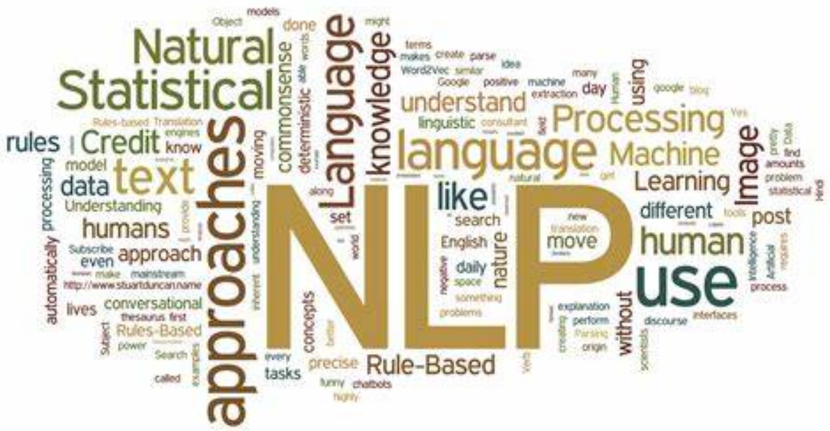
facebook
Artificial Intelligence Research

Multi-Modal Machine Learning

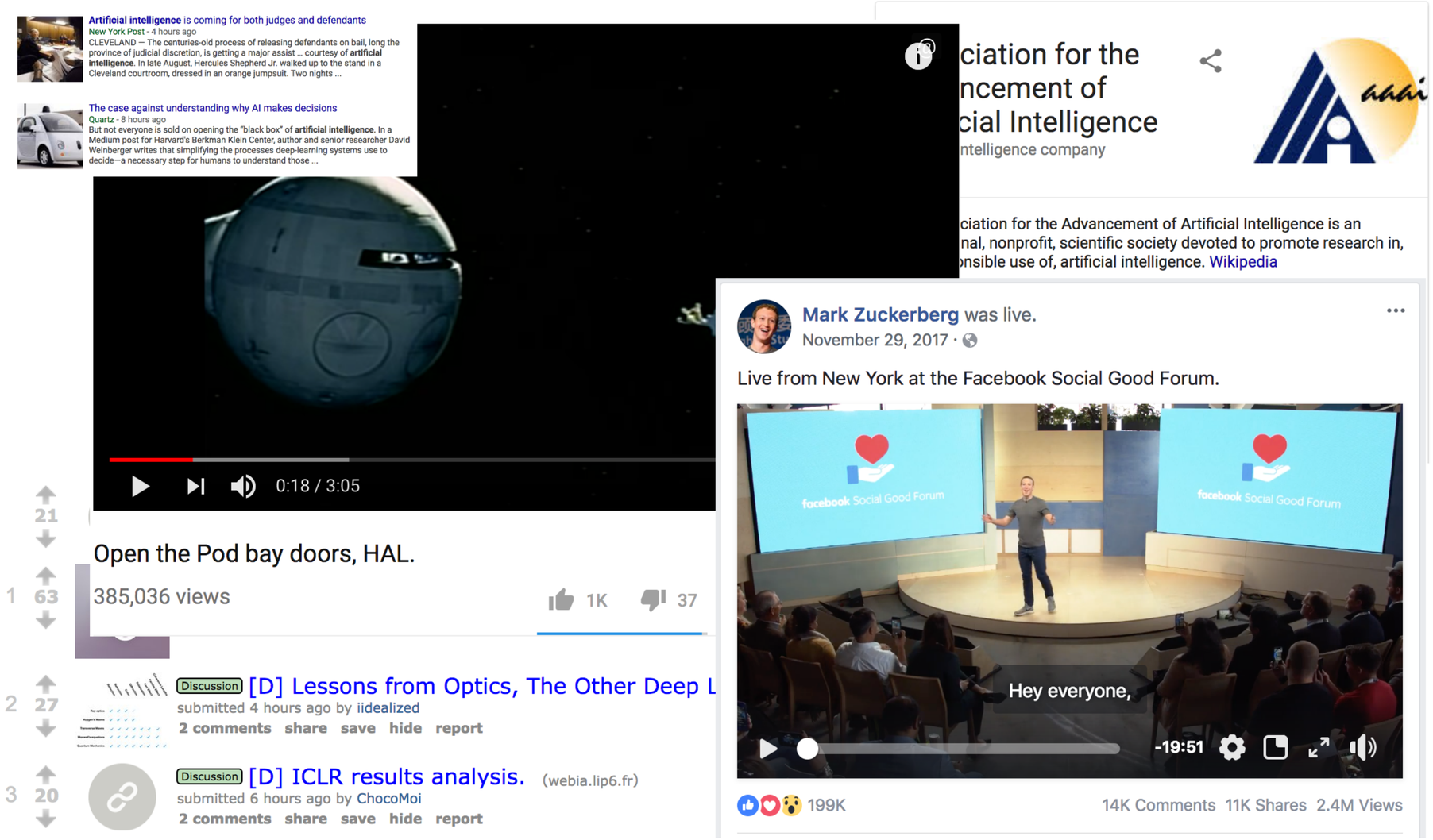
Douwe Kiela
Pronunciation guide: DOW-uh KEE-lah
Facebook AI Research (FAIR)
New York

Massive Progress in Machine Learning

… but it has mostly been unimodal: speech, text, vision.

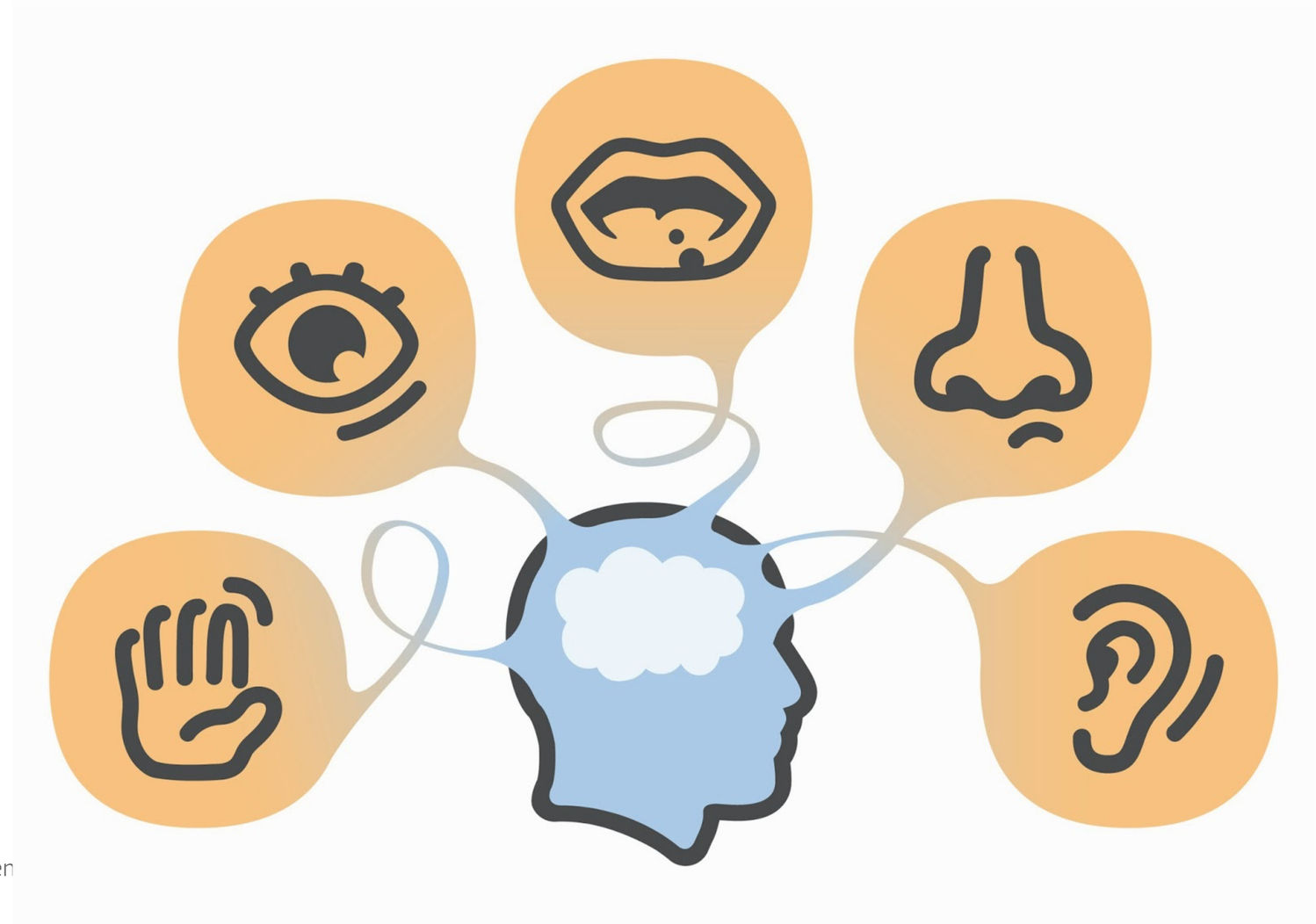


The Internet is Multi-Modal



More Importantly: The *World* is Multi-Modal

… think of animal intelligence, human intelligence.



Outline

What am I going to be talking about?

- **Background:** The meaning of Multi-modal
- **Core Concepts:** Representation Learning, Attention, Fusion
- **Applications:** Image/video captioning, VQA, Representation Learning, Visual Dialogue, Robotics
- **In practice** How to use this in practice
- **The Future:** Open Problems

Outline

- **Background:** The meaning of Multi-modal
- **Core Concepts:** Representation Learning, Attention, Fusion
- **Applications:** Image/video captioning, VQA, Representation Learning, Visual Dialogue, Robotics
- **In practice** How to use this in practice
- **The Future:** Open Problems

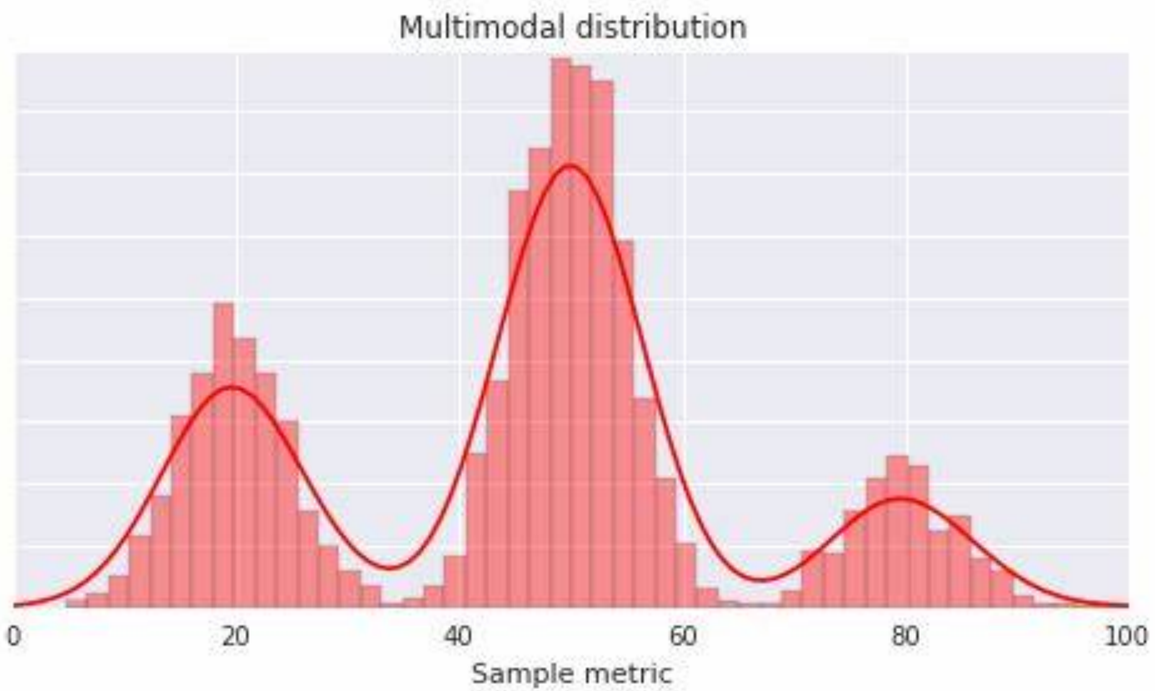
The Meaning of Multi-Modal

But what does it mean

- **Modality:**
  a particular mode in which something exists or is experienced or expressed.

- **Mode:** point where probability mass function takes its maximum value

- Also related to **Multimedia:**
  content that uses information of different content forms
  such as text, audio, images, animations, video and interactions.



Multi-Modality in the Brain

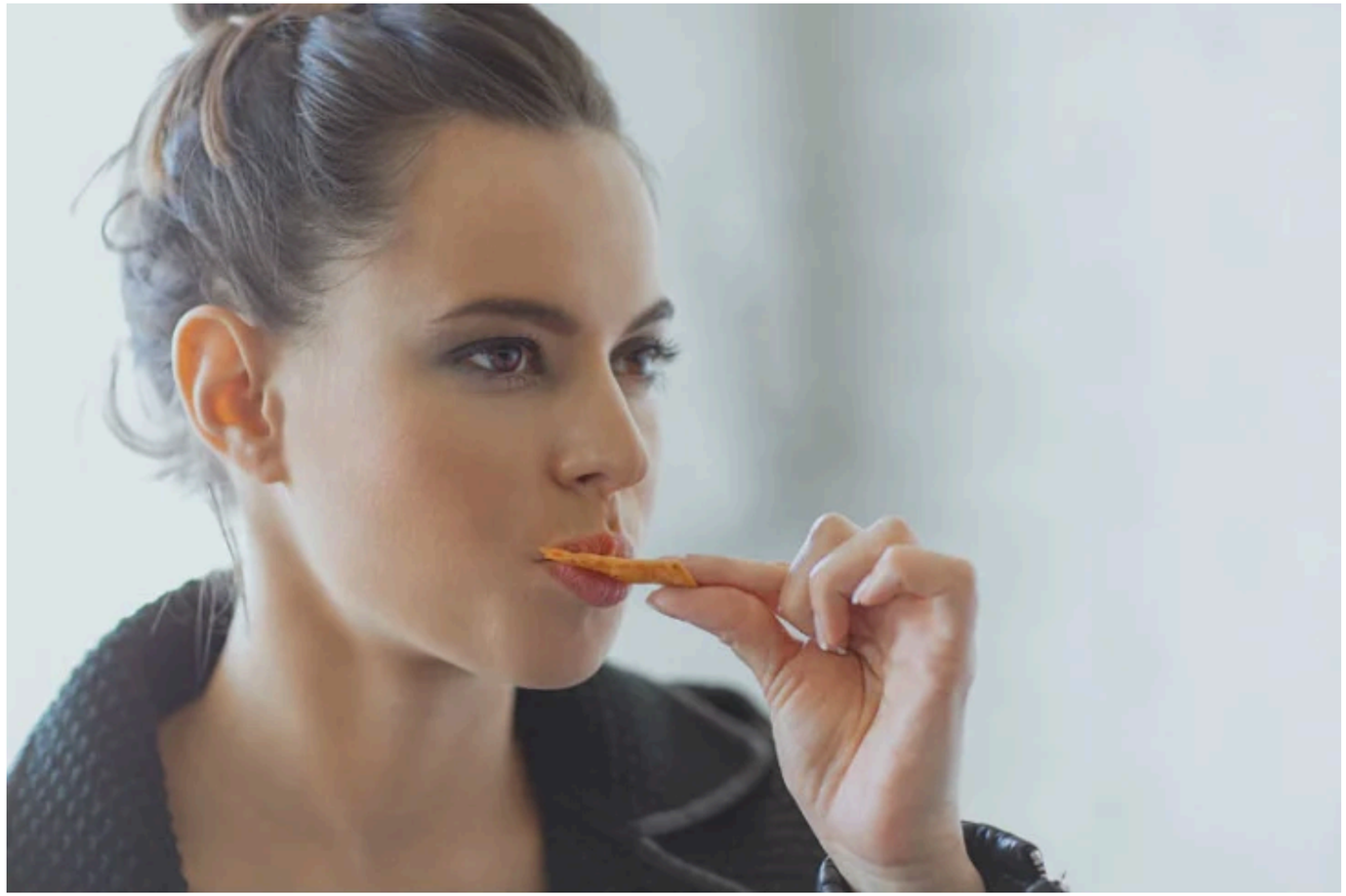
McGurk Effect

- **Our brains** are very **well known to be multi-modal**
- In **the McGurk effect, what you see** determines **what you hear.**
- **Example: https://www.youtube.com/watch?v=G-lN8vWm3m0 (@1:35)**

Multi-Modality in the Brain

Taste and Sound

- **The Role of Auditory Cues in Modulating the Perceived Crispness and Staleness of Potato Chips (Zampini & Spence)**



Multi-Modality in the Brain

Synesthesia

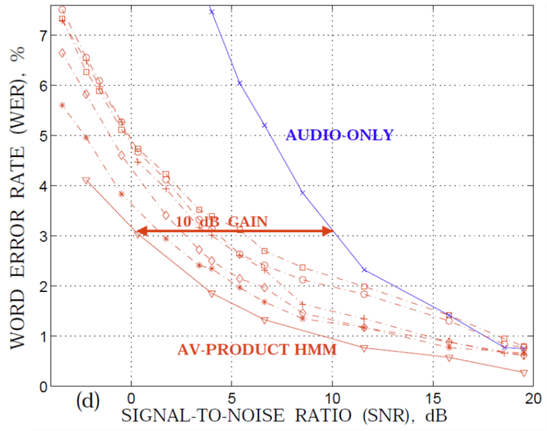
- **Some people can:**
  hear colors
  see music
  smell numbers
  taste names

- **https://www.youtube.com/watch?v=obrBAysVef0 (@1:13)**

- **Stimulation of one sensory or cognitive pathway leads to automatic, involuntary experiences in a second sensory or cognitive pathway.**

Multi-Modality in ML

Long history. Here, focus on deep learning

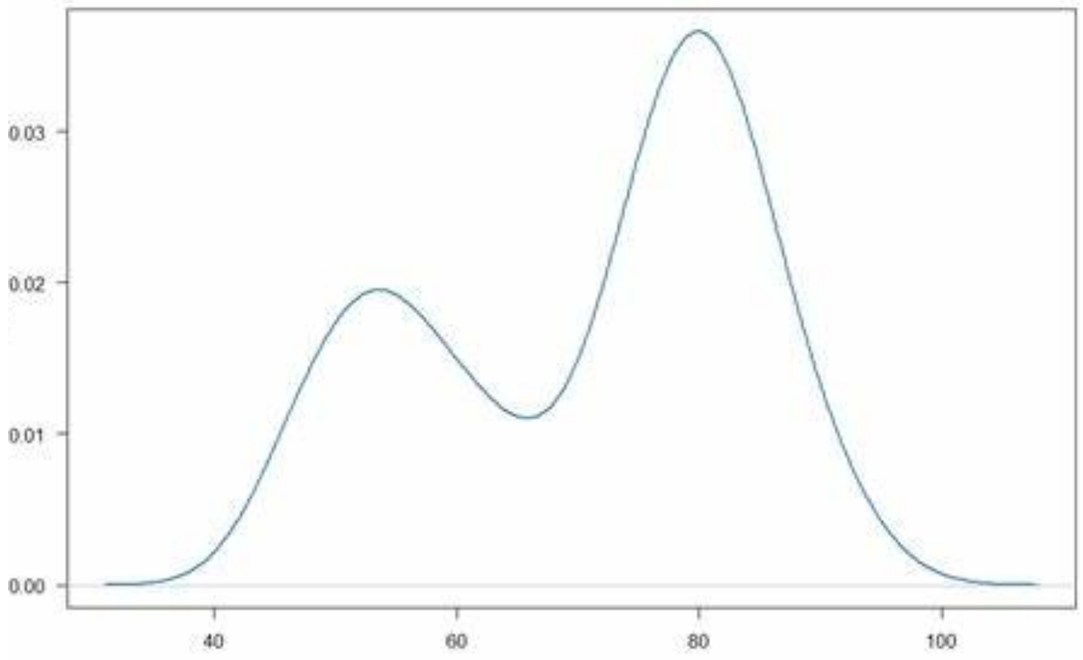


- **Multi-Modality has a long history in ML**
- **Human Computer Interaction and Affective Computing <- became its own research community**
- **Papers:**
  - Audio-Visual Speech Recognition: Automatic lipreading to enhance speech recognition (Petajan, 1986)
  - Glove-Talk: a neural network interface between a data-glove and a speech synthesizer (Fels & Hinton, 1993)
  - An Asynchronous Hidden Markov Model for Audio-Visual Speech Recognition (Bengio, 2003)
  - Multimodal Deep Learning (Ngiam et al, 2011)
  - Multimodal Learning with Deep Boltzmann Machines (Srivastava and Salakhutdinov, 2014)

Why isn't every system multi-modal?

Not always easy

- One modality can **dominate** other modalities
- Additional modalities can **induce noise**
- **Availability** of information from other modalities is not always guaranteed.
- Preference to solve one thing at a time (understandable).

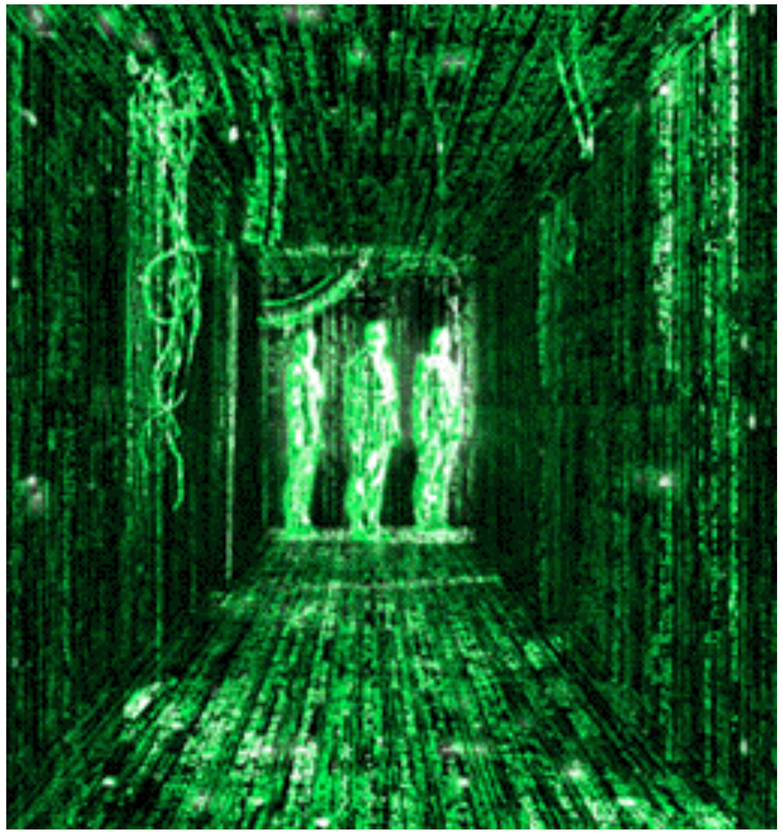


Why this is exciting

Lots of interesting problems; relatively unexplored outside of some mainstream topics

- **To summarize:**
  - The internet is multi-modal
  - The world is multi-modal
  - Our brain is multi-modal

- **The MVP (most valuable problem) of AI (Hassabis):**
  how is it that perceptual input leads to conceptual representations that can be reasoned with?

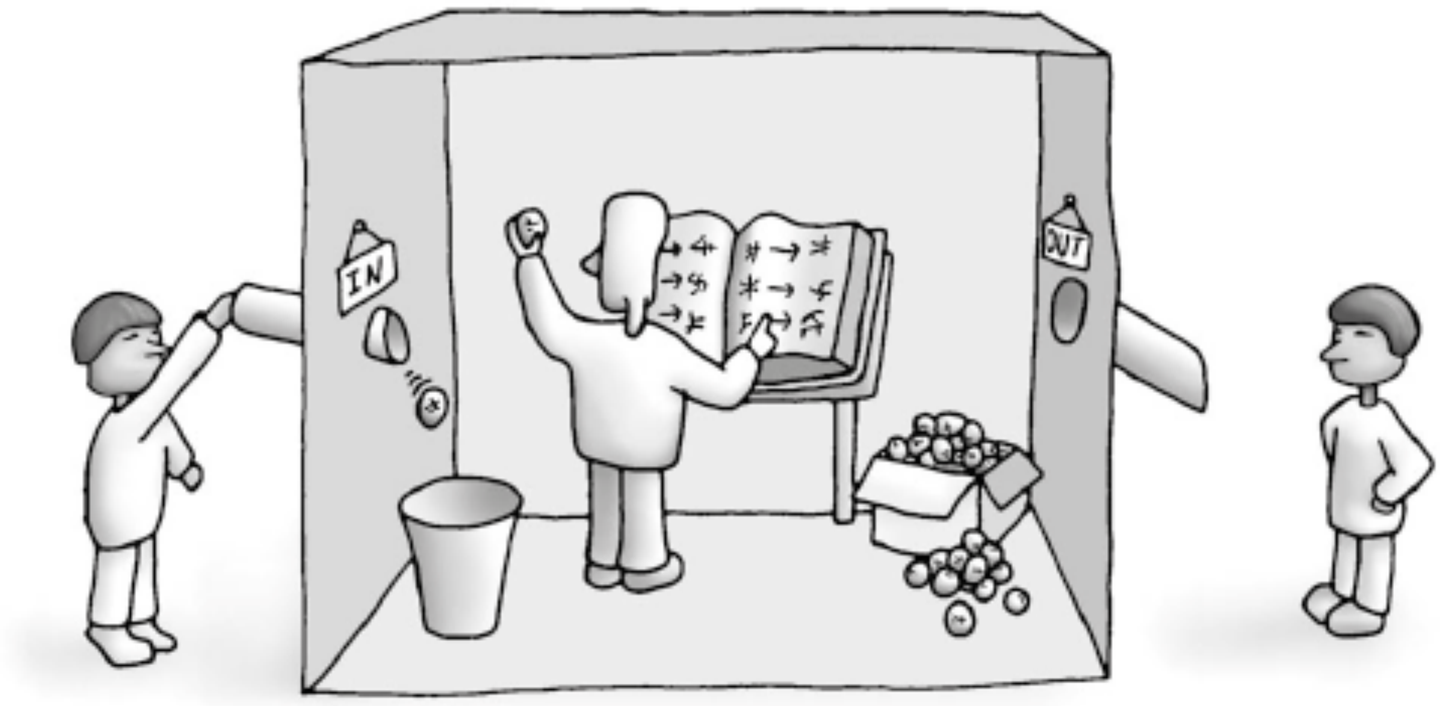


The holy grail of AI: meaning

Meaning is the holy grail not only of linguistics, but also of philosophy, psychology, and neuroscience -- Ray Jackendoff, 2002

The grounding problem - the origin of meaning

How can you know the meaning of symbols if they are defined only in terms of other symbols



The grounding problem - the origin of meaning

How can you know the meaning of symbols if they are defined only in terms of other symbols

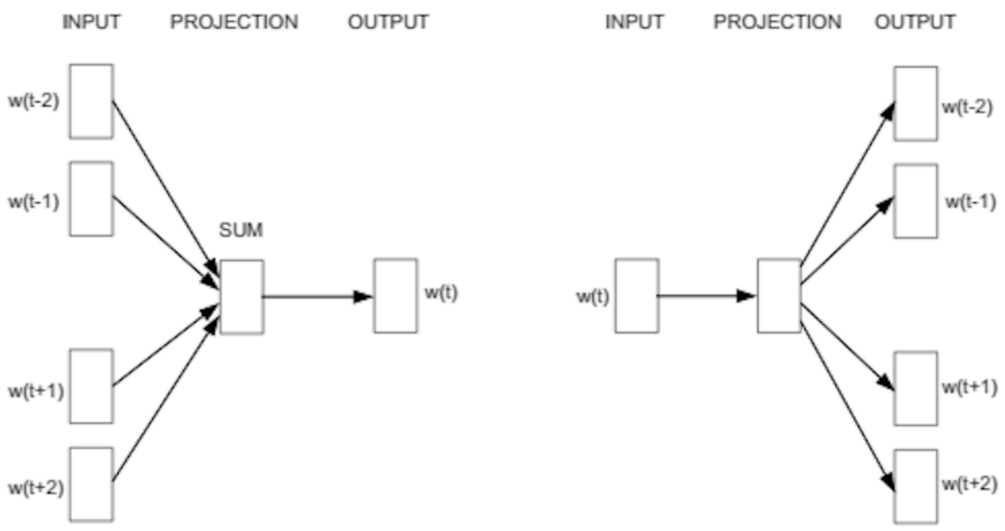
Outline

- **Background:** The meaning of Multi-modal
- **Core Concepts:** Representation Learning, Attention, Fusion
- **Applications:** Image/video captioning, VQA, Representation Learning, Visual Dialogue, Robotics
- **In practice** How to use this in practice
- **The Future:** Open Problems

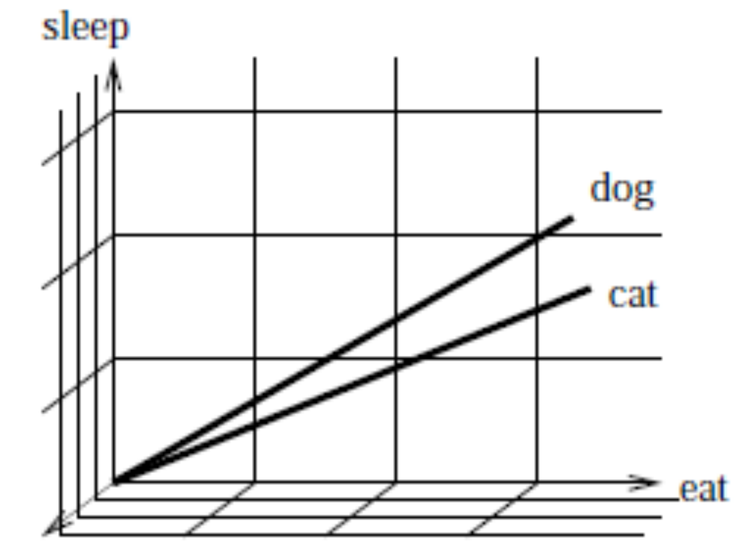
Embeddings

Unimodal

- In mathematics, an **embedding** (or imbedding) is one instance of some mathematical structure contained within another instance.
- That's not very helpful. Better: a projection from one space to another
- **Vector space models** and their **low-rank approximations.**
- **Function mapping from one-hot to real-valued.**



CBOW Skip-gram

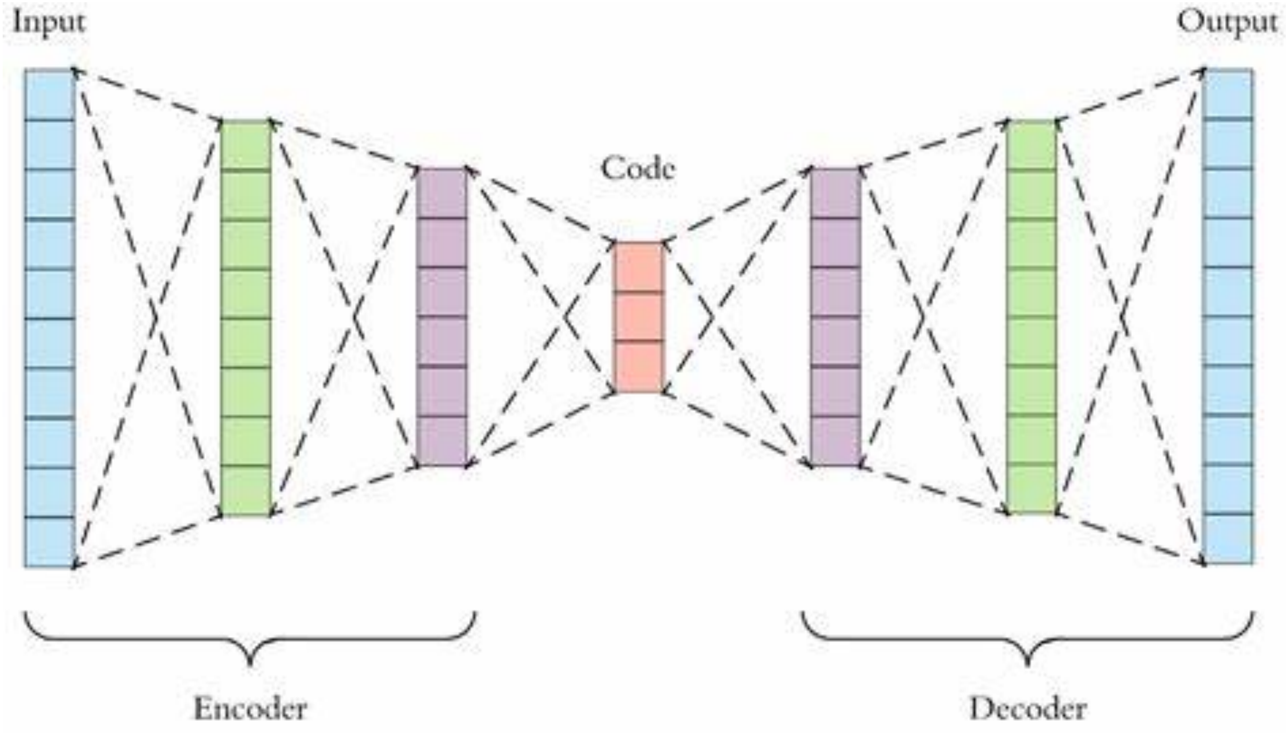


(Clark '14)

Encoders and Decoders

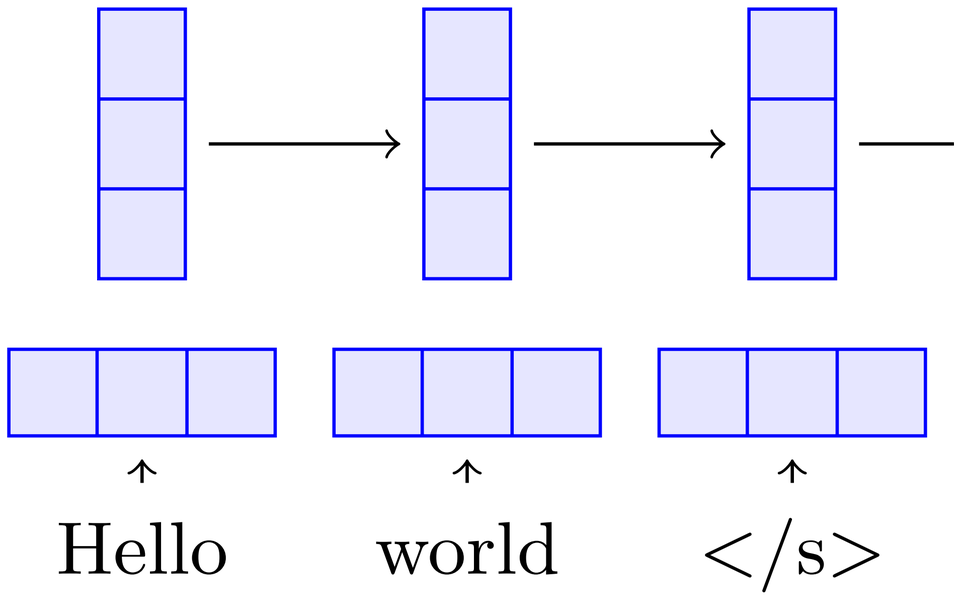
Unimodal

- Embeddings are a way to **encode** meaning of words.
- What about sentences?
- If we can also **decode**, we have an **autoencoder.**



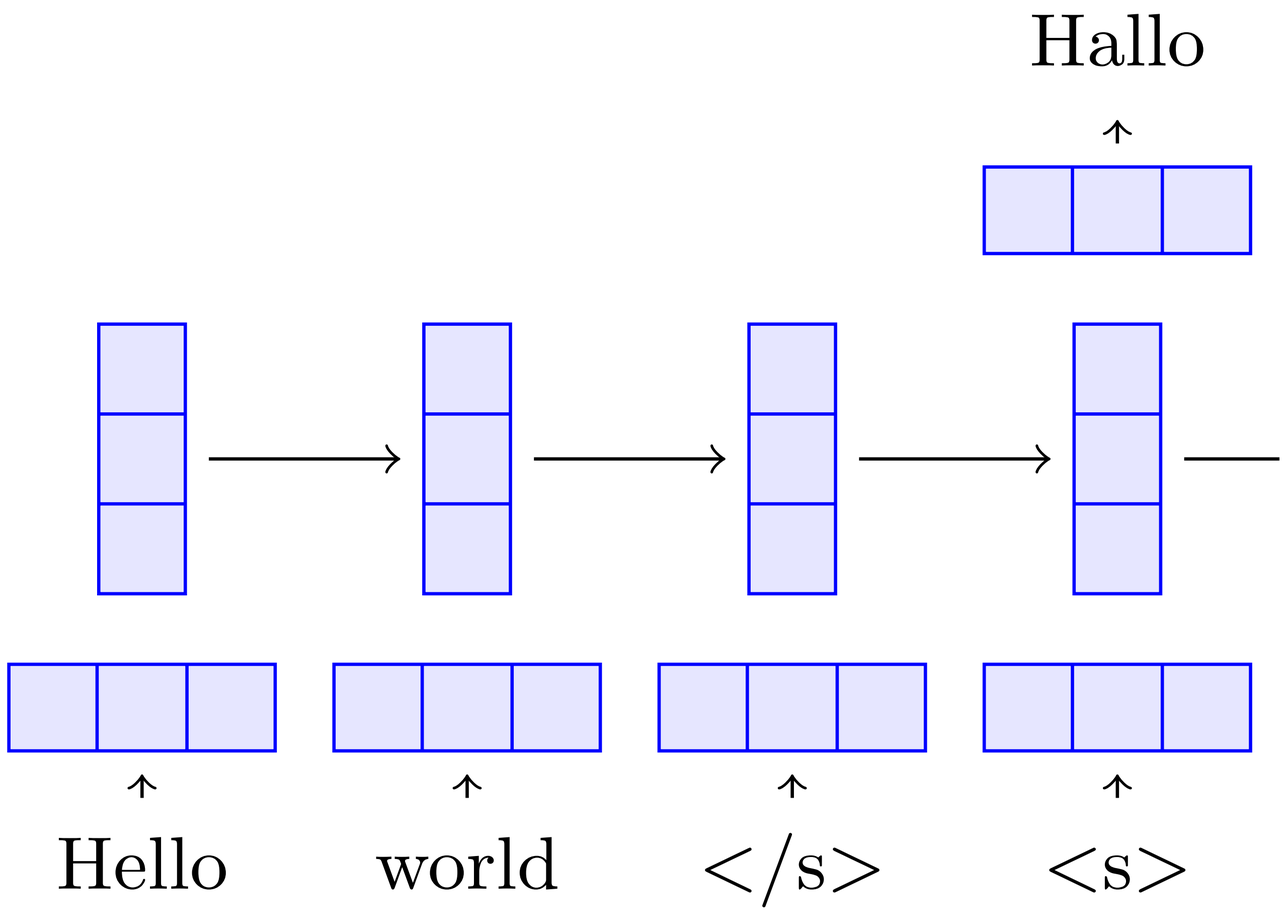
Encoder-Decoder (Seq2Seq)

Neural machine translation



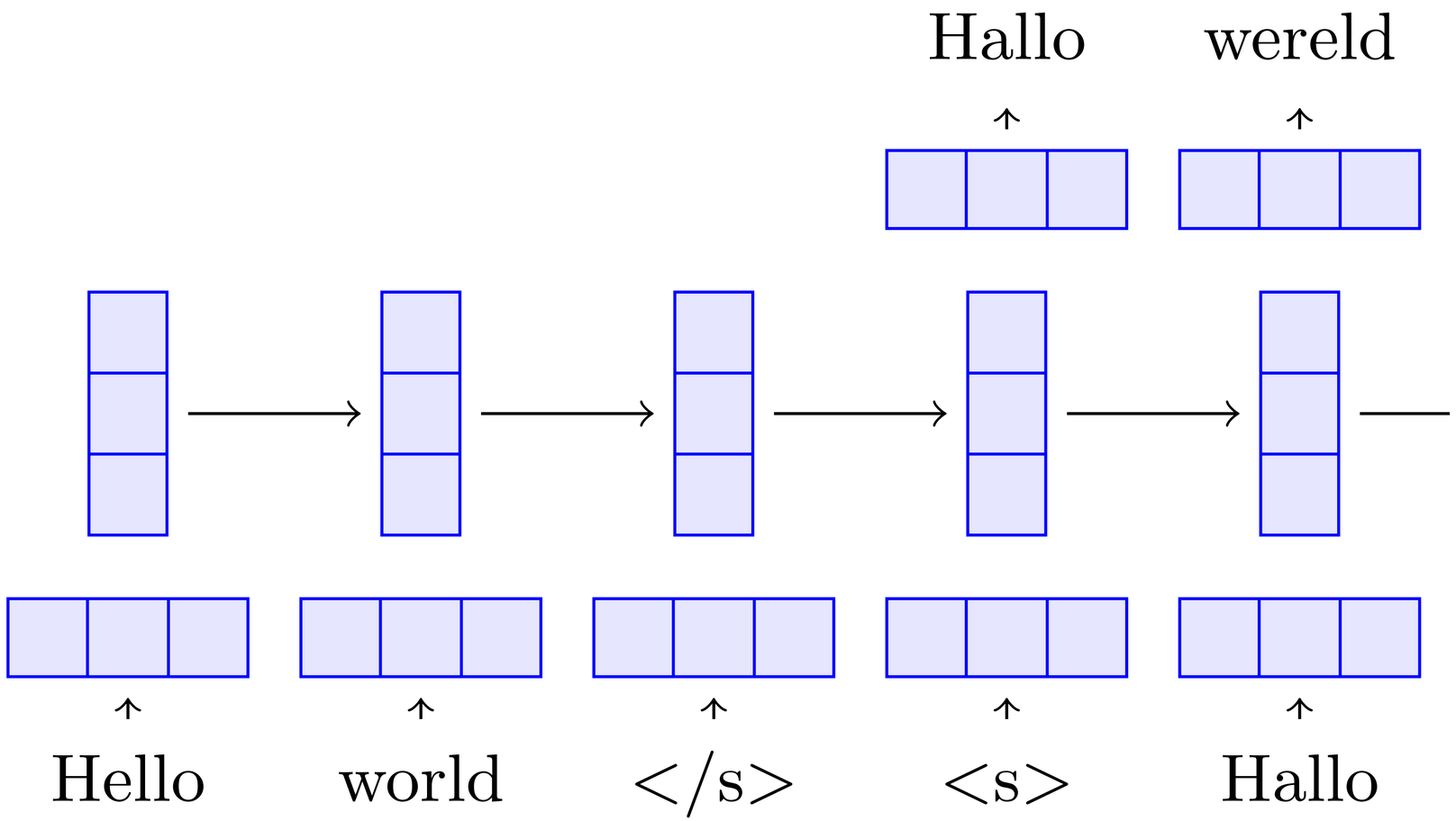
Encoder-Decoder (Seq2Seq)

Neural machine translation



Encoder-Decoder (Seq2Seq)

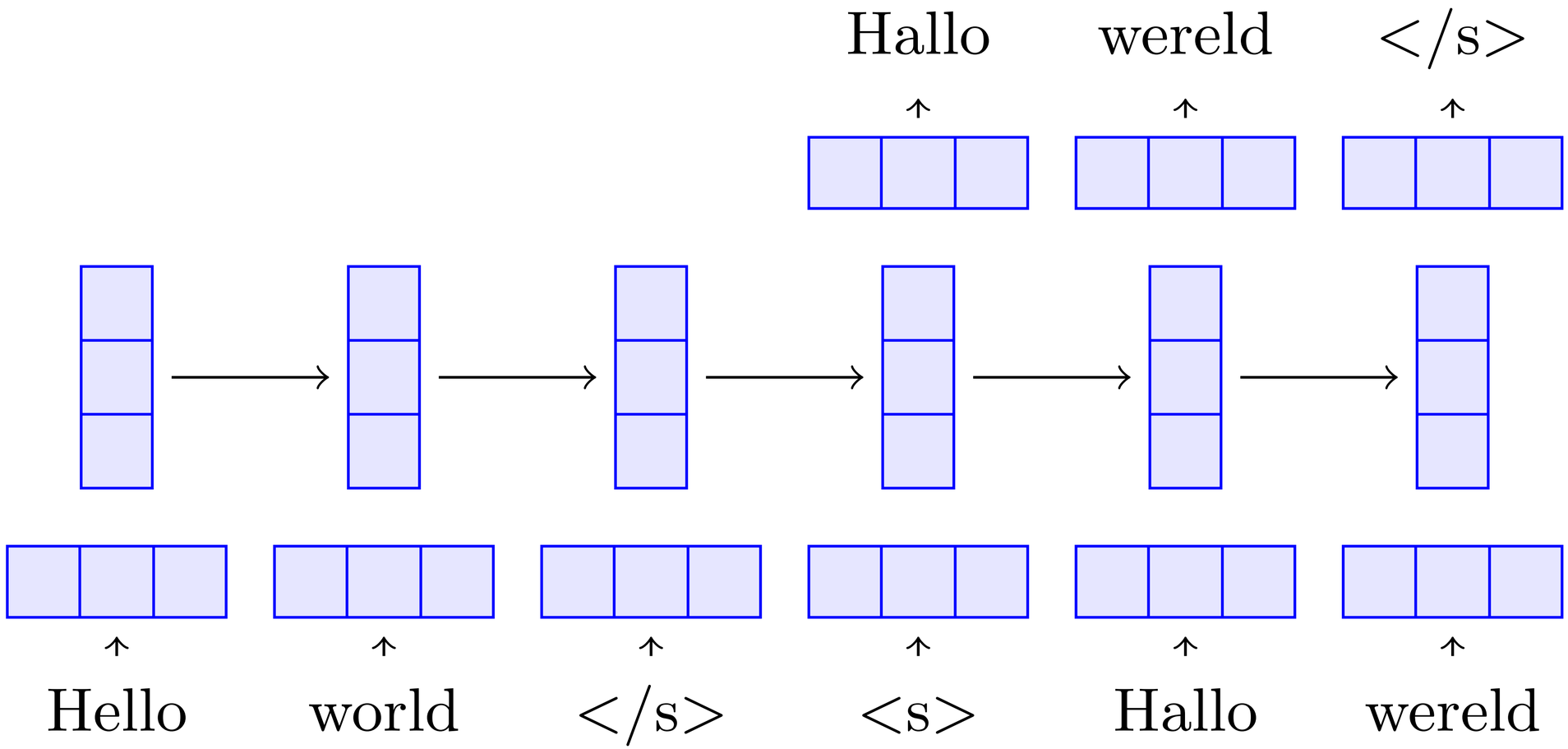
Neural machine translation



Encoder-Decoder (Seq2Seq)

$$\texttt{argmax}_\theta \sum_n^N \log p(y^n|x;\theta)$$

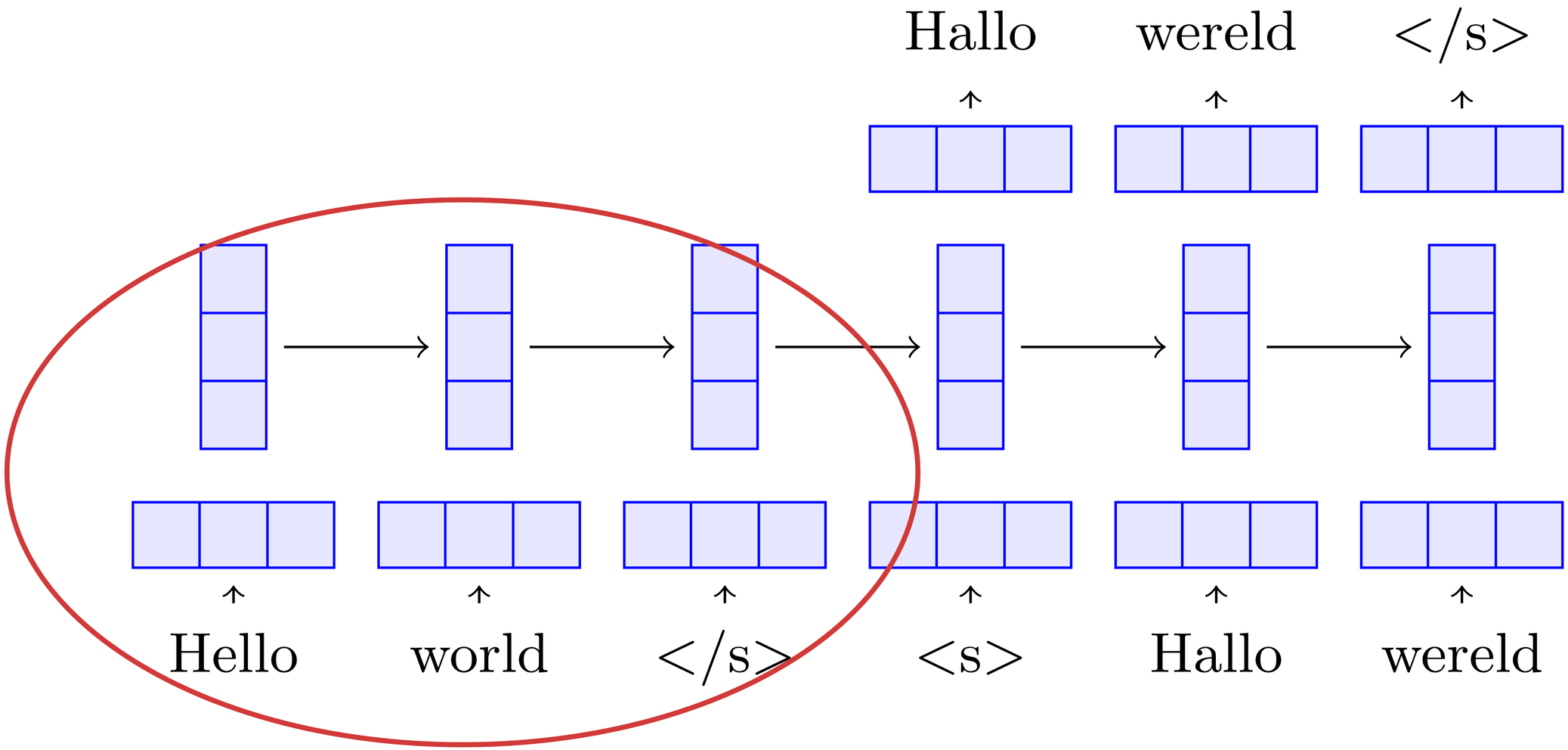
Neural machine translation



Encoder-Decoder (Seq2Seq)

$$\texttt{argmax}_\theta \sum_n^N \log p(y^n|x;\theta)$$

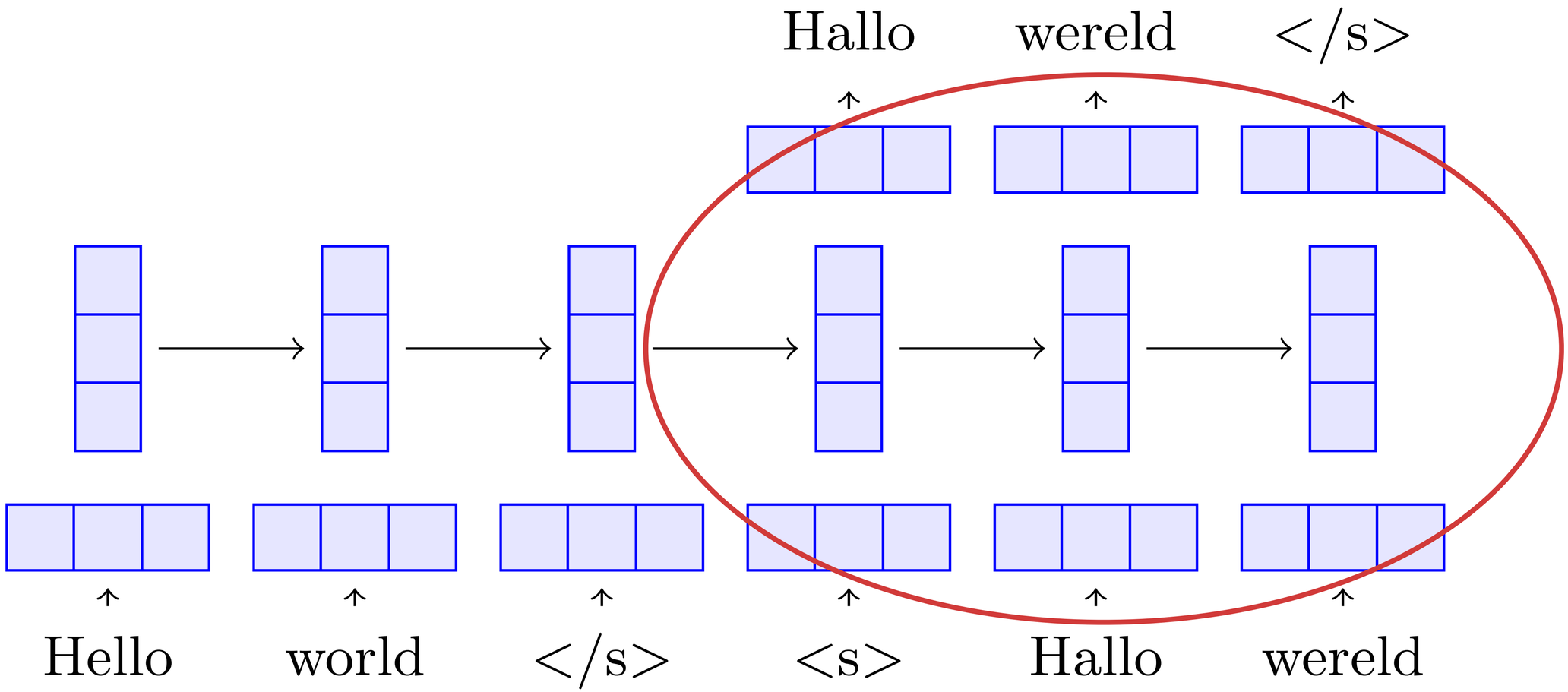
Encoder



Encoder-Decoder (Seq2Seq)

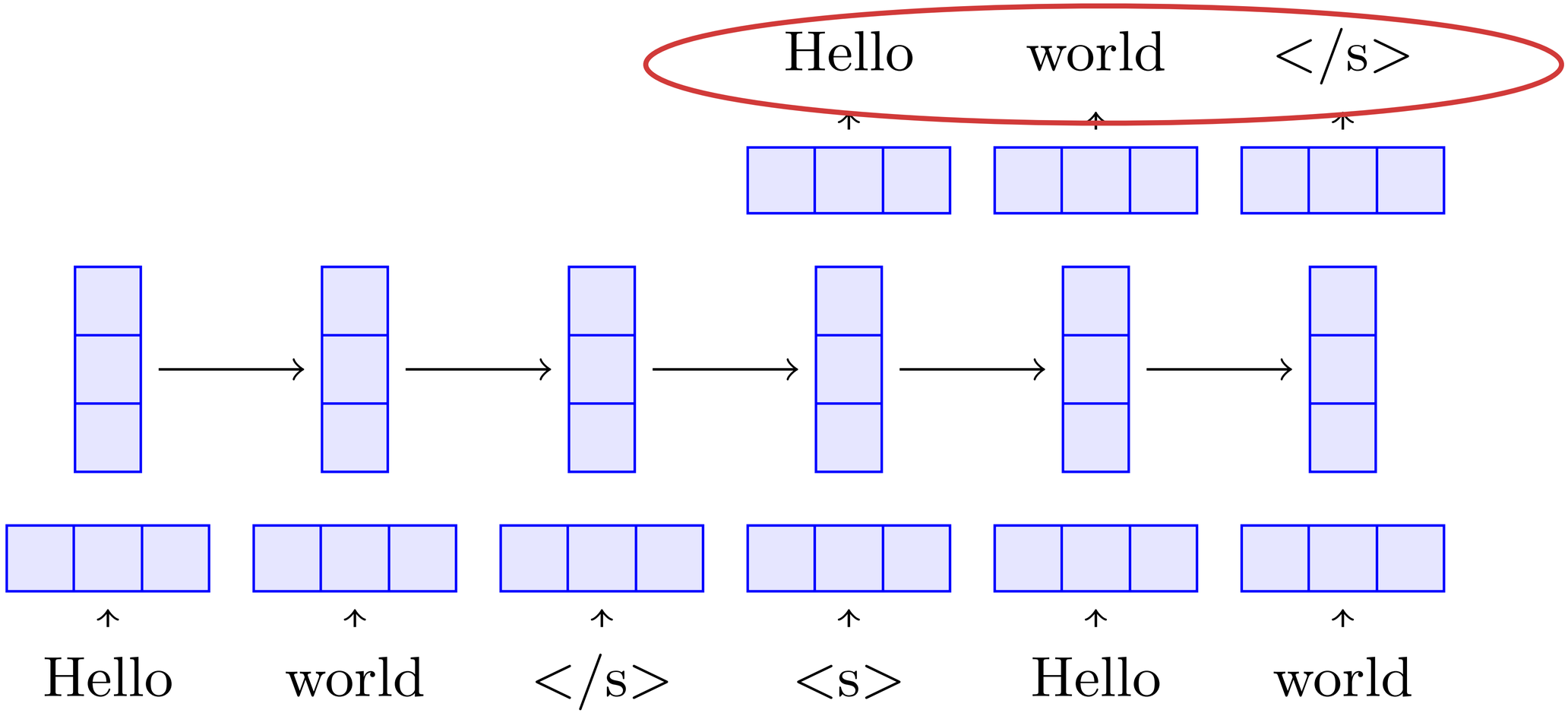
$$\texttt{argmax}_\theta \sum_n^N \log p(y^n|x;\theta)$$

Decoder



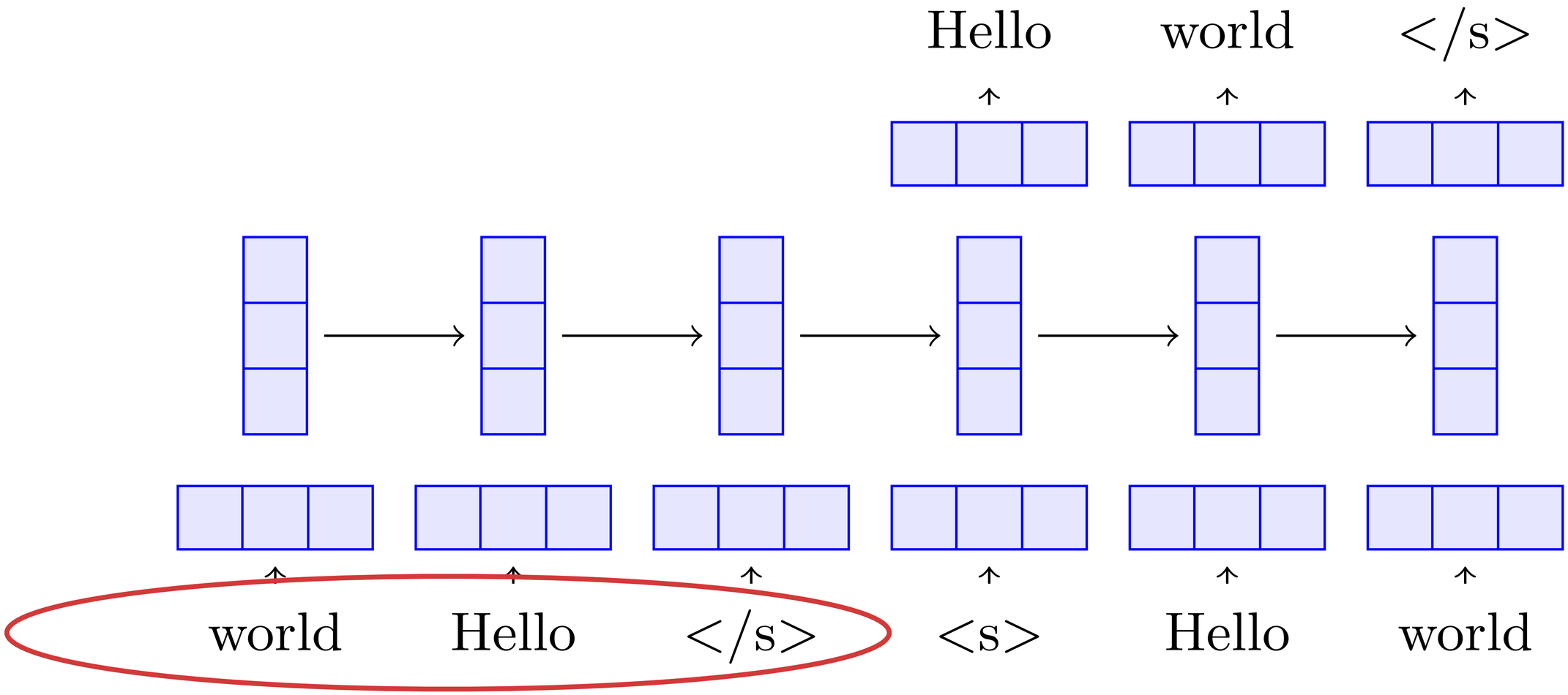
Sentence Auto-Encoder

Sentence auto encoder



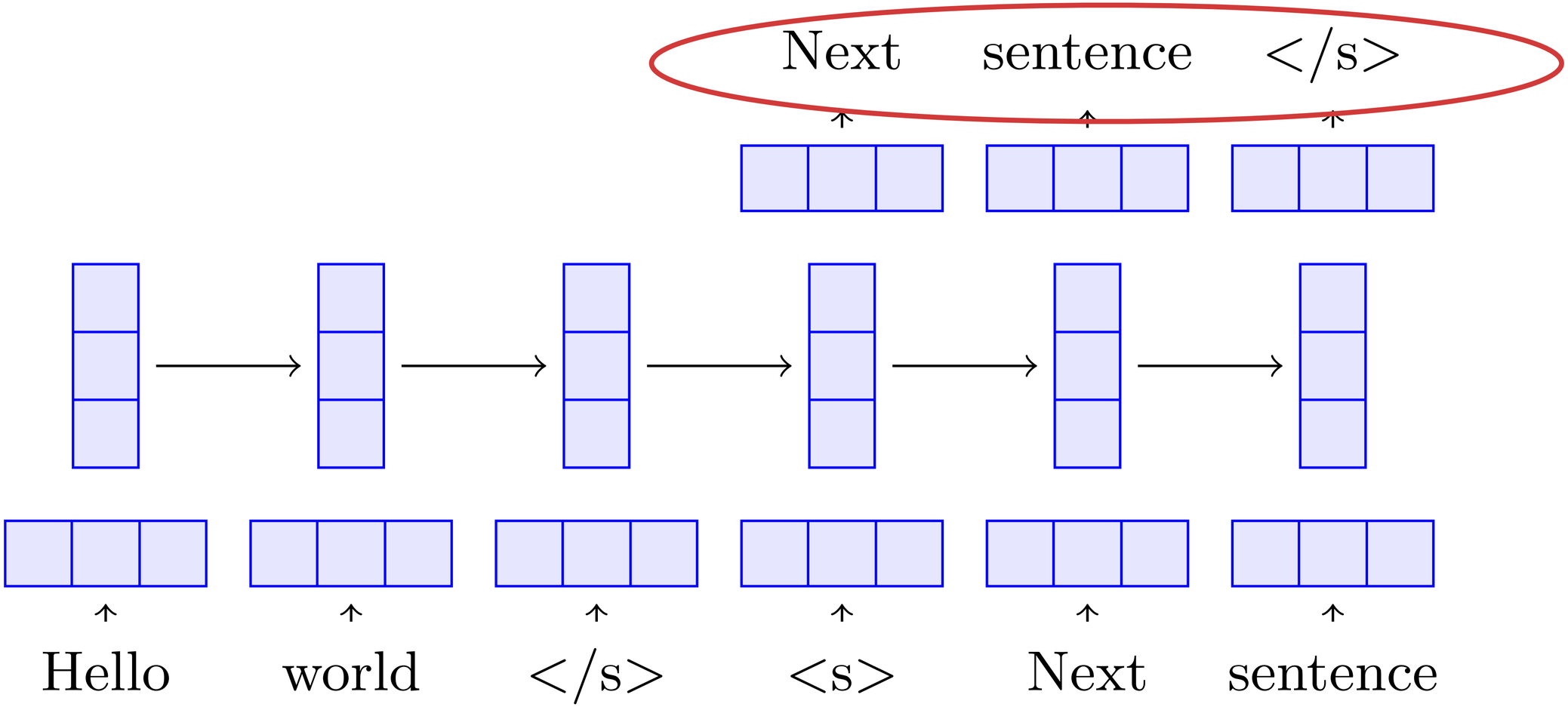
De-Noising Auto-Encoder

Sequential de-noising auto encoder (Hill et al) - Similar to BERT



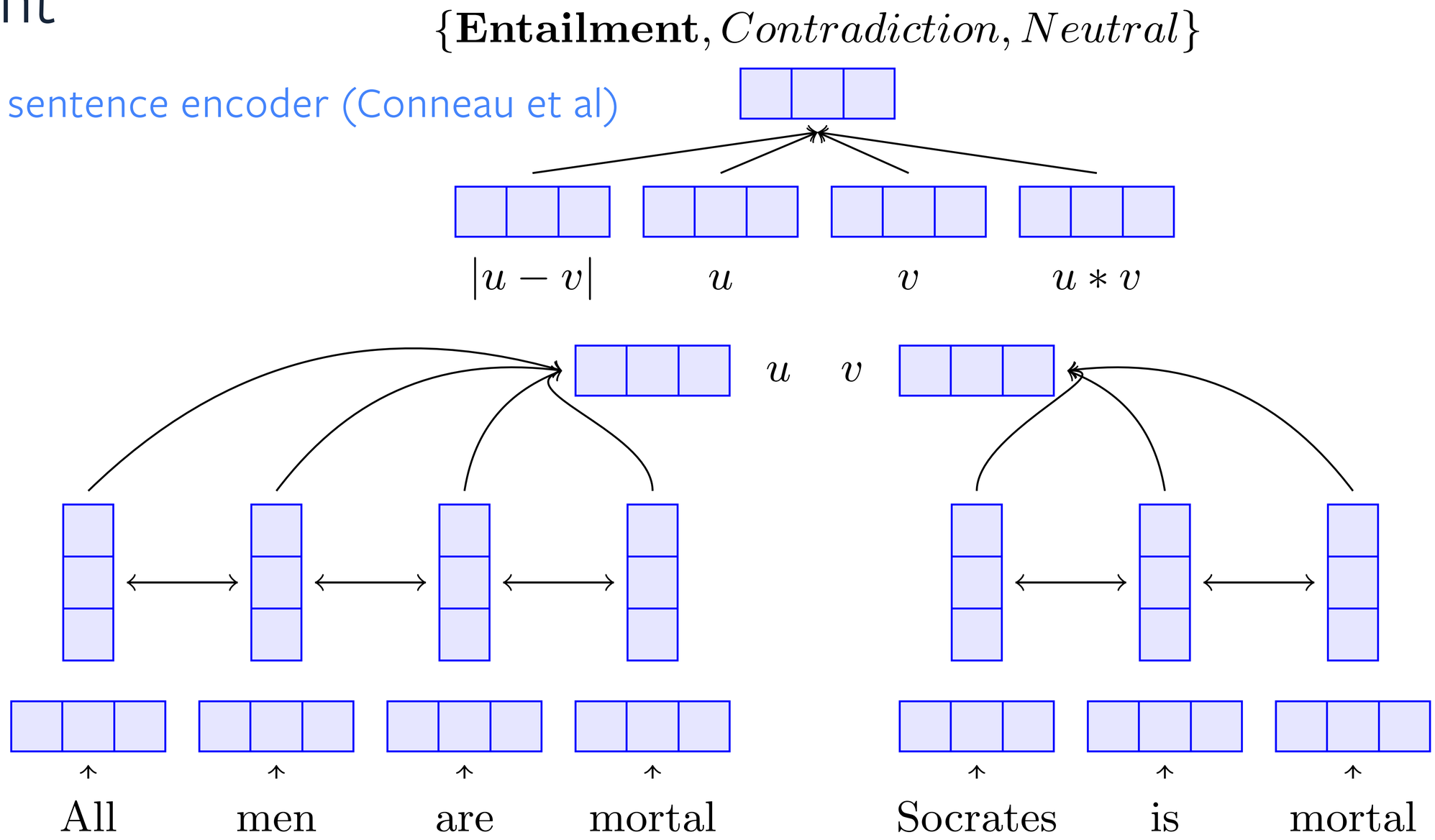
SkipThought

Unsupervised sentence encoder (Kiros et al)



InferSent

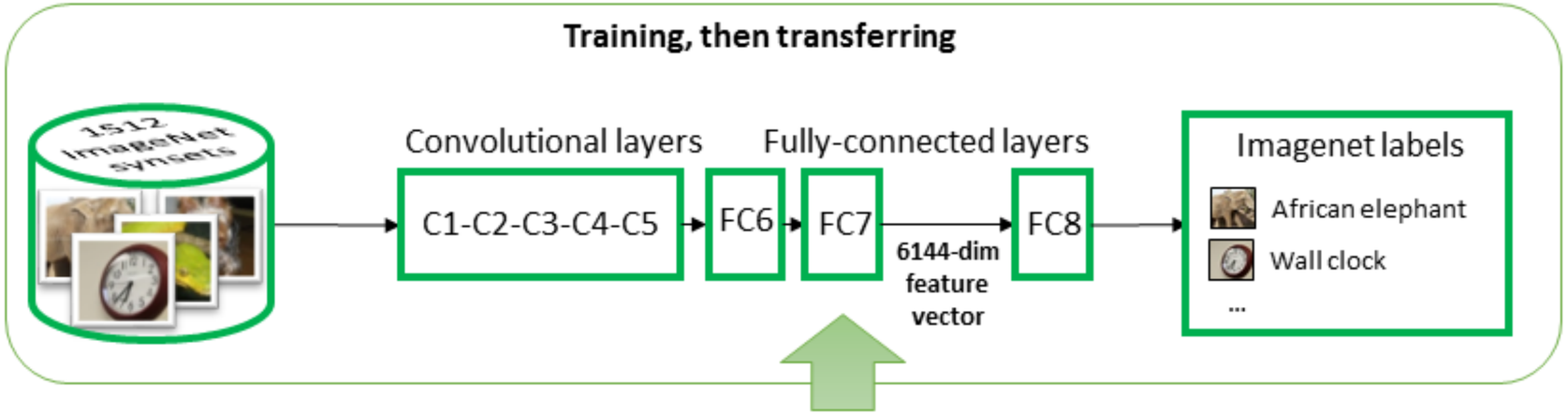
Supervised sentence encoder (Conneau et al)



Pre-Training and Transfer Learning

Image representations

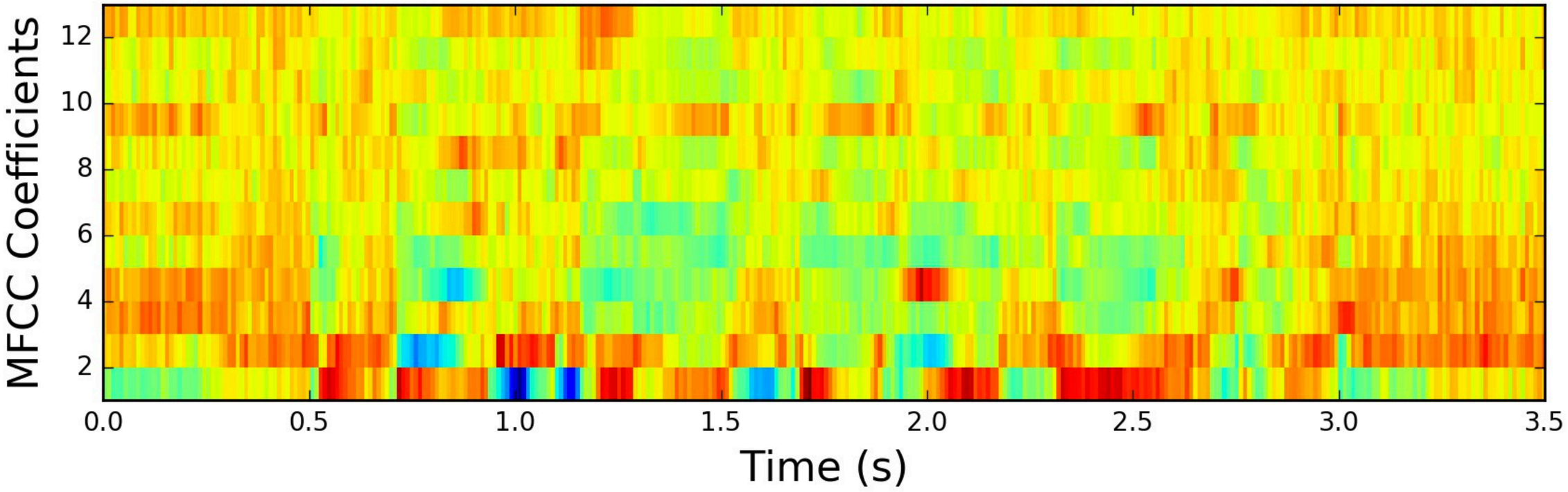
- **Off-the-shelf CNN features** as supervised **image representations** (Razavian et al.; Oquab et al.)
- **Pre-train** on **ImageNet**; then **transfer** to other task**.**
- **AlexNet** (Krizhevsky et al), **VGGNet** (Simonyan et al), **GoogLeNet** (Szegedy et al), **ResNet** (He et al)



Encoding audio

MFCC features, sequence modeling

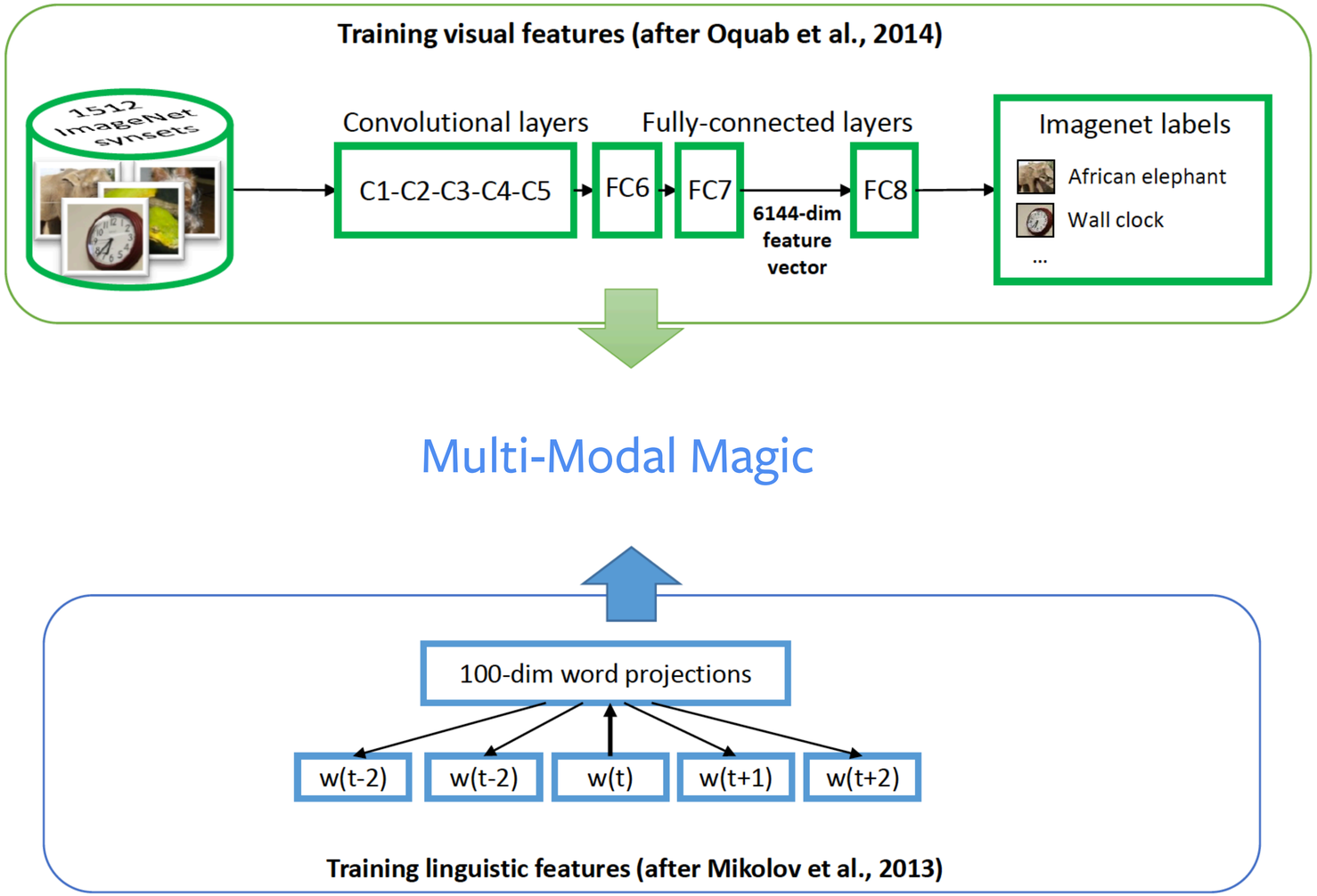
- **Audio and speech are similar to text, can even use convnet on spectrograms.**



Example: Encoding images and text

Word embeddings on one hand, pre-trained AlexNet on the other

- **How do we combine information?**
  - Cross-modal tasks
  - Multi-modal tasks



Multi-Modal Fusion

How do we combine information?

- **Similarity**
  - Inner product: $\mathbf{u}\mathbf{v}$
- **Pooling**
  - Concatenation: $W[\mathbf{u}, \mathbf{v}]$
  - Sum pooling: $W\mathbf{u} + V\mathbf{v}$
  - Max-pooling: $\max(W\mathbf{u}, V\mathbf{v})$
  - Attention: $\alpha W\mathbf{u} + \beta V\mathbf{v}$

Multi-Modal Fusion

How do we combine information?

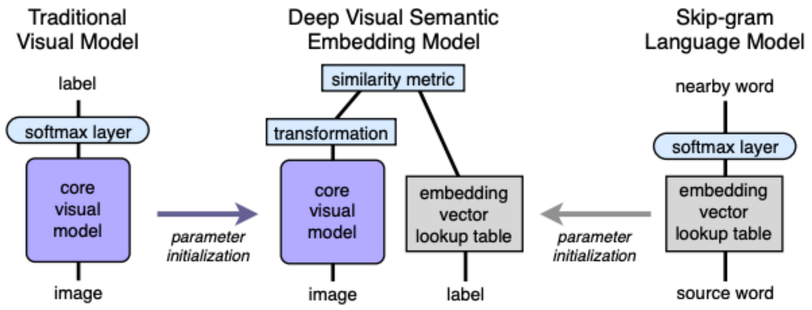
- **Multiplicative**
  - Multiplication: $W\mathbf{u} \odot V\mathbf{v}$
  - Gating: $W\mathbf{u} \odot V\mathbf{v}$
  - LSTM-style: $\tanh(W\mathbf{u}) \odot V\mathbf{v}$

- **Bilinear**
  - Bilinear: $\mathbf{u}W\mathbf{v}$
  - Bilinear gating: $\mathbf{u}W\sigma(\mathbf{v})$
  - Low-rank bilinear: $\mathbf{u}U^TV\mathbf{v} = P(U\mathbf{u} \odot V\mathbf{v})$
  - Compact bilinear: $FFT^{-1}(FFT(\Psi(x, h_1, s_1)) \odot FFT(\Psi(x, h_2, s_2)))$

Joint embedding spaces: cross-modal retrieval

Project unimodal embeddings to joint space

- **WSABI (Web Scale Annotation by Image Embedding; Weston et al)**
  - WARP (Weighted Approximate-Rank Pairwise) Loss
- **DeViSE (Deep Visual-Semantic Embedding; Frome et al)**
  - Hinge rank Loss
- **VSE (Visual-Semantic Embeddings; Kiros et al)**
  - Triplet hinge rank loss
  - LSTM encoder for sentences
- **VSE++ (Faghri et al)**



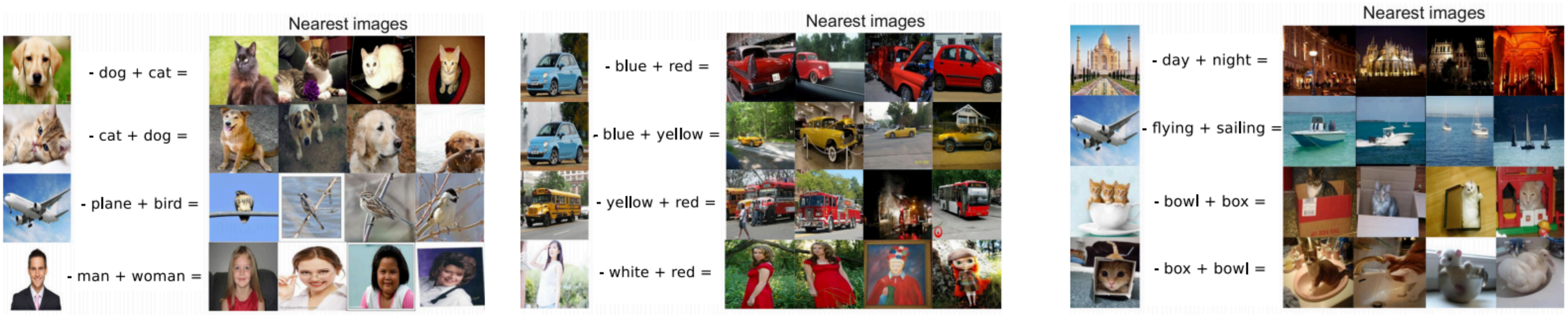
$$loss(image, label) = \sum_{j \neq label} \max[0, margin - \vec{t}_{label} M \vec{v}(image) + \vec{t}_j M \vec{v}(image)]$$

$$\min_{\boldsymbol{\theta}} \sum_{\mathbf{x}} \sum_{k} \max\{0, \alpha - s(\mathbf{x}, \mathbf{v}) + s(\mathbf{x}, \mathbf{v}_k)\} + \sum_{\mathbf{v}} \sum_{k} \max\{0, \alpha - s(\mathbf{v}, \mathbf{x}) + s(\mathbf{v}, \mathbf{x}_k)\}$$

$$\ell_{MH}(i,c) = \max_{c'} \left[\alpha + s(i,c') - s(i,c)\right]_+ + \max_{i'} \left[\alpha + s(i',c) - s(i,c)\right]_+$$

Joint embedding spaces

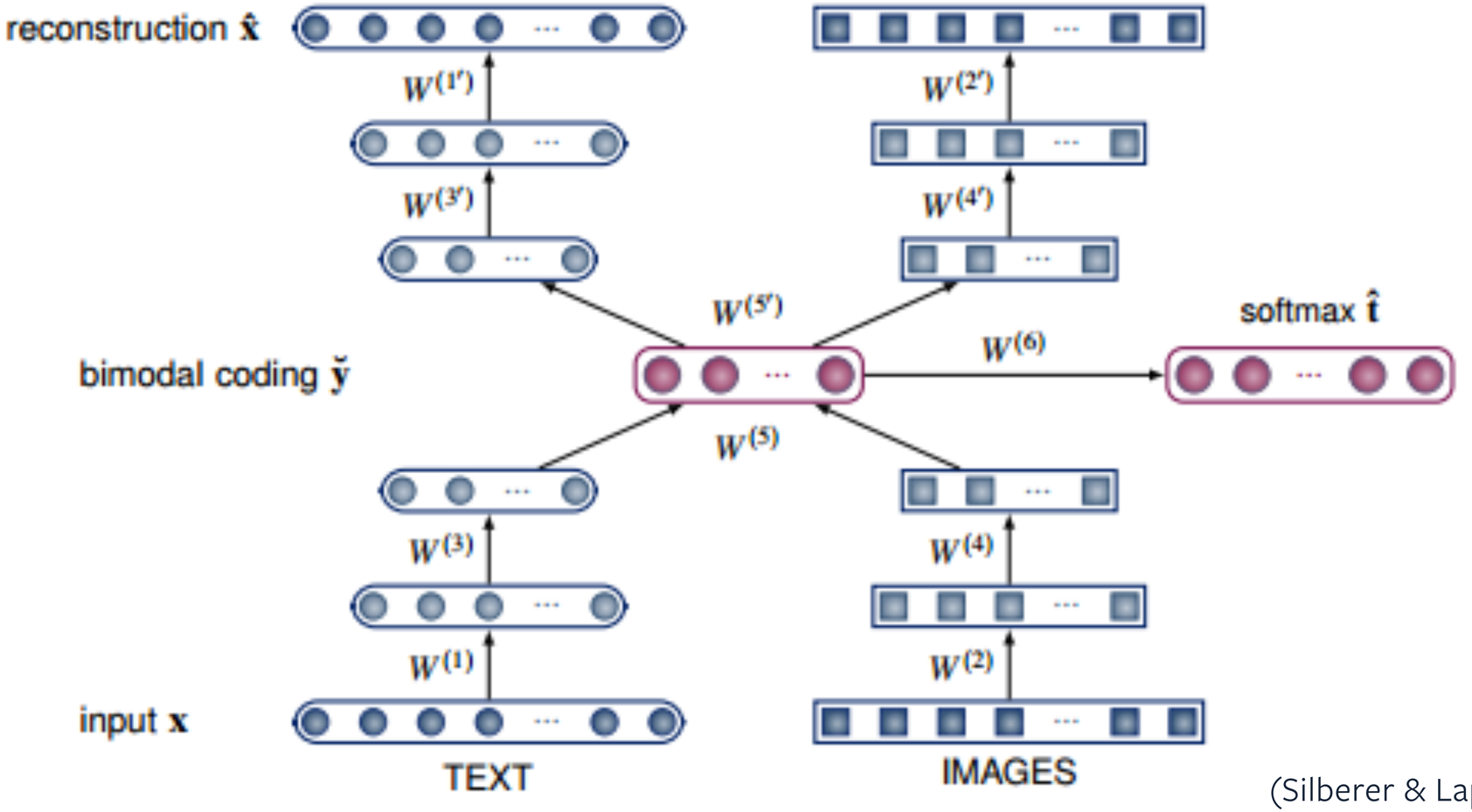
Vector arithmetic (Kiros et al)



(a) Simple cases (b) Colors (c) Image structure

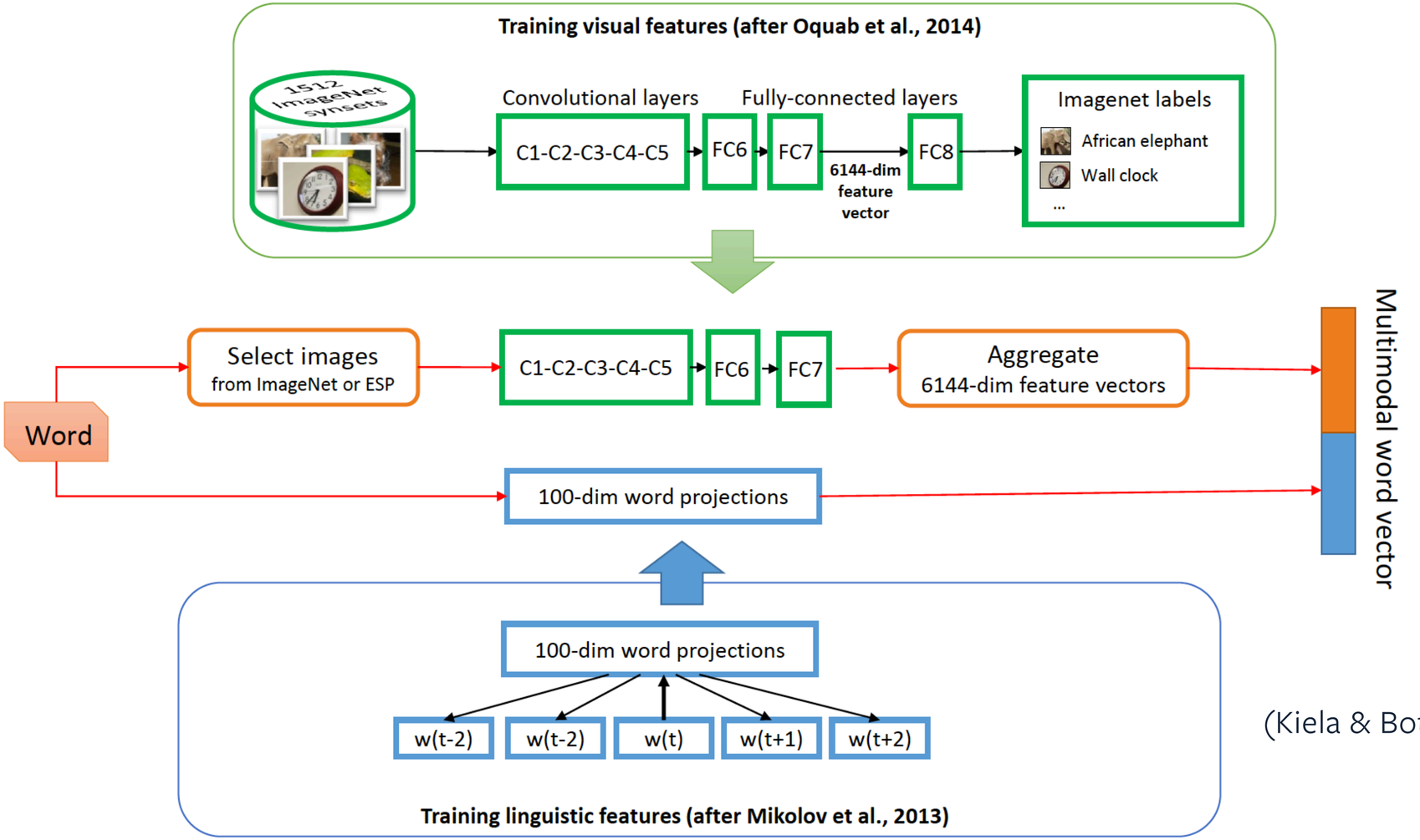
Lexical grounded representation learning

Representation Learning: Multi-Modal Auto-encoder



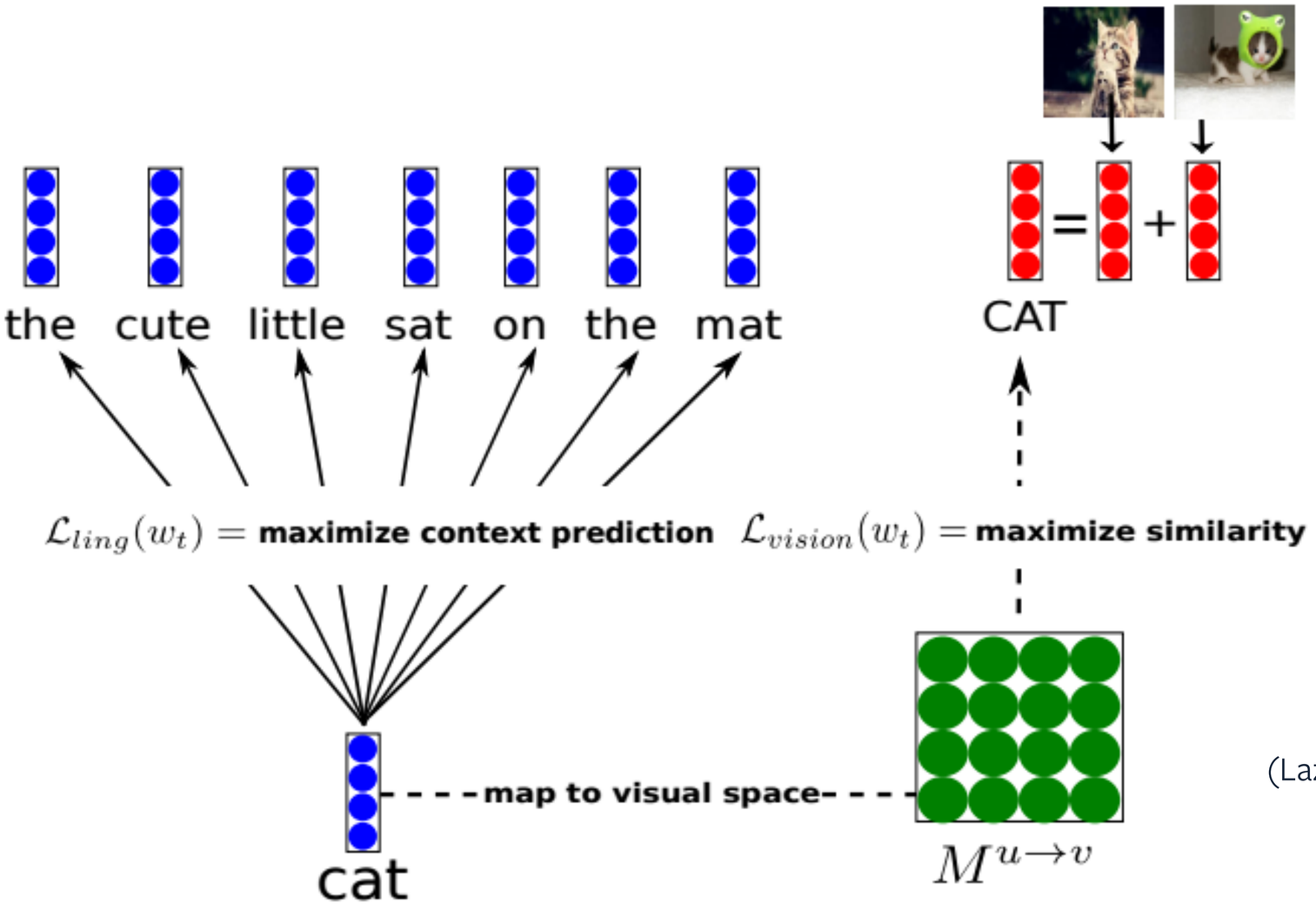
(Silberer & Lapata, 2014)

Representation Learning: Pre-trained fusion



(Kiela & Bottou, 2014)

Representation Learning: Multi-Modal Skip-Gram



(Lazaridou et al, 2015)

Early, Middle and Late Fusion

How do we combine information?

- **Classifier:** ϕ**; e.g.** $\sigma(W\mathbf{x} + b)$
- **Fusion model:** f
- **Early fusion:** $\phi(f(u, v))$; u, v **jointly learned**
  Jointly optimized over multiple modalities
- **Middle fusion:** $\phi(f(u, v))$; u, v **independent**
  Independently optimized modalities that are subsequently combined.
- **Late fusion:** $f(\phi(u), \phi(v))$
  Combined unimodal outputs.

Fusion by Attention

E.g. GRU hidden states with pre-trained AlexNet

- **Base case: Modulating modalities** $[\alpha\mathbf{u}, (1-\alpha)\mathbf{v}]$

- **Attention between modalities**
  - Instead of having one vector $\mathbf{u}$ and one vector $\mathbf{v}$, what if we have k hidden states $\mathbf{u}_k$?

    We can learn attention weights: $(\sum_i^n \alpha_i \mathbf{u}_k)\mathbf{v}$

    where $\alpha = softmax(\mathbf{U}\mathbf{v})$ where every row in $\mathbf{U}$ is a $\mathbf{u}_k$

- **What if both are sequences, e.g. hidden states and various pooling outputs of a convnet, or bounding boxes?**
  - That's where it becomes interesting (examples later).

Manaal will discuss in more detail tomorrow ->

Outline

- **Background:** The meaning of Multi-modal
- **Core Concepts:** Representation Learning, Attention, Fusion
- **Applications:** Image/video captioning, VQA, Representation Learning, Visual Dialogue, Robotics
- **In practice** How to use this in practice
- **The Future:** Open Problems

Image Captioning

Provided an image, give a caption

- **Dataset availability was very important: MS COCO (Lin et al., 2014), Flickr30k (Young et al, 2014).**
- Two ways to set this up: **retrieval** (typically rank loss) or **generative** (typically NLL of output sequence)
- Lots of papers came out all around the same time that did image captioning: timely idea, facilitated by Encoder-Decoder architectures and transfer learning of ConvNet features.

Image Captioning: Encoder-Decoder

CNN Encoder + Language Model Decoder = Show and Tell (Vinyals et al., 2014)

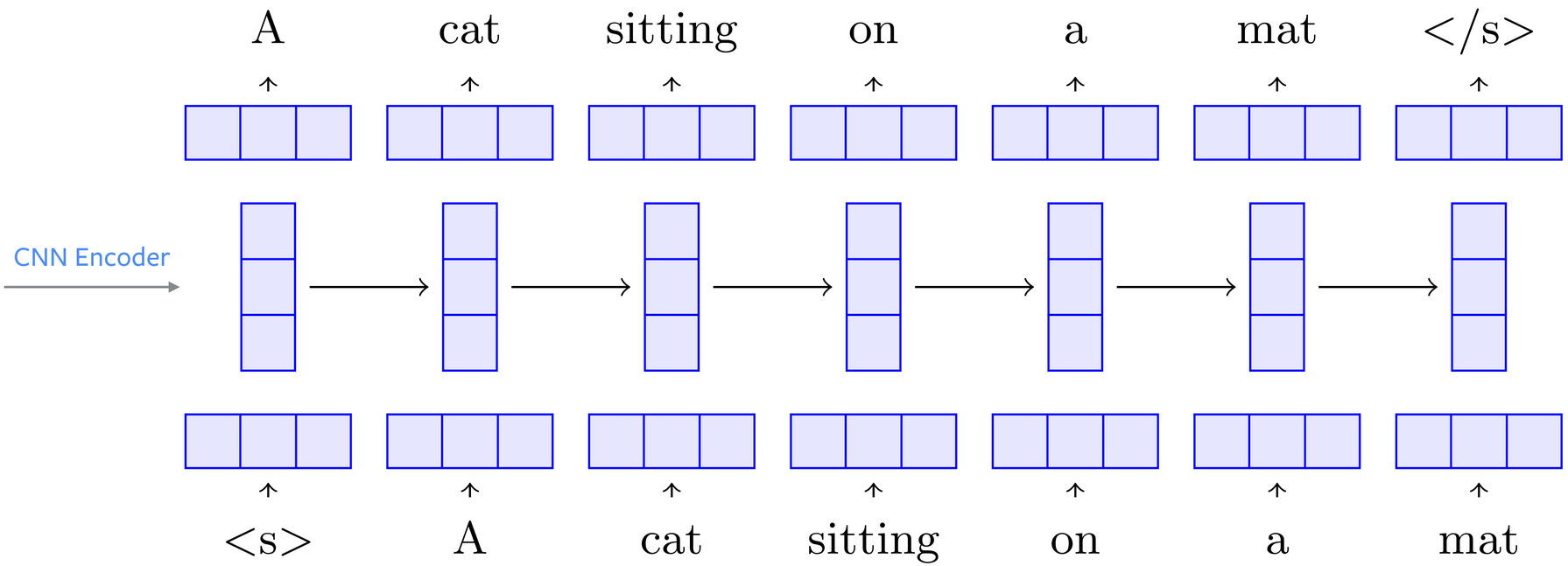


Image Captioning: Encoder-Decoder

Alternatives (Mao et al; Donahue et al; and others)

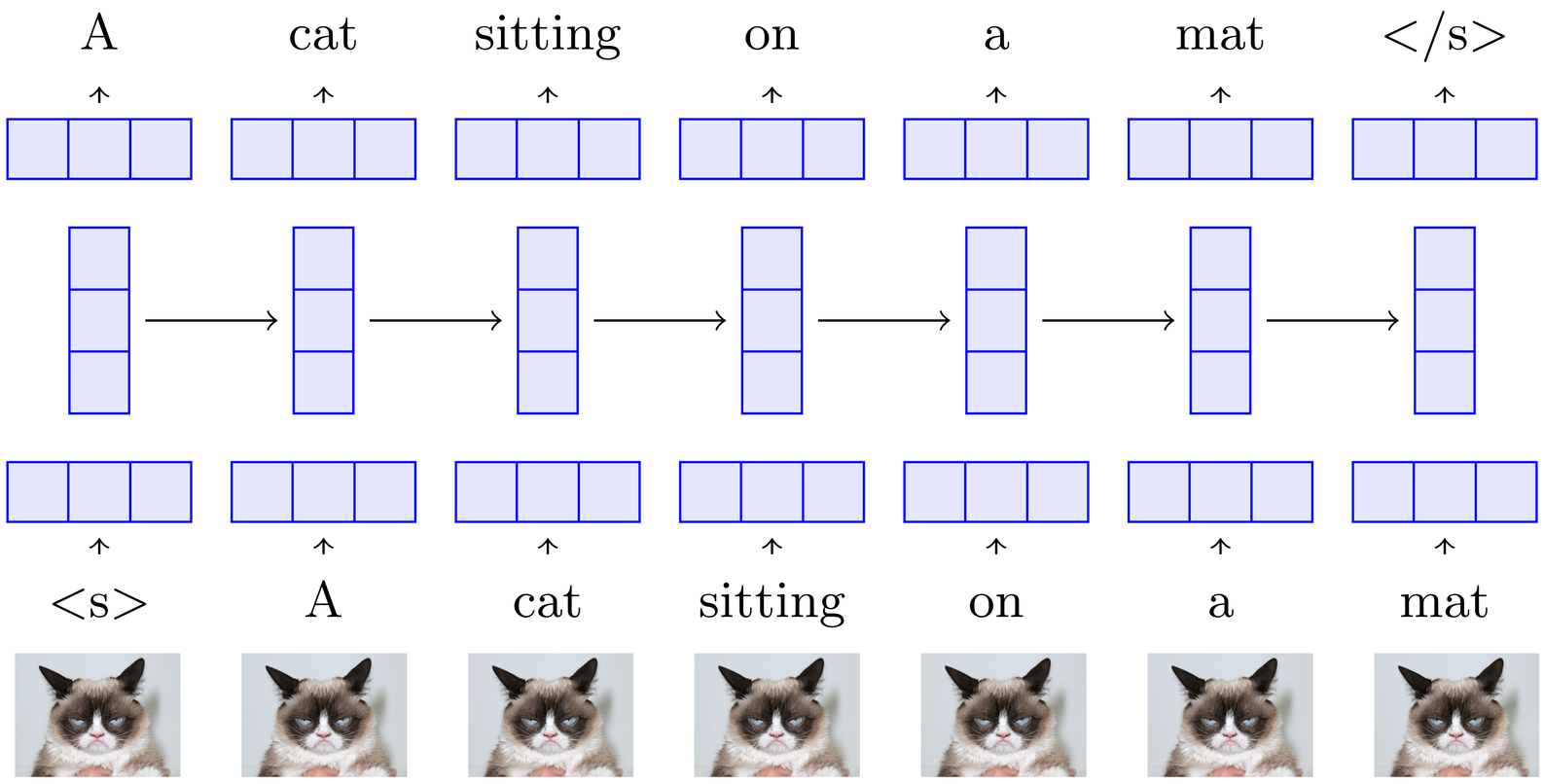


Image Captioning: Encoder-Decoder

Alternatives (Mao et al; Donahue et al; and others)

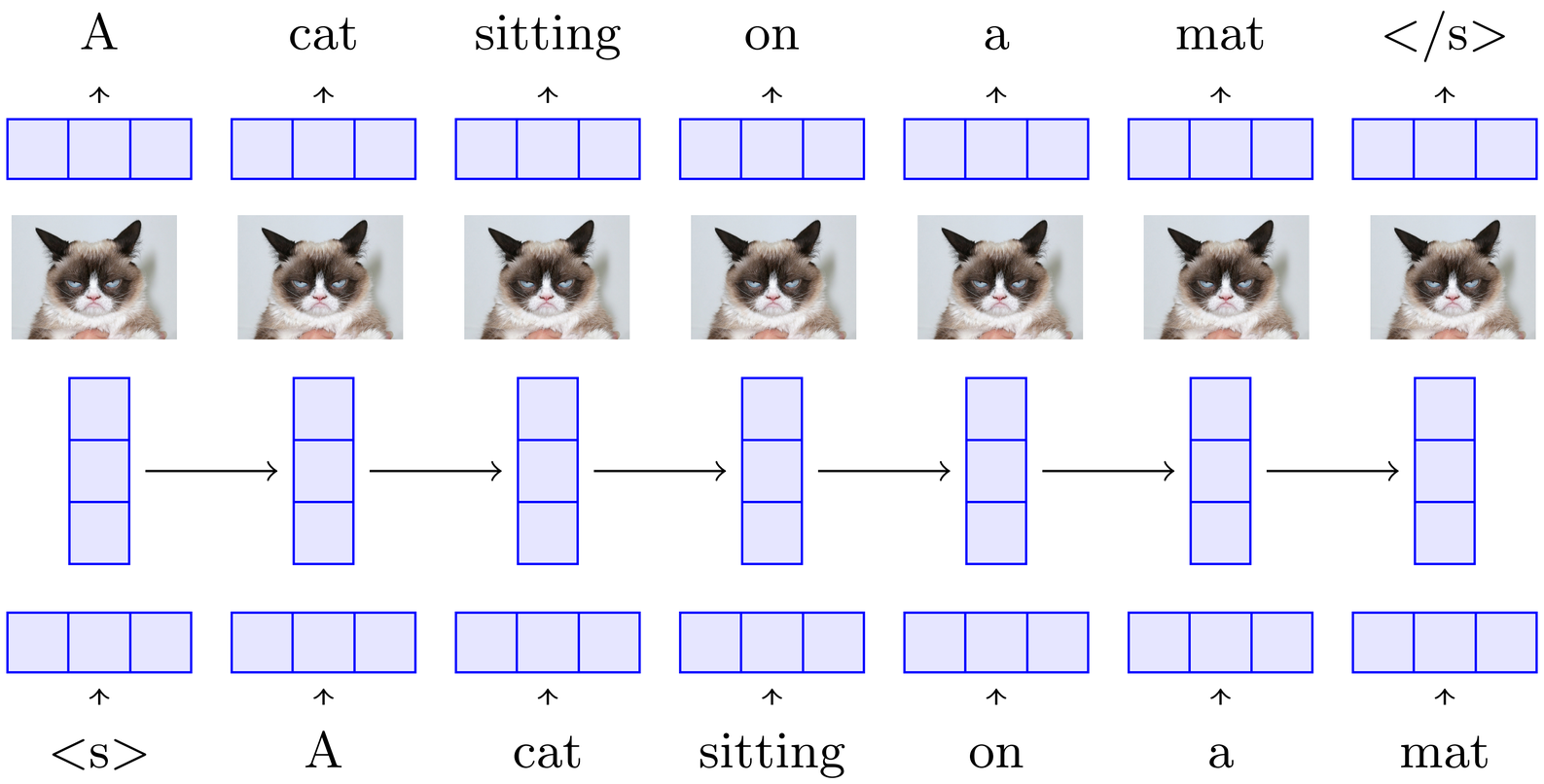


Image Captioning: Deep Visual-Semantic Alignments

Making use of bounding boxes (Karpathy & Li, 2015)

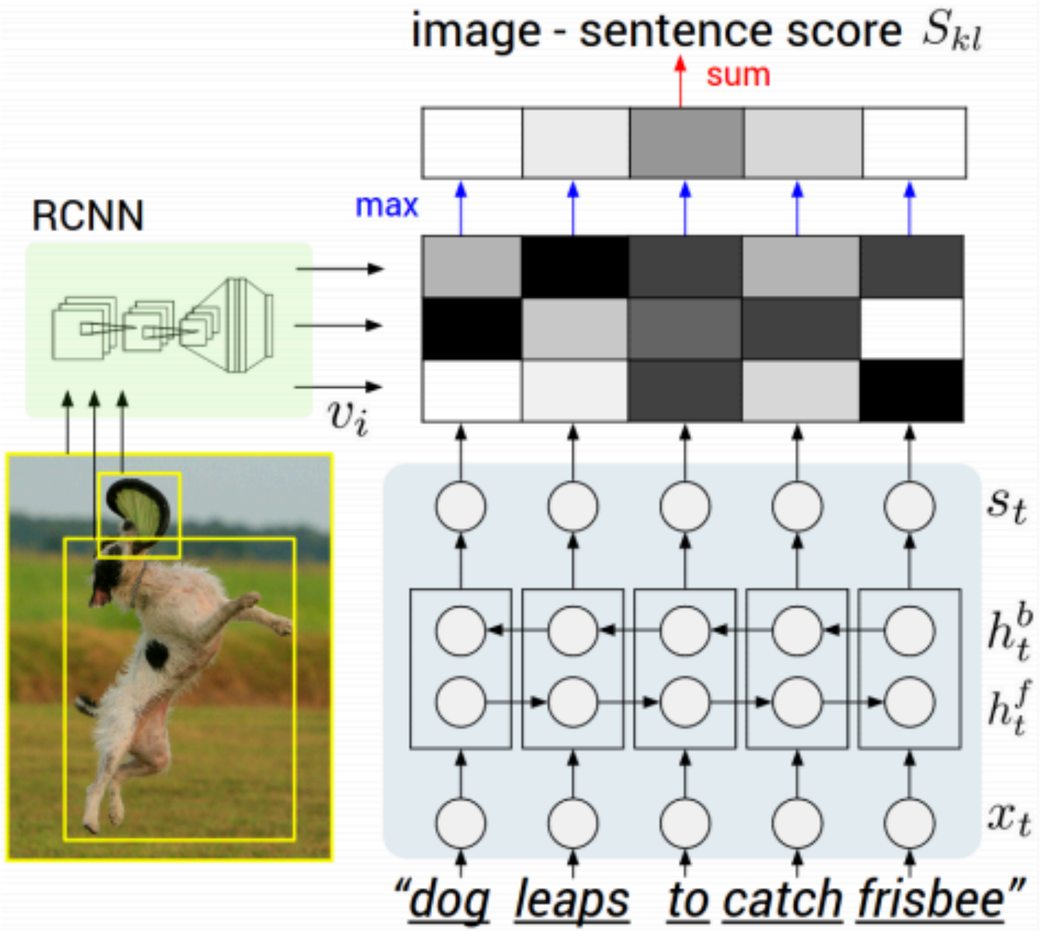


Image Captioning: Show, Attend and Tell (1)

Incorporating Attention into Image Captioning (Xu et al., 2016)

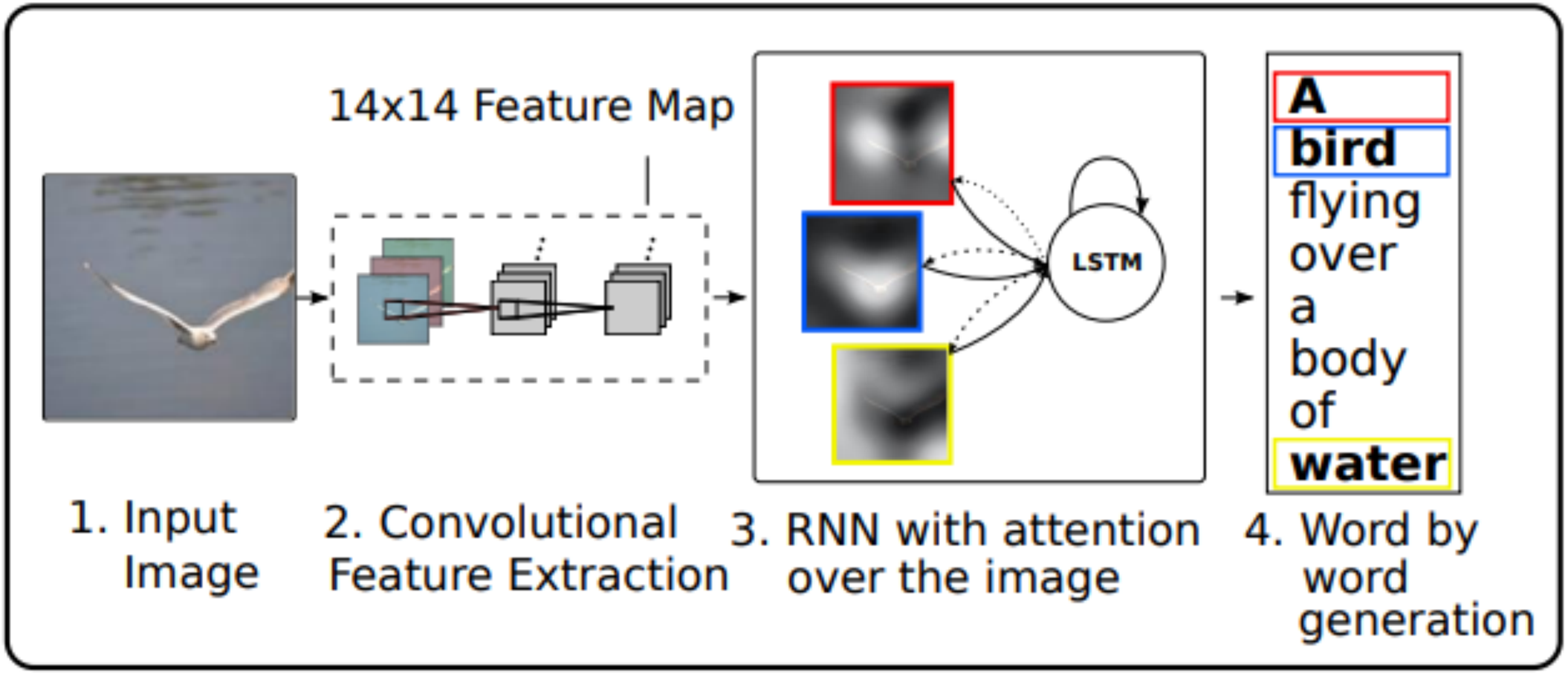


Image Captioning: Show, Attend and Tell (2)

Examples

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.

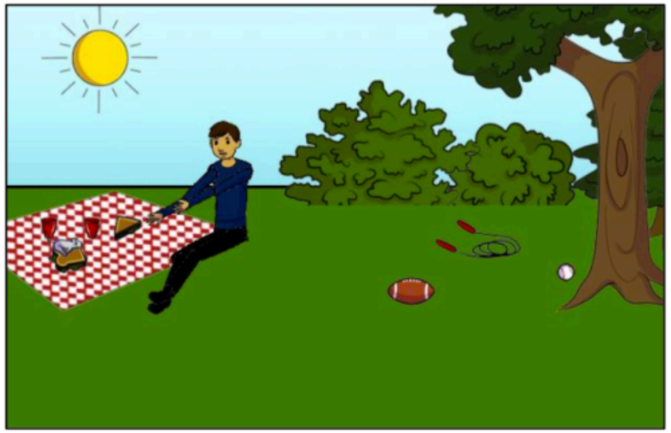
Visual Question Answering

A Visual "Turing Test"

- **But captioning showed limitations**
  - Heavily biased
- **Visual Question Answering is supposed to be harder and to really test for visual reasoning.**
- **Example tasks:**
  - Closed question
  - Abstract reasoning
  - Open question

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

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

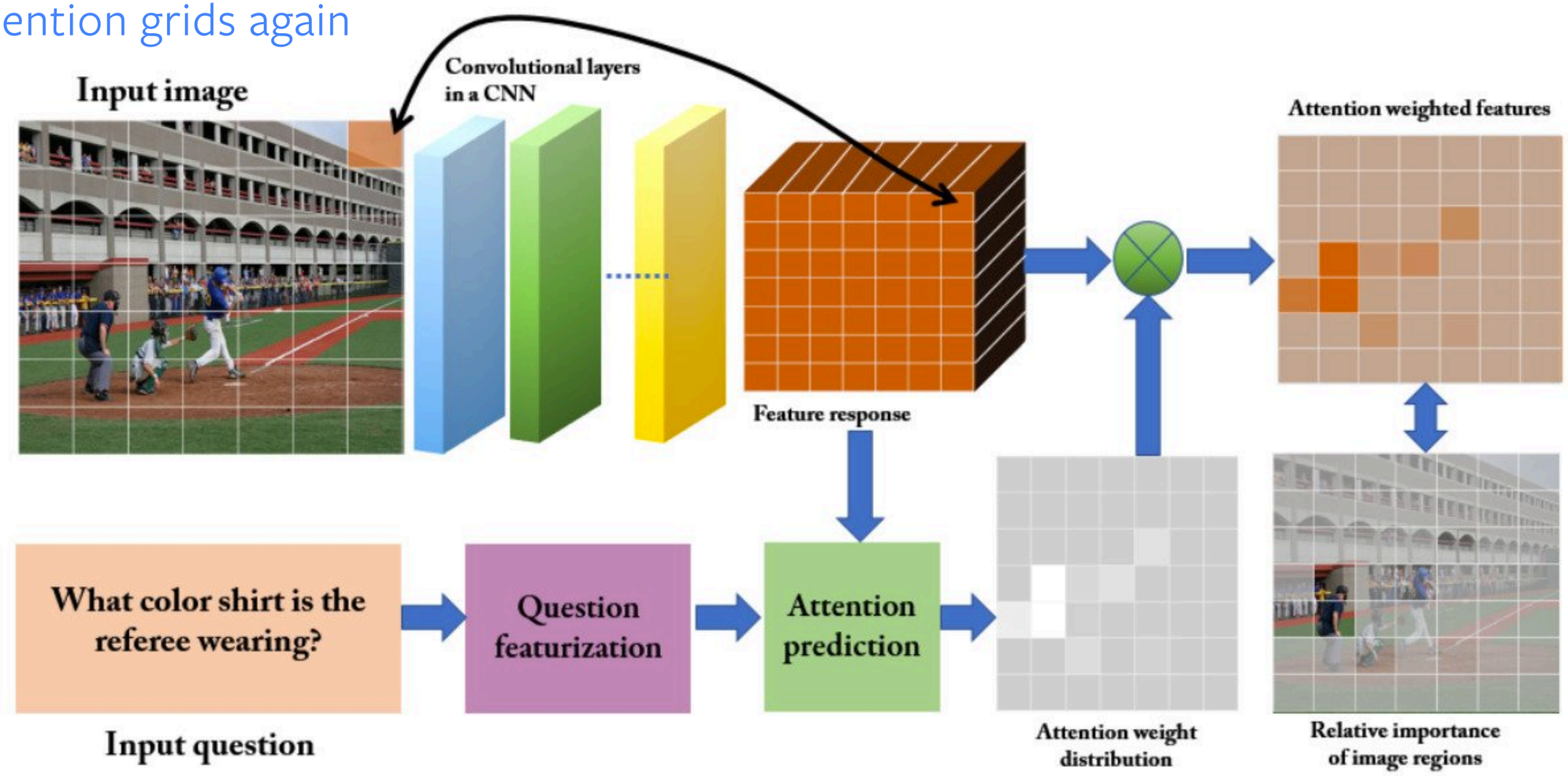


Is this person expecting company?
What is just under the tree?

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

Visual Question Answering

Attention grids again



Bilinear Attention Networks

A more technical example

- **Attention over sequences on both sides**
  - Low-rank bilinear model: $f_i = \mathbf{x}^T \mathbf{W}_i \mathbf{y} \approx \mathbf{x}^T \mathbf{U}_i \mathbf{V}_i^T \mathbf{y} = \mathbf{1}^T (\mathbf{U}_i^T \mathbf{x} \odot \mathbf{V}_i^T \mathbf{y})$

    Low-rank bilinear pooling: $\mathbf{f} = \mathbf{P}^T (\mathbf{U}^T \mathbf{x} \odot \mathbf{V}^T \mathbf{y})$

    Here, we have $\mathbf{X}$ and $\mathbf{Y}$ instead, so:

    Bilinear attention map: $\mathbf{f}'_k = (\mathbf{X}^T \mathbf{U}')_k^T \mathbf{A} (\mathbf{Y}^T \mathbf{V}')_k = \sum_{i=1}^{\rho} \sum_{j=1}^{\phi} A_{i,j} (\mathbf{X}_i^T \mathbf{U}'_k)(\mathbf{V}_k'^T \mathbf{Y}_j) = \sum_{i=1}^{\rho} \sum_{j=1}^{\phi} A_{i,j} \mathbf{X}_i^T (\mathbf{U}'_k \mathbf{V}_k'^T) \mathbf{Y}_j$

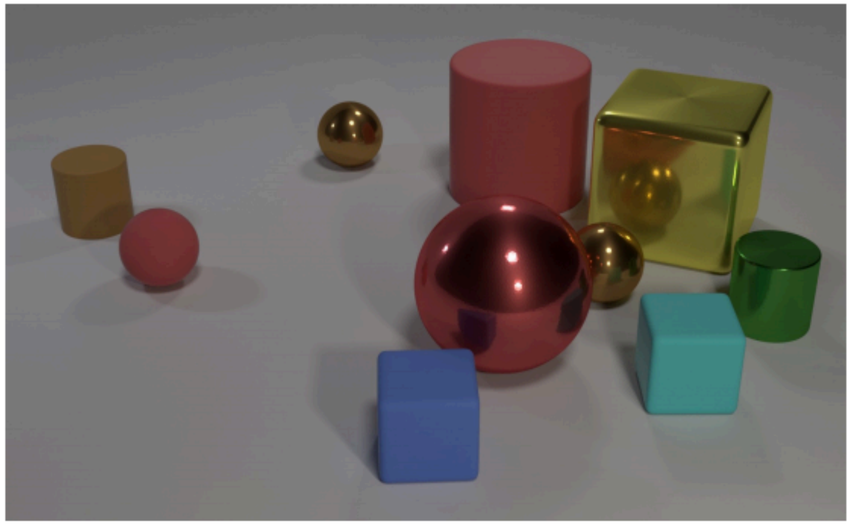
    Attention logits pre-softmax: $A_{i,j} = \mathbf{p}^T ((\mathbf{U}^T \mathbf{X}_i) \odot (\mathbf{V}^T \mathbf{Y}_j))$

- **This can be extended arbitrarily, using very basic concepts: attention and bilinear models.**

Visual Reasoning - CLEVR

Compositional Language and Elementary Visual Reasoning

- **But VQA datasets have problems also, e.g. strong biases on Q side.**
- **Meant as a diagnostic dataset, showed models were far from perfect.**



Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

Figure 1. A sample image and questions from CLEVR. Questions test aspects of visual reasoning such as attribute identification, counting, comparison, multiple attention, and logical operations.

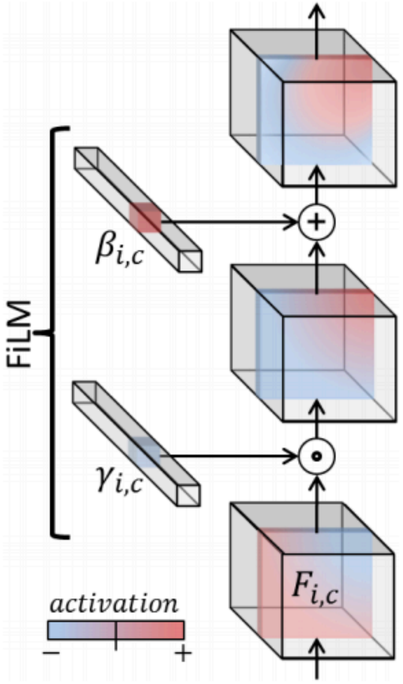
FiLM

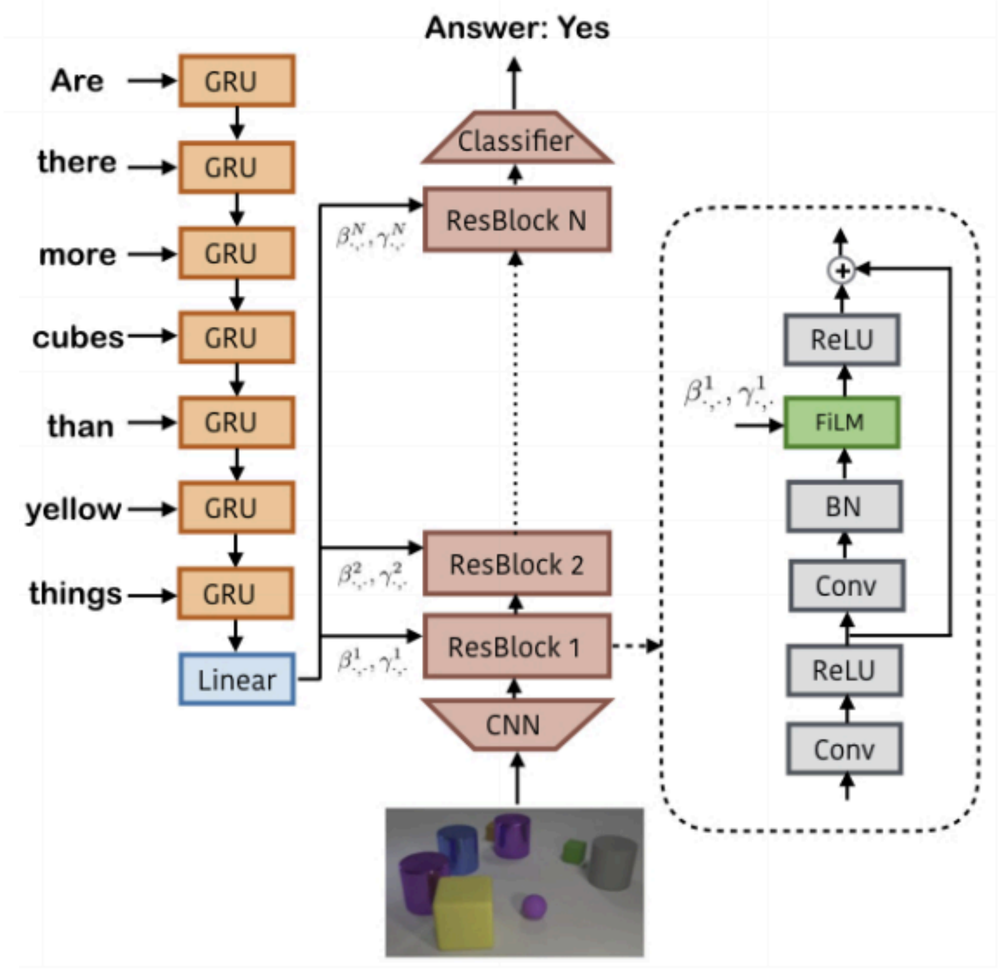
Feature-wise Linear Modulation (Perez et al)

- **We can learn to modulate one modality from another one:**

$$\gamma_{i,c} = f_c(\mathbf{x}_i)$$
$$\beta_{i,c} = h_c(\mathbf{x}_i)$$
$$FiLM(\mathbf{F}_{i,c} \mid \beta_{i,c}, \gamma_{i,c}) = \gamma_{i,c}\mathbf{F}_{i,c} + \beta_{i,c}$$

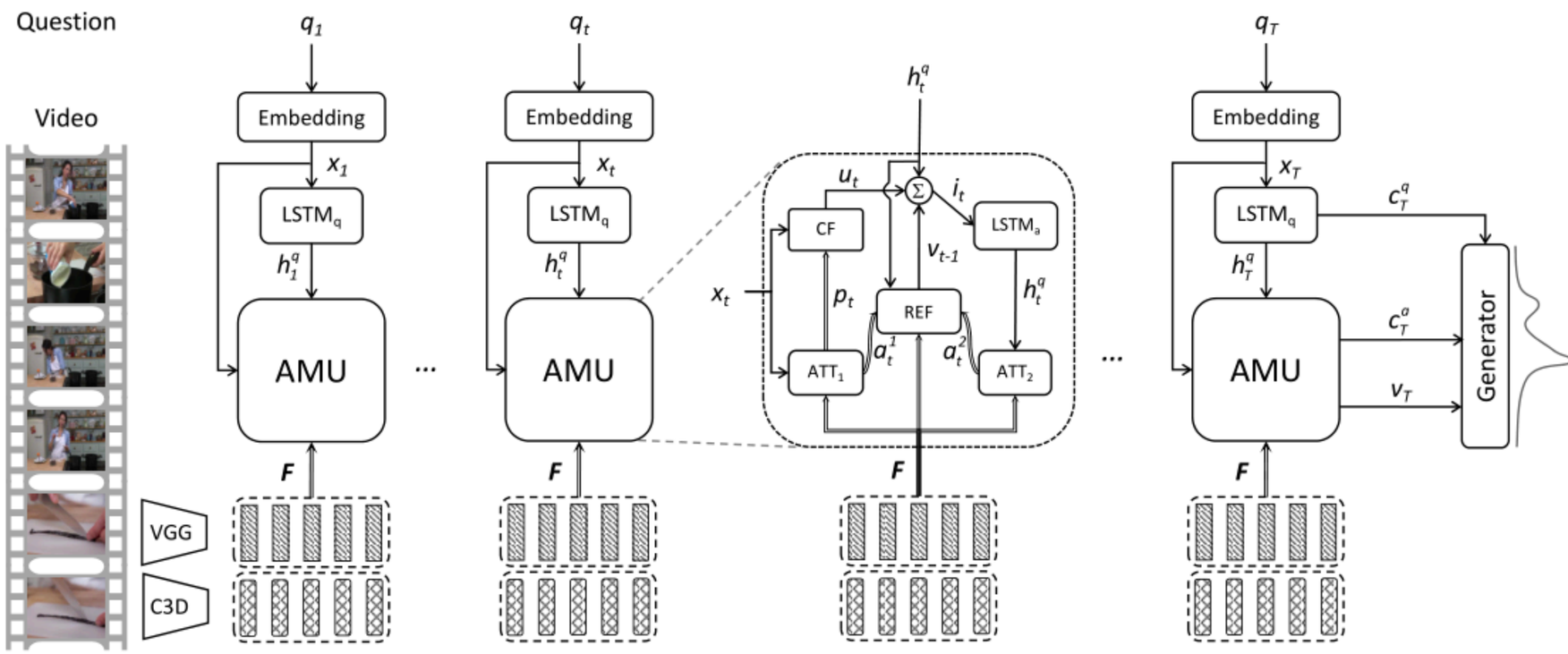
- **Impressive results on CLEVR**





Video

= Image Feature(s)
+ Audio Features + Other

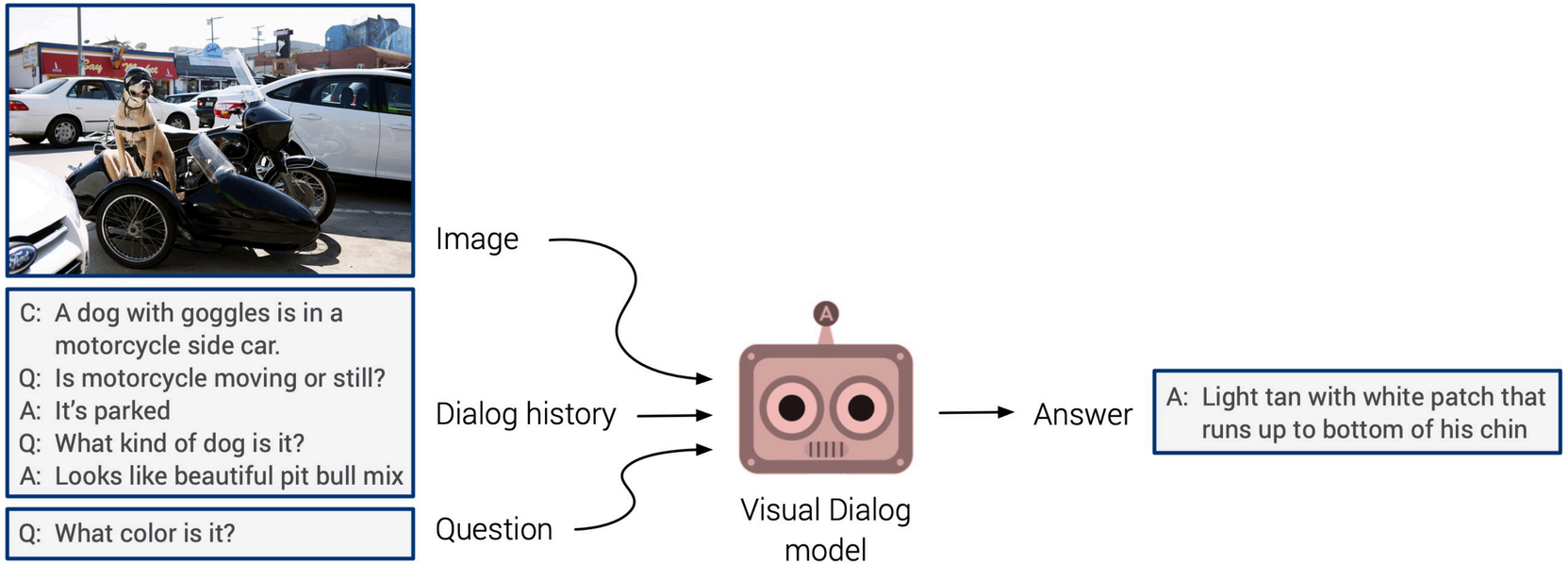


- **Audio-Visual Speech Recognition ("Captioning")**
- **Video content description (Alternative form of Captioning)**
- **Video QA**

Visual Dialogue

Grounded Dialogue (Das et al)

- **Next utterance prediction conditioned not only on dialogue history (and question) but also image.**



Beyond Vision

Olfactory, Gustatory, Tactile, anything goes..

- **Bag of chemical compounds (Kiela et al 2015)**

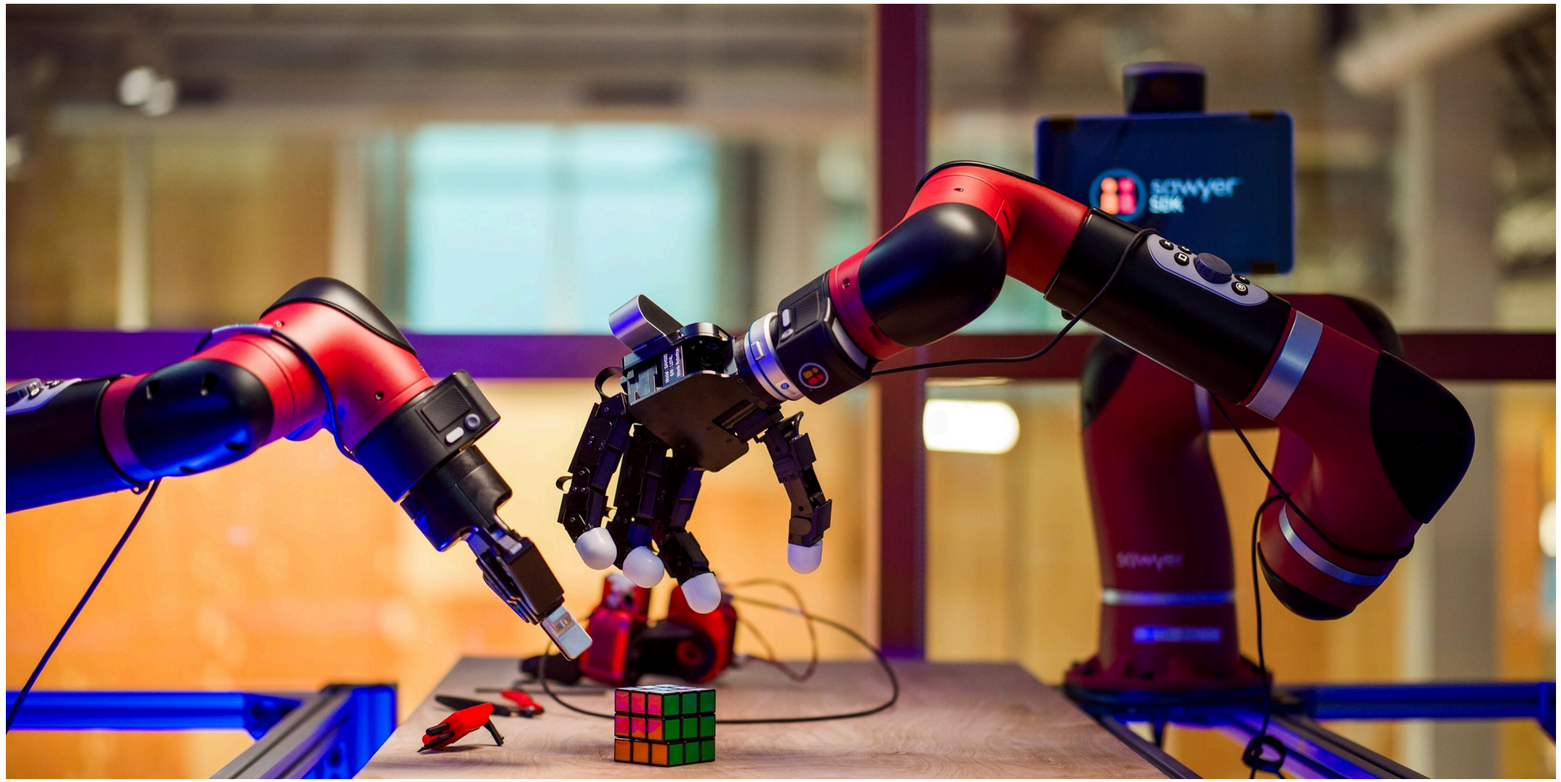
Smell label	Chemical Compound: Phenethyl acetate	Isoamyl butyrate	Anisyl butyrate	Myrcene	Syringaldehyde
Melon	✓	✓			
Pineapple	✓				✓
Licorice			✓		
Anise			✓	✓	
Beer				✓	✓

apple	bacon	brandy	cashew
pear	smoky	rum	hazelnut
banana	roasted	whiskey	peanut
melon	coffee	wine-like	almond
apricot	mesquite	grape	hawthorne
pineapple	mossy	fleshy	jam
chocolate	lemon	cheese	caramel
cocoa	citrus	grassy	nutty
sweet	geranium	butter	roasted
coffee	grapefruit	oily	maple
licorice	tart	creamy	butterscotch
roasted	floral	coconut	coffee

Robotics

Obviously multimodal

- **Lots of old parallel work**
- **Lots of potential**
- **Lots of modalities**
- **BUT: Physical constraints**

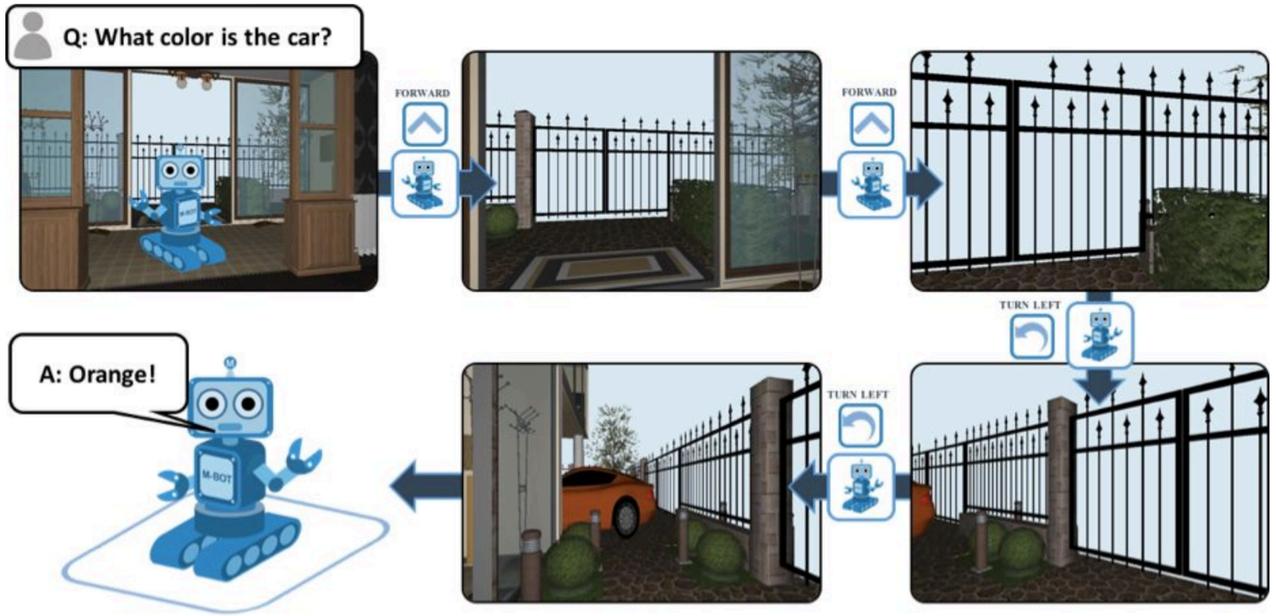


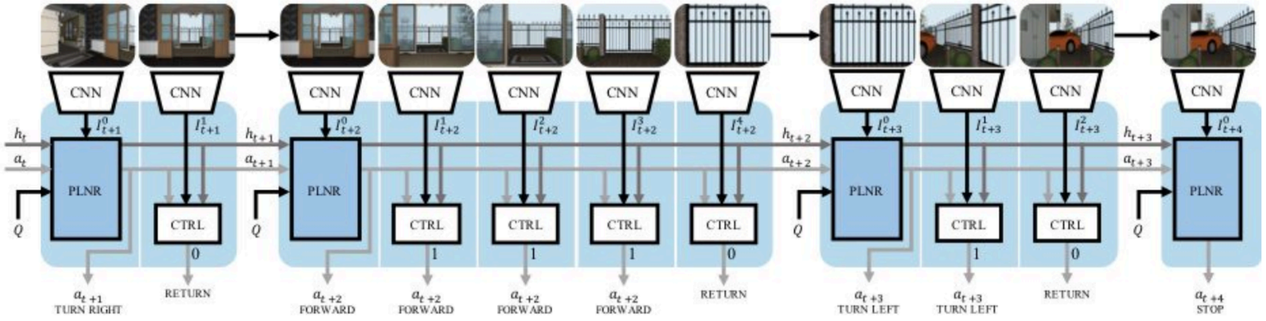
Embodied QA

What about virtual embodiment (Das et al)

- **An "embodied Turing test"?**
  Question answering in virtual worlds.

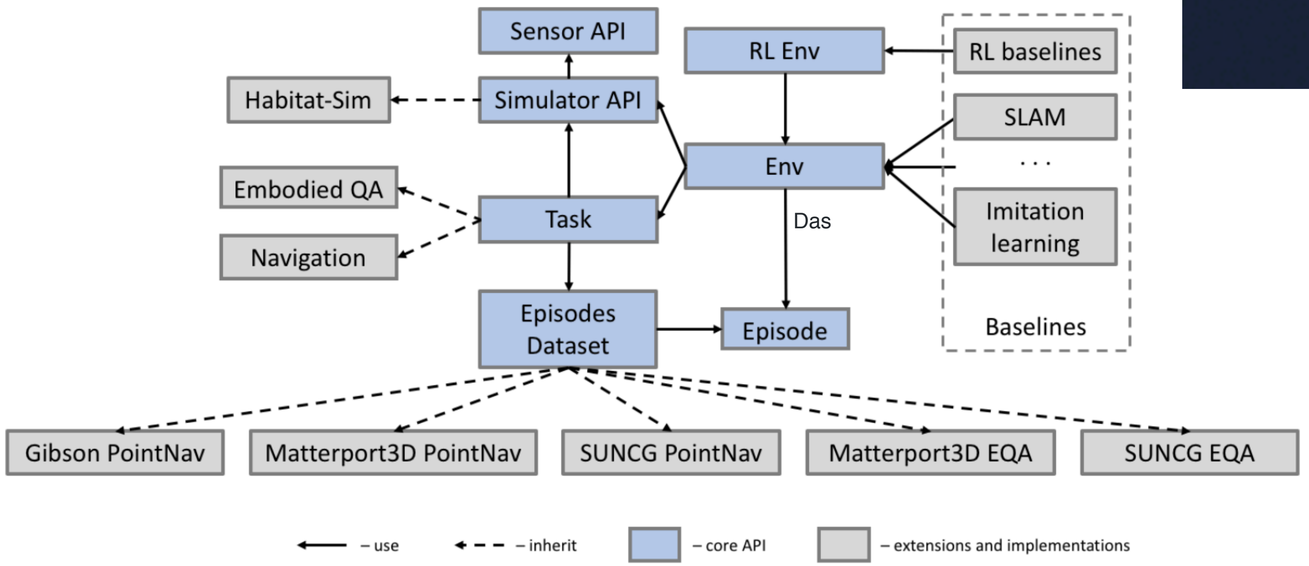
- **What roboticists call Sim2Real**





Habitat

Virtual worlds for sim2real



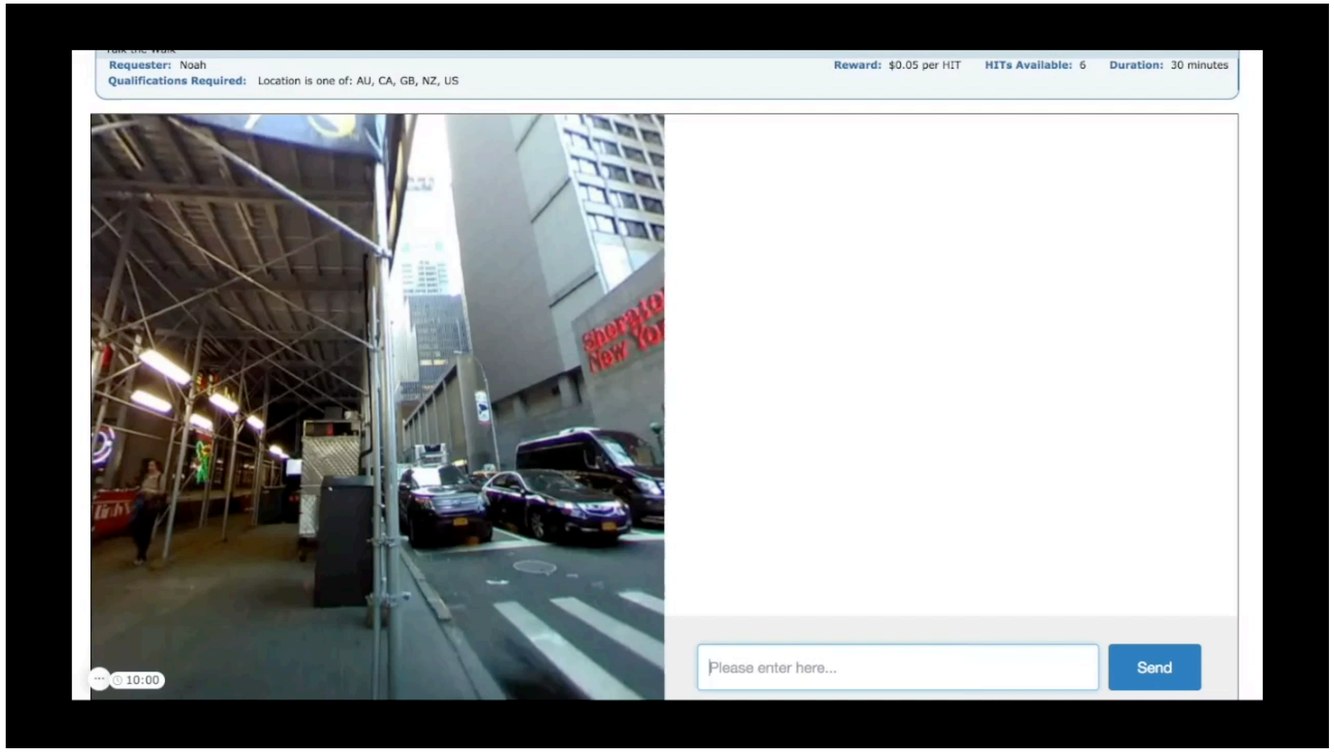
Talk The Walk

Dataset for Perception, Action & Interaction

- **Tourist and Guide work together to navigate NYC's grid system via dialogue.**

- **"Next Frontier" dataset:**
  - Combines 360 images (perception), navigation (action) and dialogue (interaction) into one single task.

- **Partially observable "navigation game": communication *required*.**



Talk The Walk

Dataset for Perception, Action & Interaction

- **In the paper:**
  - Perfect perception
  - Emergent language upper-bounds
  - Random walk navigation
  - Natural language baselines

- **A LOT OF WORK TO BE DONE.**
  https://github.com/facebookresearch/talkthewalk/

THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO MORE

TECH \ FACEBOOK \ ARTIFICIAL INTELLIGENCE \

Facebook sets a new task for AI: guide a virtual tourist around New York

The challenge is designed to spur research in AI language and perception

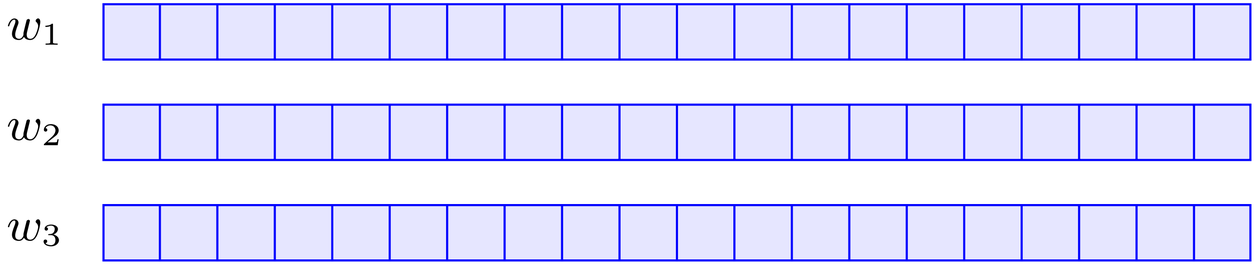
Outline

- **Background:** The meaning of Multi-modal
- **Core Concepts:** Representation Learning, Attention, Fusion
- **Applications:** Image/video captioning, VQA, Representation Learning, Visual Dialogue, Robotics
- **In practice** How to use this in practice
- **The Future:** Open Problems

In practice

Discretization

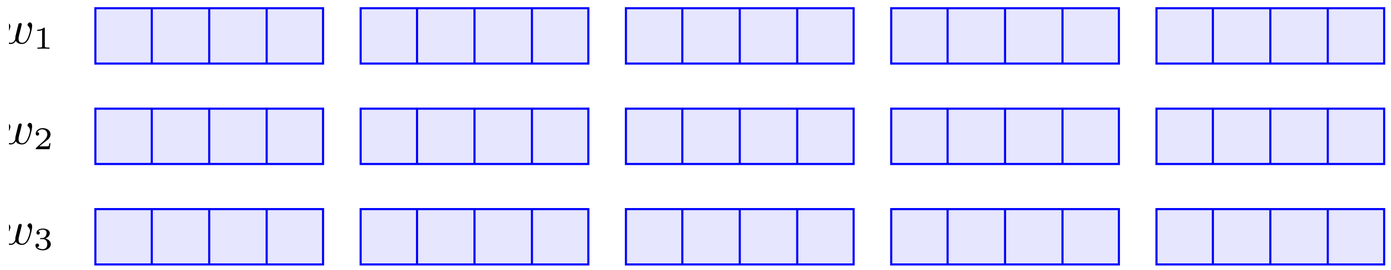
- **ConvNets are expensive, running GPUs (esp. for inference) is not always possible.**
- **How do we scale?**
- **Discretization. E.g. production quantization.**



In practice

Discretization

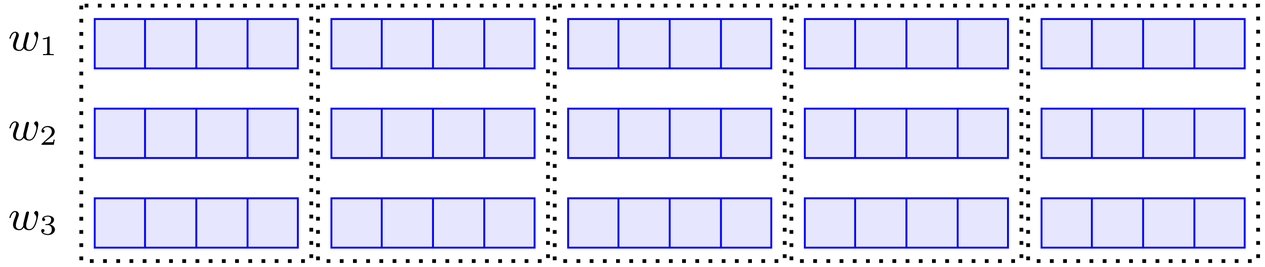
- **ConvNets are expensive, running GPUs (esp. for inference) is not always possible.**
- **How do we scale?**
- **Discretization.**



In practice

Discretization

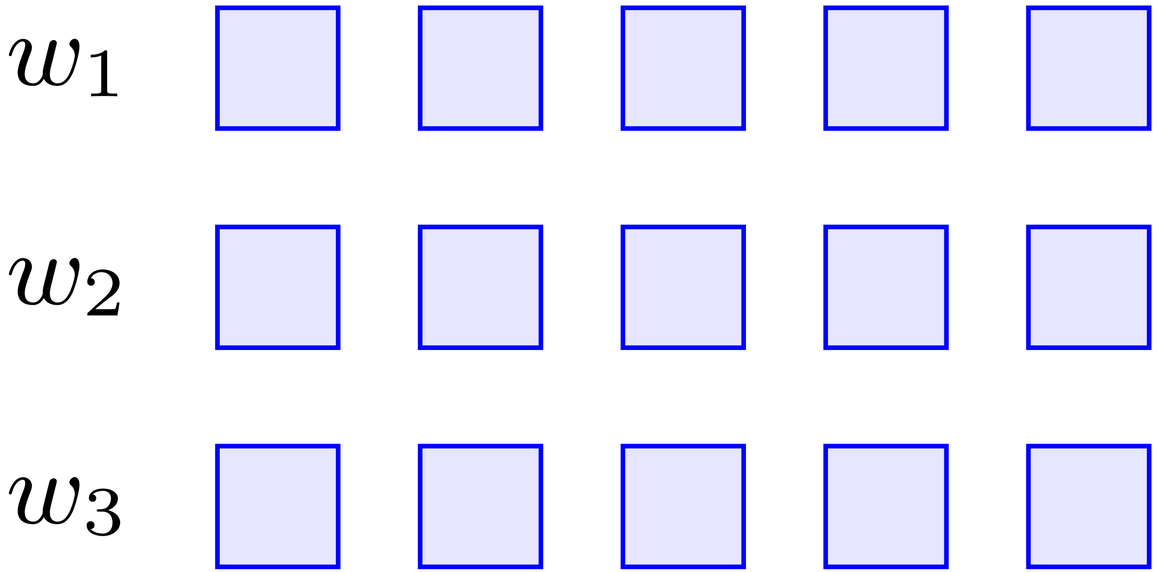
- **ConvNets are expensive, running GPUs (esp. for inference) is not always possible.**
- **How do we scale?**
- **Discretization.**



In practice

Discretization

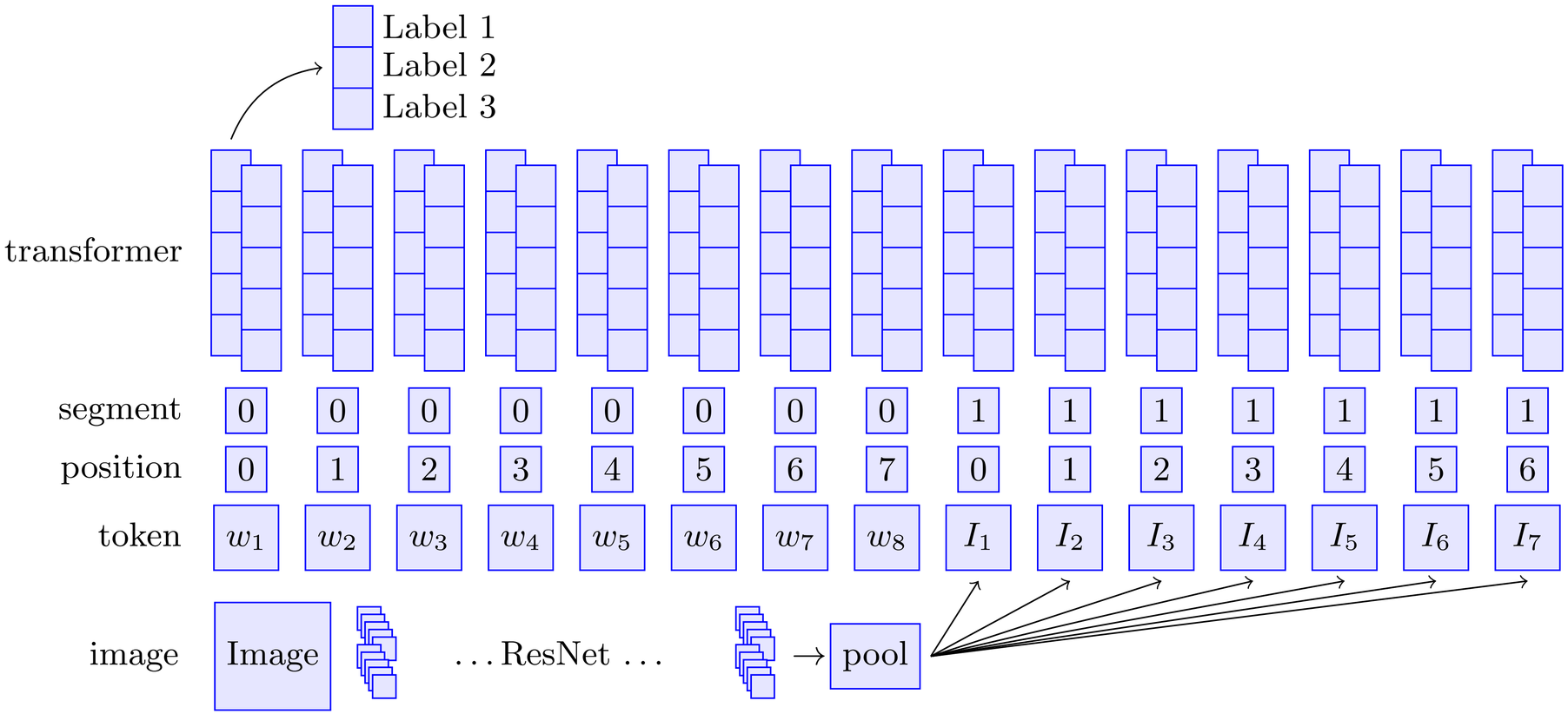
- **ConvNets are expensive, running GPUs is not always possible.**
- **How do we scale?**
- **Discretization.**
- **Image 1 -> ”__q__1_26 __q__13_123” etc.**



Outline

- **Background:** The meaning of Multi-modal
- **Core Concepts:** Representation Learning, Attention, Fusion
- **Applications:** Image/video captioning, VQA, Representation Learning, Visual Dialogue, Robotics
- **In practice** How to use this in practice
- **The Future:** Open Problems

Self-supervised learning



Three Pillars of Meaning

Essential Components of Real Natural Language Semantics

Perception.

We have a shared sensorimotor experience of the world.

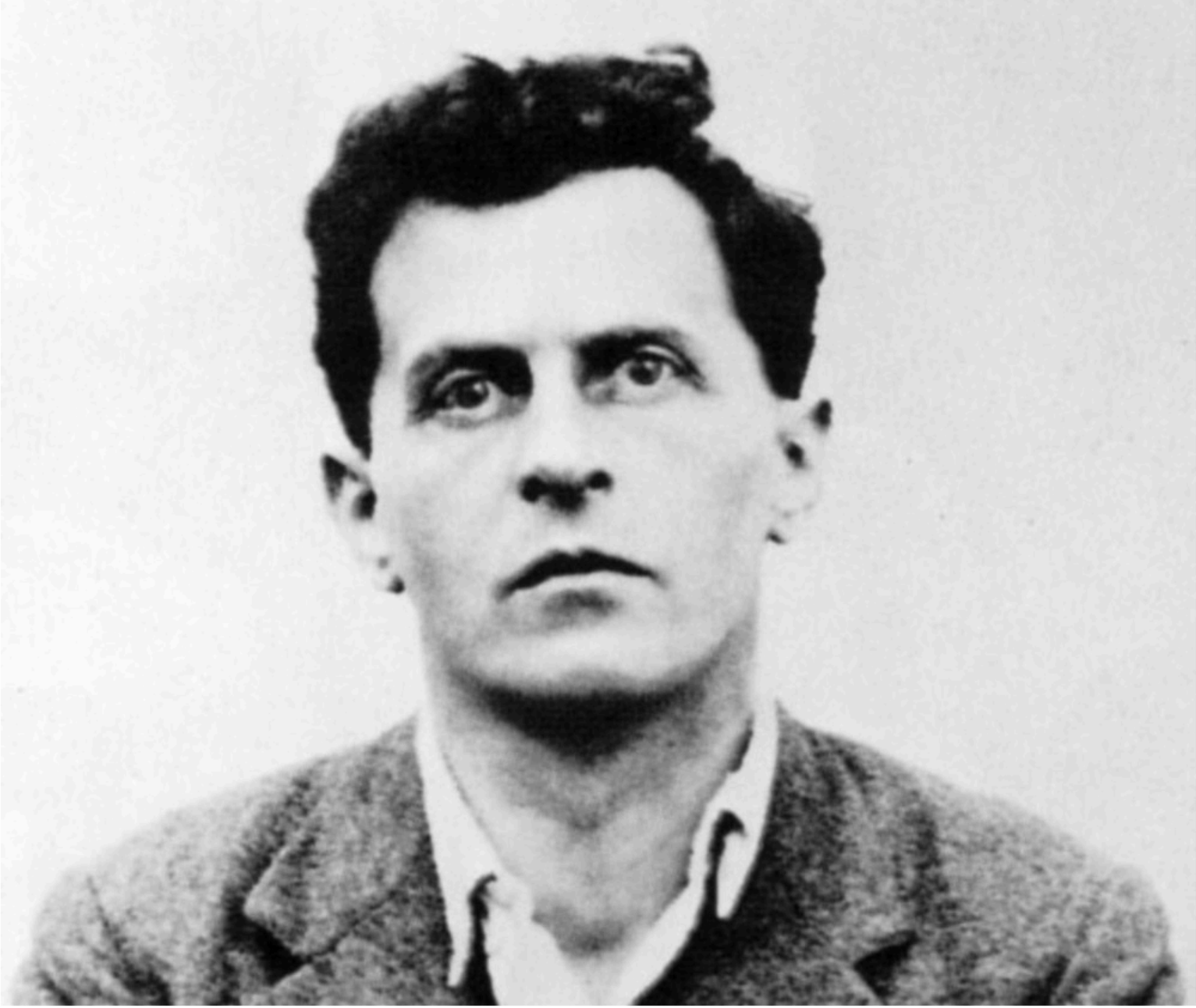
Action.

We take actions to manipulate and change the world around us.

Interaction.

We interact with other agents playing "language games".

- **See also, Virtual Embodiment: A Scalable Long-Term Strategy for Artificial Intelligence Research (Kiela et al. 2016)**



Wittgenstein

Meaning = Use

Philosophische Untersuchungen (1950)

"the meaning of words lies in their use"

But what is the meaning of the word "five"?—No such thing was in question here, only how the word "five" is used.

The Meaning of "Meaning"

What the Philosophical Investigations say

(7) We can also think of the whole process of using words in (2) as one of those games by means of which children learn their native language. I will call these games "language-games" and will sometimes speak of a primitive language as a language-game. [...] **I shall also call the whole, consisting of language and the actions into which it is woven, the "language-game".**

(43) For a large class of cases—though not for all—in which we employ the word "meaning" it can be defined thus: **the meaning of a word is its use in the language**. And the meaning of a name is **sometimes explained by pointing** to its bearer.

The Meaning of "Meaning"

What the Philosophical Investigations *really* Say (According to... Me)

(43) **For a large class of cases**—though not for all—in which we employ the word "meaning" it can be defined thus: **the meaning of a word is its use in the language**. And the meaning of a name is sometimes explained by pointing to its bearer.

The meaning of language **mostly depends on active usage** in a community of language speakers playing "language games".

The Meaning of "Meaning"

What the Philosophical Investigations *really* Say (According to… Me)

(43) For a large class of cases—**though not for all**—in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language. **And the meaning of a name is sometimes explained by pointing to its bearer.**

The meaning of language depends in part on **language-to-world grounding.**

The Meaning of "Meaning"

What the Philosophical Investigations *really* Say (According to... Me)

(43) **For a large class of cases**—**though not for all**—in which we employ the word "meaning" it can be defined thus: **the meaning of a word is its use in the language**. **And the meaning of a name is sometimes explained by pointing to its bearer.**

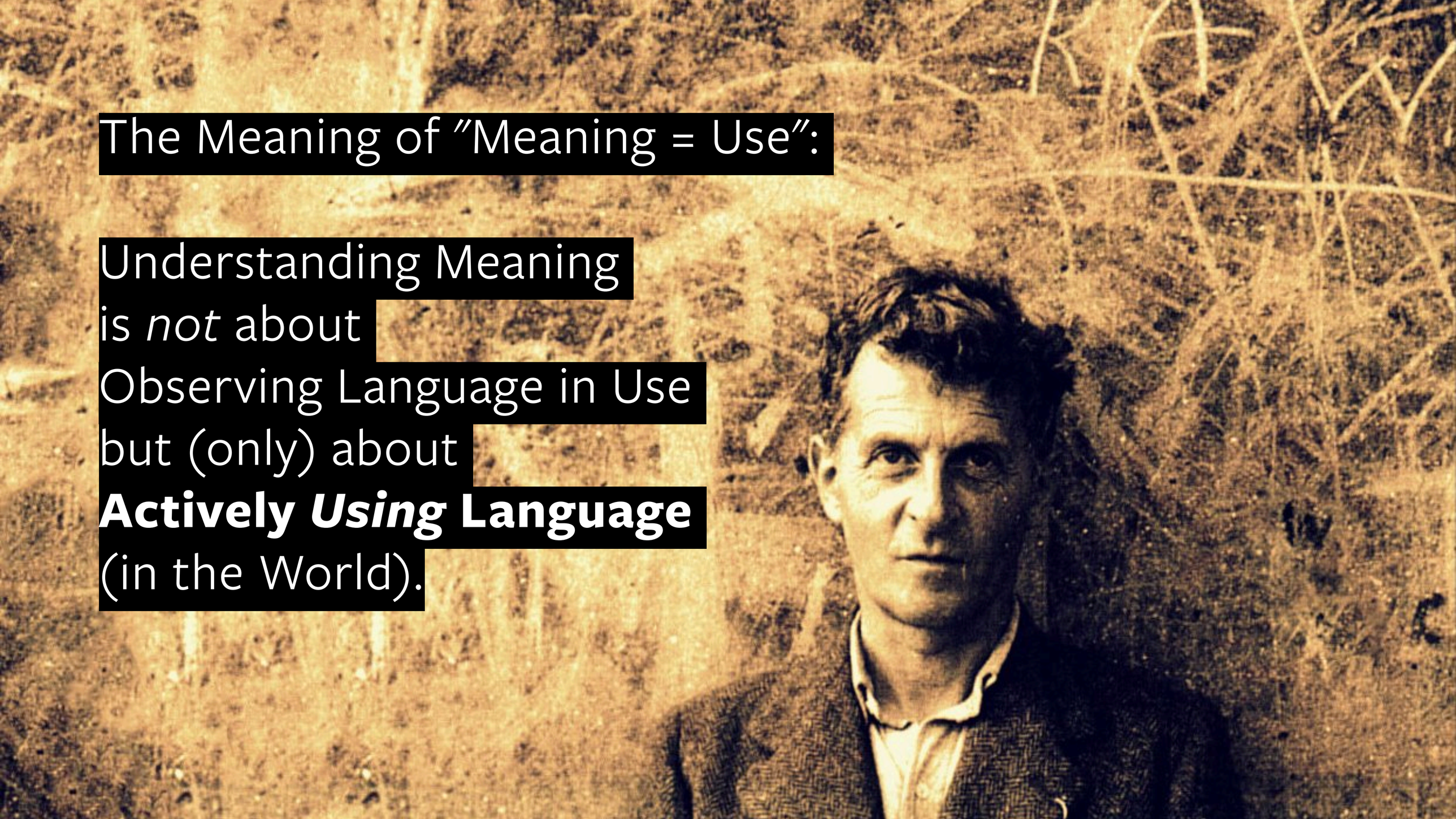
The meaning of language depends in part on **language-to-world grounding** and **mostly depends on active usage** in a community of language speakers playing "language games".

^ THIS IS NOT AT ALL WHAT NLP IS DOING AT THE MOMENT.

Has NLP become lazy?

"All we need is more data, more compute."

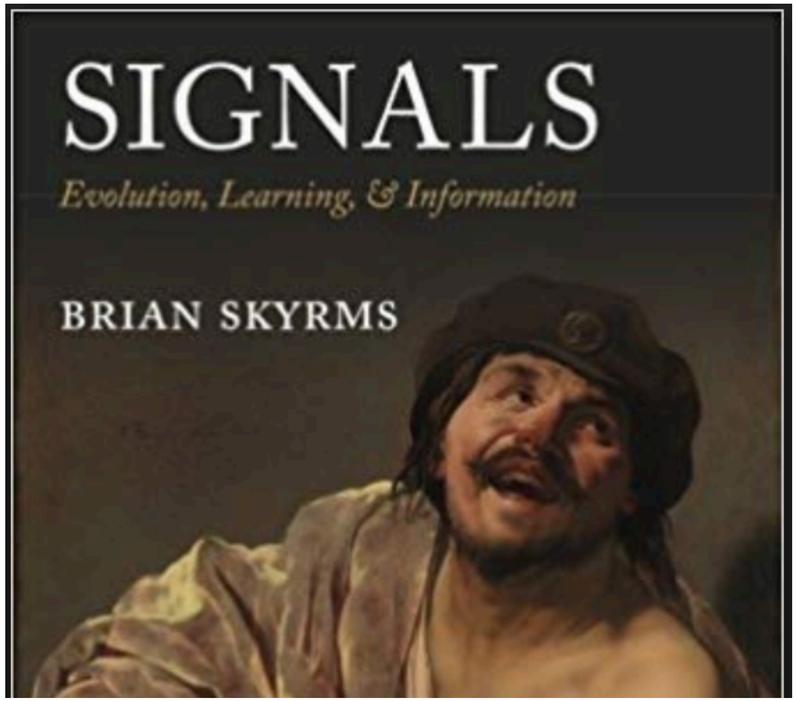
(which, to be fair, has been an extremely successful strategy)

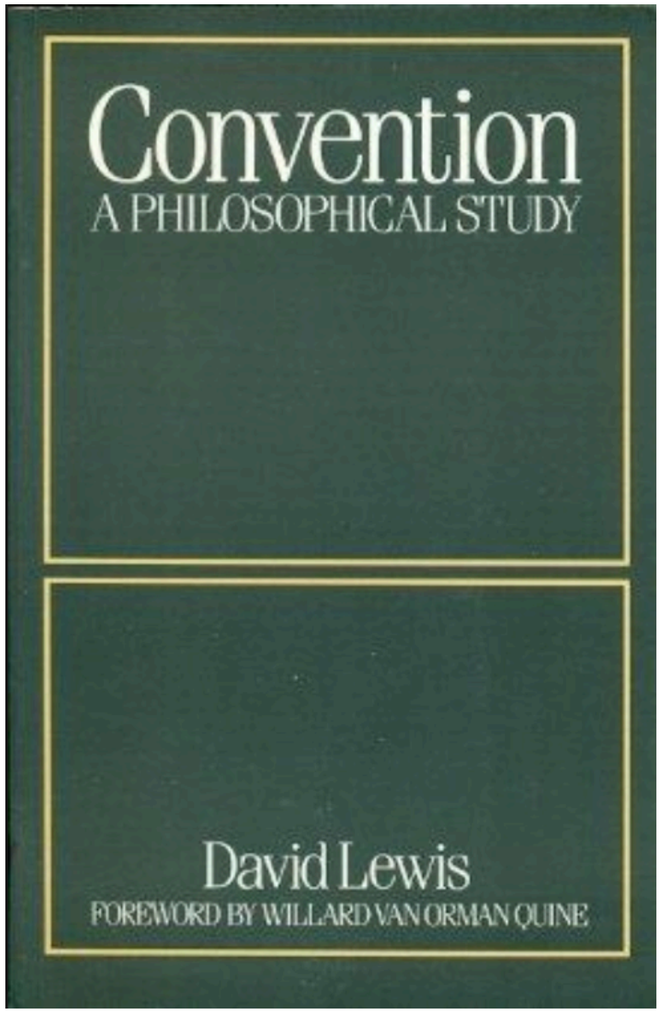

The Meaning of "Meaning = Use":
Understanding Meaning
is *not* about
Observing Language in Use
but (only) about
Actively *Using* Language
(in the World).

Language games

Multi-Agent Cooperation and the Emergence of (Natural) Language

Angeliki Lazaridou[1]*, **Alexander Peysakhovich**[2], **Marco Baroni**[2,3]
[1]Google DeepMind, [2]Facebook AI Research, [3]University of Trento
angeliki@google.com, {alexpeys,mbaroni}@fb.com





Emergent Translation in Multi-Agent Communication

Lee, Cho, Weston, Kiela. ICLR 2017.

Grounded Language Games

Multi-modal ground-up language learning

- **Perception, action, interaction**
- **The future is multi-modal**

Thanks!

Thanks for listening

- **The future is multi-modal**