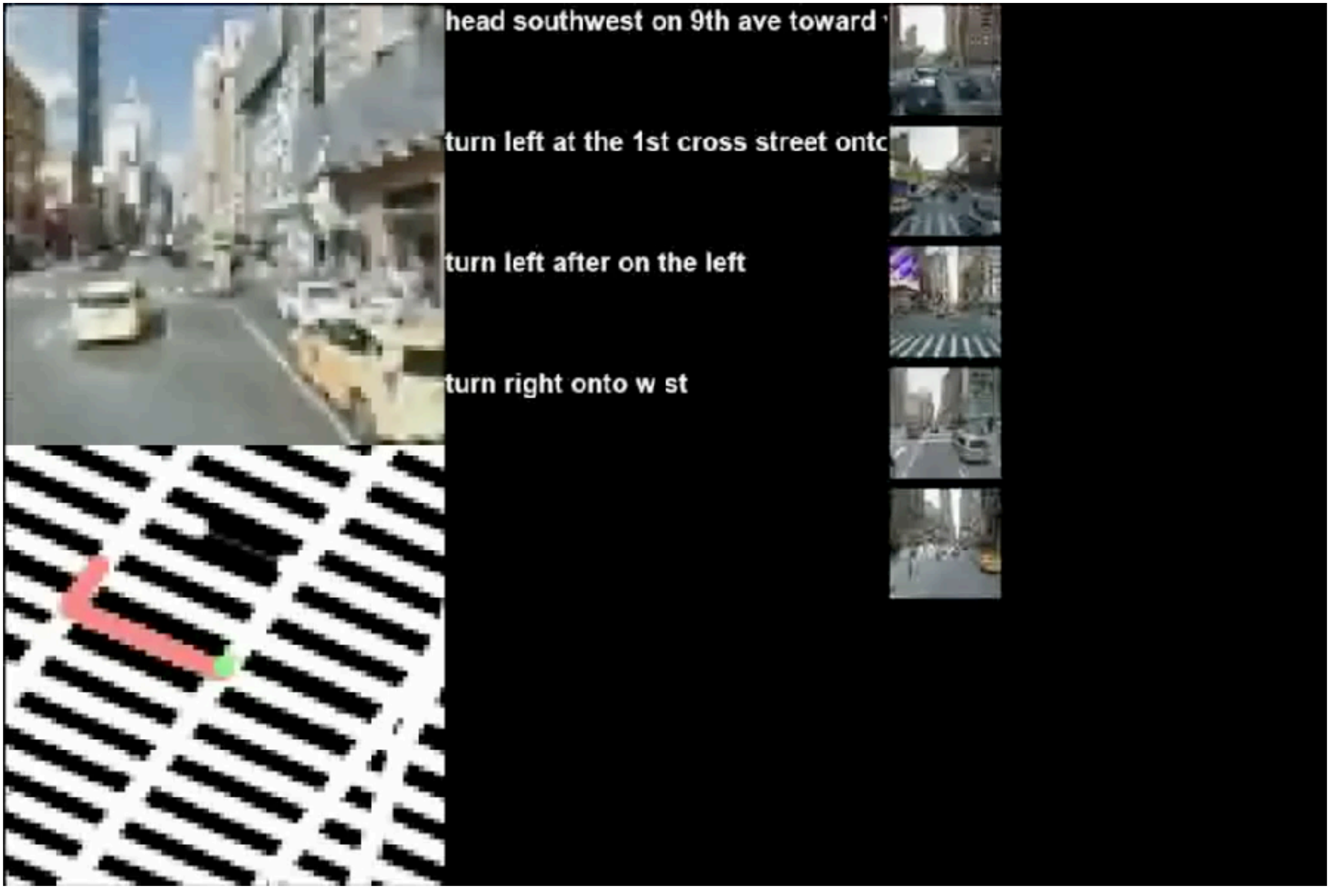
Malinowski, Hermann, Mirowski et al. “Learning To Follow Directions in Street View”

Bias in StreetLearn



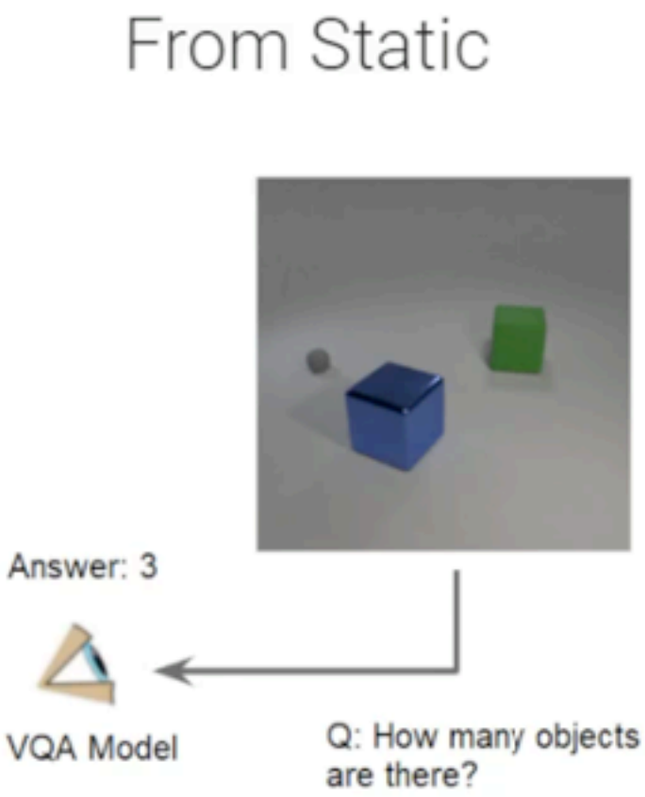
Malinowski, Hermann, Mirowski et al. “Learning To Follow Directions in Street View”

Bias in StreetLearn



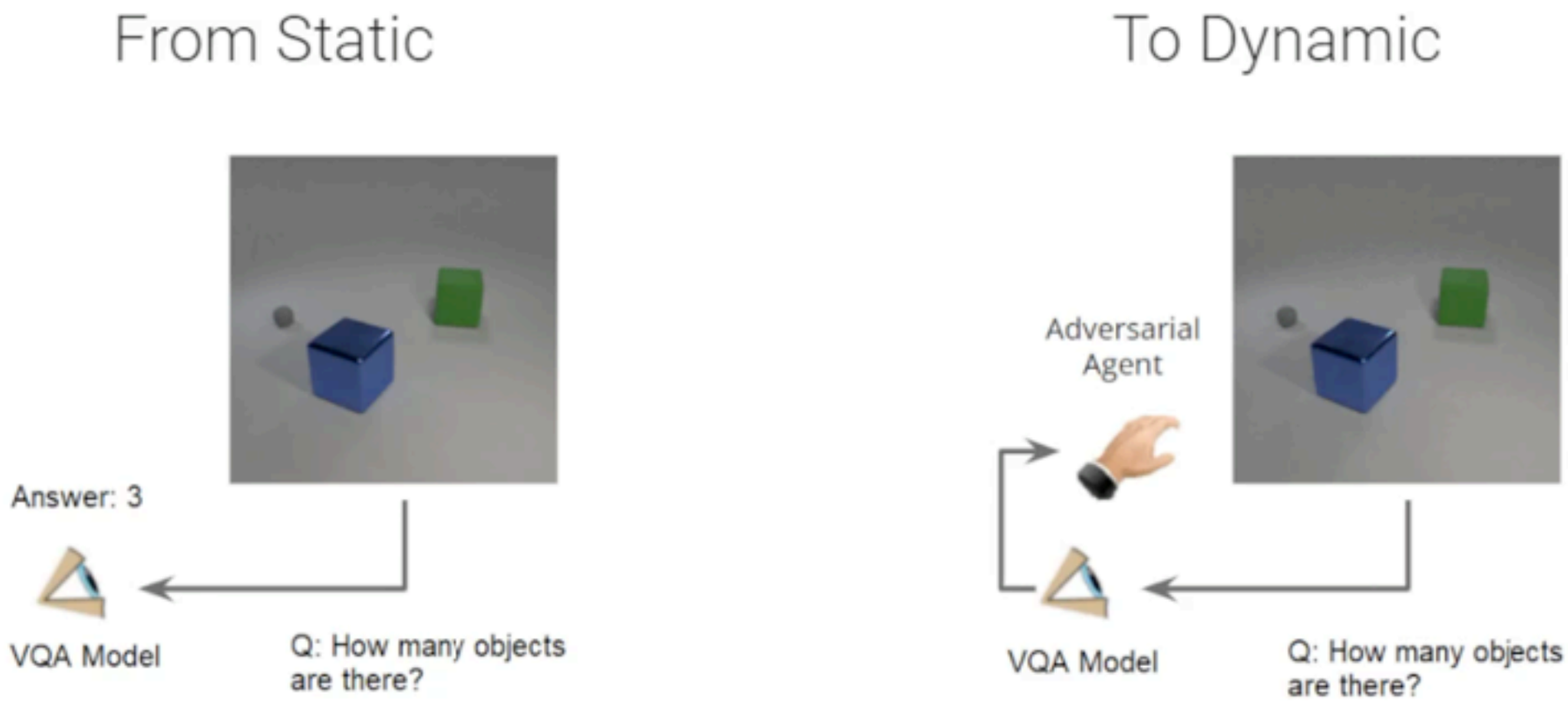
Malinowski, Hermann, Mirowski et al. “Learning To Follow Directions in Street View”

Dynamic & adversarial environments



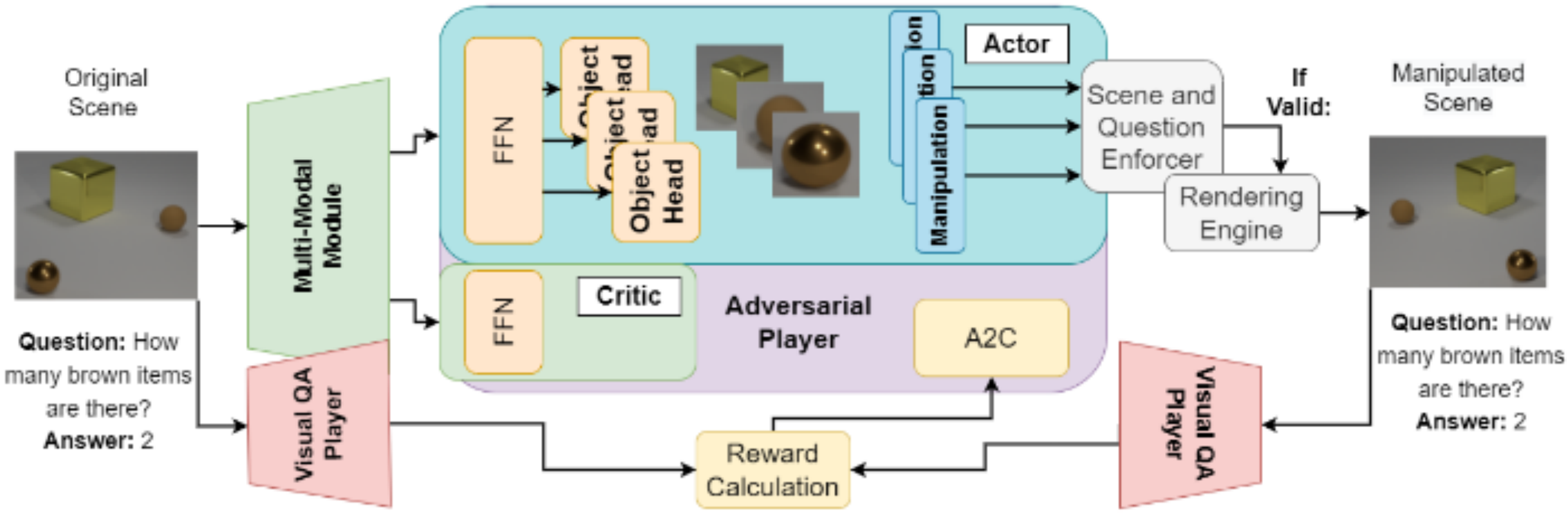
Mouselinos et al. “Measuring CLEVRness: Blackbox testing of Visual Reasoning Models”

Dynamic & adversarial environments



Mouselinos et al. “Measuring CLEVRness: Blackbox testing of Visual Reasoning Models”

Two-agent system

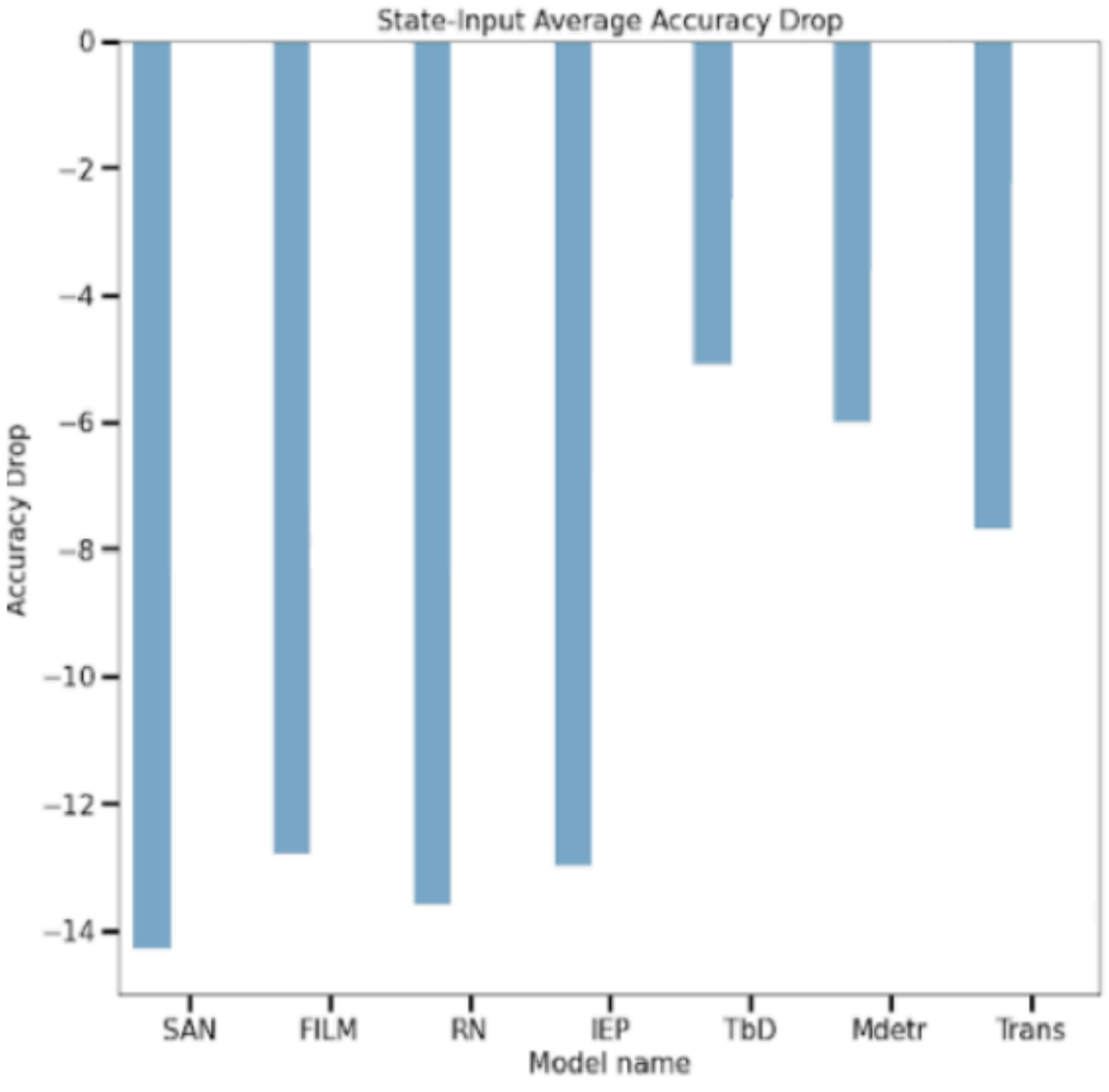


Mouselinos et al. “Measuring CLEVRness: Blackbox testing of Visual Reasoning Models”

“Super-human” performance on CLEVR

SAN (Yang et al.)	72.1	Raw Signals
FILM (Perez et al.)	96.2	
RN (ours)	95.5	
IEP (Johnson et al.)	96.9	Program
TbD (Mascharka et al.)	99.1	
MDetr (Kamath et al.)	99.7	States

Drop in performance



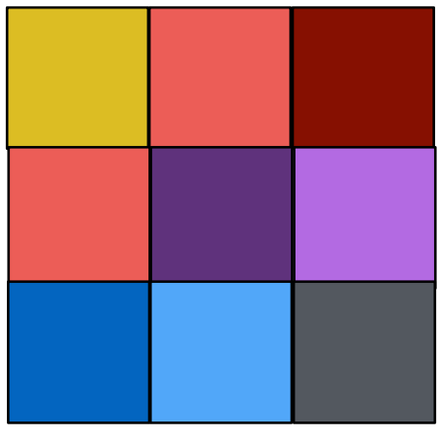
Mouselinos et al. “Measuring CLEVRness: Blackbox testing of Visual Reasoning Models”
More under this link: https://www.measuringclevrness.com

DeepMind

AI Creativity

Mateusz Malinowski

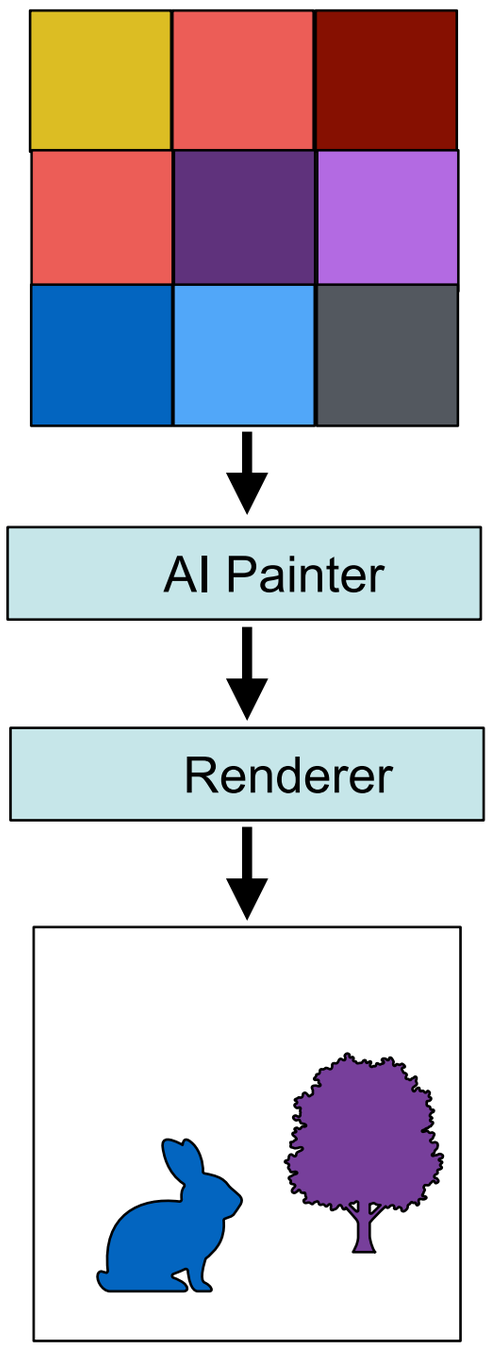
AI Art



“A blue rabbit and a purple tree”

Open AI “CLIP: Connecting Text and Images”
C. Fernando et al. “Arnheim 2.0”

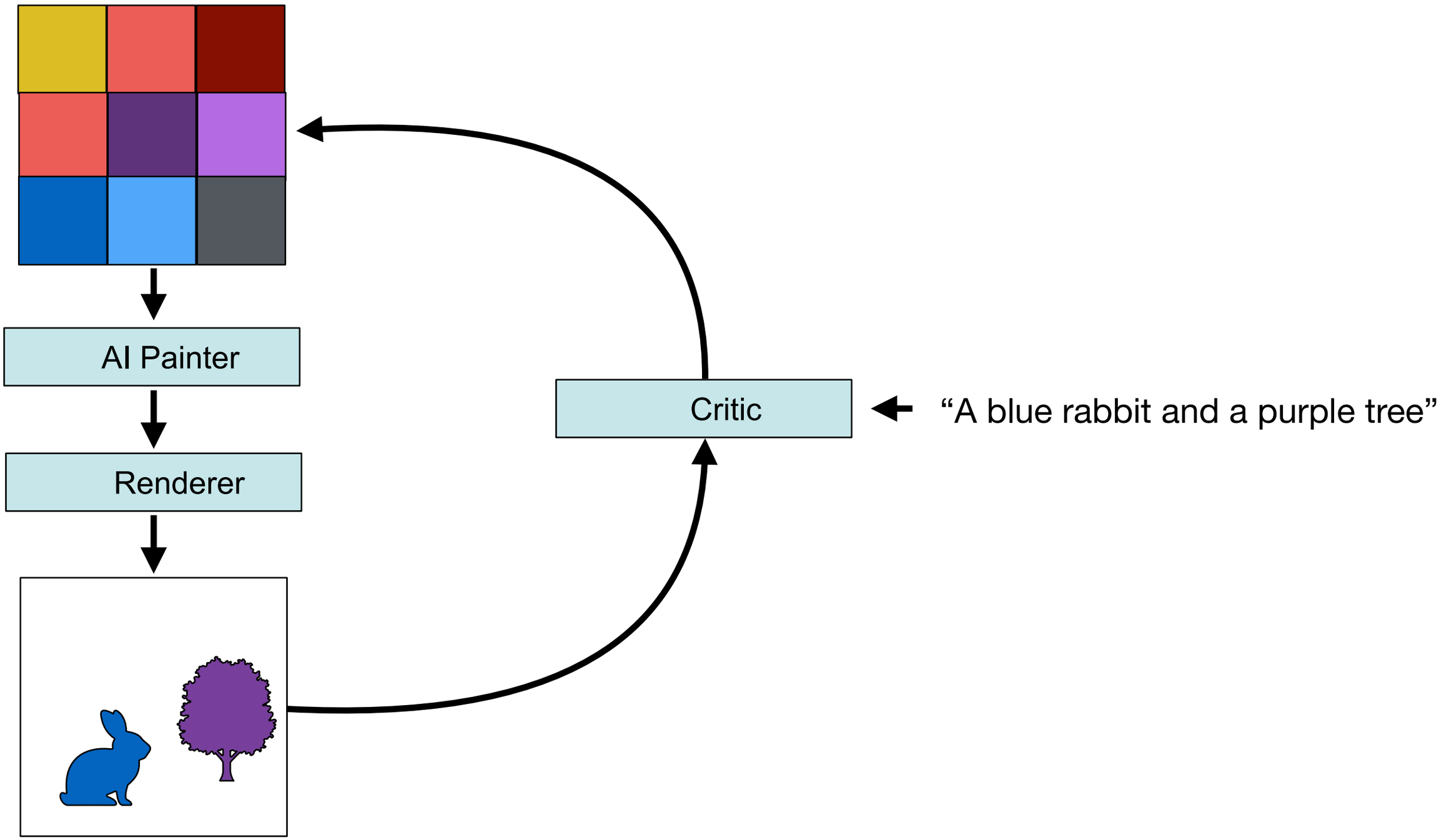
AI Art



“A blue rabbit and a purple tree”

Open AI “CLIP: Connecting Text and Images”
C. Fernando et al. “Arnheim 2.0”

AI Art



Open AI “CLIP: Connecting Text and Images”
C. Fernando et al. “Arnheim 2.0”

Role of Art in our society

Documenting events

C. Stanfield “Battle of Trafalgar”

Role of Art in our society

Documenting events

C. Stanfield “Battle of Trafalgar”

Amplifying social norms

J. Pałucha “Polish Casino”

Role of Art in our society

Documenting events

C. Stanfield “Battle of Trafalgar”

Amplifying social norms

J. Pałucha “Polish Casino”

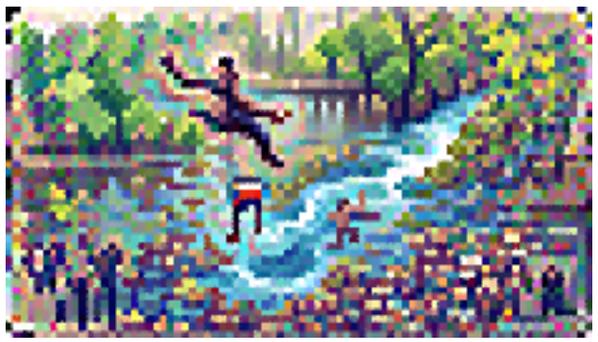
Investigating space and time

G. Balla “Dynamism of a dog on a leash”

Verb understanding

A man jumping into a river

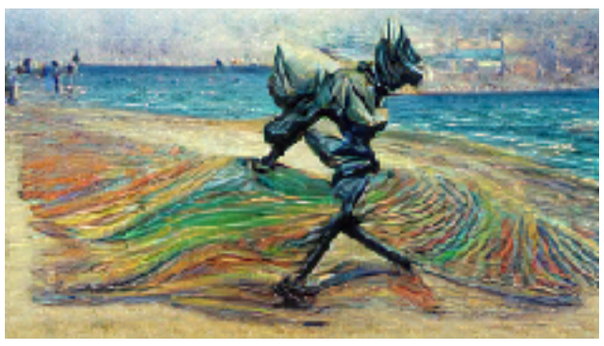


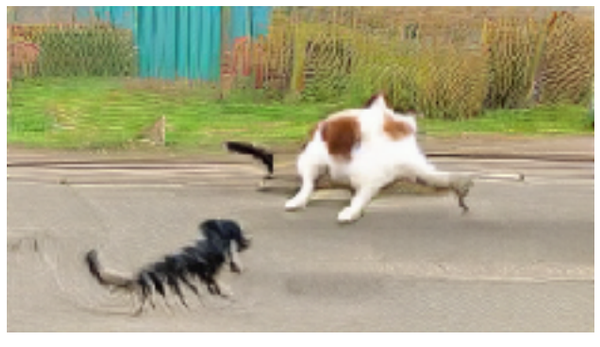
A person jogs on the beach



Umberto Boccioni



A cat chasing a dog




A dog chasing a cat


Based on “Pixray PixelDraw”

Multiplicity

CLIP-VQGAN “A cat chasing a dog”
AI-creativity

Based on “Pixray PixelDraw”

G. Balla “Dynamism of a dog on a leash”. Futurism.

The real thing.

Preposition understanding

A cup under a table

#Lucio Fontana, Paul Gauguin

Based on “Pixray PixelDraw”

Preposition understanding

A cup under a table

#Lucio Fontana, Paul Gauguin

Based on “Pixray PixelDraw”

Intuitive physics

A ball falling off the table

Based on “Pixray PixelDraw”

Intuitive physics, combinatorial creativity

A ball falling off the table

Half-car half-bike

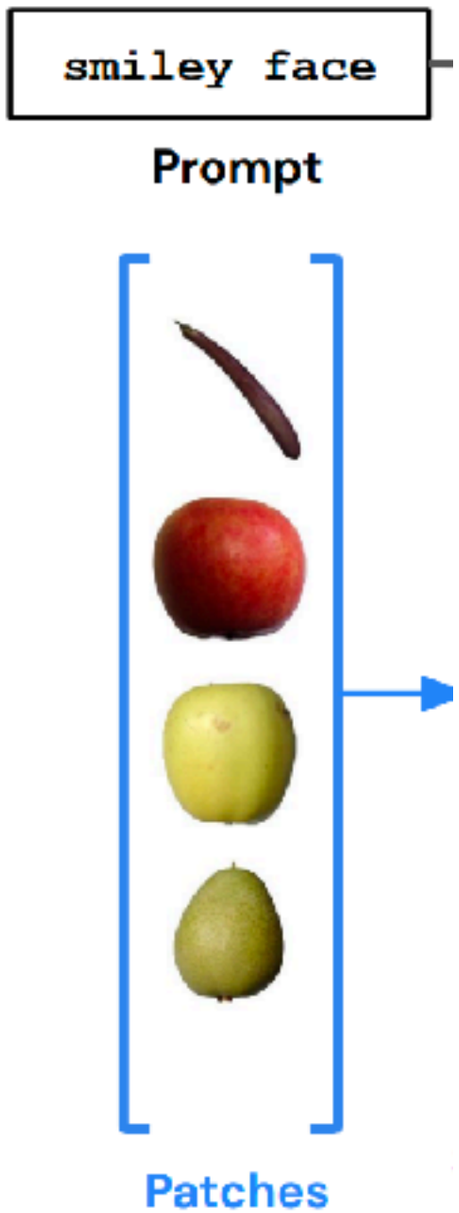
Cubical ball

Based on “Pixray PixelDraw”

Colorless green ideas sleep furiously. Grounded!

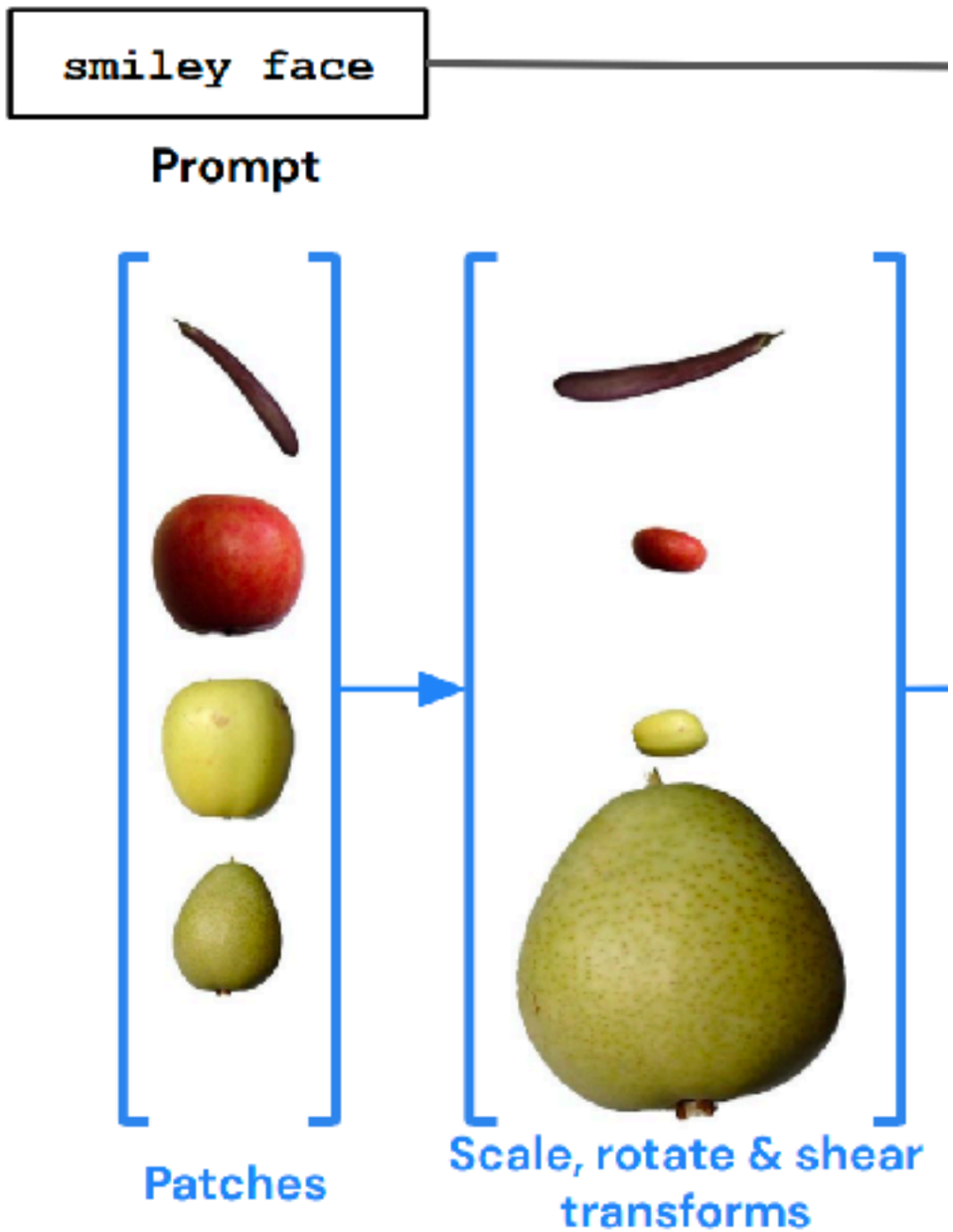
Based on “Pixray PixelDraw”

“Collage”



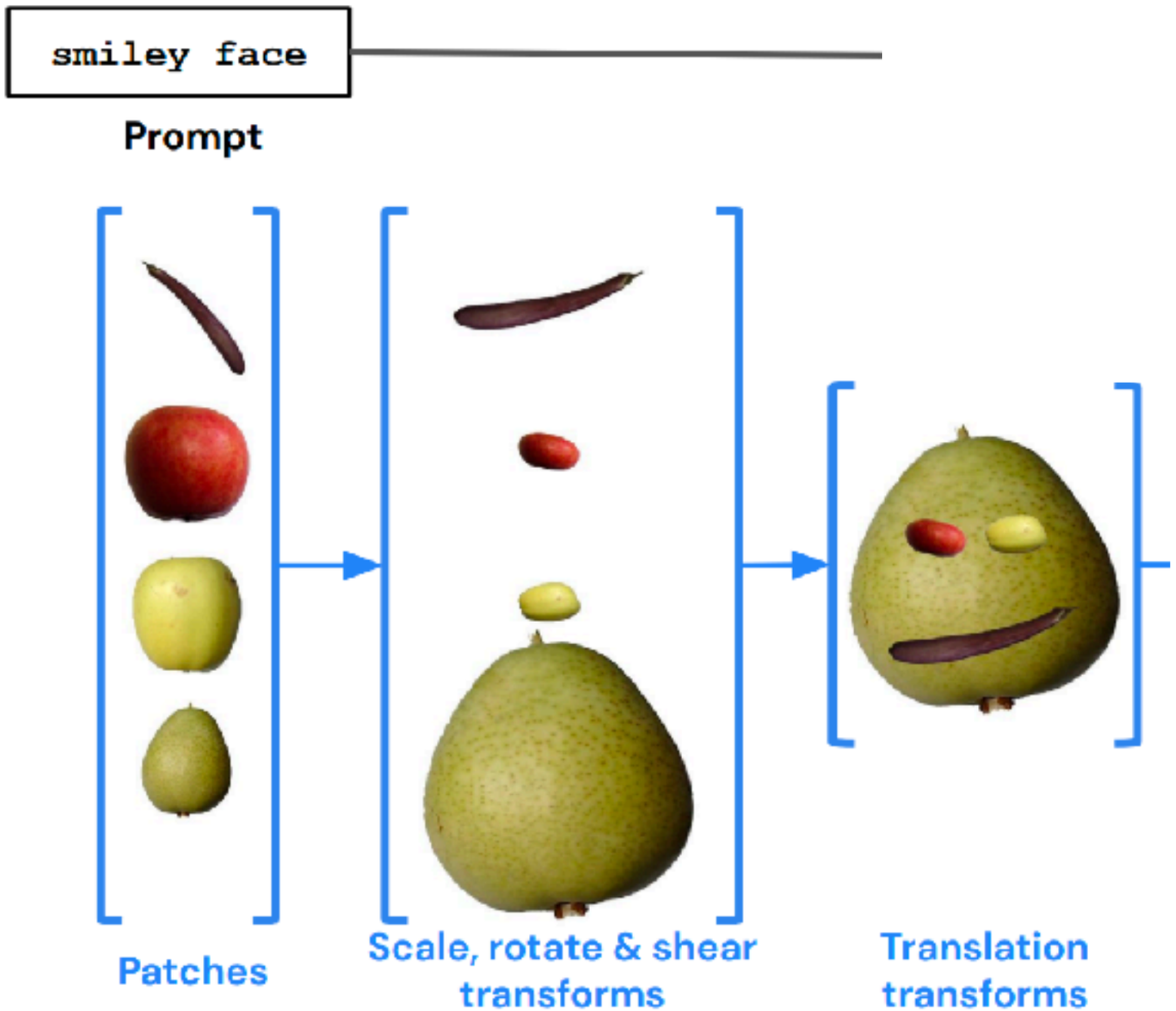
P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

“Collage”



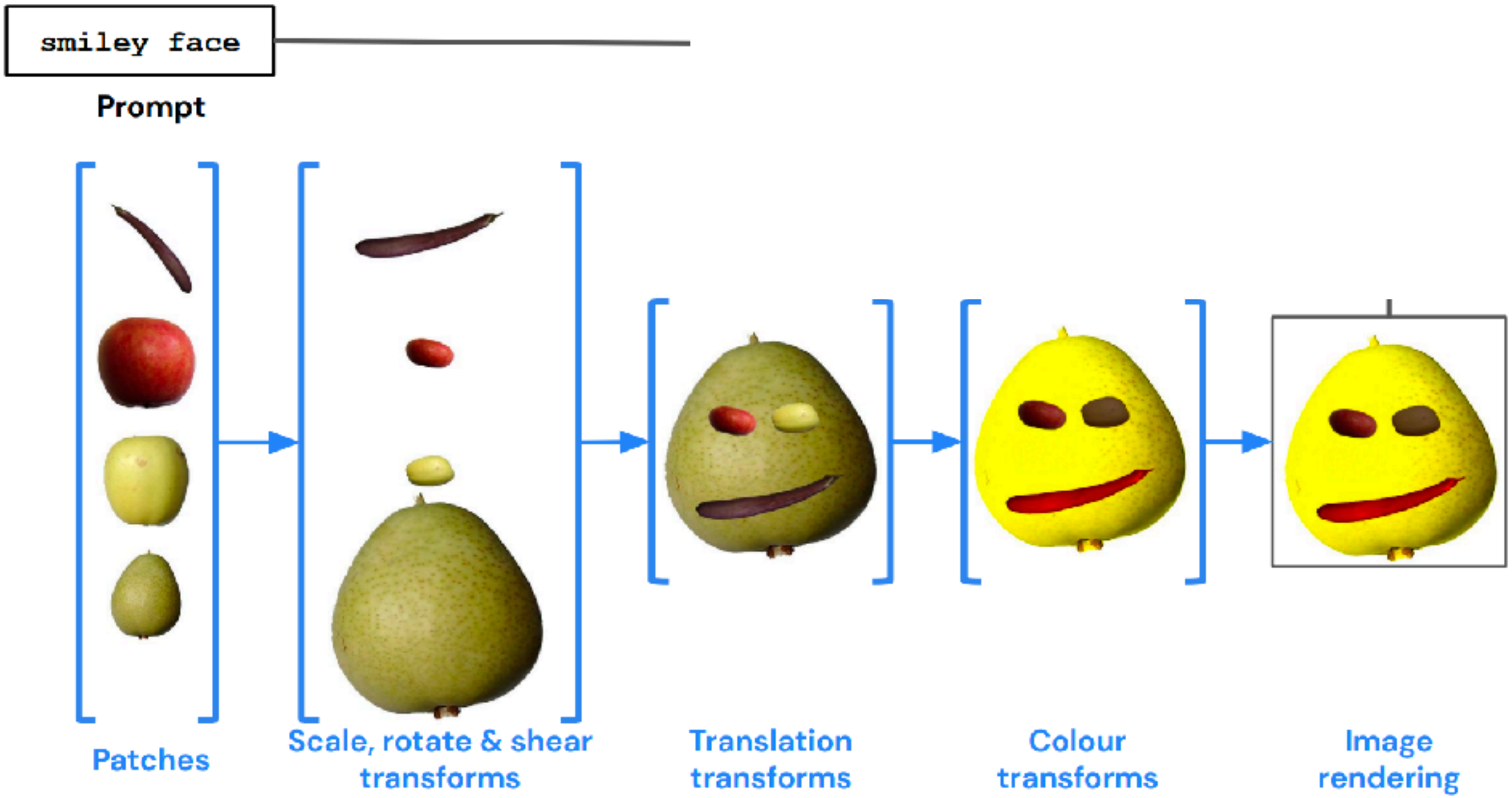
P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

“Collage”



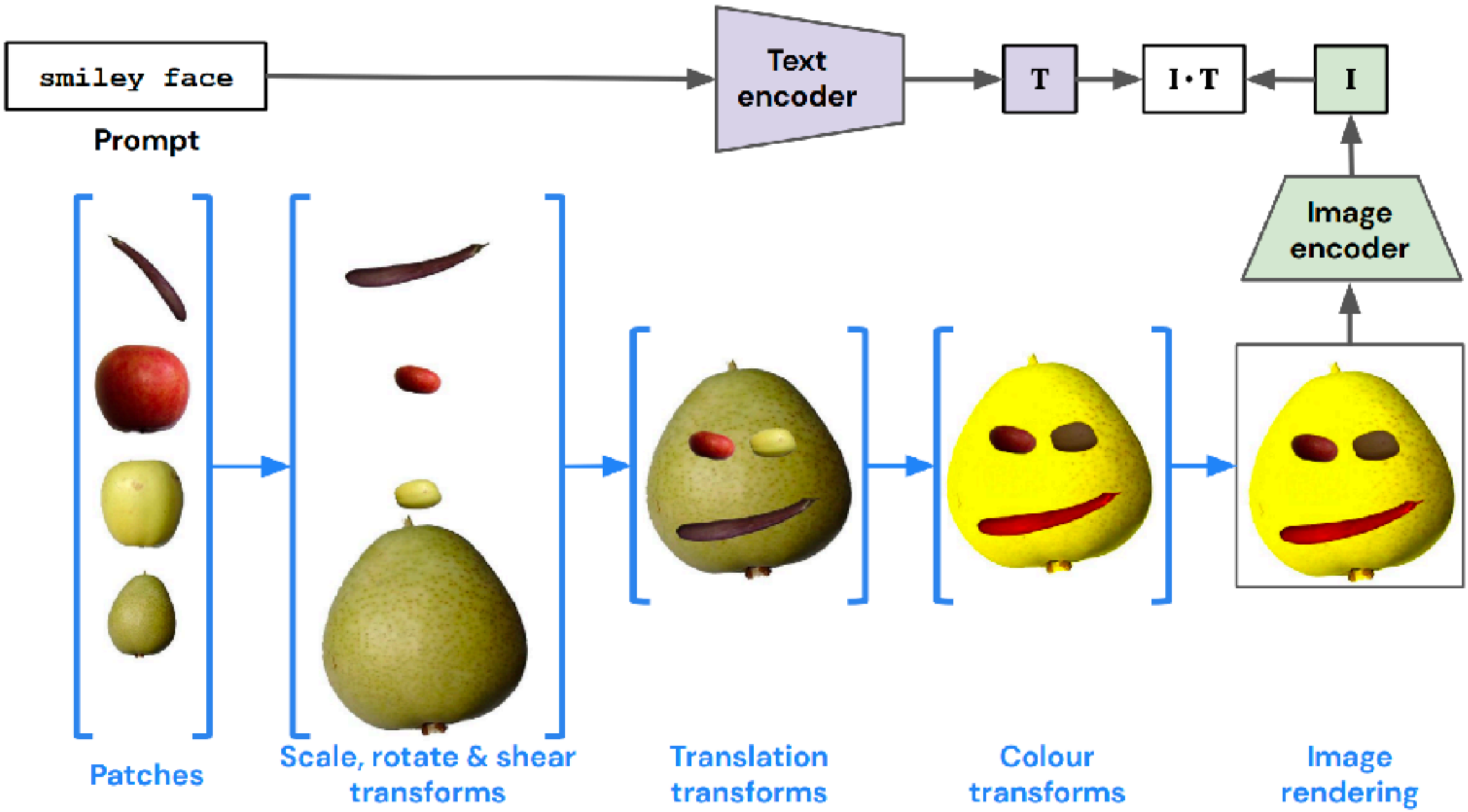
P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

“Collage”



P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

“Collage”



P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

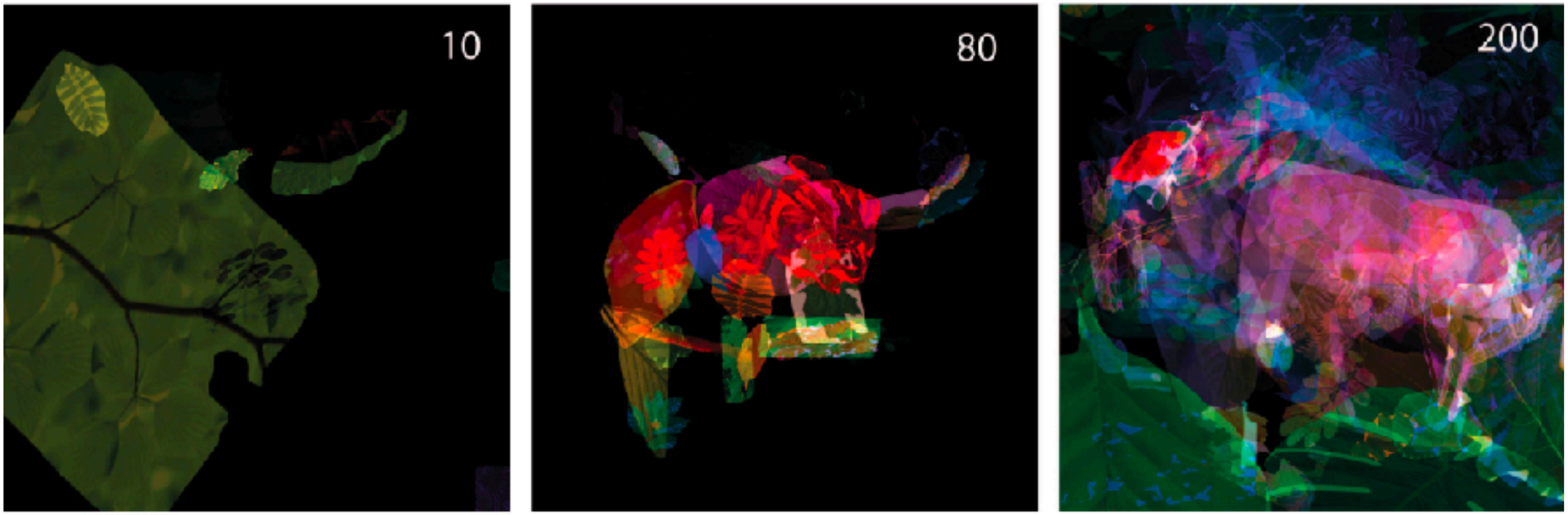
The Fall of the Damned after Rubens and Eaton

P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

Underwater coral

P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

Level of abstraction (Bull)



P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

Different assets (animals, fruits, animals)

P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

Different assets (animals, fruits, animals)

P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

Turing (animals, broken plates)

P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

DALLE-2 (Open AI)

An astronaut riding a horse in a photorealistic style

DeepMind

Summary

Mateusz Malinowski

Why do we need explainable AI?

Why do we need explainable AI?

- Neural Nets are black-box and thus hard to understand
  - And so to debug

Why do we need explainable AI?

- Neural Nets are black-box and thus hard to understand
- Can we fully trust systems that we don’t understand?
  - It is about building trust between us

What should we do?

- No single test, so we might need to rely on suite of various tests

What should we do?

- No single test, so we might need to rely on suite of various tests
- Each test is imperfect
  - Hard to quantify “explanation” (e.g. attention)
  - We might loose interpretability at different layers of the network
  - Often we need special datasets to quantify explainability