Neural Network Tools and Principles, part 3

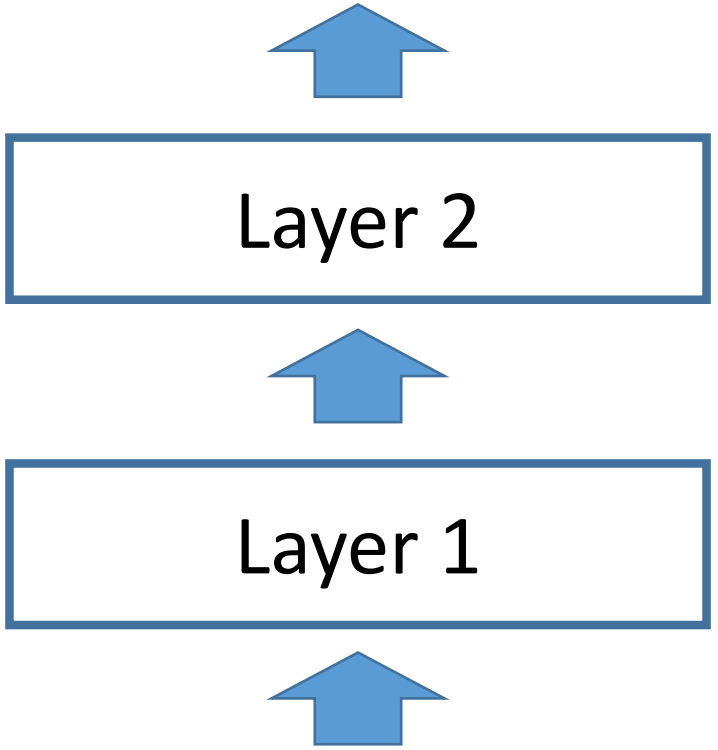
Sequential problems

Computational Intelligence in Games, Spring 2018
Prof. Perttu Hämäläinen
Aalto University

Sequence modeling

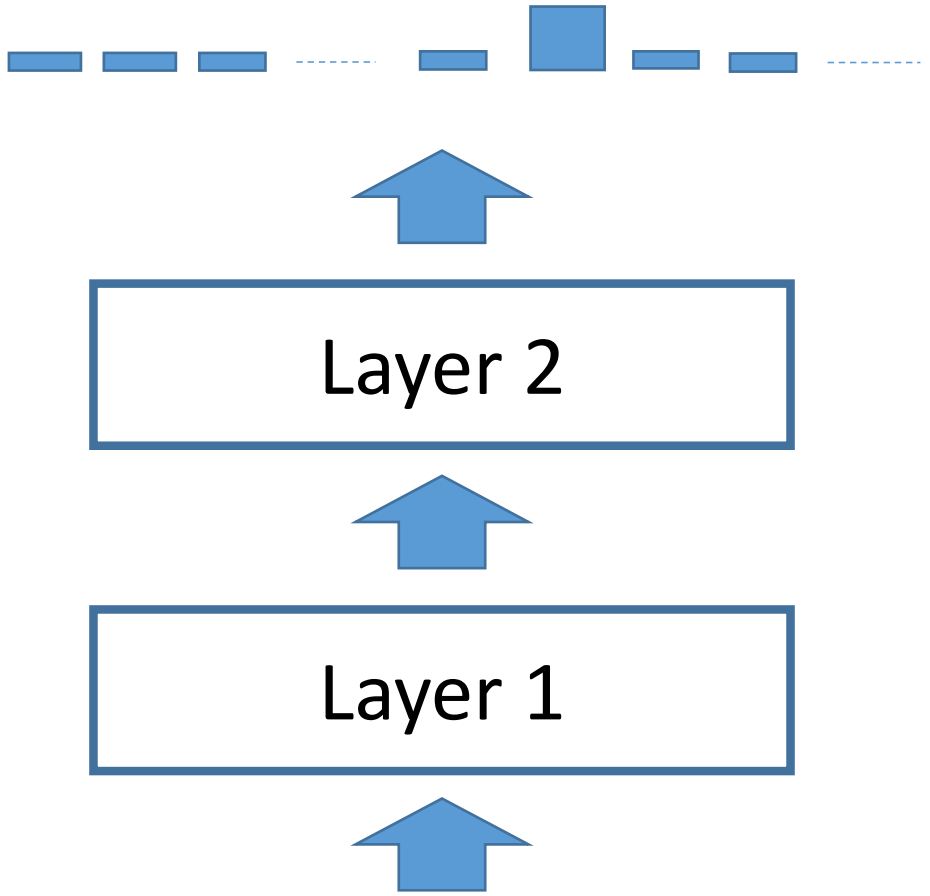
- Predictive text input
- Generating text
- Generating audio
- Language translation (sequence-to-sequence mapping)

A B C … M N O P …



P R E D I C T I

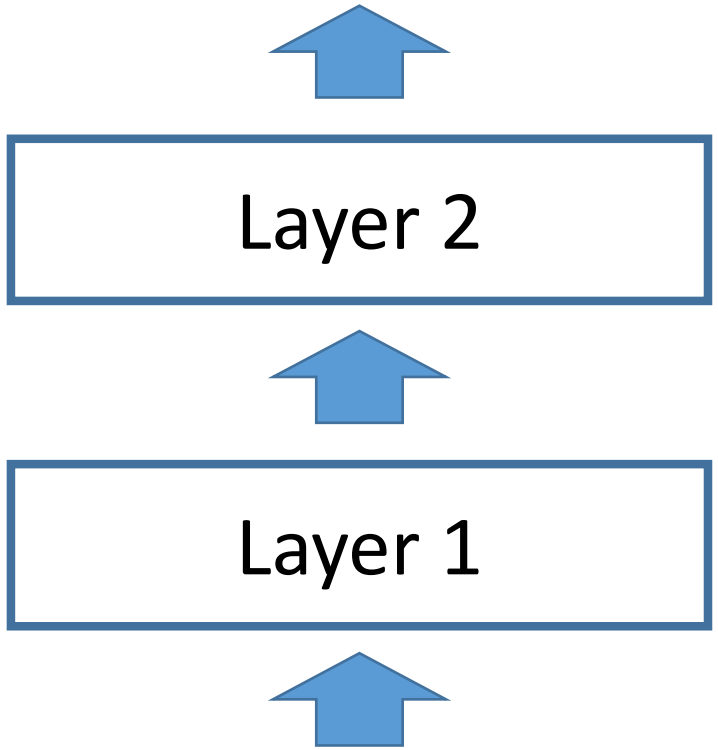
A B C … M N O P …



I A M P R E D I C T I

Infinite input sequence length = infinite computing cost

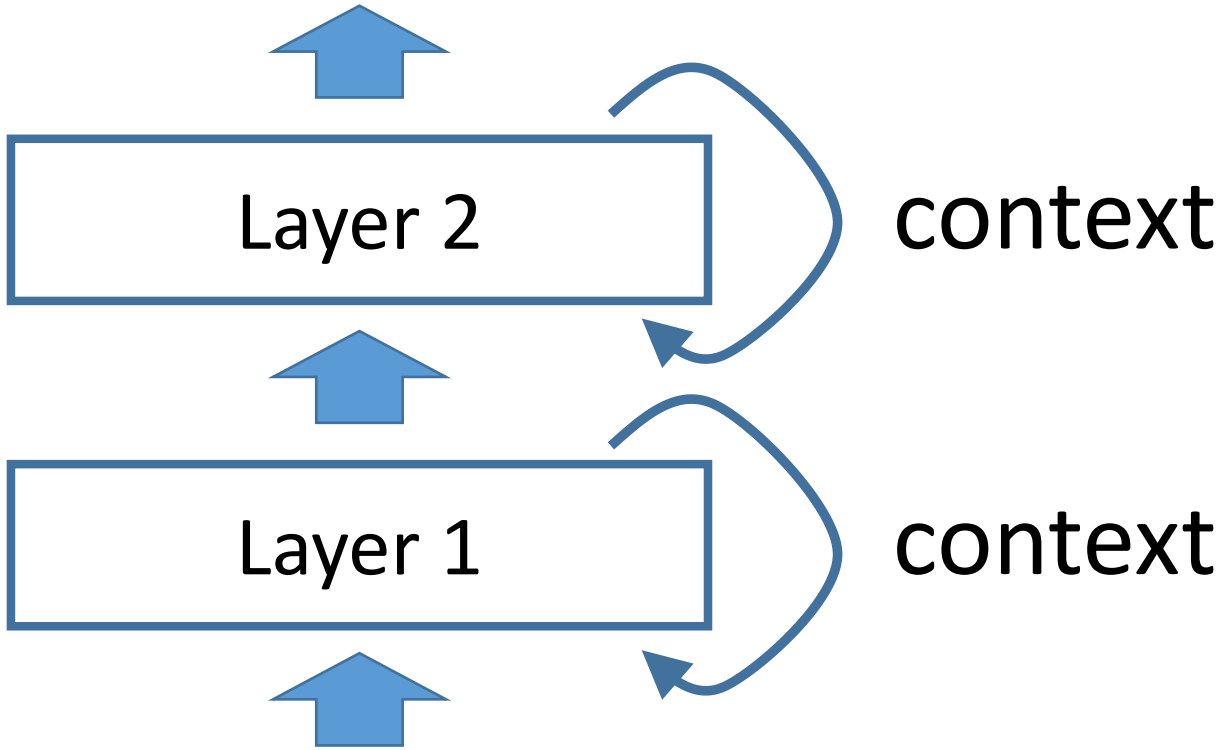
A B C … M N O P …



I

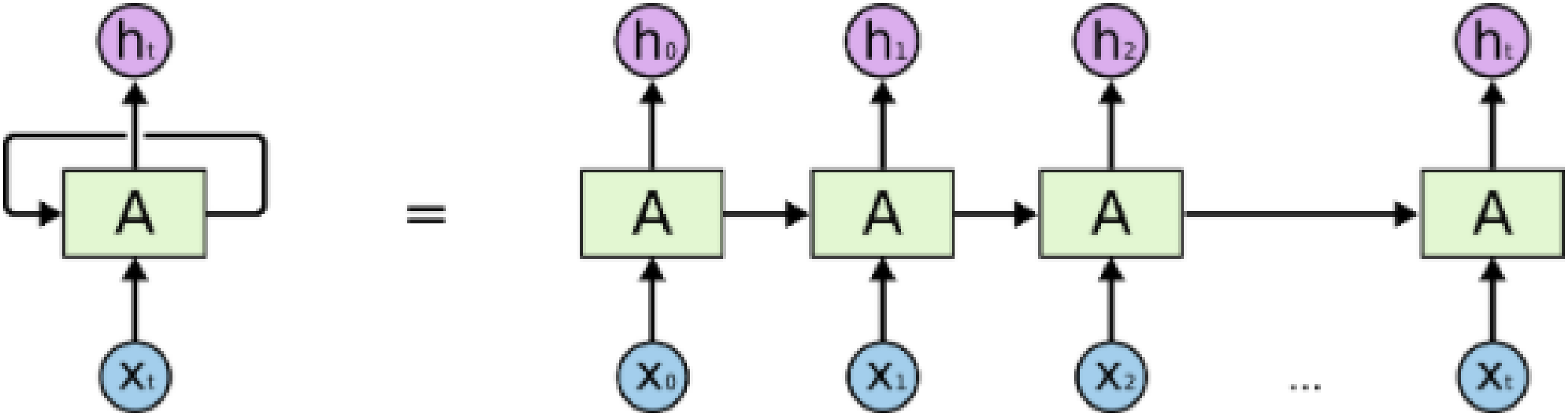
How to make the network build and maintain context internally, when trained one symbol at a time?

A B C … M N O P …



I

Unrolling the compute graph



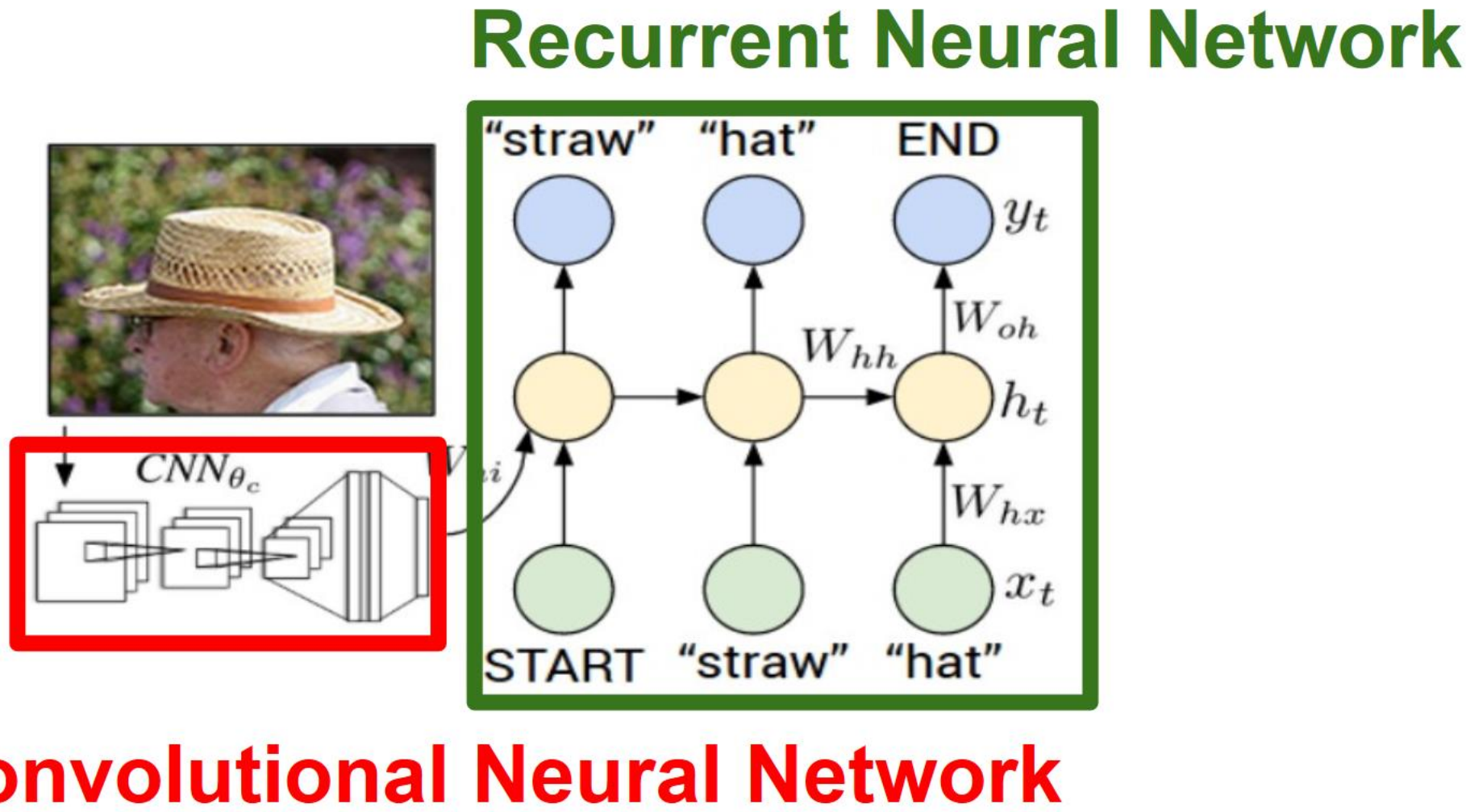
What can it do?

- http://karpathy.github.io/2015/05/21/rnn-effectiveness/
- Tensorflow version: https://github.com/sherjilozair/char-rnn-tensorflow/blob/master/model.py

Naturalism and decision for the majority of Arab countries' capitalide was grounded by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal in the [[Protestant Immineners]], which could be said to be directly in Cantonese Communication, which followed a ceremony and set inspired prison, training. The emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known in western [[Scotland]], near Italy to the conquest of India with the conflict. Copyright was the succession of independence in the slop of Syrian influence that was a famous German movement based on a more popular servicious, non-doctrinal and sexual power post. Many governments recognize the military housing of the [[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]], that is sympathetic to be to the [[Punjab Resolution]]
(PJS)[http://www.humah.yahoo.com/guardian.
cfm/7754800786d17551963s89.htm Official economics Adjoint for the Nazism, Montgomery was swear to advance to the resources for those Socialism's rule,
was starting to signing a major tripad of aid exile.]]

What can it do?

- https://cs.stanford.edu/people/karpathy/sfmltalk.pdf




a man standing in front of a store with a woman in the background
HOT-DOGS

RNN benefits

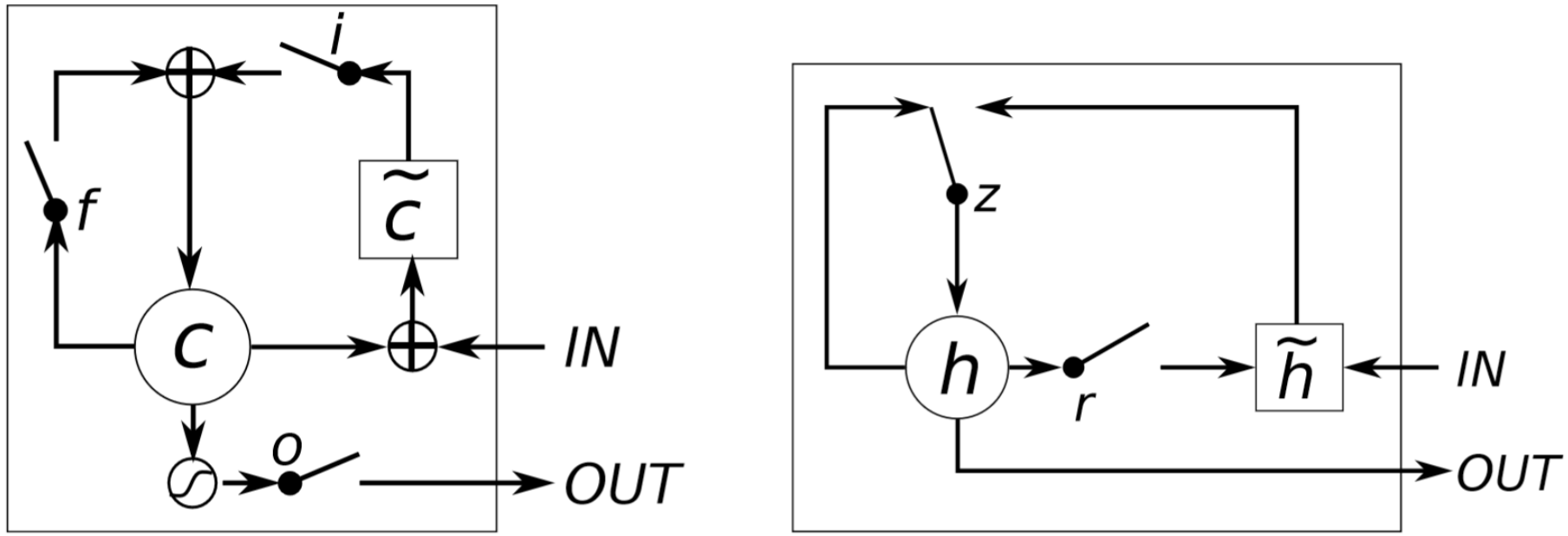
- Input sequence length not fixed
- More computationally efficient for long sequences in runtime

RNN Problems

- Training may be unstable. How to init internal state for each unrolled sequence?
- Longer unrolling improves results but uses more memory
- Very long memory does not emerge in practice

LSTM, GRU

- Solve the long-term memory problem using *gates*

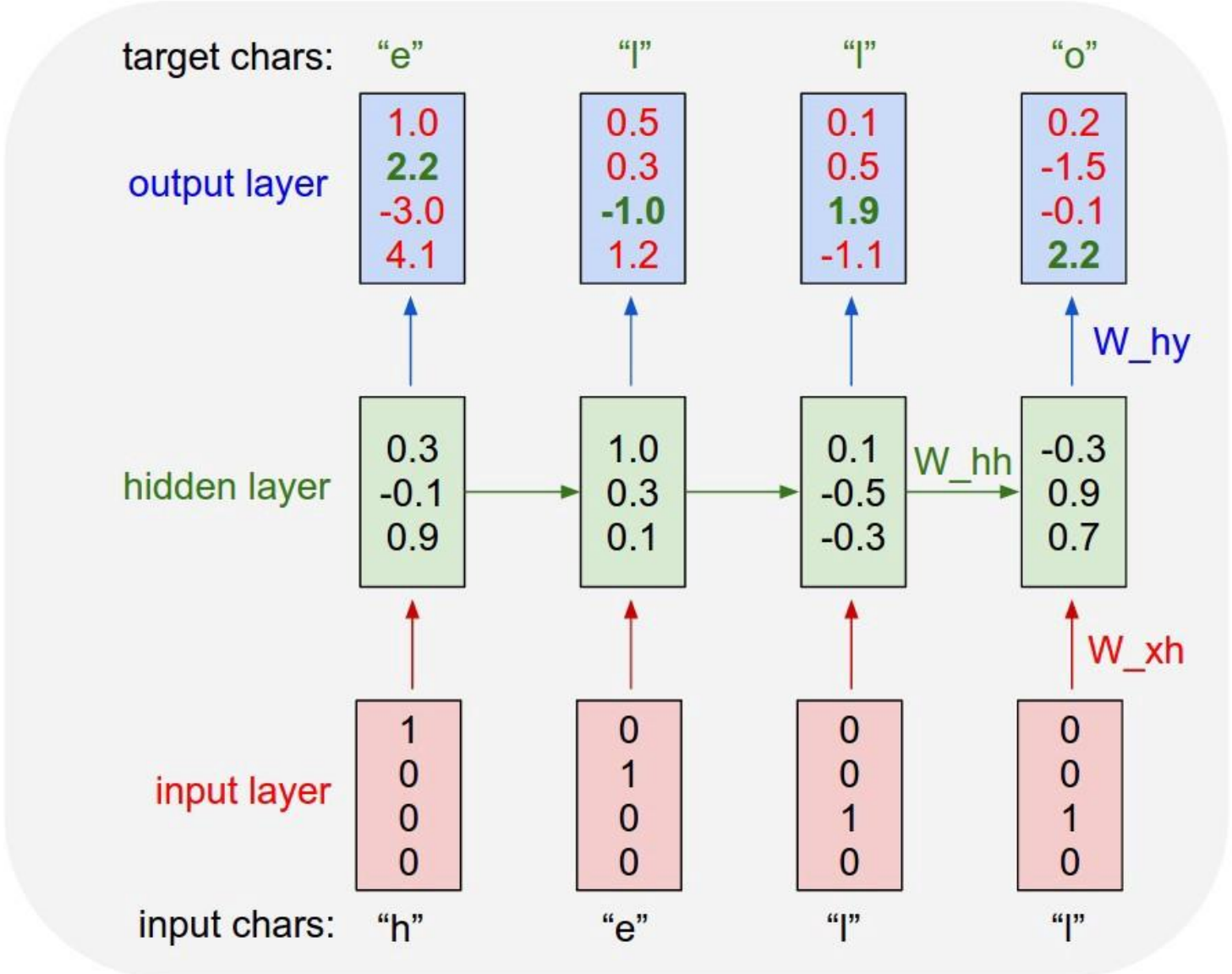


(a) Long Short-Term Memory

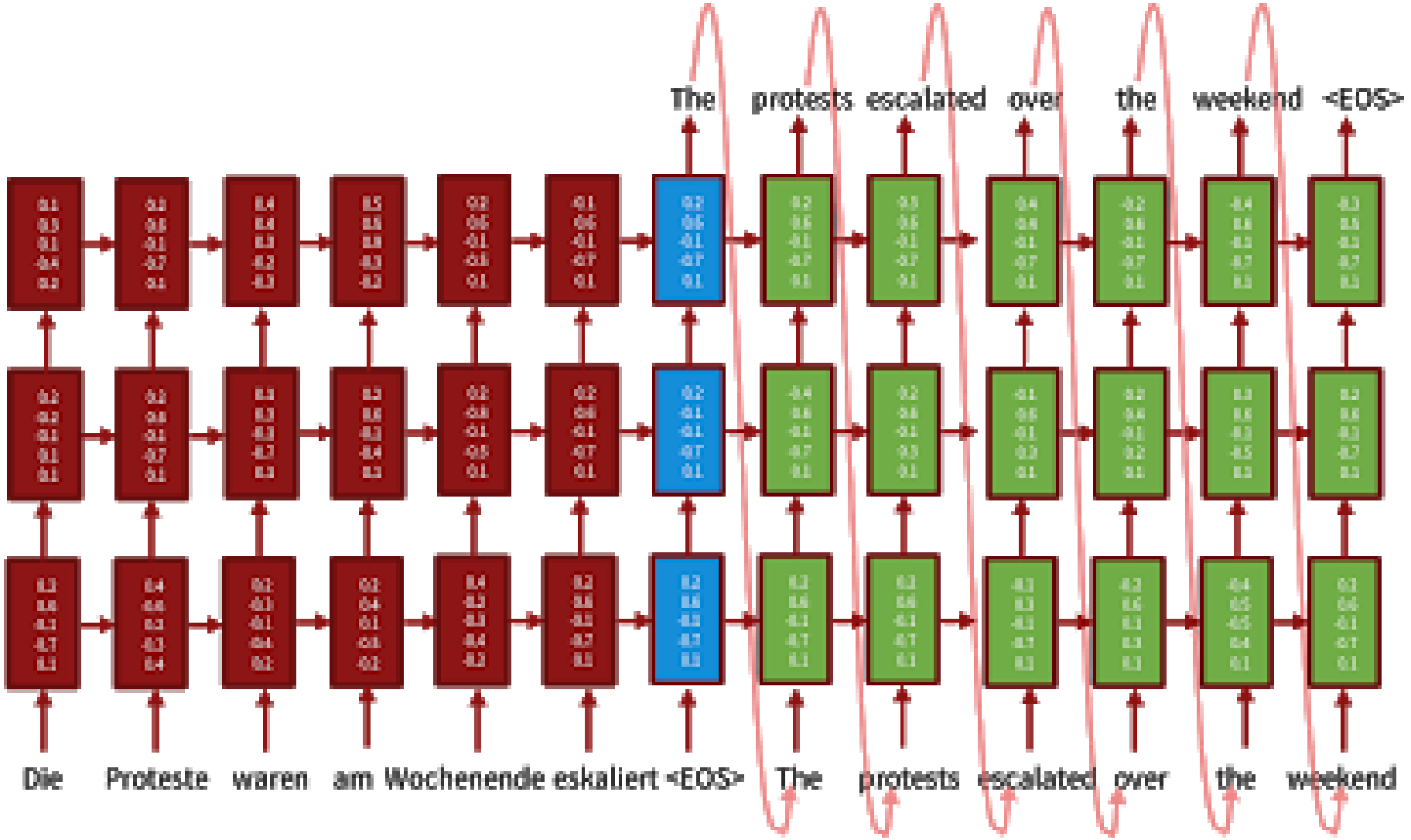
(b) Gated Recurrent Unit

Figure 1: Illustration of (a) LSTM and (b) gated recurrent units. (a) i, f and o are the input, forget and output gates, respectively. c and $\tilde{c}$ denote the memory cell and the new memory cell content. (b) r and z are the reset and update gates, and h and $\tilde{h}$ are the activation and the candidate activation.

More details: one-hot encoding



Neural machine translation



Advanced

- https://distill.pub/2016/augmented-rnns/ (Neural Turing machines etc.)

Case study: Relation networks

- Combining convolutional image processing and LSTM or GRU to allow answering natural language questions about images

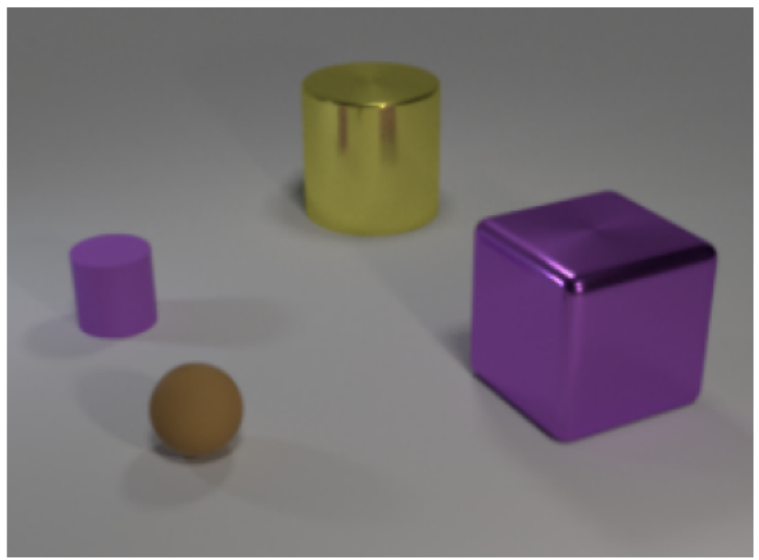
A simple neural network module for relational reasoning

Santoro, Raposo, Barrett, Malinowski, Pascanu, Battaglia, Lillicrap

Submitted to arxiv on 5 Jun 2017

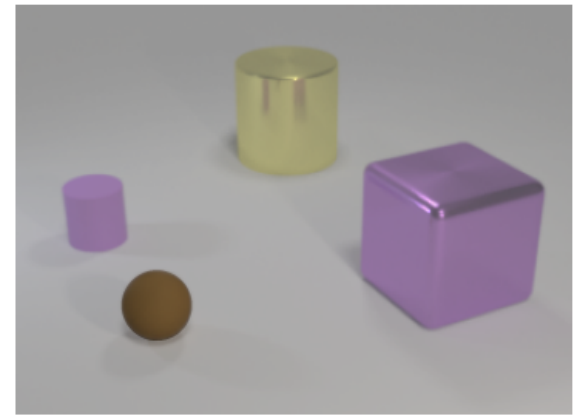
This summary by Perttu Hämäläinen

Original Image:



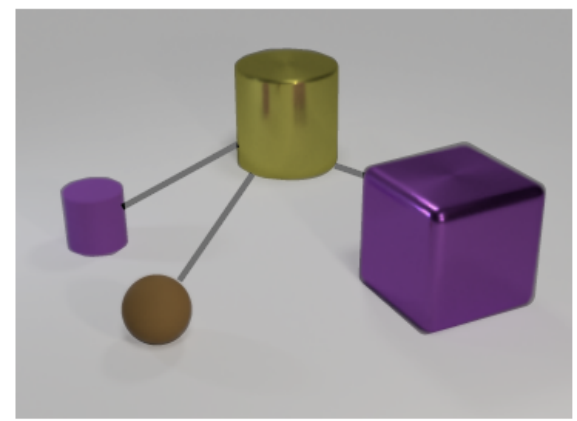
Non-relational question:

What is the size of the brown sphere?



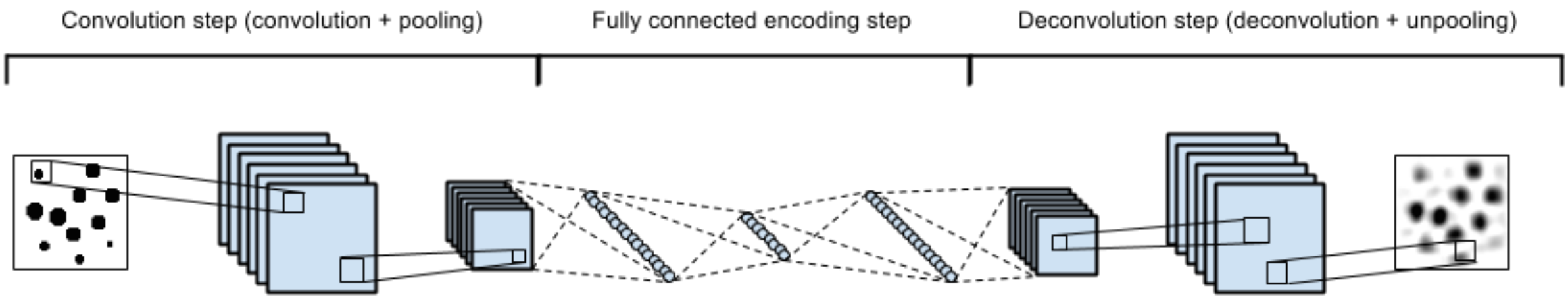
Relational question:

Are there any rubber things that have the same size as the yellow metallic cylinder?

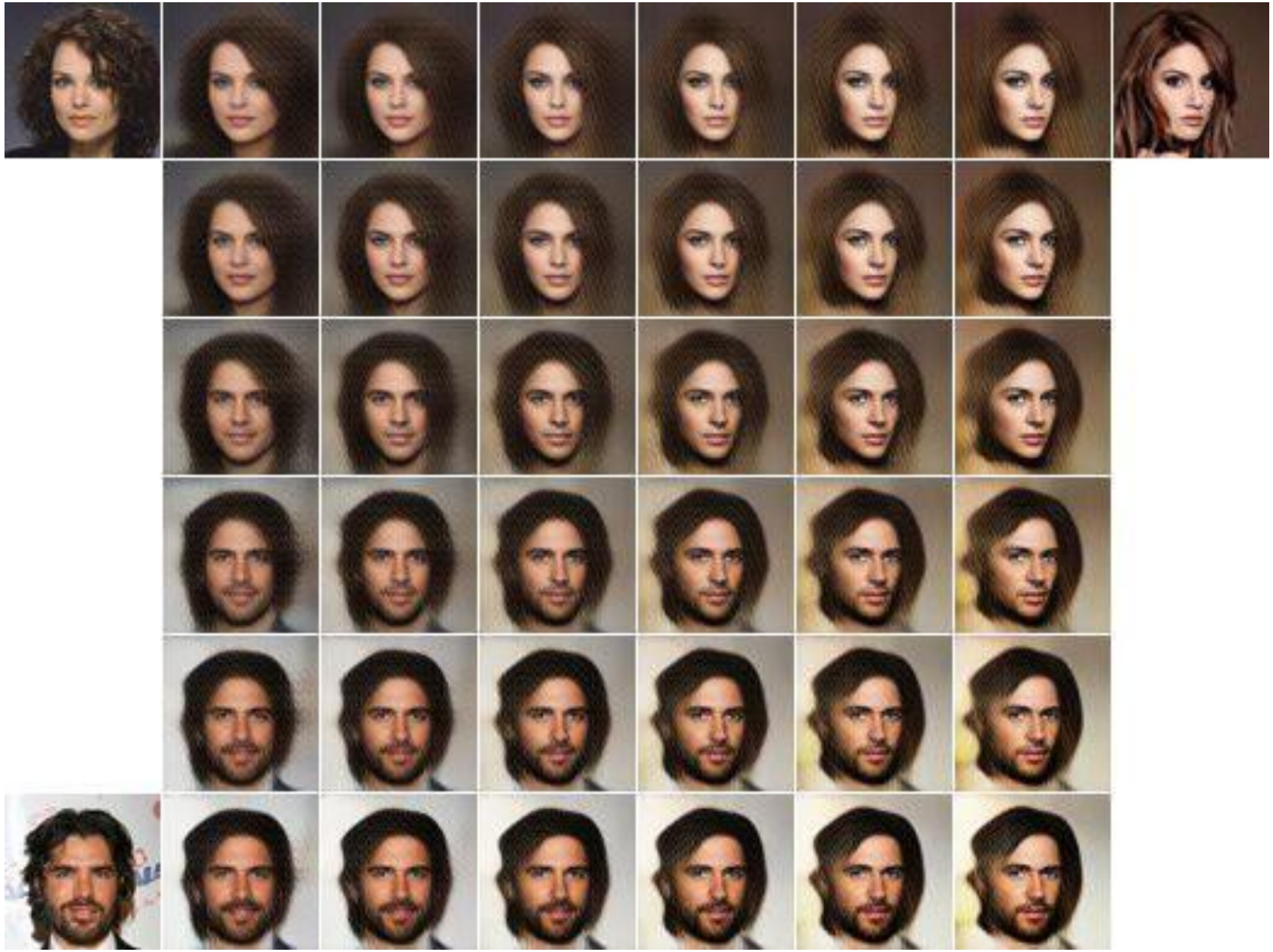


How to represent **relations**?

- In standard network architectures, unit output represents presence/absence/strength of a feature (line, corner, eye, face, feminity/masculinity, bald vs hair)
- Outputs of all units of a layer represent the input in some feature space (latent space)
- But how to go from features to relations, especially in complex inputs such as images?

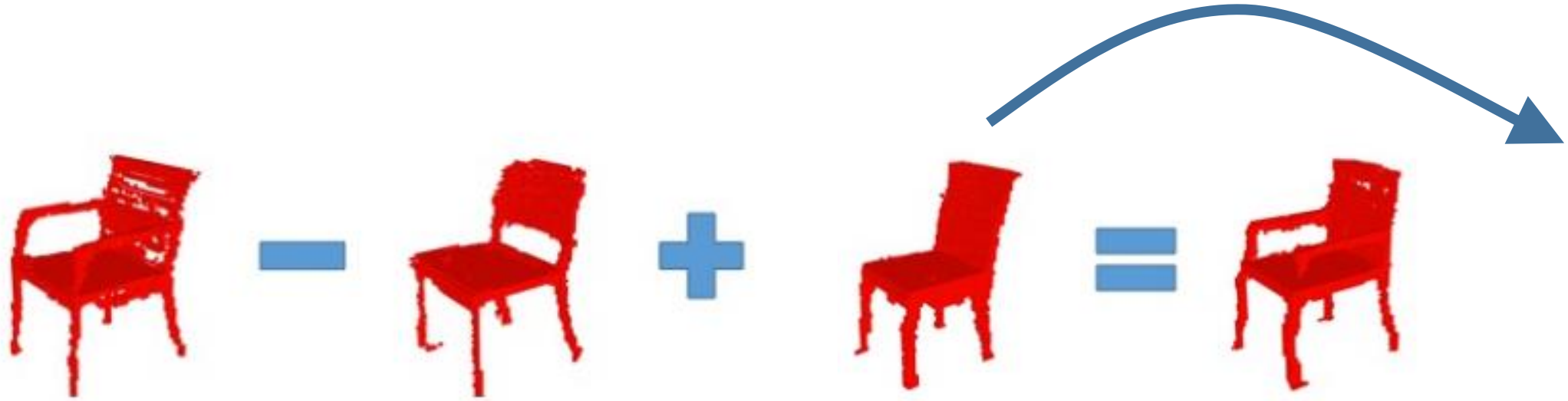


https://swarbrickjones.wordpress.com/2015/04/29/convolutional-autoencoders-in-pythontheanolasagne/

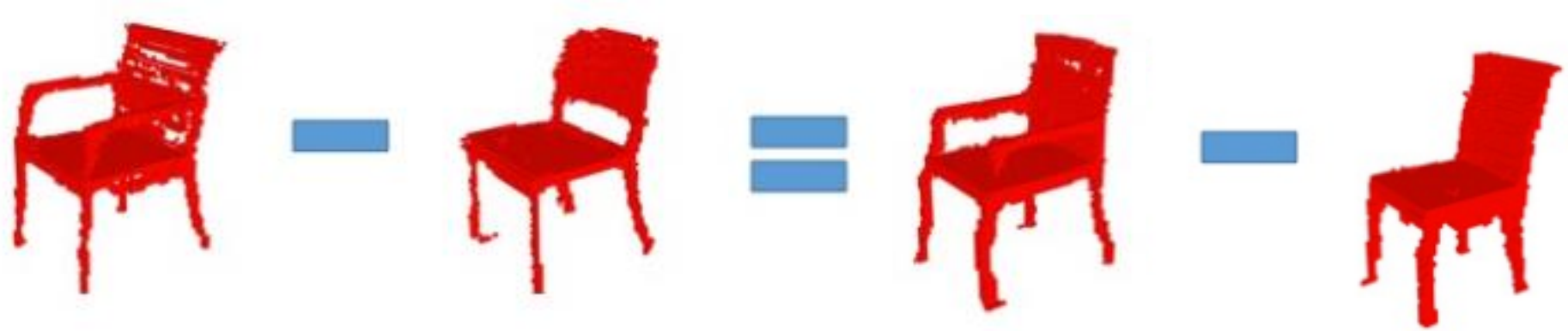


https://medium.com/@juliendespois/latent-space-visualization-deep-learning-bits-2-bd09a46920df

Wu et al., Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling, NIPS 2016

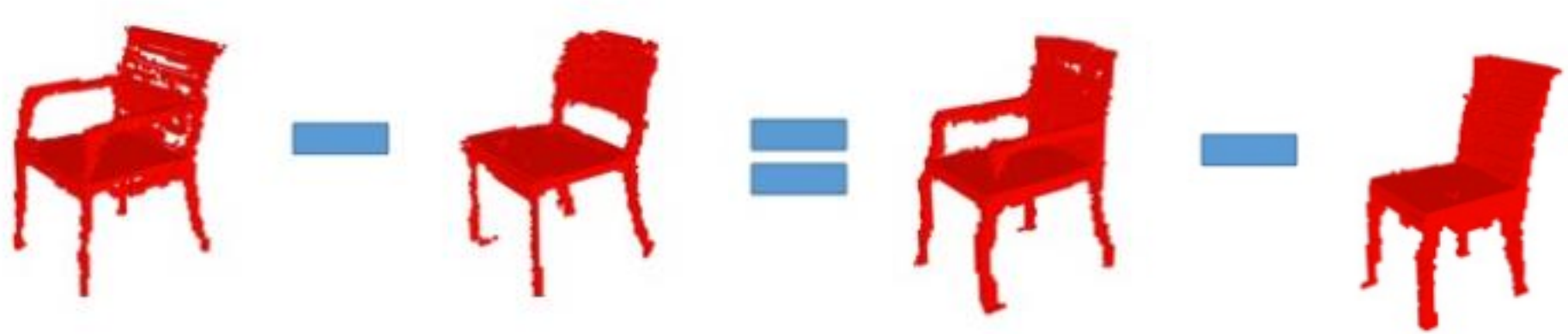


Wu et al., Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling, NIPS 2016



Wu et al., Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling, NIPS 2016

This is to that like this is to that



Wu et al., Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling, NIPS 2016

How to answer questions about relations?

How to answer questions about relations?

- In many problems, selecting the correct network architecture is key – the architecture imposes a strong prior on the learning
- Convnets: same features/structures can occur anywhere spatially, spatial connectivity is sparse
- This paper: what kind of architecture allows the units to represent relations efficiently?
- Answer: Relational Network (RN)

$$\mathrm{RN}(O) = f_\phi \left(\sum_{i,j} g_\theta(o_i, o_j) \right)$$

Objects

$$\mathrm{RN}(O) = f_\phi \left(\sum_{i,j} g_\theta(o_i, o_j) \right)$$

MLPs Objects

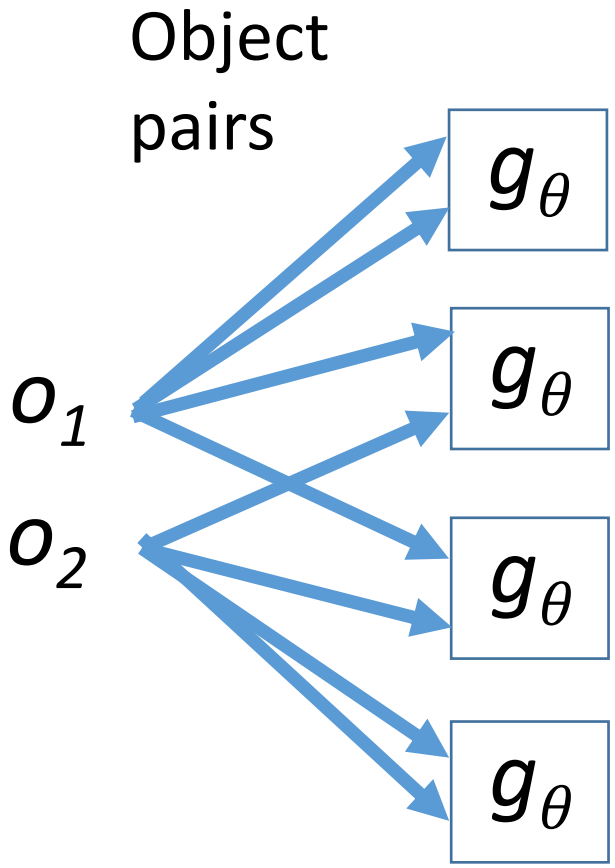
$$\mathrm{RN}(O) = f_\phi \left(\sum_{i,j} g_\theta(o_i, o_j) \right)$$

$$\mathrm{RN}(O) = f_\phi \left(\sum_{i,j} g_\theta(o_i, o_j) \right)$$

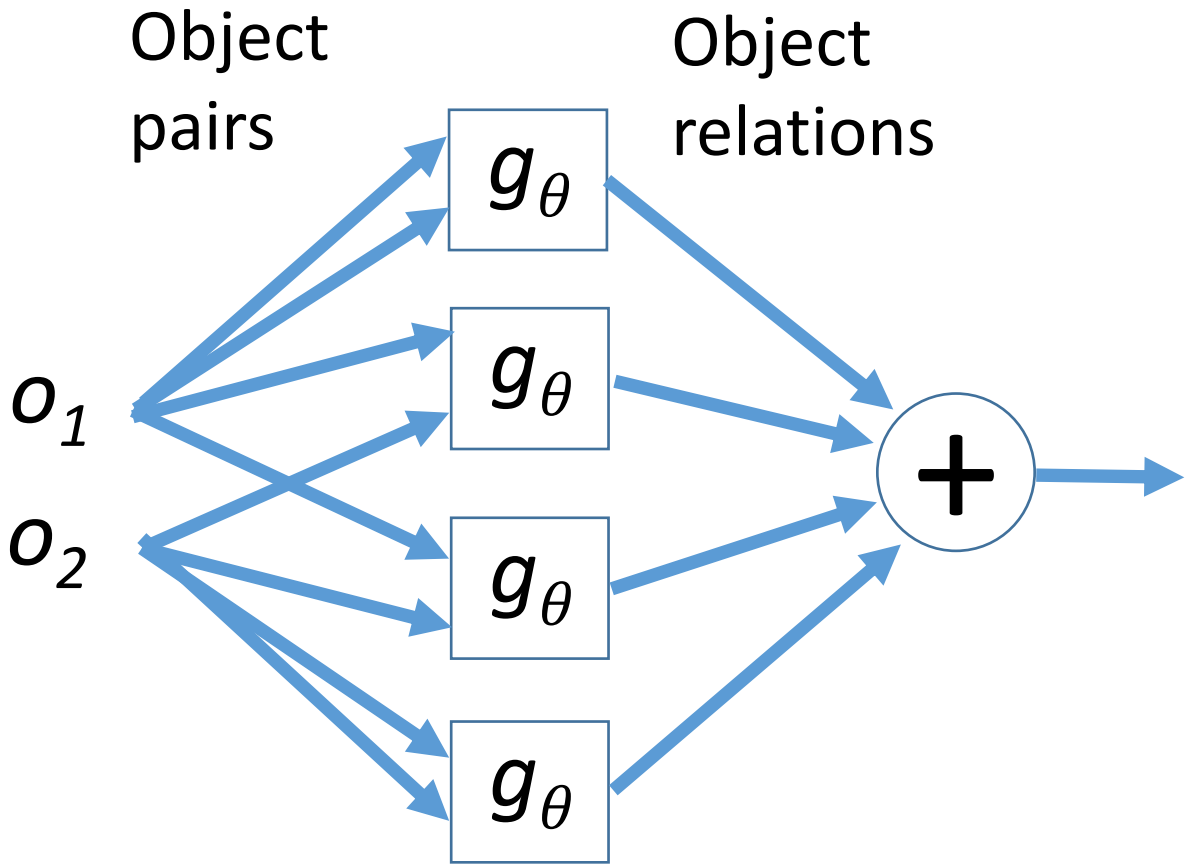
o_1

o_2

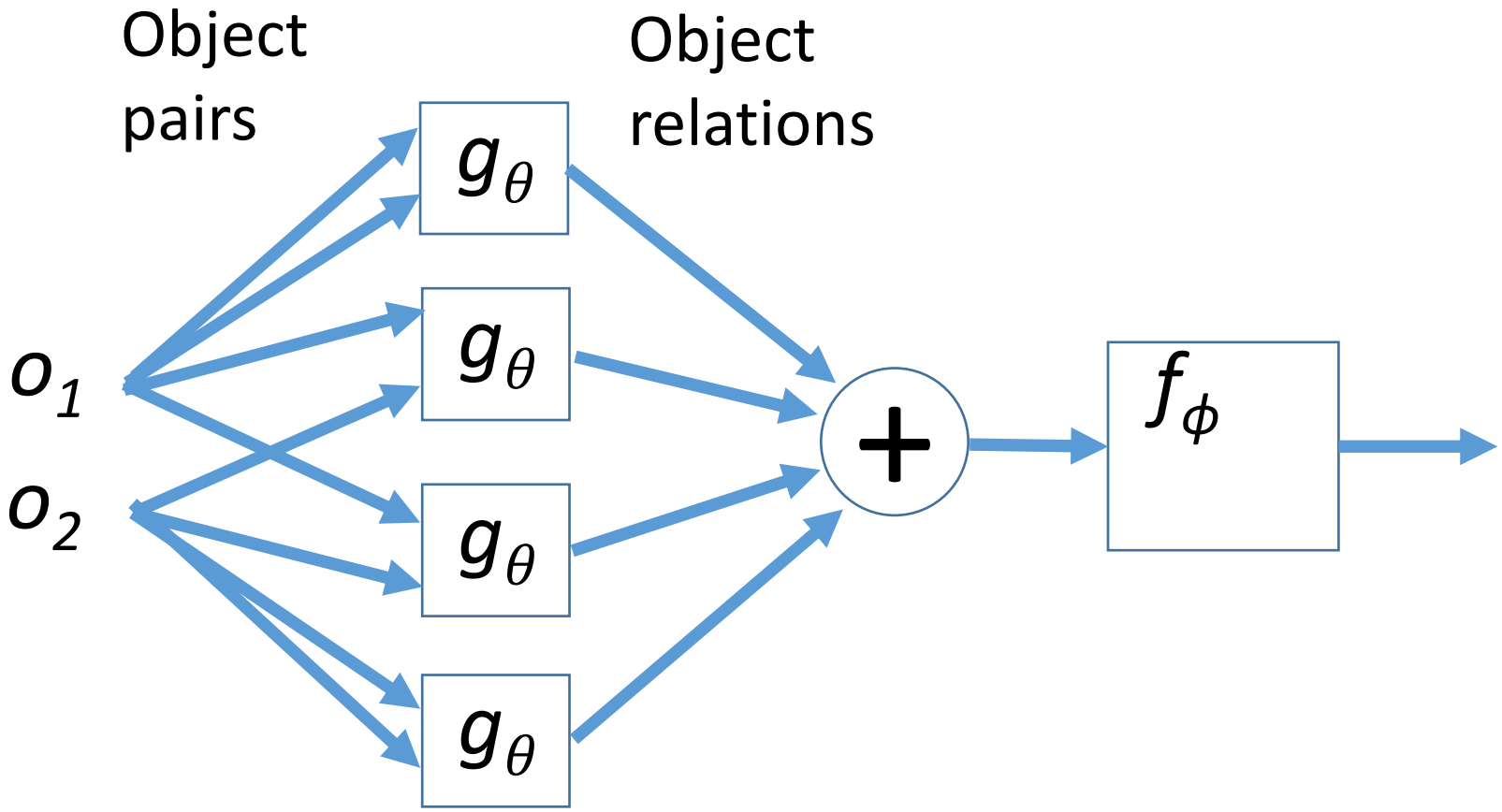
$$\mathrm{RN}(O) = f_\phi \left(\sum_{i,j} g_\theta(o_i, o_j) \right)$$



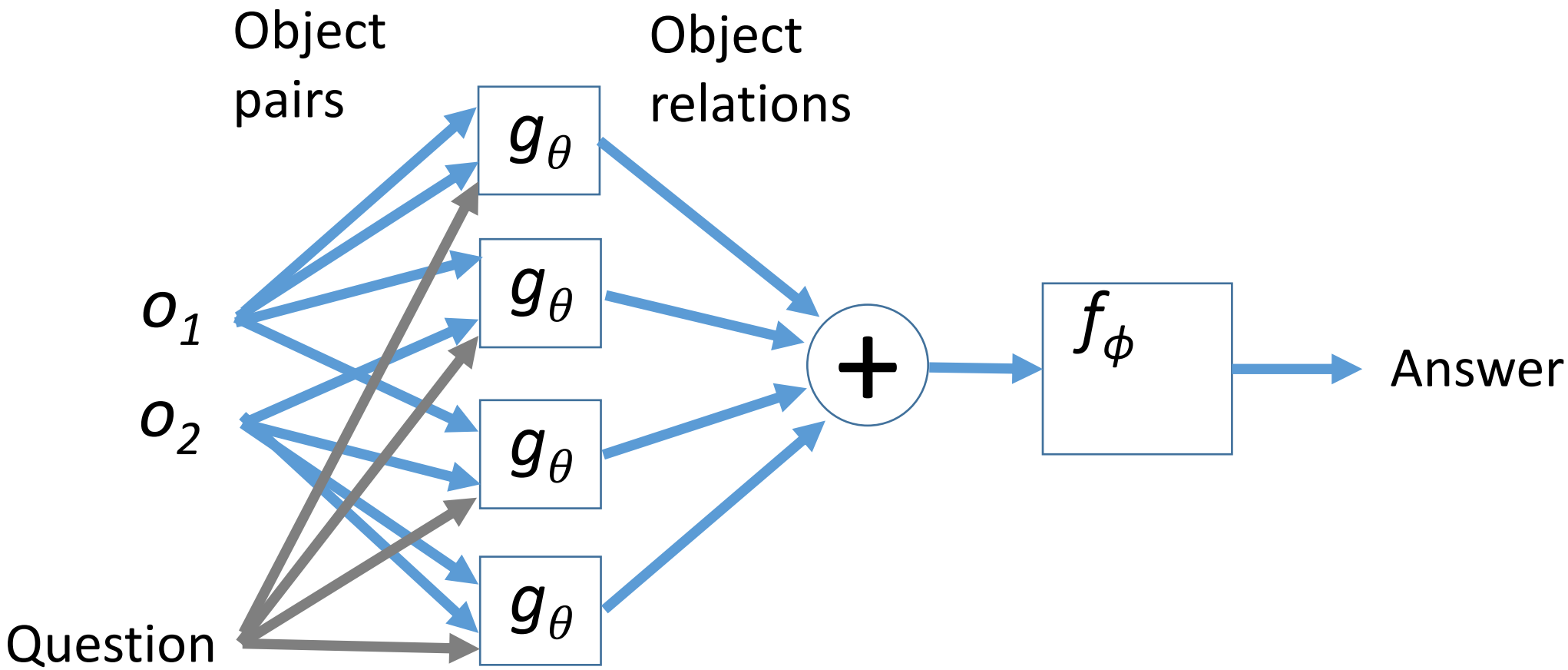
$$\mathrm{RN}(O) = f_\phi \left(\sum_{i,j} g_\theta(o_i, o_j) \right)$$

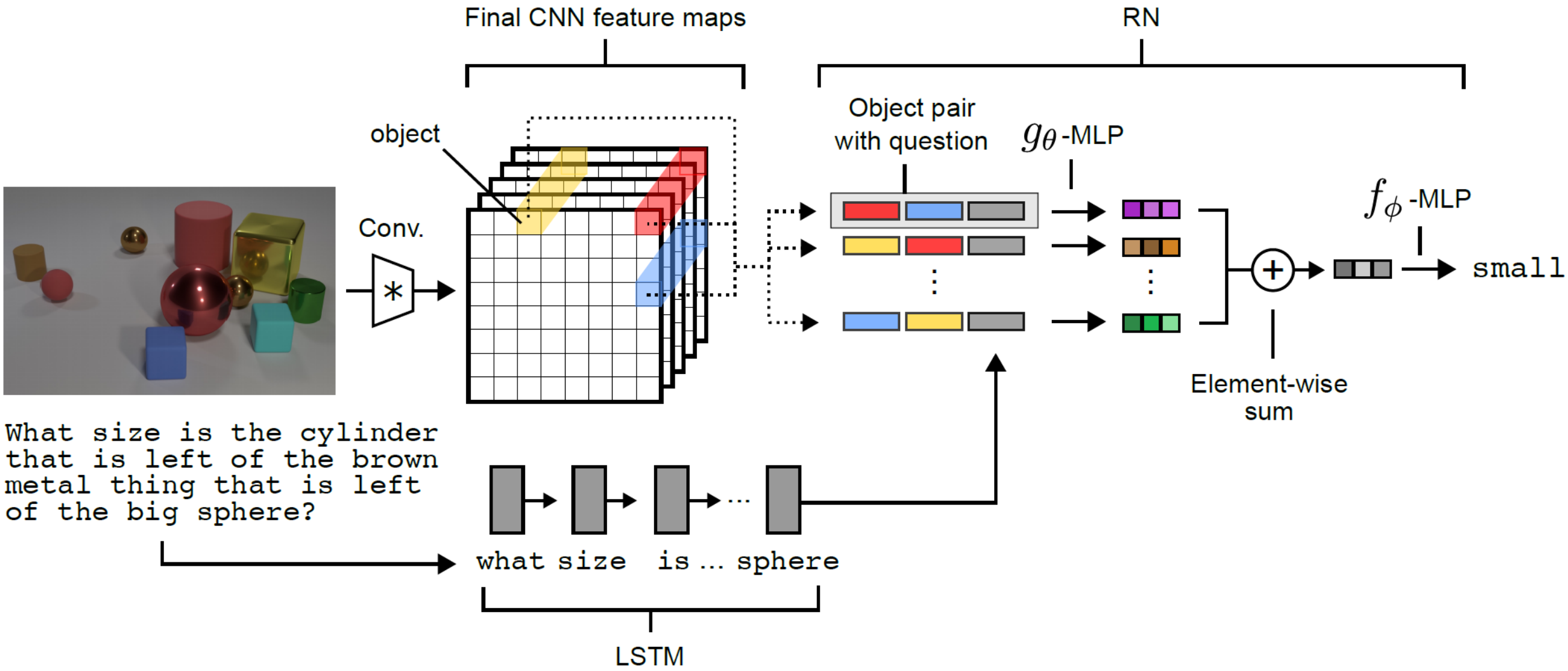


$$\mathrm{RN}(O) = f_\phi \left(\sum_{i,j} g_\theta(o_i, o_j) \right)$$

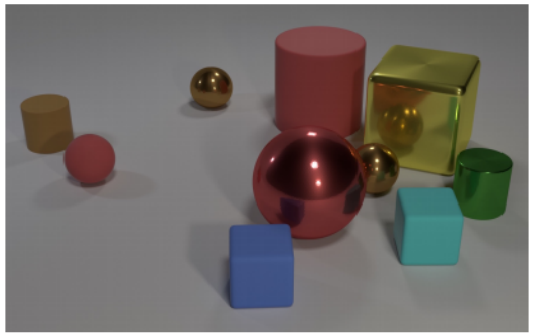


$$\mathrm{RN}(O) = f_\phi \left(\sum_{i,j} g_\theta(o_i, o_j) \right)$$




Final CNN feature maps
RN
object
Conv.
Object pair
with question
g_θ -MLP
f_ϕ -MLP
small
Element-wise
sum
What size is the cylinder
that is left of the brown
metal thing that is left
of the big sphere?
what size is ... sphere
LSTM

Conv. & pooling not continued down to 1x1 pixel output



Conv.

Final CNN feature maps

object

RN

Object pair with question

g_θ-MLP

f_ϕ-MLP

small

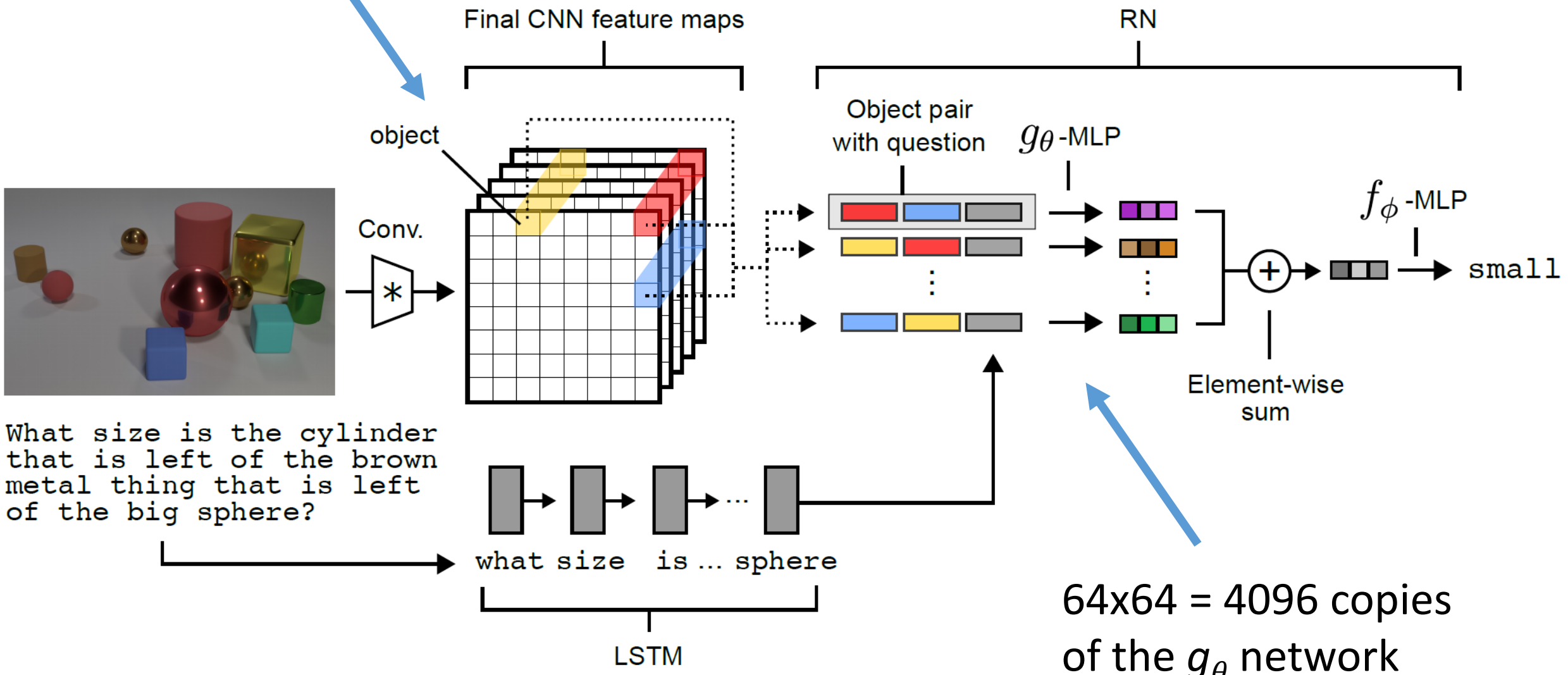
Element-wise sum

What size is the cylinder that is left of the brown metal thing that is left of the big sphere?

what size is ... sphere

LSTM

Conv. & pooling not continued down to 1x1 pixel output



64x64 = 4096 copies
of the g_θ network

Benchmarks

- CLEVR dataset with 14 tasks (image-based question answering)
- Sort-of-CLEVR (separated relational and non-relational questions)
- bAbI (facebook research text understanding and reasoning)
- Dynamic physical systems (infer invisible springs between some moving objects)
- Superhuman performance in over 90% of the benchmarks

Model	**Overall**	Count	Exist	Compare Numbers	Query Attribute	Compare Attribute
Human	92.6	86.7	96.6	86.5	95.0	96.0
Q-type baseline	41.8	34.6	50.2	51.0	36.0	51.3
LSTM	46.8	41.7	61.1	69.8	36.8	51.8
CNN+LSTM	52.3	43.7	65.2	67.1	49.3	53.0
CNN+LSTM+SA	68.5	52.2	71.1	73.5	85.3	52.3
CNN+LSTM+SA*	76.6	64.4	82.7	77.4	82.6	75.4
CNN+LSTM+RN	**95.5**	**90.1**	**97.8**	**93.6**	**97.9**	**97.1**

* Our implementation, with optimized hyperparameters and trained fully end-to-end.

Table 1: **Results on CLEVR from pixels.**

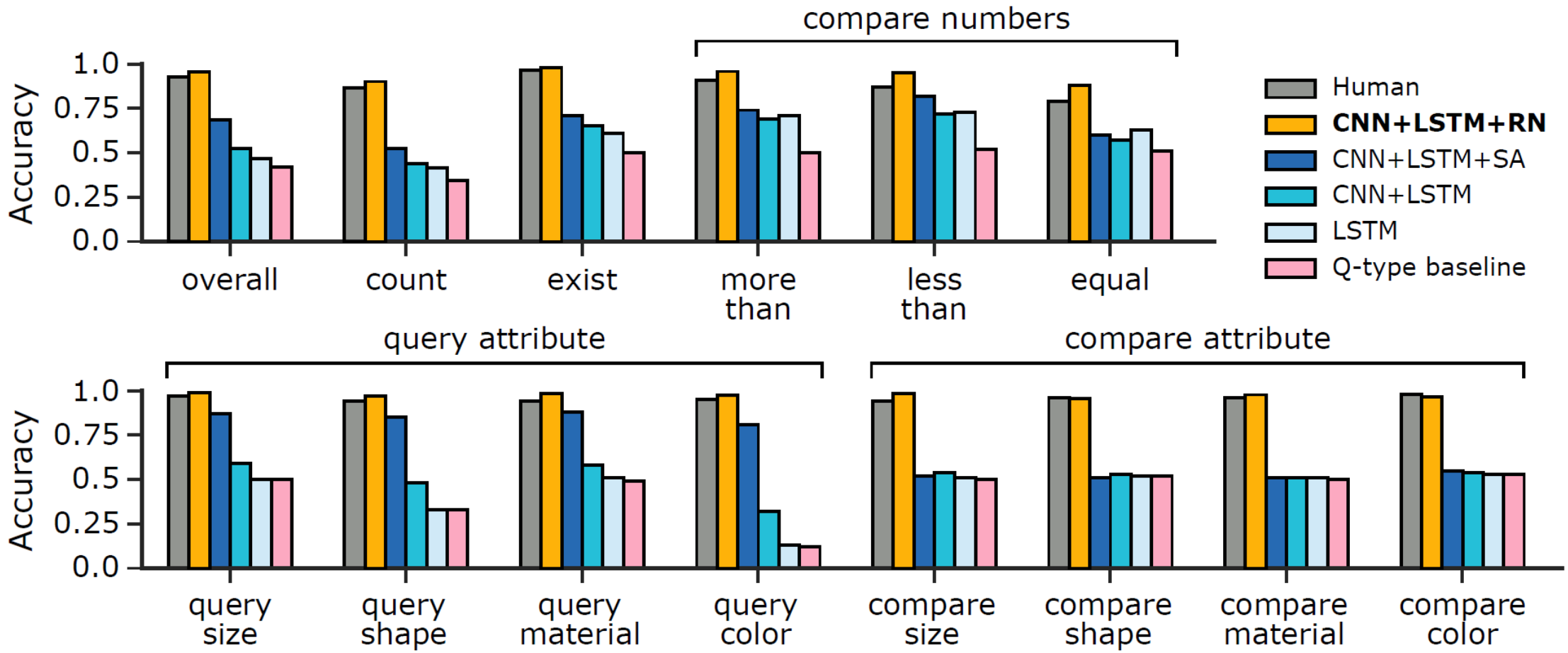
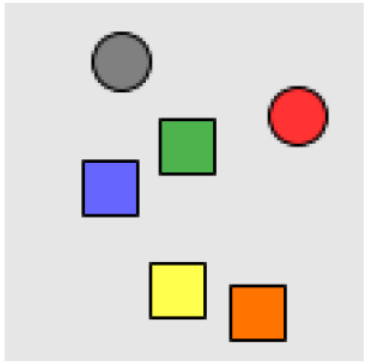


Figure 3: **Results on CLEVR from pixels.**



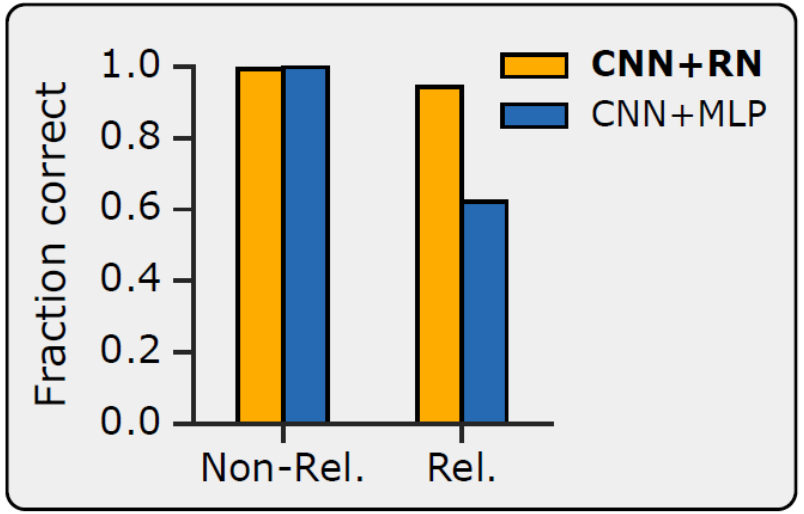
Image

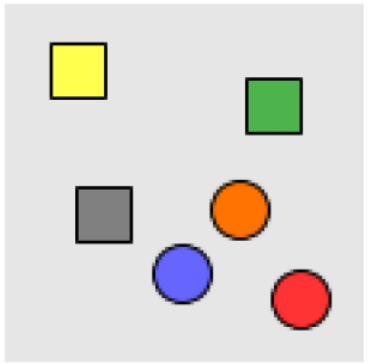
Non-relational question

Q: What is the shape of the gray object?
A: circle

Relational question

Q: What is the shape of the object
that is furthest from the gray object?
A: square



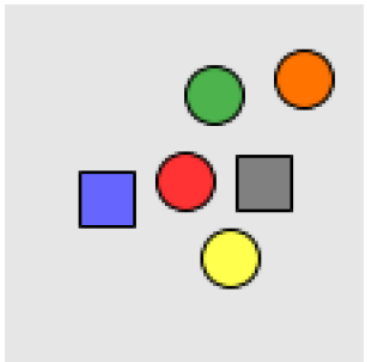


Non-relational question

Q: Is the green object on the left or on the right?
A: right

Relational question

Q: How many objects have the shape of the orange object?
A: 3

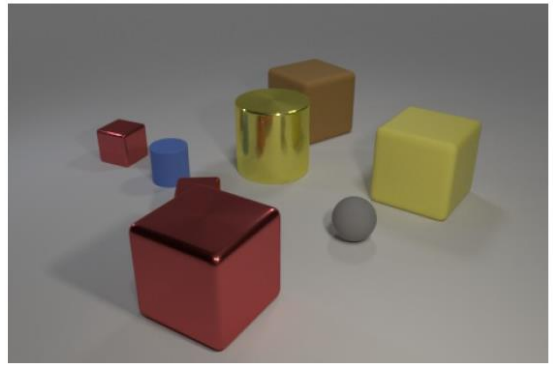


Non-relational question

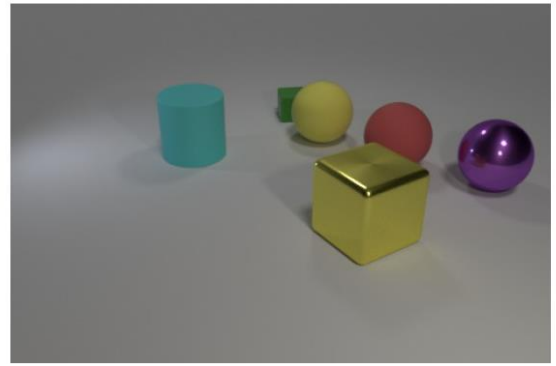
Q: Is the yellow object on the top or on the bottom?
A: bottom

Relational question

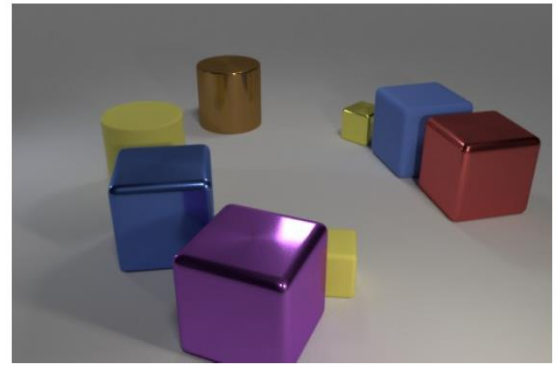
Q: What is the color of the object that is closest to the blue object?
A: red



What shape is the small object that is in front of the yellow matte thing and behind the gray sphere?

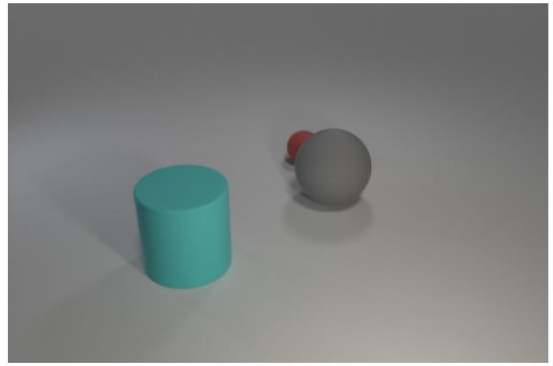


What number of things are either tiny green rubber objects or shiny things that are behind the big metal block?

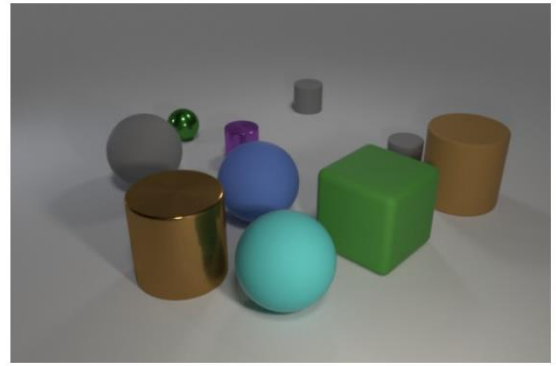


What number of objects are blocks that are in front of the large red cube or green balls?

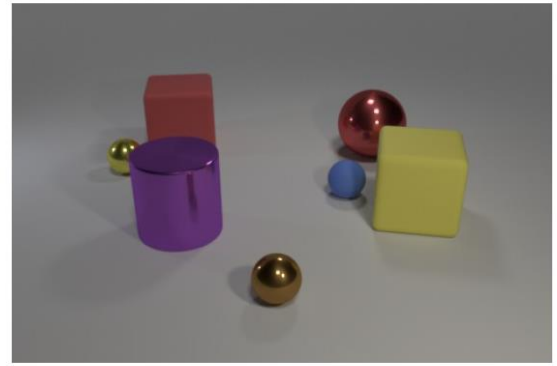
RN:	cylinder	1	2
GT:	cube	2	3



Is the shape of the small red object the same as the large matte object that is right of the small rubber ball?



How many gray objects are in front of the tiny green shiny ball and right of the big blue matte thing?



What number of objects are big red matte cubes or things on the right side of the large red matte block?

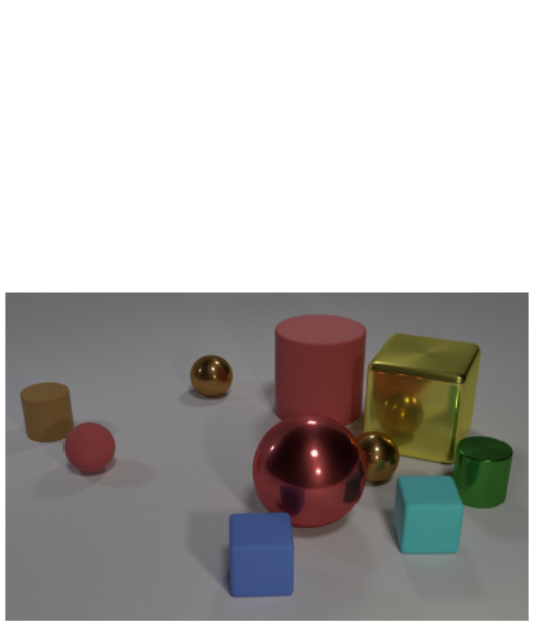
RN:	no	0	5
GT:	yes	1	6

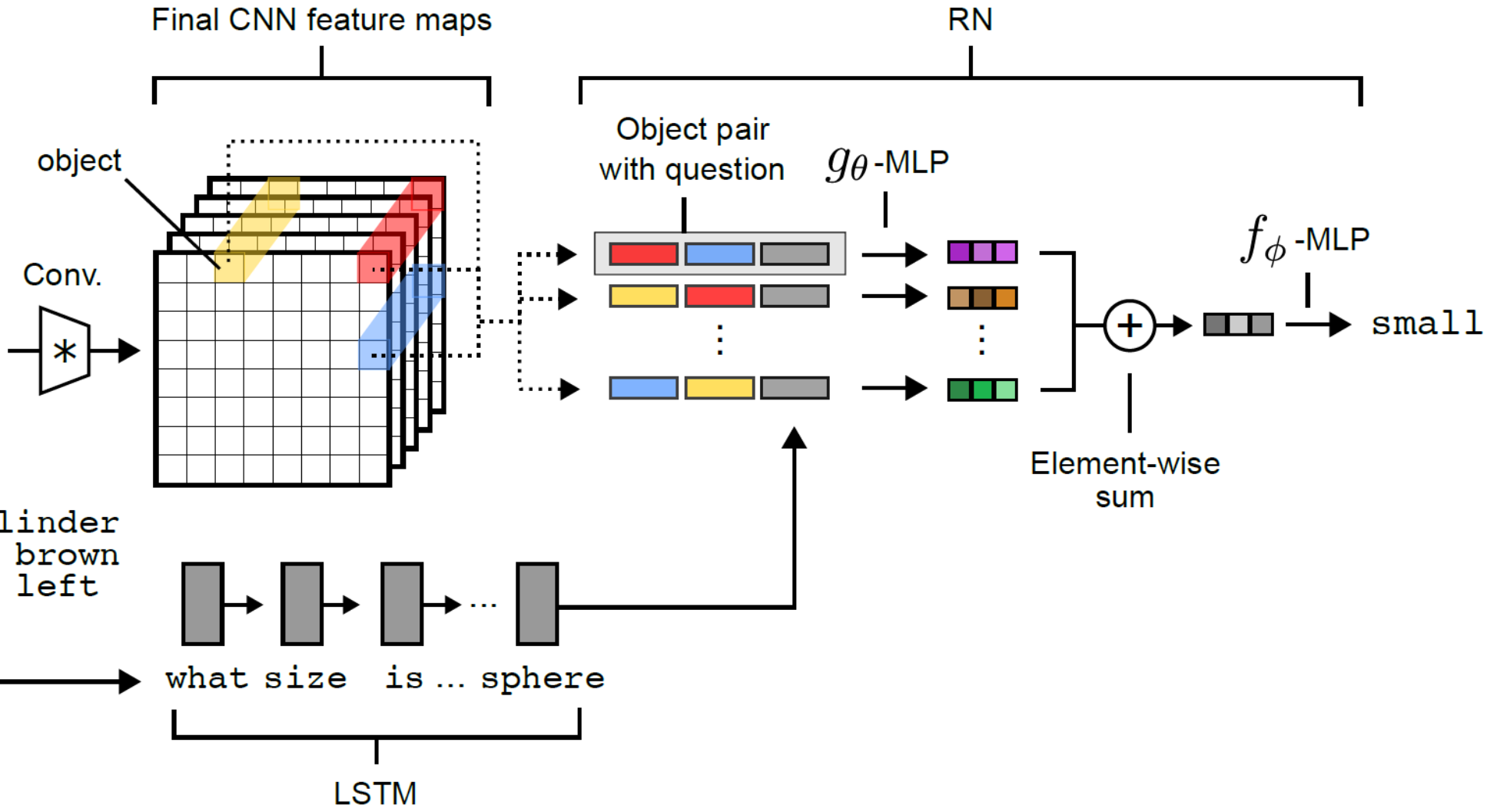
Future work, applications

- Rich scene understanding in RL agents
- Social network modeling
- Abstract problem solving
- ” Relation Networks are a simple and powerful approach for learning to perform rich, structured reasoning in complex, real-world domains.”

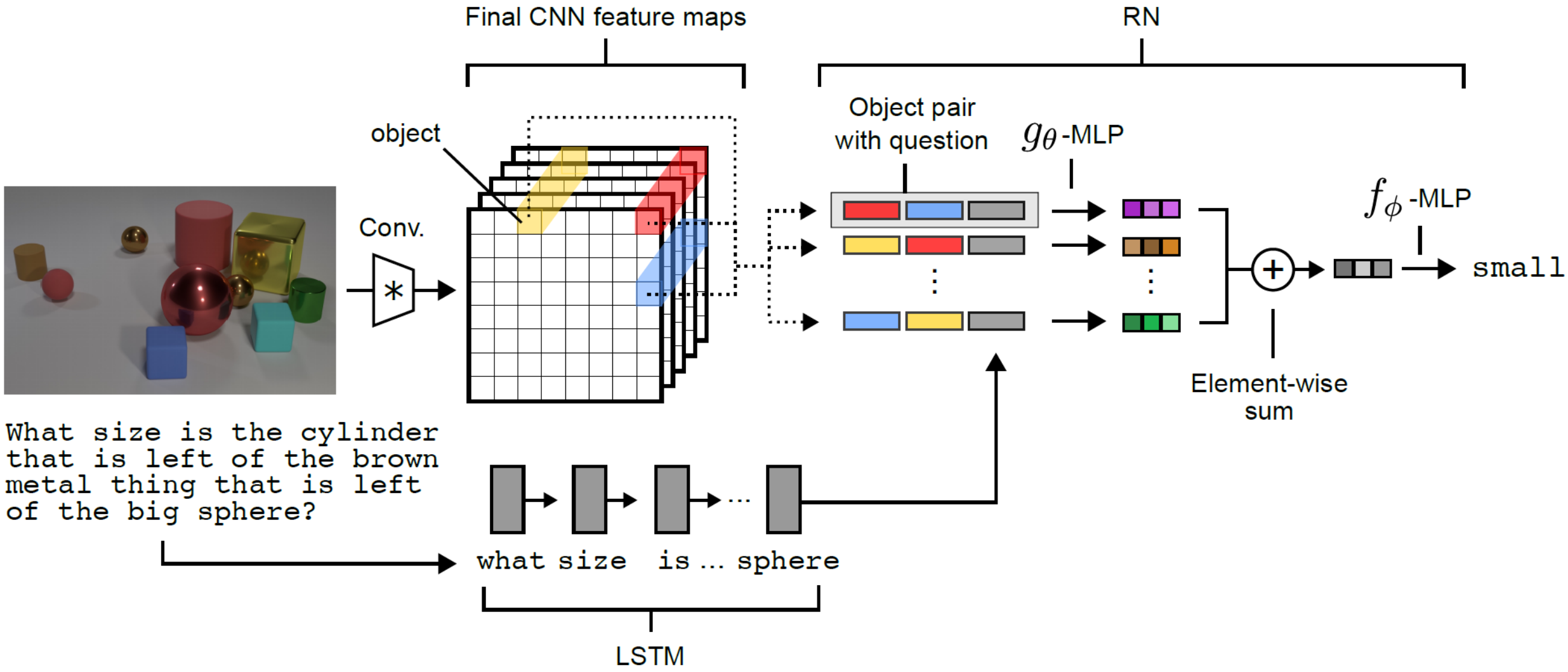
A more critical view

- The architecture exploits the weakness of the CLEVR benchmarks?
- Need a benchmark with large & complex objects, e.g., human crowds?



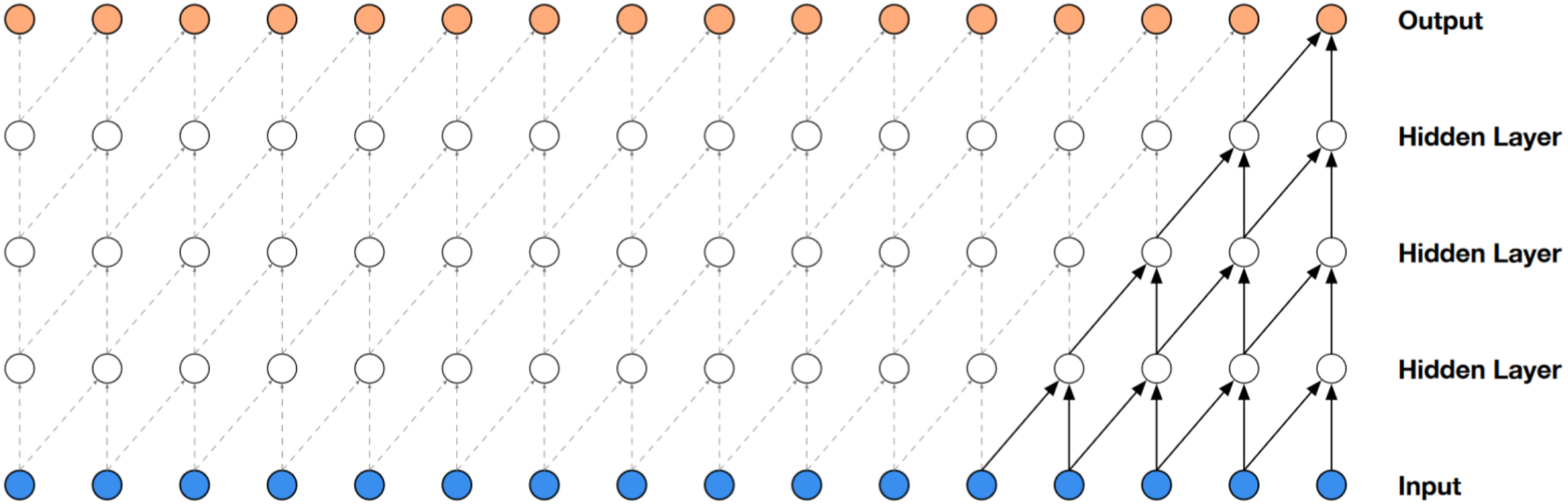


Questions, discussion?



Alternative to recurrent networks: Dilated causal convolution

- Popularized by Deepmind Wavenet: https://deepmind.com/blog/wavenet-generative-model-raw-audio/



Audio synthesis

- Tacotron & Tacotron 2 build on WaveNet, Tacotron 2 used in Google's voice assistant
- Online tool for synthesizing speech :

https://acapela-box.com/AcaBox/index.php

An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling

Shaojie Bai [1] **J. Zico Kolter** [2] **Vladlen Koltun** [3]

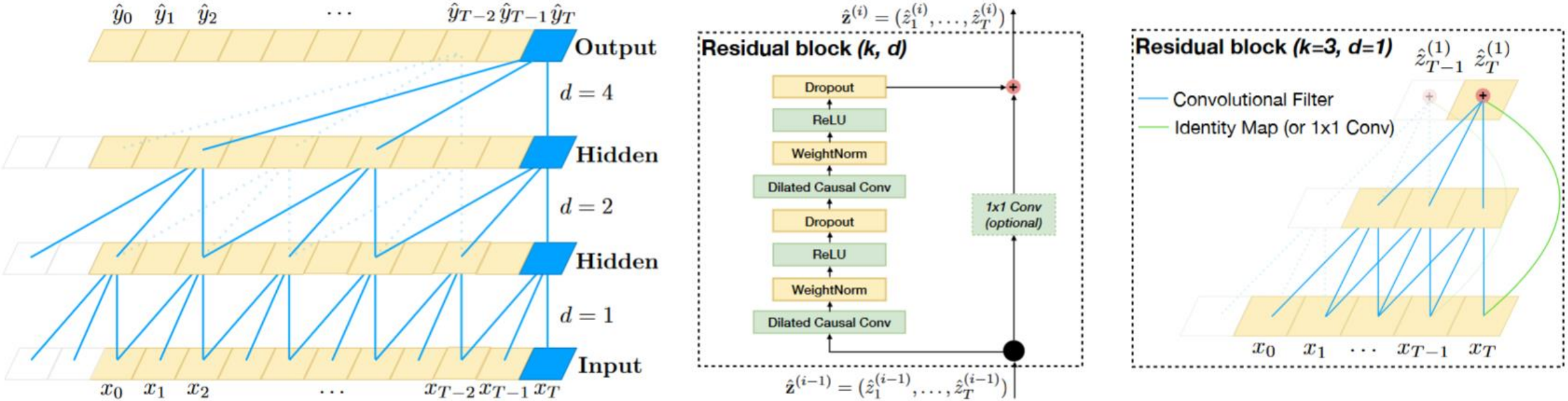
Abstract

For most deep learning practitioners, sequence modeling is synonymous with recurrent networks. Yet recent results indicate that convolutional architectures can outperform recurrent networks on tasks such as audio synthesis and machine translation. Given a new sequence modeling task or dataset, which architecture should one use? We conduct a systematic evaluation of generic convolutional and recurrent architectures for sequence modeling. The models are evaluated across a broad range of standard tasks that are commonly used to benchmark recurrent networks. Our results indicate that a simple convolutional architecture outperforms canonical recurrent networks such as LSTMs across a diverse range of tasks and datasets, while demonstrating longer effective memory. We conclude that the common association between sequence modeling and recurrent networks should be reconsidered, and convolutional networks should be regarded as a natural starting point for sequence modeling tasks.

chine translation (van den Oord et al., 2016; Kalchbrenner et al., 2016; Dauphin et al., 2017; Gehring et al., 2017a;b). This raises the question of whether these successes of convolutional sequence modeling are confined to specific application domains or whether a broader reconsideration of the association between sequence processing and recurrent networks is in order.

We address this question by conducting a systematic empirical evaluation of convolutional and recurrent architectures on a broad range of sequence modeling tasks. We specifically target a comprehensive set of tasks that have been repeatedly used to compare the effectiveness of different recurrent network architectures. These tasks include polyphonic music modeling, word- and character-level language modeling, as well as synthetic stress tests that had been deliberately designed and frequently used to benchmark RNNs. Our evaluation is thus set up to compare convolutional and recurrent approaches to sequence modeling on the recurrent networks' "home turf".

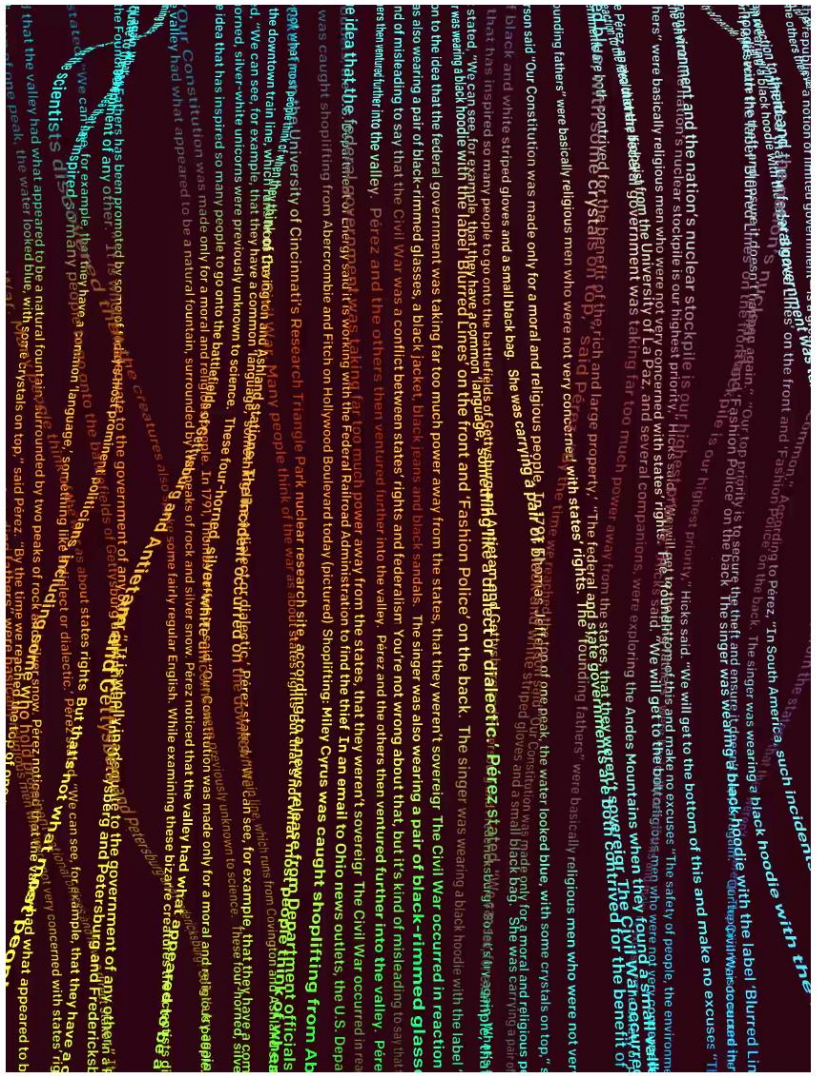
To represent convolutional networks, we describe a generic temporal convolutional network (TCN) architecture that is applied across all tasks. This architecture is informed by



(a) (b) (c)

Figure 1. Architectural elements in a TCN. (a) A dilated causal convolution with dilation factors $d = 1, 2, 4$ and filter size $k = 3$. The receptive field is able to cover all values from the input sequence. (b) TCN residual block. An 1x1 convolution is added when residual input and output have different dimensions. (c) An example of residual connection in a TCN. The blue lines are filters in the residual function, and the green lines are identity mappings.

Case study 2: OpenAI GPT-2



Better Language Models and Their Implications

We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

FEBRUARY 14, 2019
21 MINUTE READ

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Simple but powerful

- The network is trained to predict text based on preceding text
- Can be primed with a small segment of text
- Applications: text generation, of course, but also natural language question answering
- Full model not available because of enormous misuse potential for fake news generation etc.
- A ”small” 107M parameter pre-trained network available
- Results: https://openai.com/blog/better-language-models/

Key success factors

- Transformer architecture (Vaswani et al. 2017)
- Clever dataset curation: Use Reddit karma as a proxy metric for relevance.
- The network was trained on all Reddit outgoing links with over 2 karma
- Again, collecting the right data is often more important than technical innovations

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

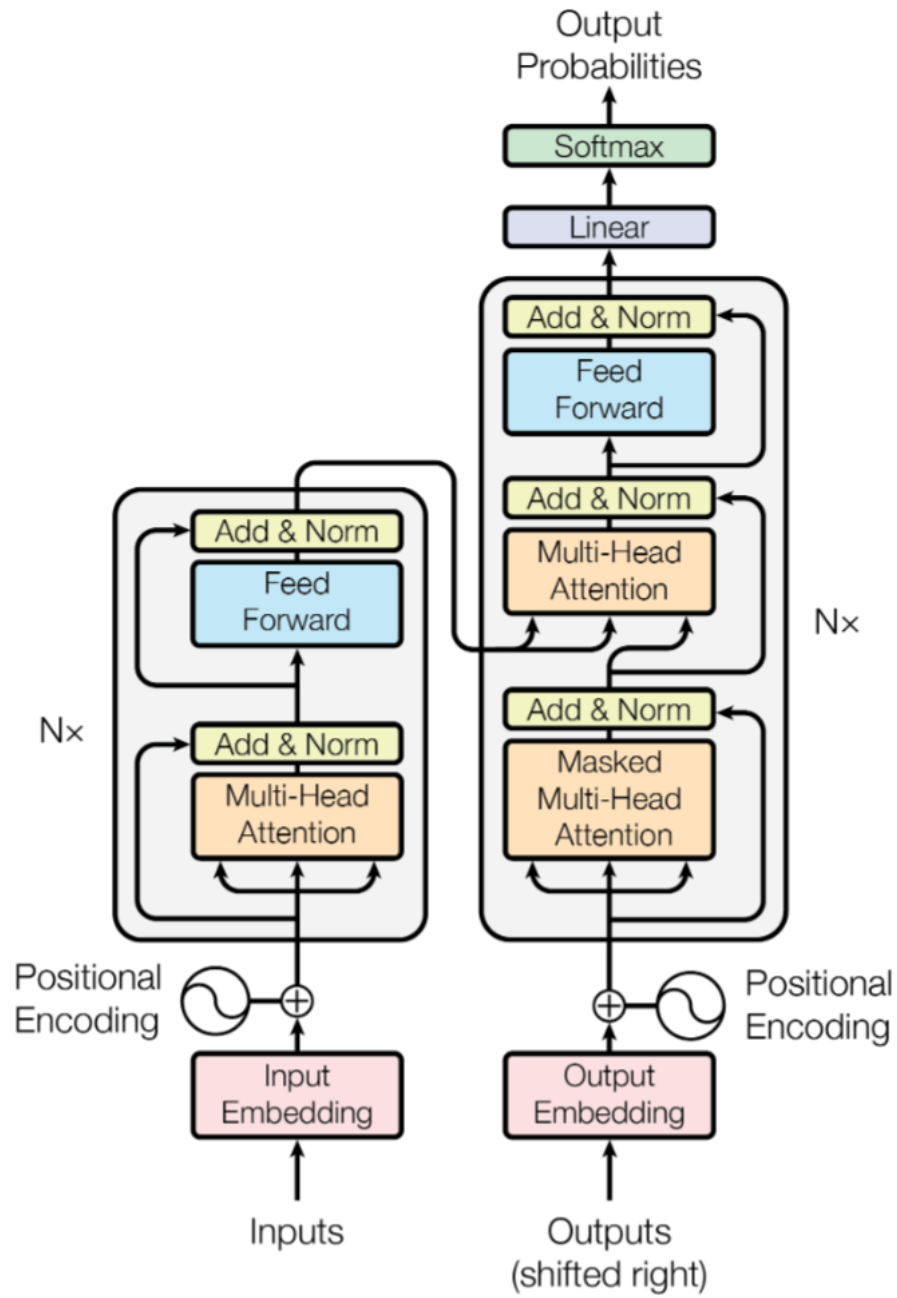


Figure 1: The Transformer - model architecture.

Summary

- Both CNNs and RNNs can be used for sequence modeling
- LSTM and GRU networks can have long memory with small computing cost
- Convolutional networks appear to perform better, with added computing cost
- CNNs and RNNs can be combined in many ways, e.g., image captioning, Relational Networks
- State of the art in text generation: Transformer networks (OpenAI GPT-2)
- Applications: text generation, image captioning, audio synthesis, animation synthesis