

WELCOME TO THE

UCL x DeepMind lecture series

In this lecture series, leading research scientists from leading AI research lab, DeepMind, will give 12 lectures on an exciting selection of topics in Deep Learning, ranging from the fundamentals of training neural networks via advanced ideas around memory, attention, and generative modelling to the important topic of responsible innovation.

Please join us for a deep dive lecture series into Deep Learning!

#UCLxDeepMind





TODAY'S SPEAKER

Felix Hill

Felix Hill has been a Research Scientist at DeepMind since 2016. He did his undergraduate and masters degrees in pure maths at the University of Oxford, and his PhD in Computer Science at the University of Cambridge. His PhD focused on unsupervised learning from text in neural networks. Since joining DeepMind, he has worked on abstract and relational reasoning, and situated models of human language learning and usage, primarily in simulated environments.





TODAY'S LECTURE

Deep Learning for Language Understanding

Since the late 1970s, a key motivation for research into artificial neural networks was as a model for the highly context-dependent nature of semantic cognition. Today, their intrinsic parallel, interactive and context-dependent processing allows deep nets to power many state-of-the-art language technology applications. In this lecture we explain why neural networks can be such an effective tool for modelling language, focusing on the Transformer, a recently-proposed architecture for sequence processing. We also dive into recent research to develop embodied neural-network 'agents' that can relate language to the world around them.



DeepMind

Deep Learning for Language Understanding

Felix Hill

UCL x DeepMind Lectures



Plan for this Lecture

Private & Confidential

01

Background:
deep learning and
language

02

The Transformer

03

Unsupervised and
transfer learning
with BERT

04

Grounded language
learning at
DeepMind: towards
language
understanding in a
situated agent



What is not covered in this lecture

Private & Confidential

01

Recurrent networks

- Seq-to-seq models and neural MT
- Speech recognition or synthesis

02

Many NLP tasks and applications

- Machine comprehension
- Question answering
- Dialogue

03

Grounding in image/video

- Visual question-answering or captioning
- CLEVR and visual reasoning
- Video captioning





1

Background: Deep learning and language



Google Cloud Text-to-Speech now has 187 voices
and 95 WaveNet voices

Microsoft's Chinese-English translation system achieves human parity

A China-US research team at software powerhouse Microsoft has developed an AI-powered system that can translate Chinese news articles into English as well as humans do.



INNOVATION / AI

OpenAI's GPT2 Now Writes Scientific Paper Abstracts

A string of tweets demonstrates the transformer neural network's incredible capabilities.

Google brings in BERT to improve its search results



Machine
Translation

Question
answering

Search /
information
retrieval

Home
assistants

Very
neural

Less /
non-neural

Speech
synthesis

Text
classification

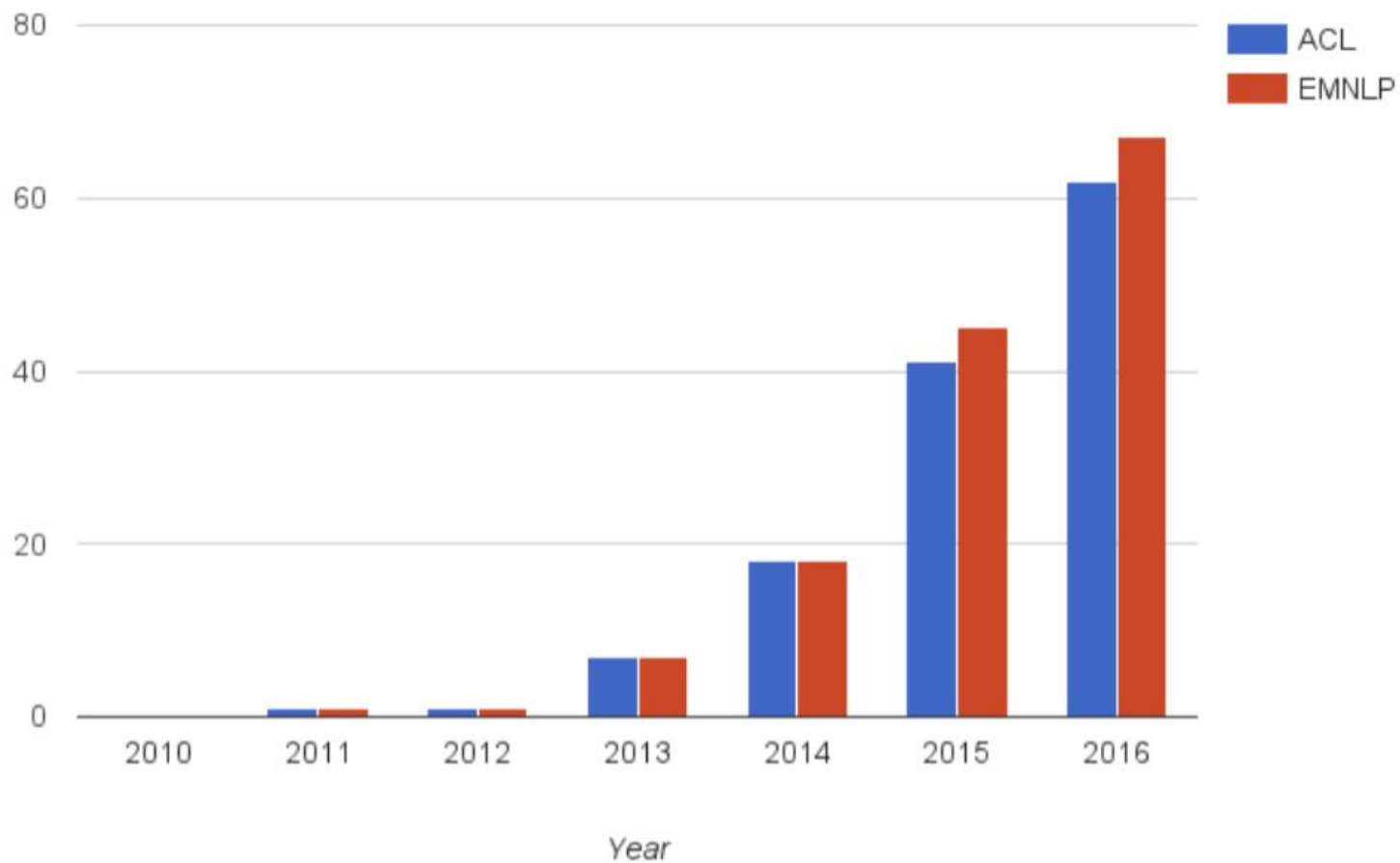
Dialogue
systems

Speech
recognition

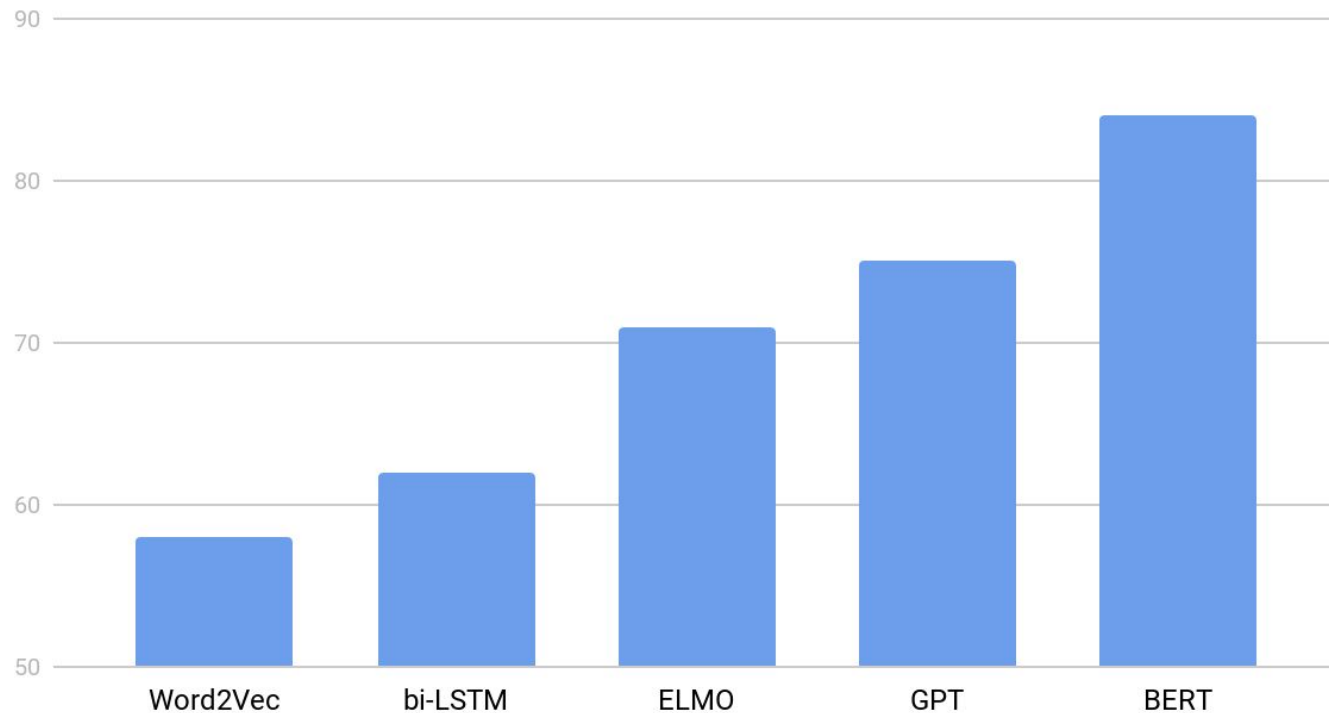
Summarization



Papers with "Deep" or "Neural" in title



Performance on GLUE benchmark (11 tasks) 2018-19



Why is Deep Learning such an effective tool for language processing?



Why is Deep Learning such an effective tool for language processing?

Can it be improved?



Some things about
language...



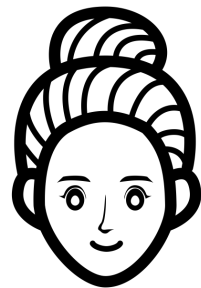
1. Words are not discrete symbols



- Did you see the look on her **face**₁ ?
- We could see the clock **face**₂ from below
- It could be time to **face**₃ his demons
- There are a few new **faces**₄ in the office today



- Did you see the look on her **face₁**?
- We could see the clock **face₂** from below
- It could be time to **face₃** his demons
- There are a few new **faces₄** in the office today



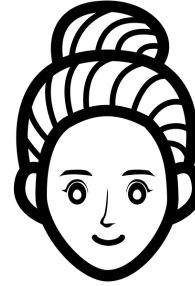
Created by Chananan
from Noun Project

Face 1

- The most important side (of the head)
- Represents you / yourself
- Used to inform / communicate
- Points forward when you address/confront something



- Did you see the look on her **face₁**?
- We could see the clock **face₂** from below
- It could be time to **face₃** his demons
- There are a few new **faces₄** in the office today



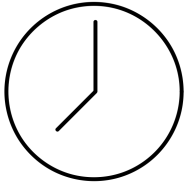
Created by Chananan
from Noun Project

Face 1

- The most important side (of the head)
- Represents you / yourself
- Used to inform / communicate
- Points forward when you address/confront something

Face 2

- The most important side (of the head)
- Used to inform / communicate



Created by Neha Tyagi
from Noun Project



- Did you see the look on her **face₁**?
- We could see the clock **face₂** from below
- It could be time to **face₃** his demons
- There are a few new **faces₄** in the office today



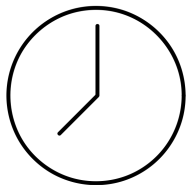
Created by Chananan
from Noun Project

Face 1

- The most important side (of the head)
- Represents you / yourself
- Used to inform / communicate
- Points forward when you address/confront something

Face 2

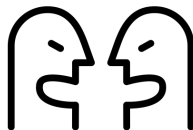
- The most important side (of the head)
- Used to inform / communicate



Created by Neha Tyagi
from Noun Project

Face 3

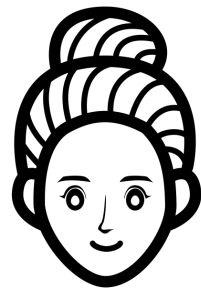
- Points forward when you address/confront something



Created by Yu luck
from Noun Project



- Did you see the look on her **face₁** ?
- We could see the clock **face₂** from below
- It could be time to **face₃** his demons
- There are a few new **faces₄** in the office today



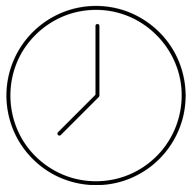
Created by Chananan
from Noun Project

Face 1

- The most important side (of the head)
- Represents you / yourself
- Used to inform / communicate
- Points forward when you address/confront something

Face 2

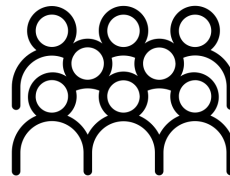
- The most important side (of the head)
- Used to inform / communicate



Created by Neha Tyagi
from Noun Project

Face 4

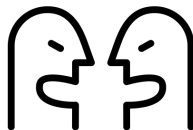
- Represents you / yourself



Created by Dániel Aczél
from Noun Project

Face 3

- Points forward when you address/confront something



Created by Yu luck
from Noun Project



1. Words are not discrete symbols

2. Disambiguation depends on context



Jack and Jill  went up the hill.

The pole vault was the last  event.



12
ABC
14



1. Words are not discrete symbols

2. Disambiguation depends on context

3. Important interactions can be non-local



The man who ate the pepper sneezed



The man who ate the pepper sneezed



The man who ate the pepper sneezed

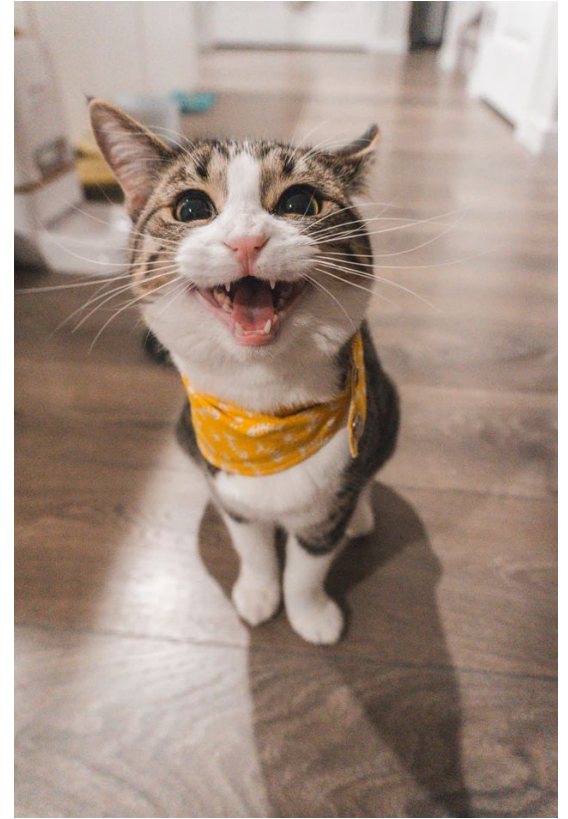


But note:

The cat who bit the dog barked



The man who ate the pepper sneezed



The cat who bit the dog barked



1. Words are not discrete symbols

2. Disambiguation depends on context

3. Important interactions can be non-local

4. How meanings combine depends on those meanings











PET

brown
white
black





PET

brown
white
black

FISH

silver grey





PET FISH

Orange
Green
Blue
Purple
Yellow





PLANT

Green
Leaves
Grows



CARNIVORE

Eats meat
Sharp teeth



CARNIVOROUS PLANT

Green
Leaves
Grows
Sharp teeth
Eats insects



1. Words have many related senses

2. Disambiguation depends on context

3. Important interactions can be non-local

4. How meanings combine depends on those meanings





2

The Transformer

Vaswani et al. 2018



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

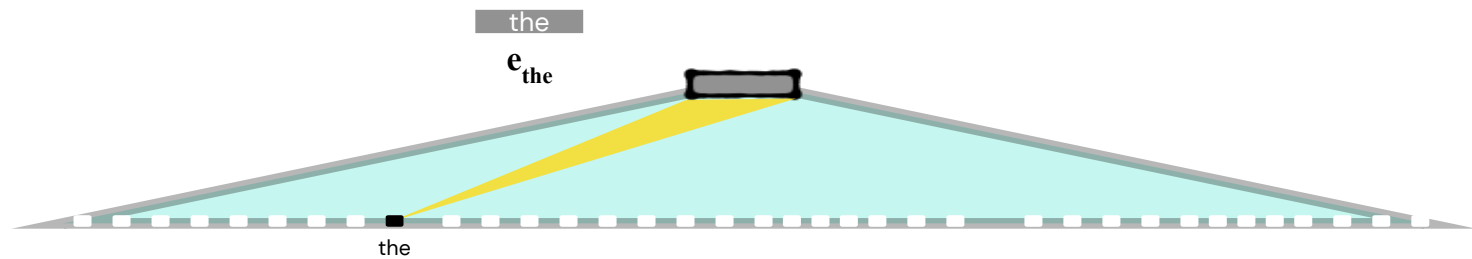
Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaizer@google.com

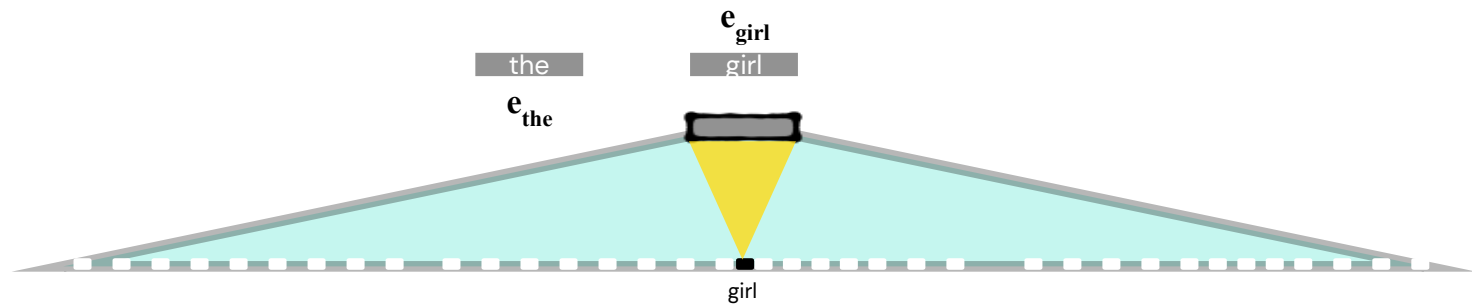
Illia Polosukhin* ‡
illia.polosukhin@gmail.com



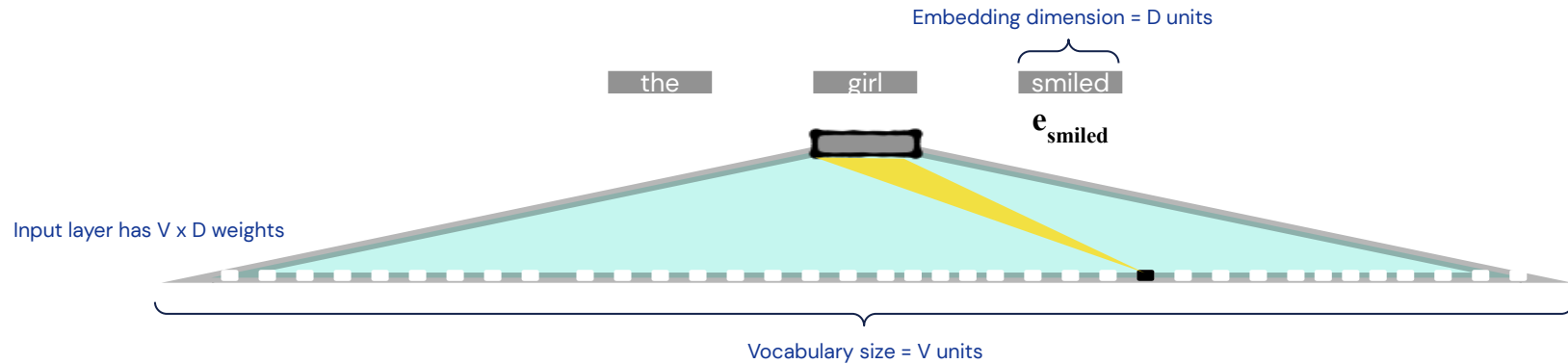
Distributed representations of words



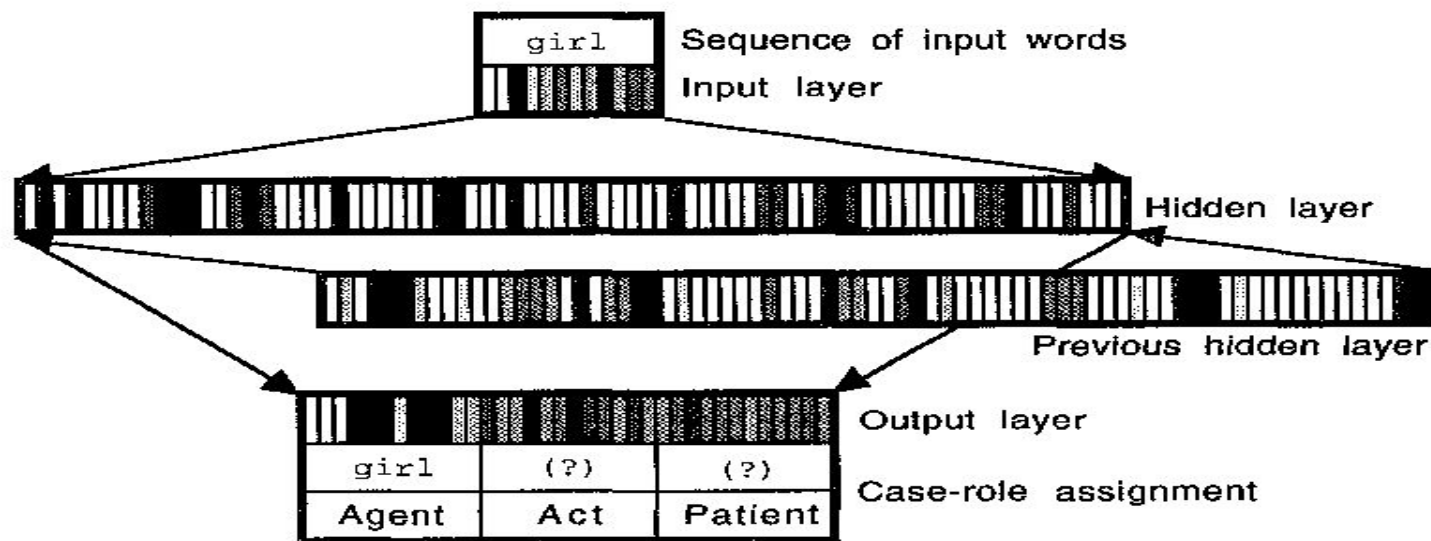
Distributed representations of words



Distributed representations of words



The Transformer builds on solid foundations

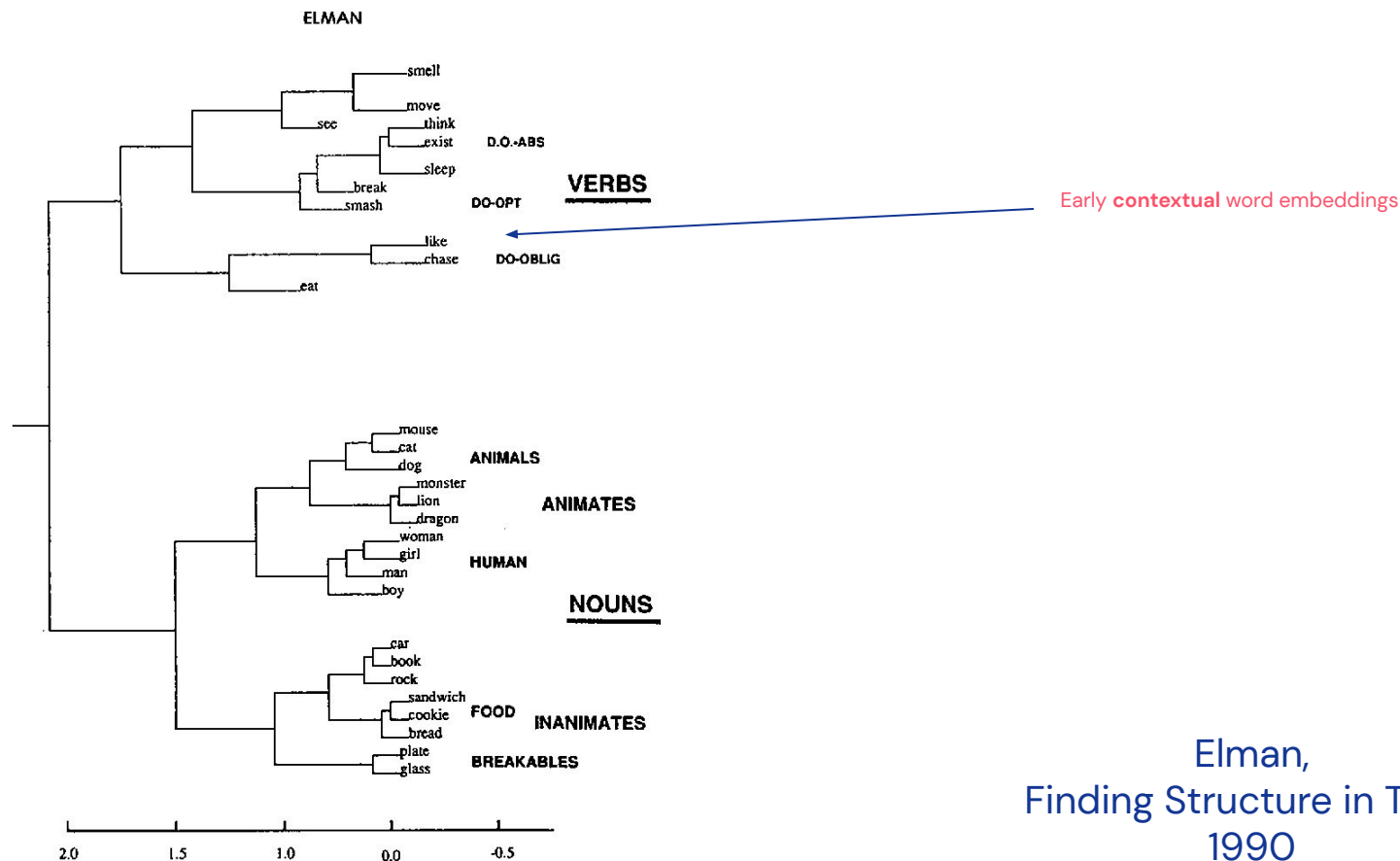


Mikkulainen & Dyer, 1991



Emergent semantic and syntactic structure in distributed representations

)



Elman,
Finding Structure in Time,
1990

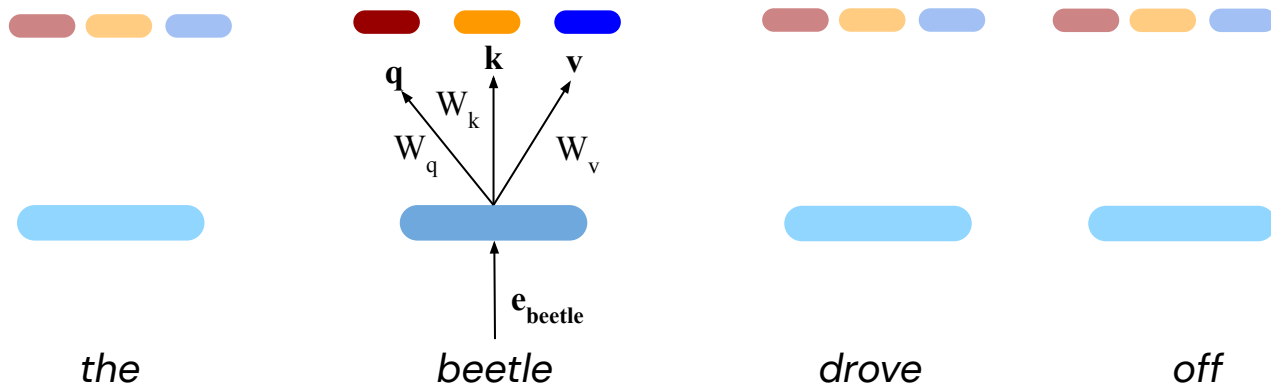


The Transformer: Self-attention over word input embeddings

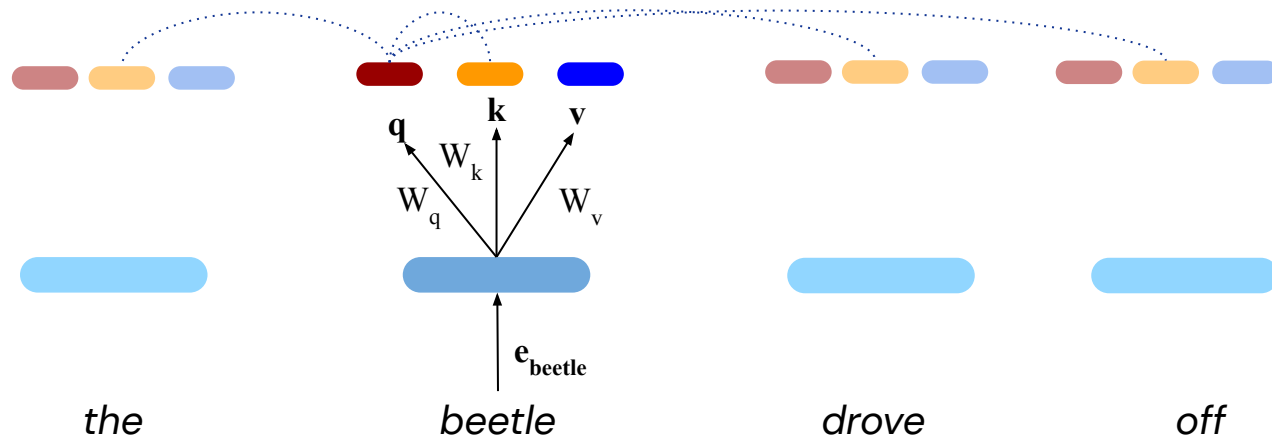
$$\mathbf{q} = \mathbf{e}_{beetle} W_q$$

$$\mathbf{k} = \mathbf{e}_{beetle} W_k$$

$$\mathbf{v} = \mathbf{e}_{beetle} W_v$$

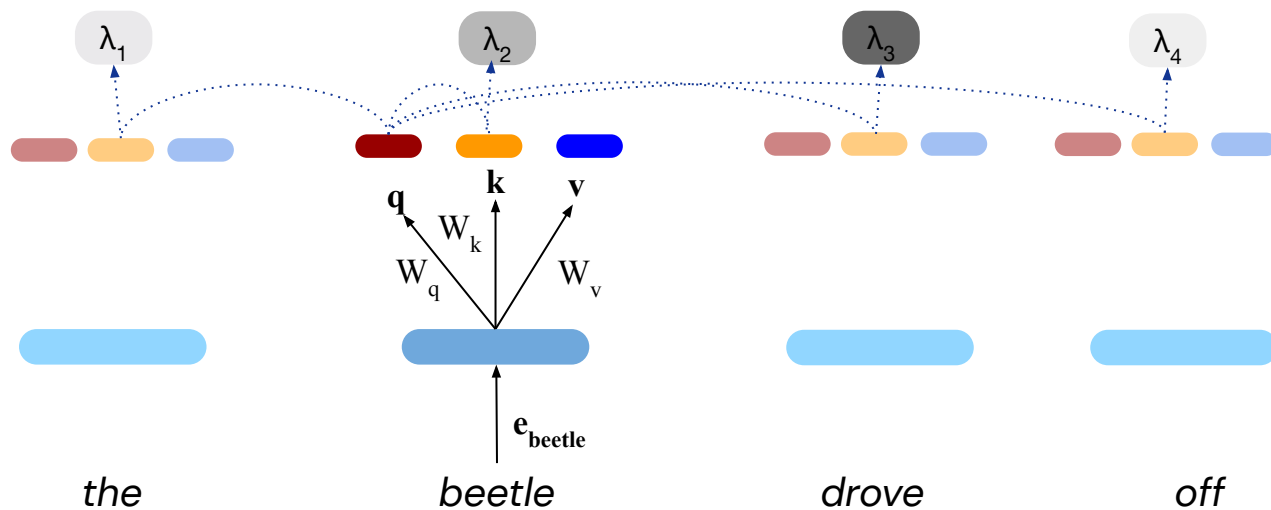


Self-attention over word input embeddings

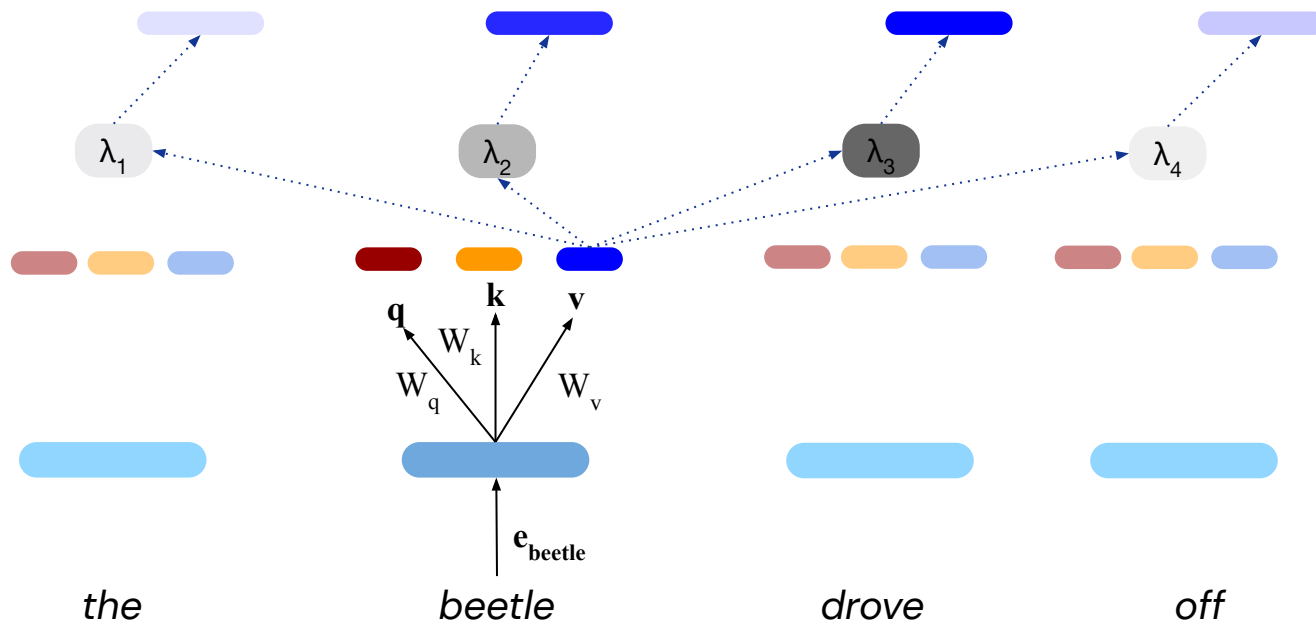


Self-attention over word input embeddings

$$\lambda_i = \frac{e^{\mathbf{q} \cdot \mathbf{k}_i}}{\sum_{i=1}^4 e^{\mathbf{q} \cdot \mathbf{k}_i}}$$

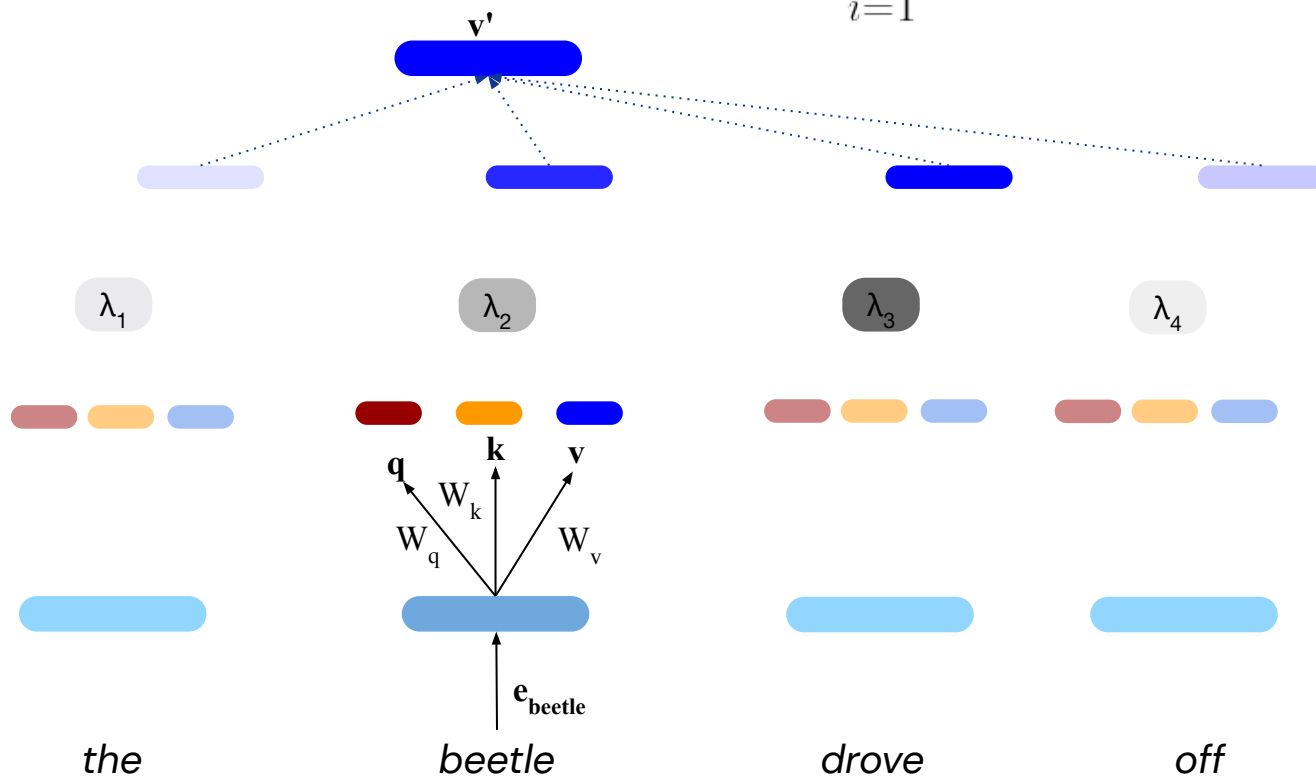


Self-attention over word input embeddings

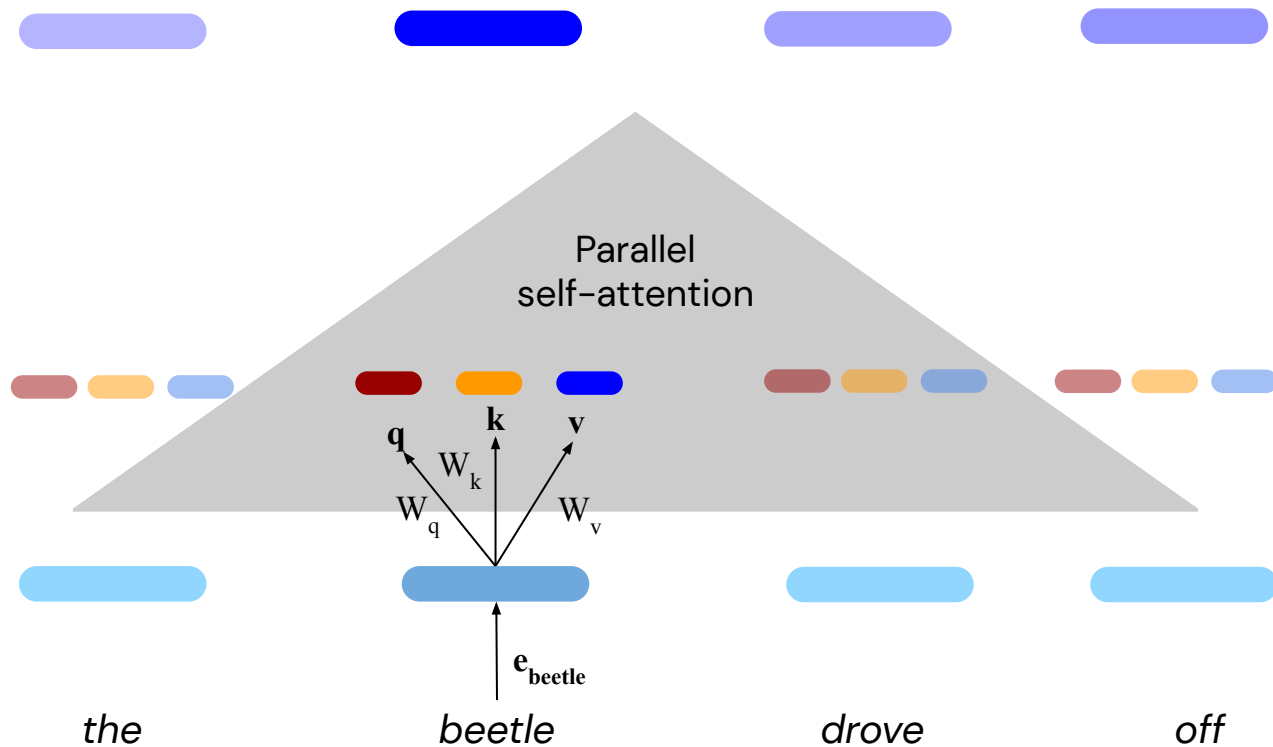


Self-attention over word input embeddings

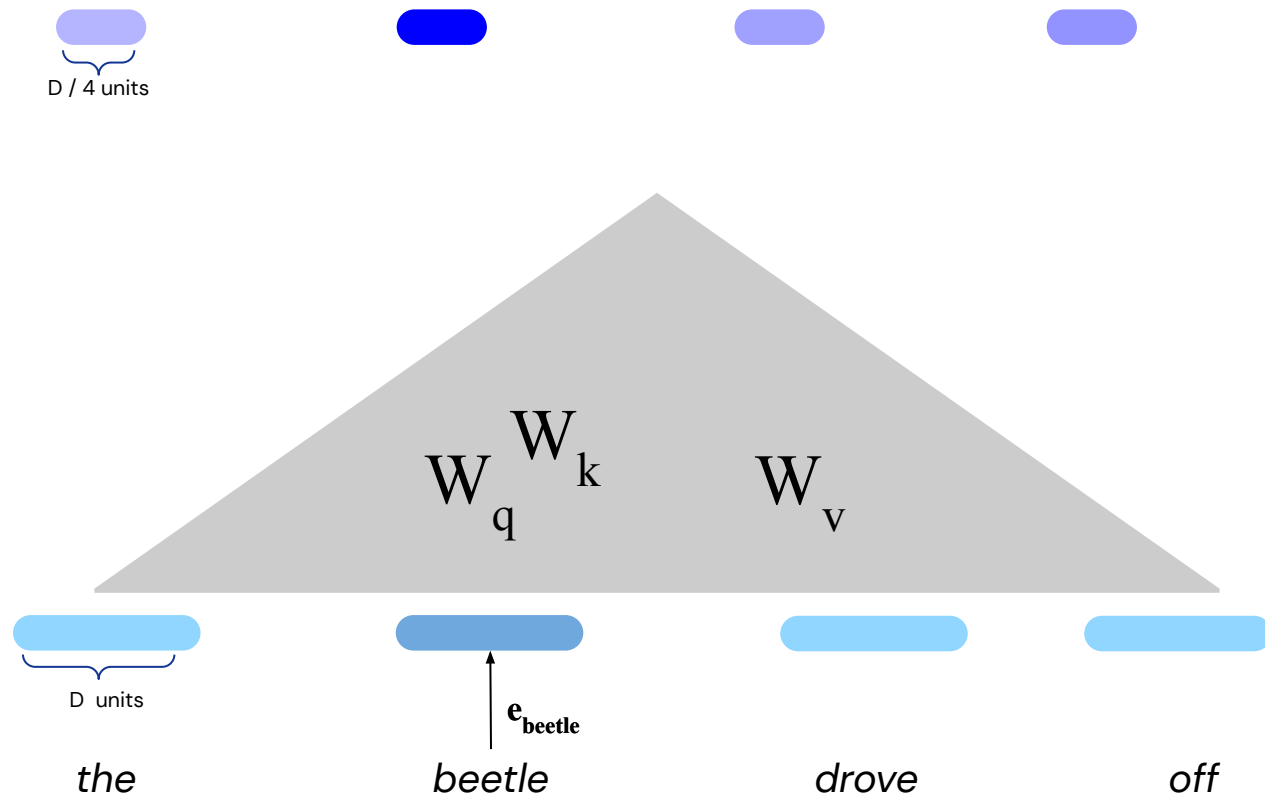
$$\mathbf{v}' = \sum_{i=1}^4 \lambda_i \mathbf{v}_i$$



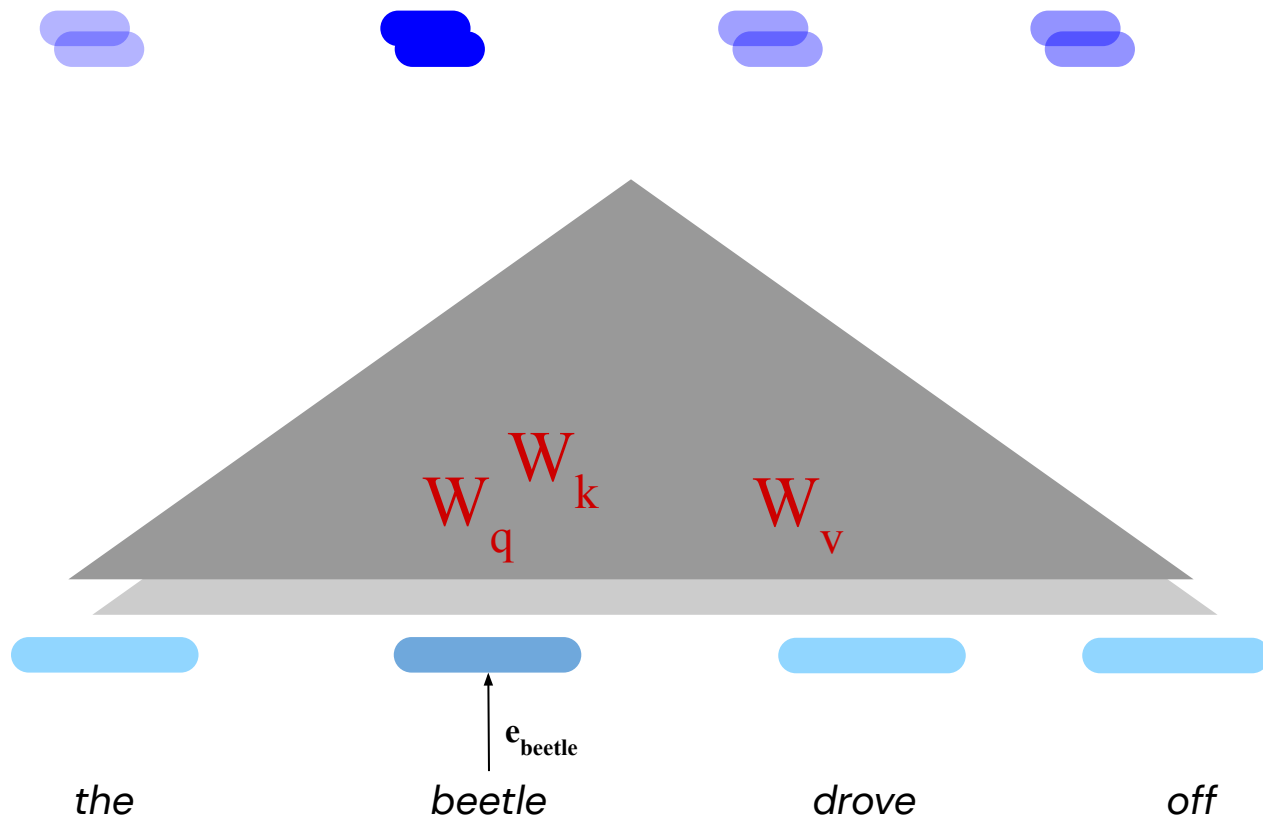
Compute self-attention for all words in input (in parallel)



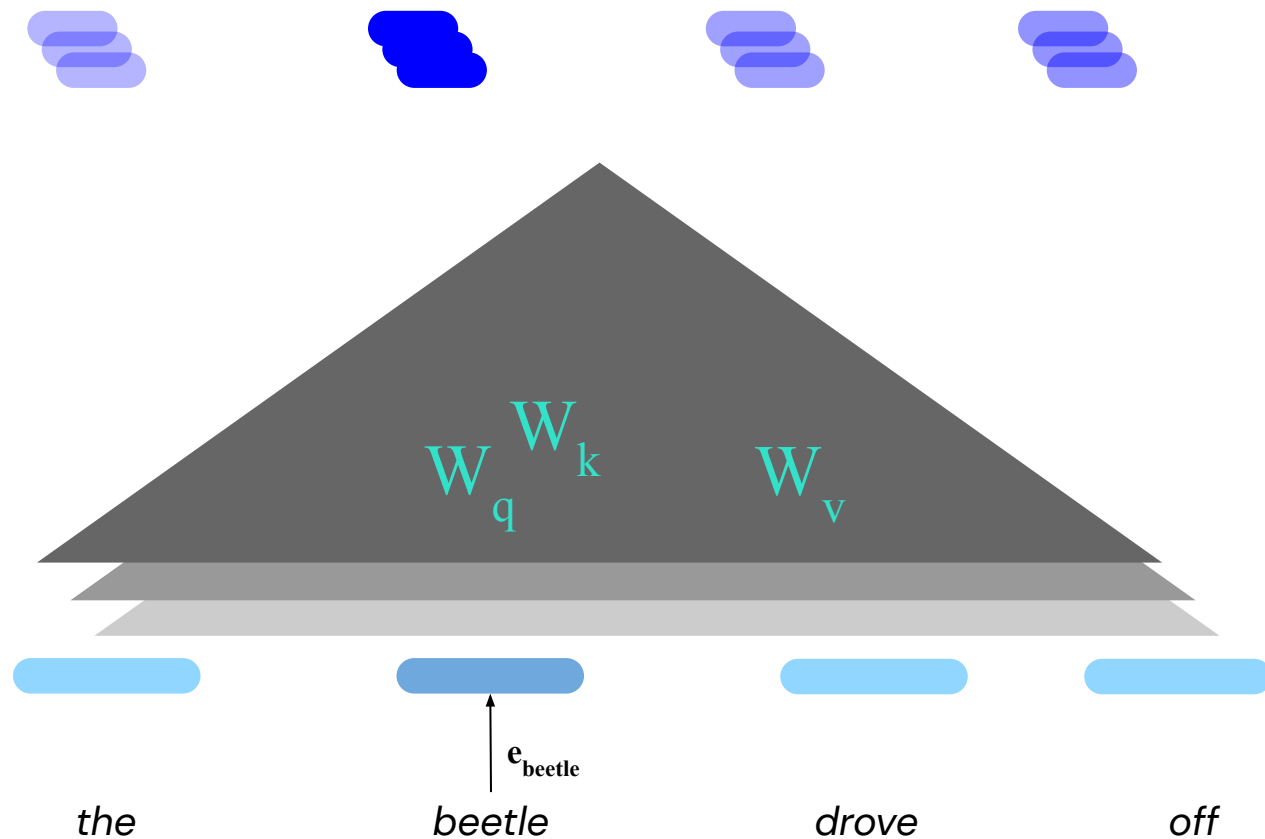
Multi-head self-attention ($H = 4$)



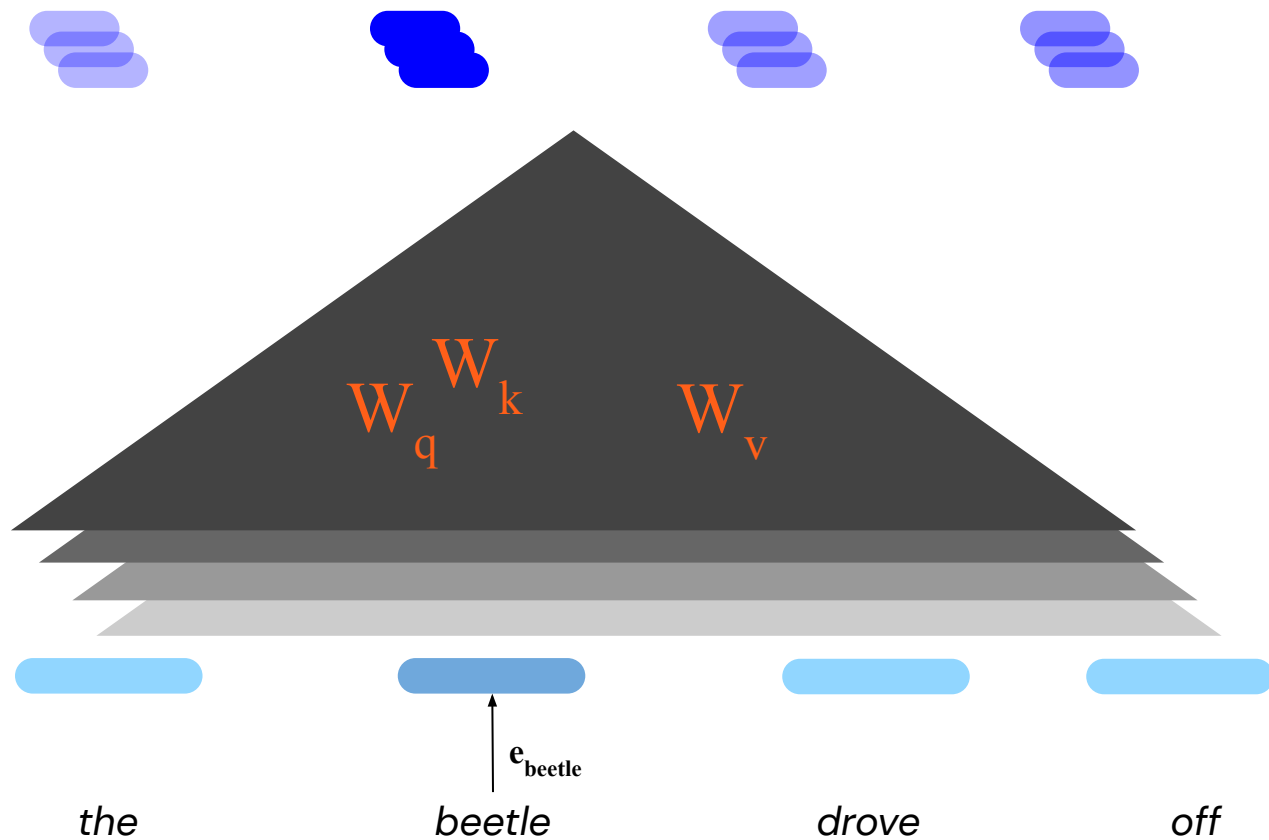
Multi-head self-attention ($H = 4$)

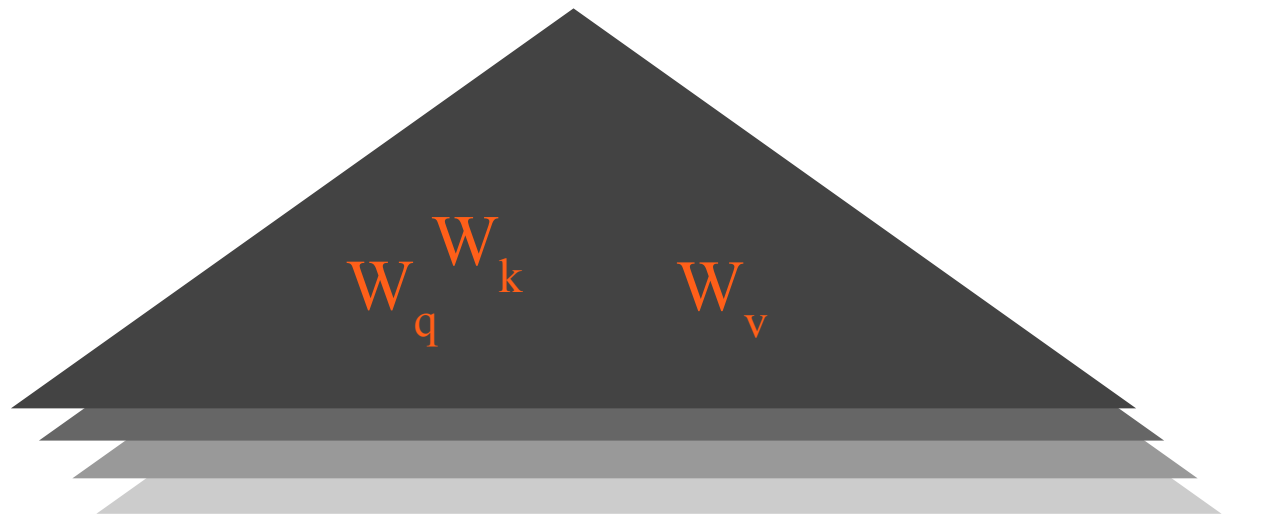
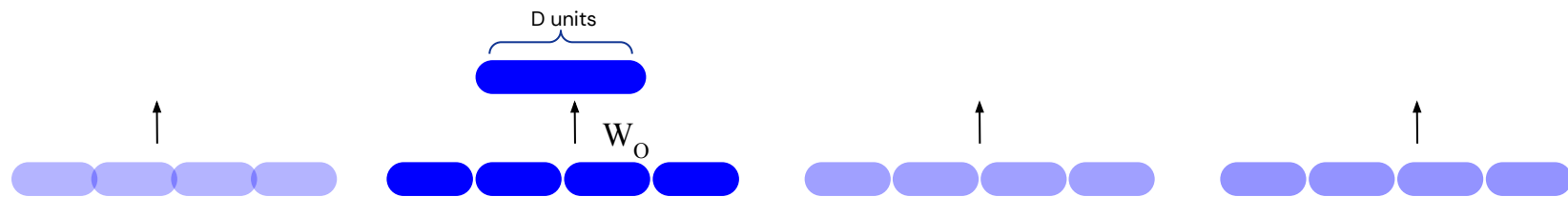


Multi-head self-attention ($H = 4$)

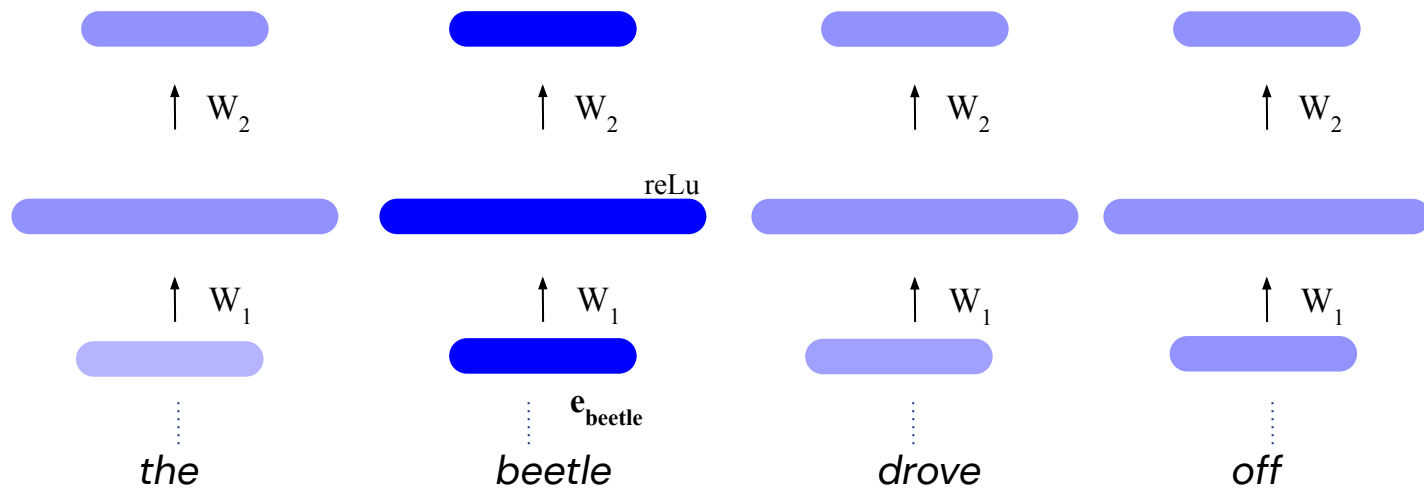


Multi-head self-attention ($H = 4$)

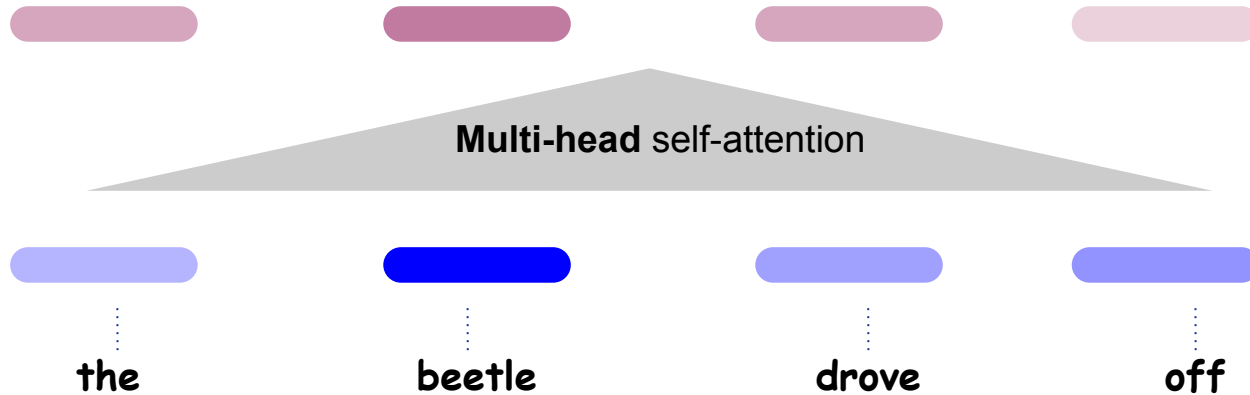




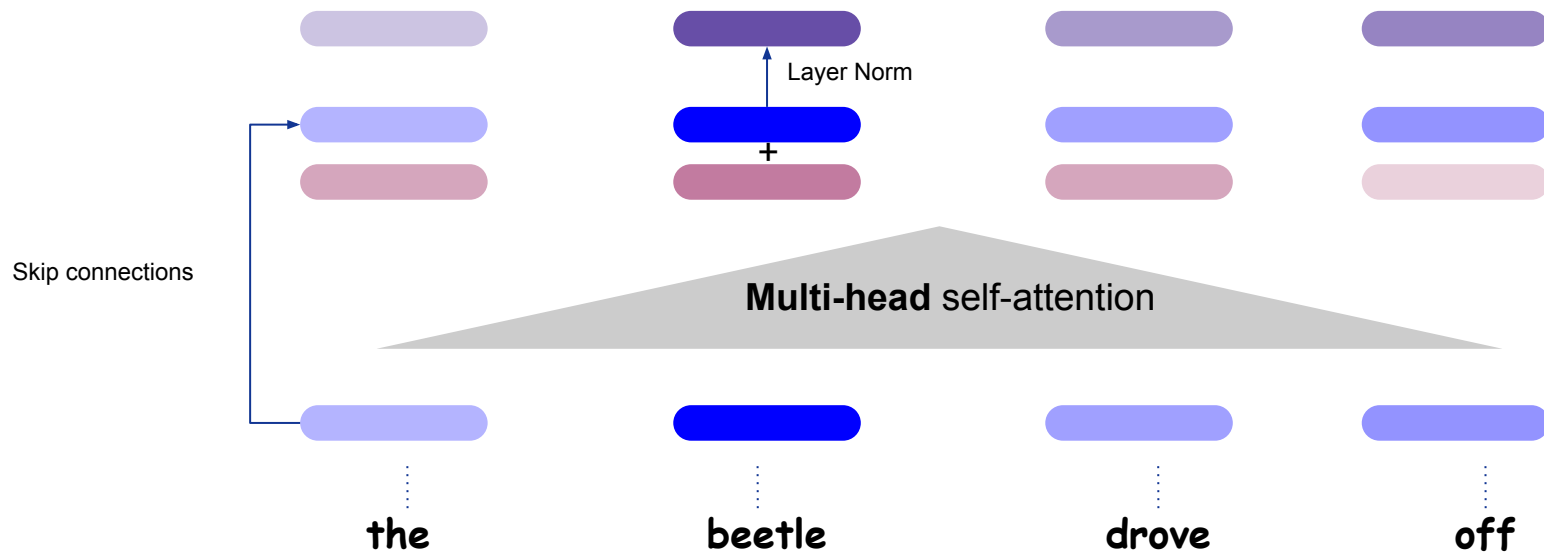
Feedforward Layer



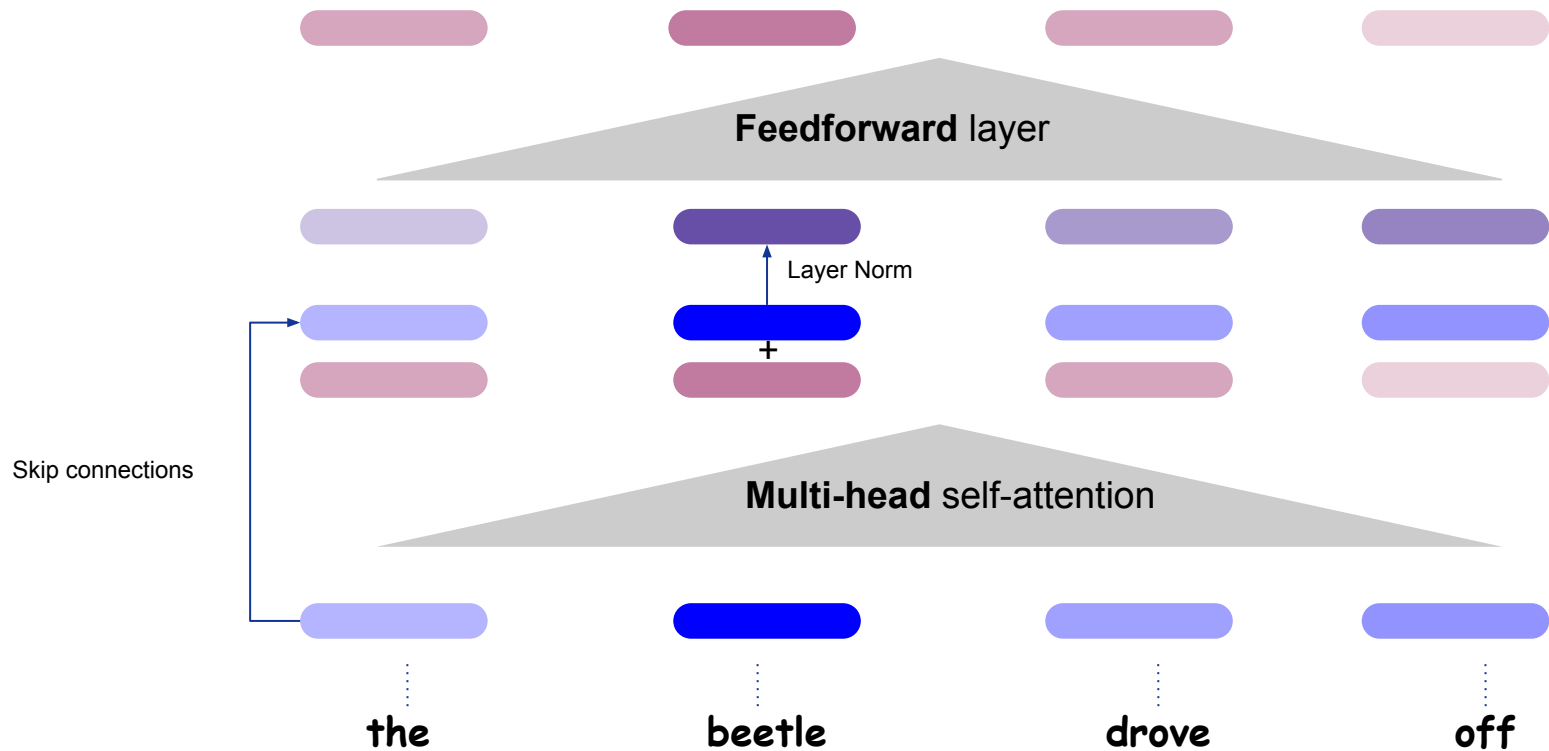
A complete Transformer block



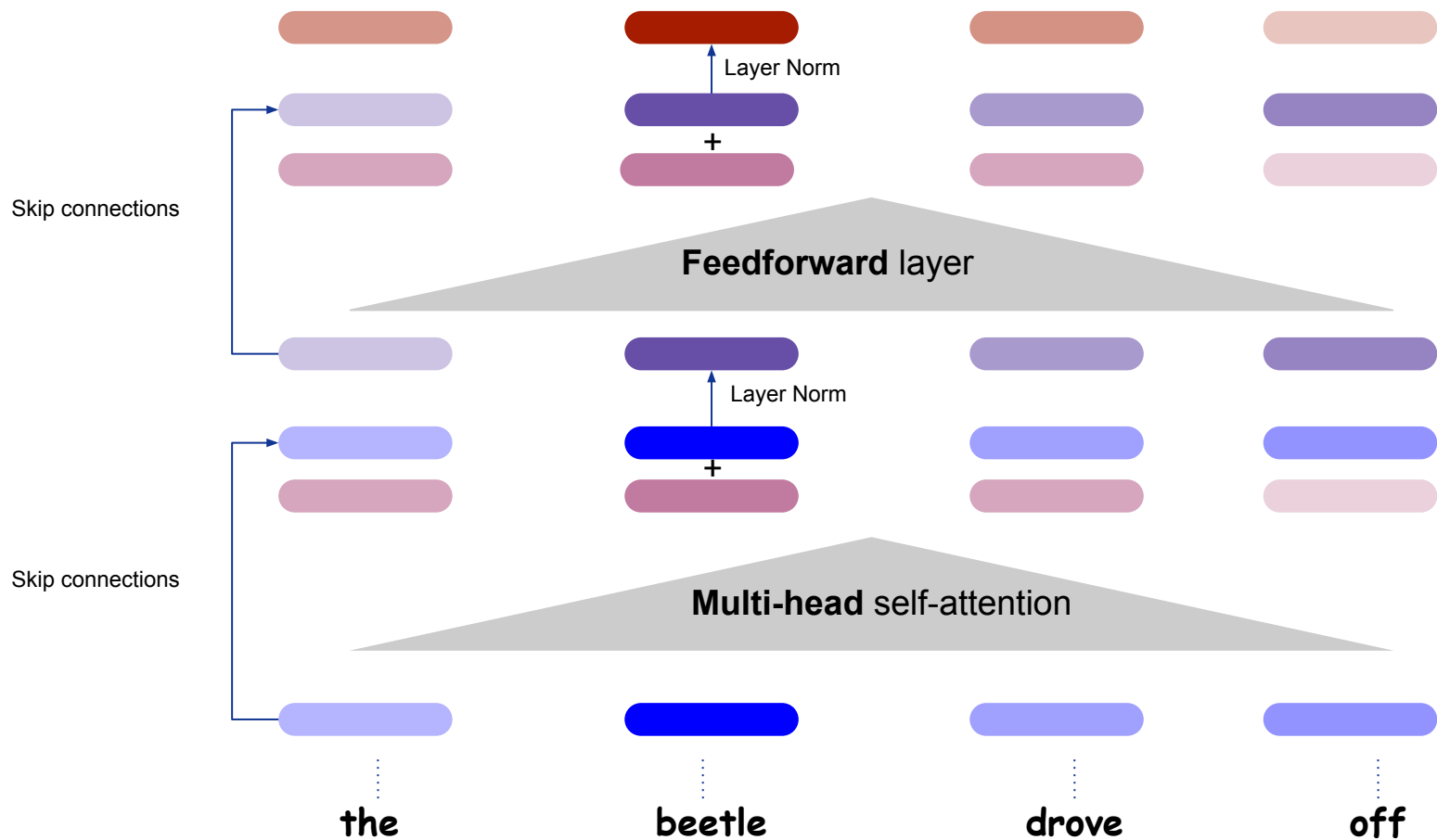
A complete Transformer block



A complete Transformer block

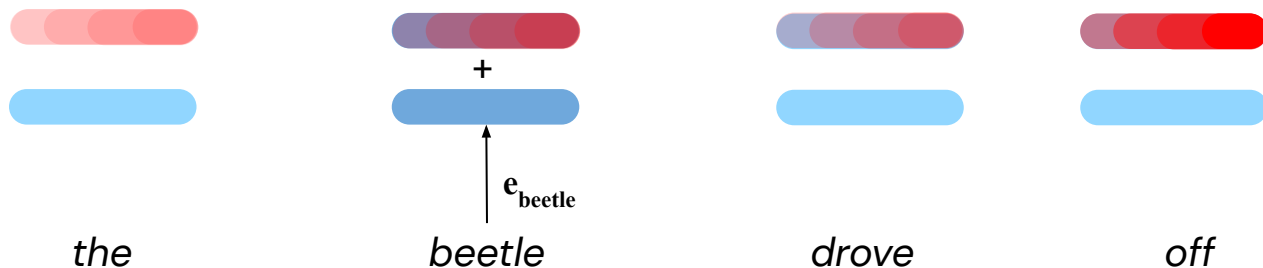


Skip-connections - for "top-down" influences



The Transformer: Position encoding of words

- Add fixed quantity to embedding activations
- The quantity added to each input embedding unit $\in [-1, 1]$ depends on:
 - The dimension of the unit within the embedding
 - The (absolute) position of the words in the input



1. Words are not discrete symbols

2. Disambiguation depends on context

3. Important interactions can be non-local

4. How meanings combine depends on those meanings



1. Words are not discrete symbols

Multi-head processing

Distributed
representations

2. Disambiguation depends on context

3. Important interactions can be non-local

4. How meanings combine depends on those meanings



1. Words are not discrete symbols

Multi-head processing

Distributed
representations

2. Disambiguation depends on context

Self-attention

3. Important interactions can be non-local

4. How meanings combine depends on those meanings



1. Words are not discrete symbols

Multi-head processing

Distributed
representations

2. Disambiguation depends on context

Self-attention

3. Important interactions can be non-local

Self-attention

Multiple layers

4. How meanings combine depends on those meanings



1. Words are not discrete symbols

Multi-head processing

Distributed
representations

2. Disambiguation depends on context

Self-attention

3. Important interactions can be non-local

Self-attention

Multiple layers

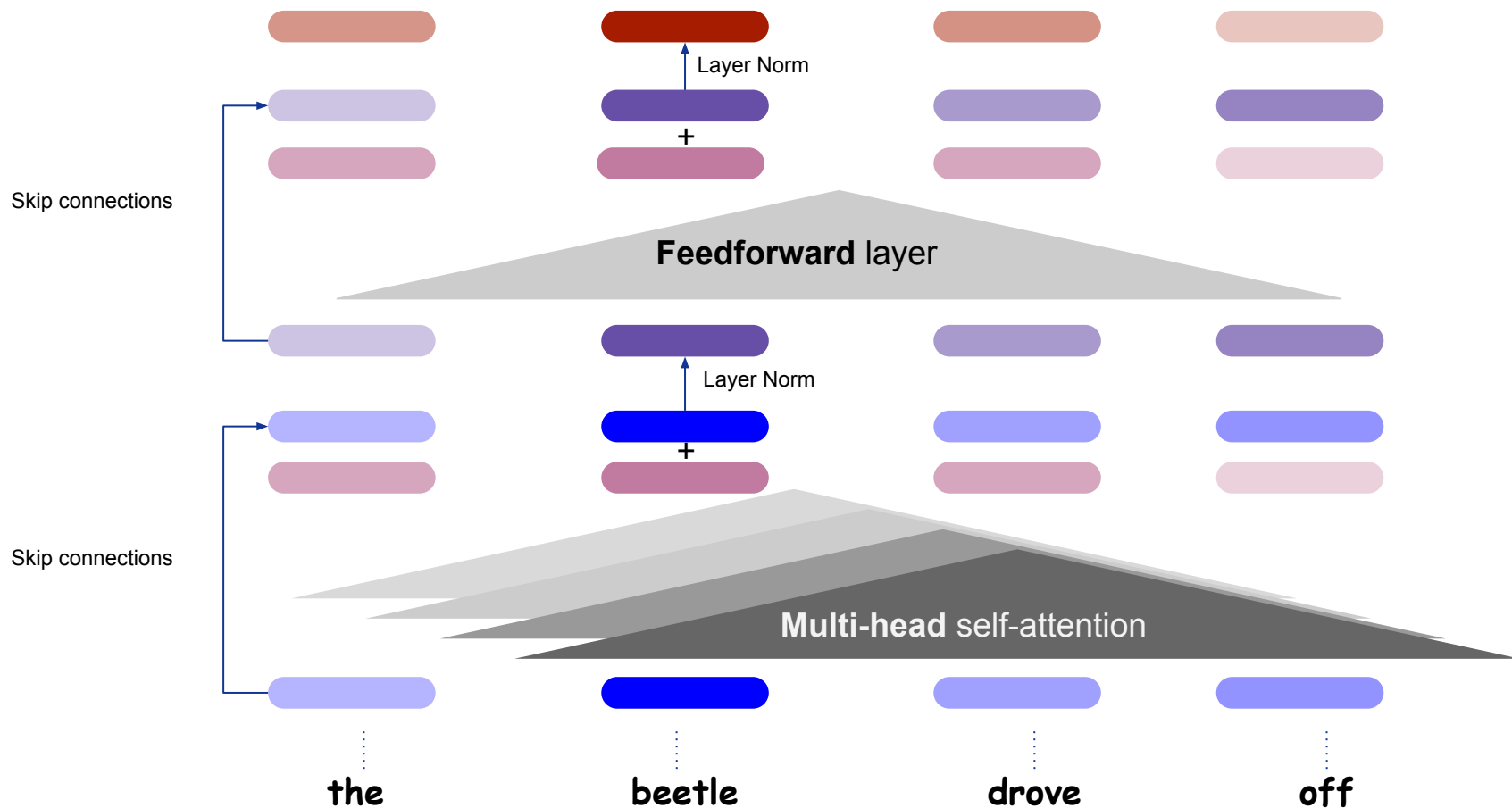
4. How meanings combine depends on those meanings

Skip connections

Distributed
representations



Skip-connections - for "top-down" influences



The background features abstract geometric shapes outlined in a light blue color. On the left, there is a large, curved shape resembling a stylized 'C' or a sector of a circle. On the right, there is a 3D wireframe cube. The overall design is minimalist and modern.

3

Unsupervised Learning With Transformers (BERT)



Time flies like an arrow



Time flies like an arrow

Fruit flies like a banana

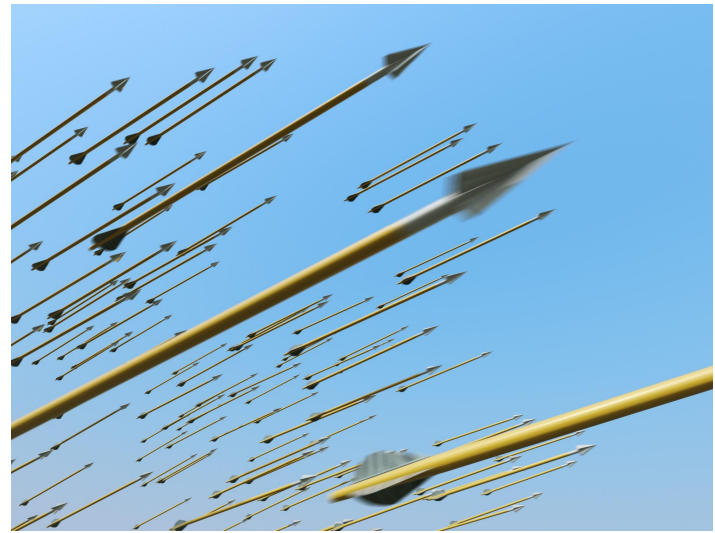
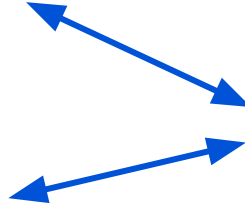


Time flies like an arrow



John works like a trojan

The trains run like clockwork



Fruit flies like a banana

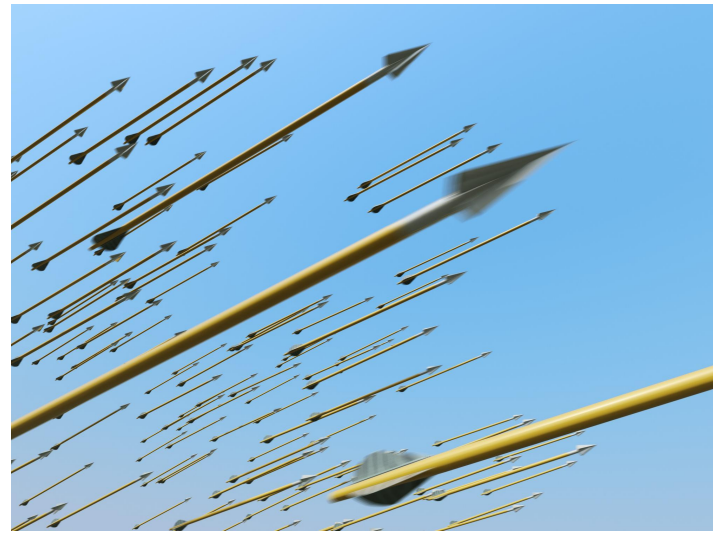
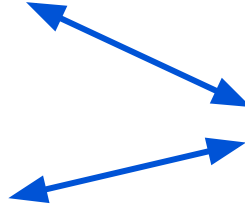


Time flies like an arrow



John works like a trojan

The trains run like clockwork



Fruit flies like a banana

Fido likes having his tummy rubbed

Grandma likes a good cuppa



1. Words have many related senses

2. Disambiguation depends on context

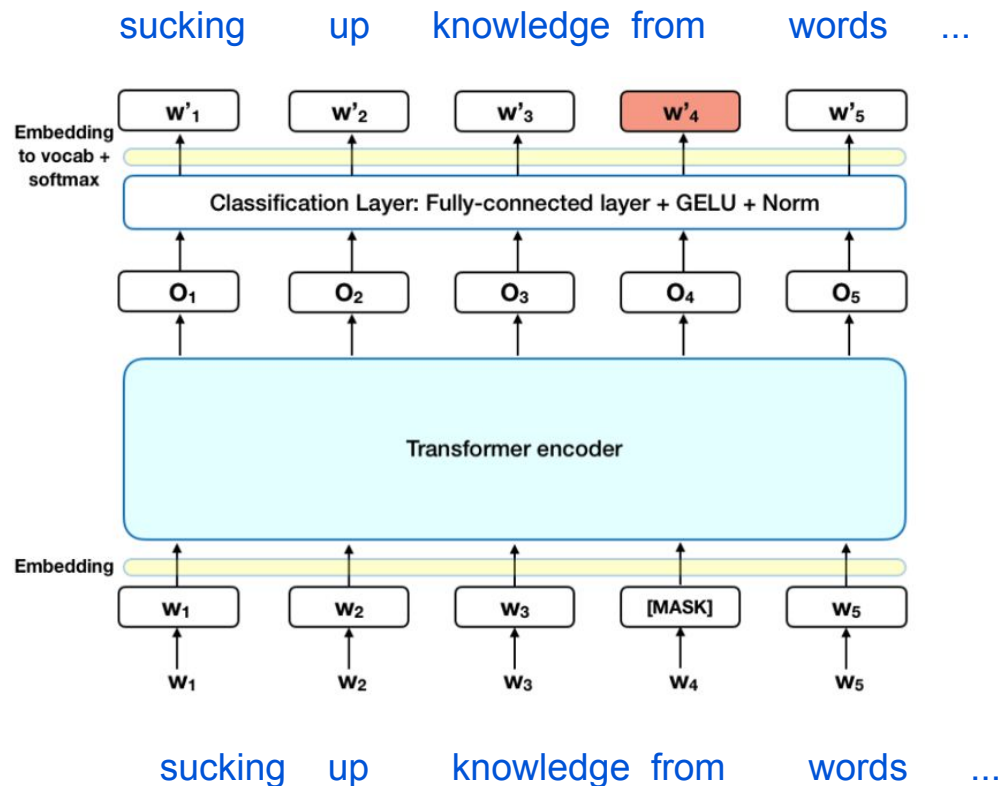
3. Relevant context can be non-local

4. 'Composition' depends on what words mean

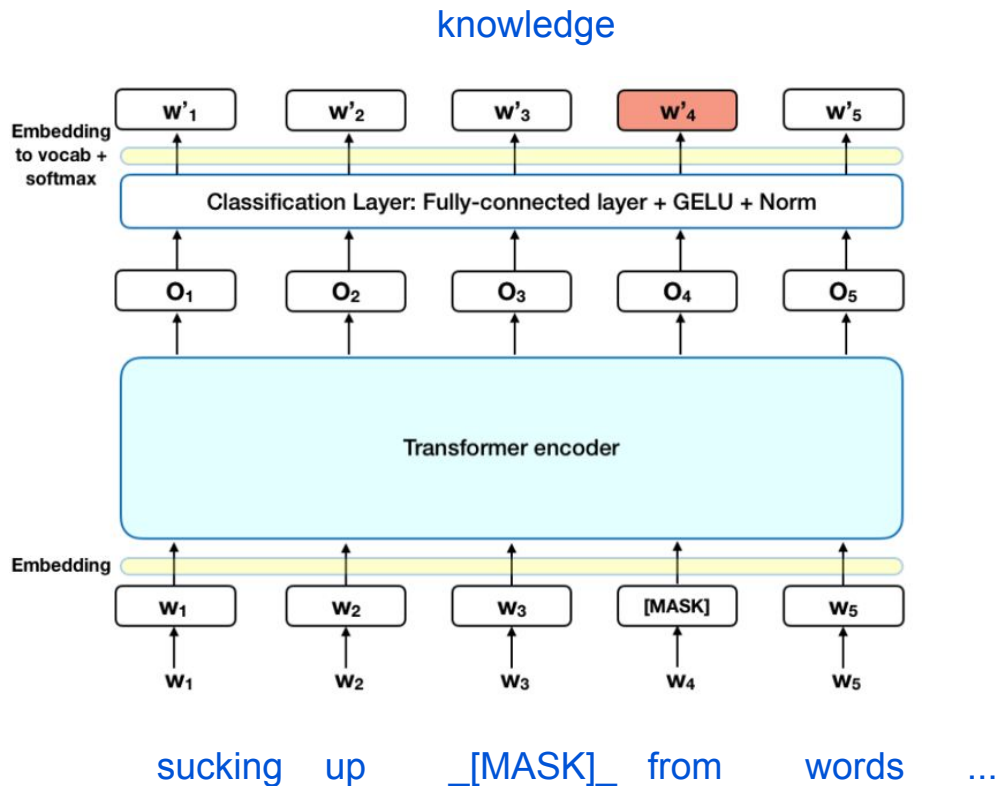
5. Understanding is balancing input with knowledge



BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding (Devlin et al. 2019)



Masked language model pretraining

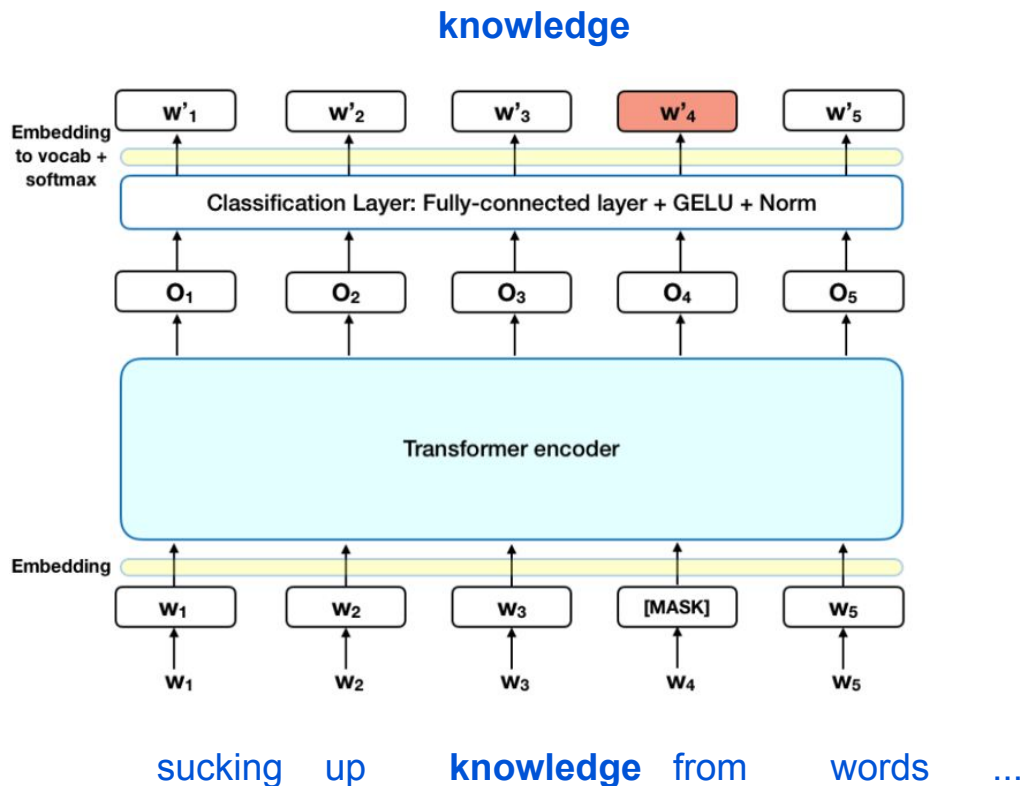


BERT (Devlin et al. 2019)

15% of words masked - predict them!



Masked language model pretraining



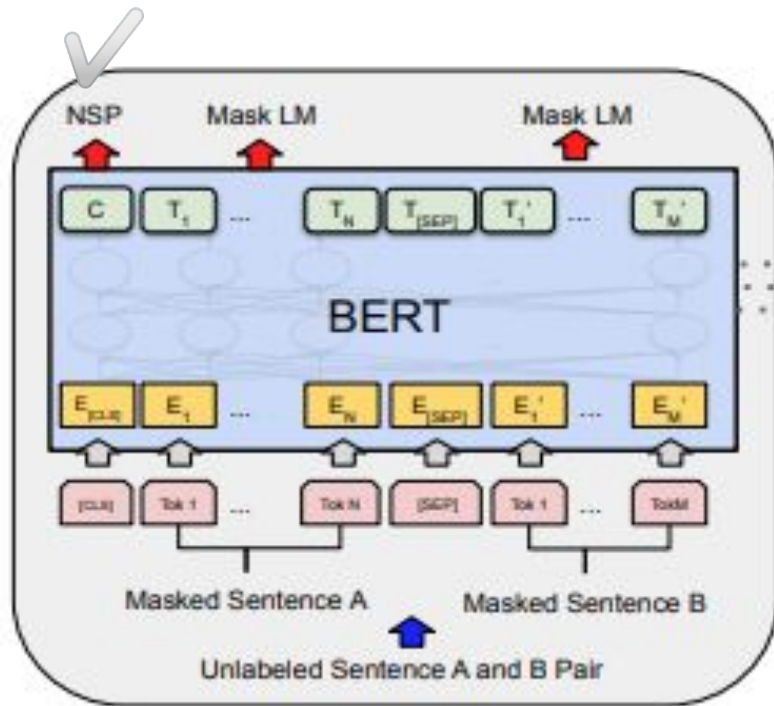
BERT (Devlin et al. 2019)

15% of words masked - predict them!

10% of these instances - leave word in place



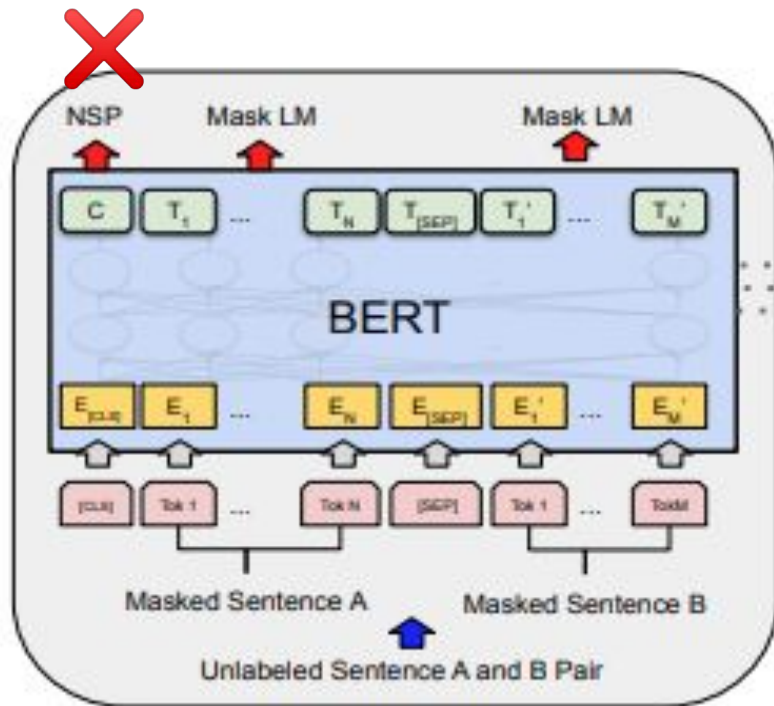
Next sentence prediction pretraining



[CLS] Sid went outside . [SEP] It began to rain .



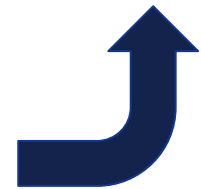
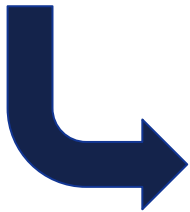
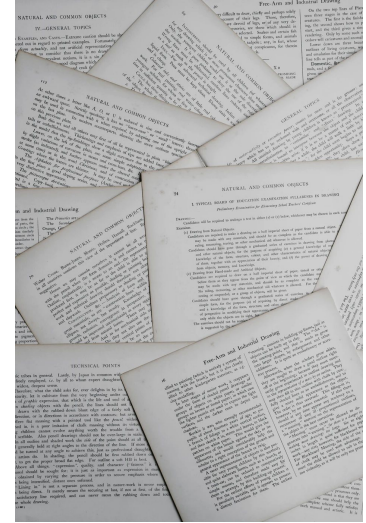
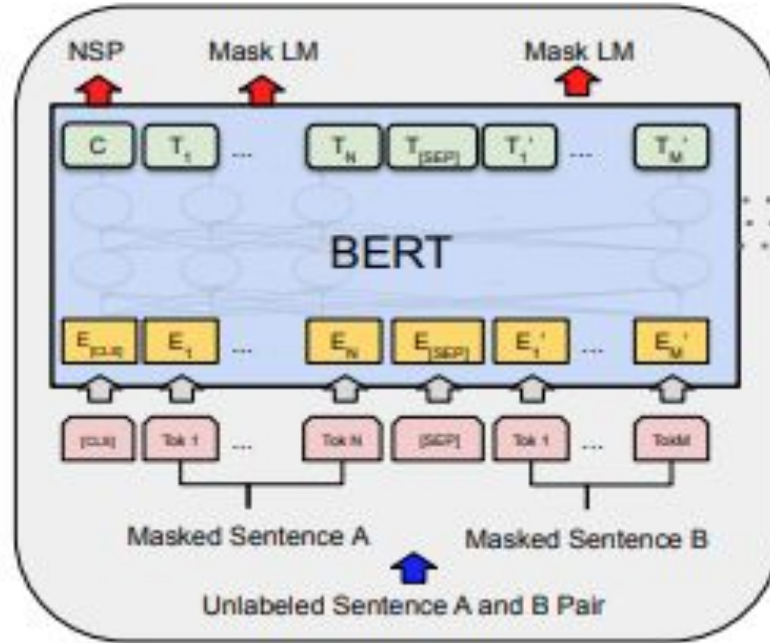
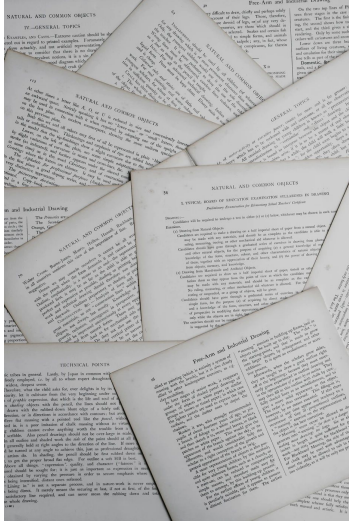
Next sentence prediction pretraining



[CLS] Sid went outside . _[SEP]_ Unfortunately it wasn't



BERT pretraining

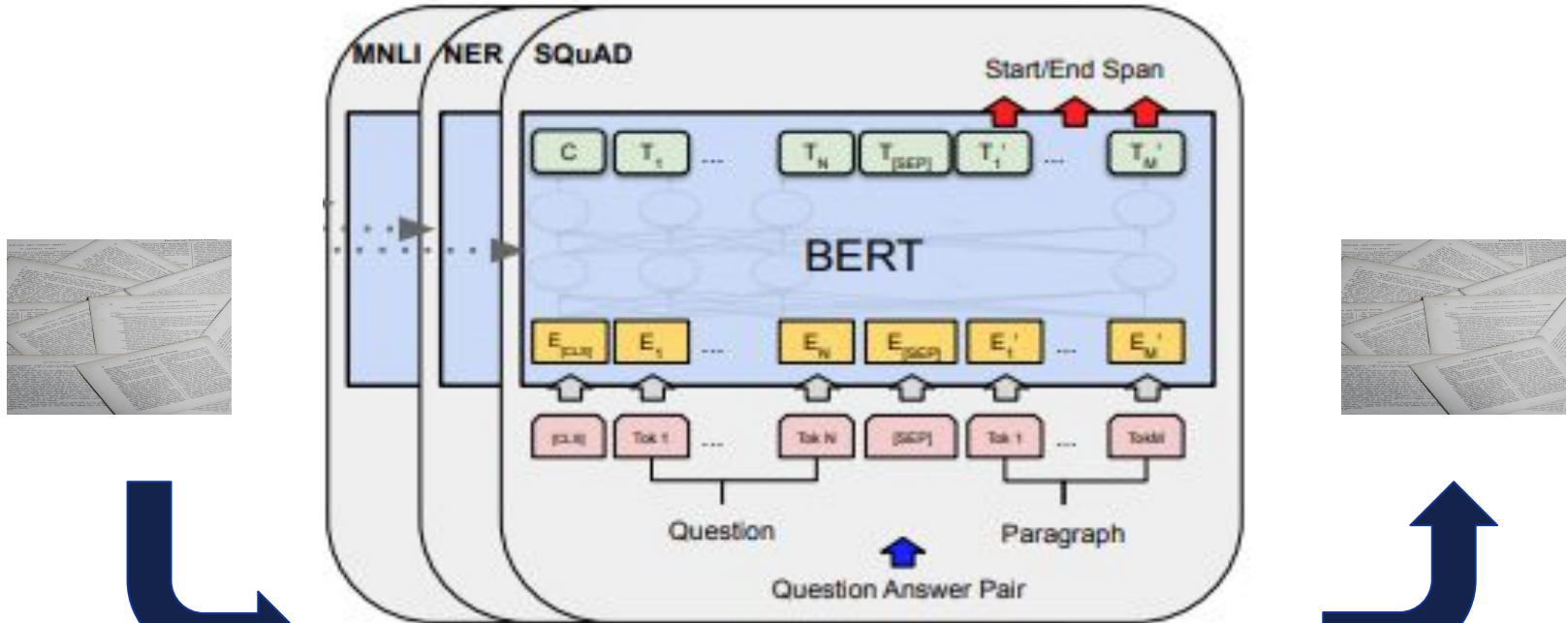


[CLS] Masses of text . [SEP] From the internet .

BERT (Devlin et al. 2019)



BERT fine-tuning



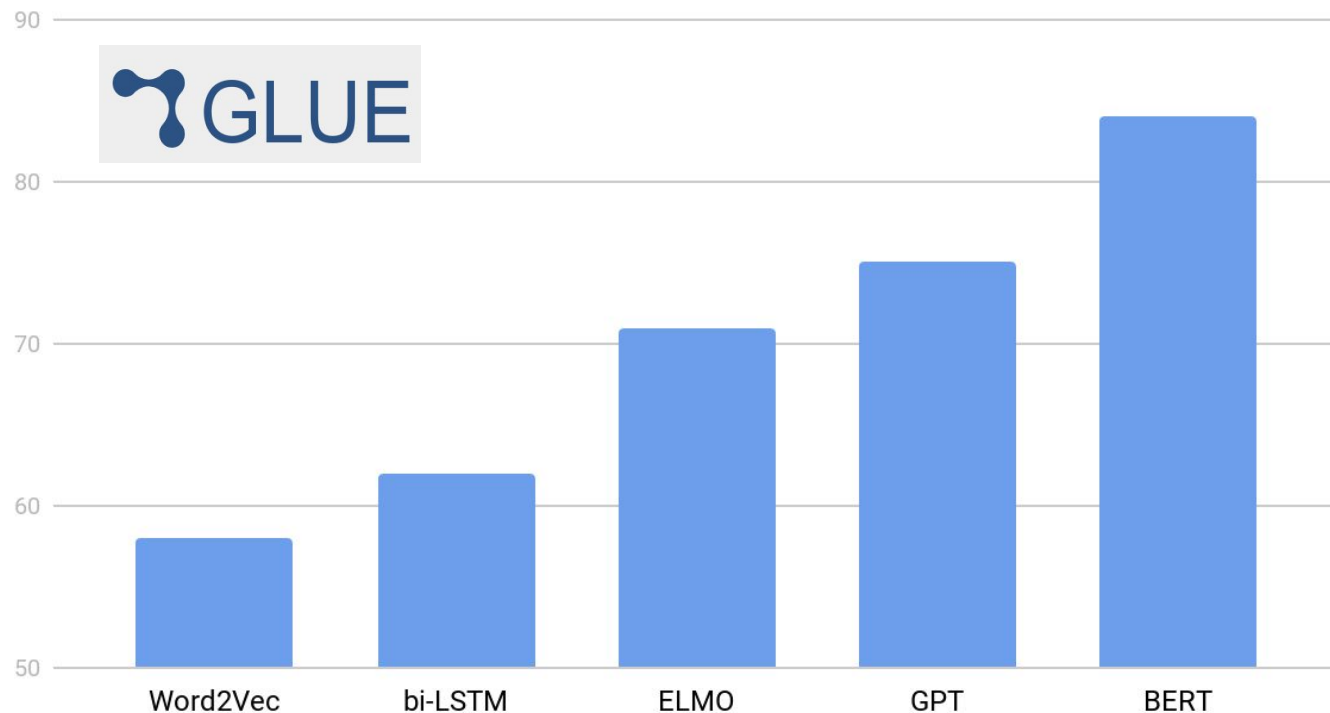
`_[CLS]_` A small amount of `_[SEP]_` Task-specific data

BERT (Devlin et al. 2019)



BERT supercharges transfer learning

Performance on GLUE benchmark (11 tasks) since 2018



1. Words have many related senses

2. Disambiguation depends on context

3. Relevant context can be non-local

4. 'Composition' depends on what words mean

5. Understanding is balancing input with knowledge



1. Words have many related senses

2. Disambiguation depends on context

3. Relevant context can be non-local

4. 'Composition' depends on what words mean

5. Understanding is balancing input with knowledge

Transfer with
unsupervised learning



The background is a solid dark blue. Overlaid on this are several thin, light blue lines that form abstract geometric shapes. On the left, a large, sweeping arc curves upwards and to the right. To the right of this, there are several straight lines that intersect to form a series of connected, angular shapes, resembling a stylized 'Z' or a series of steps. These lines are thin and light blue, providing a subtle structural framework for the text.

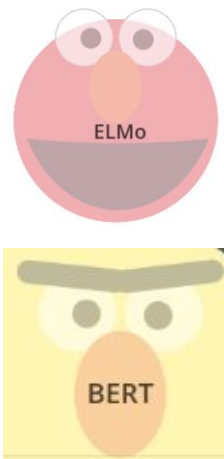
4

Extracting
language-relevant
knowledge from
the environment



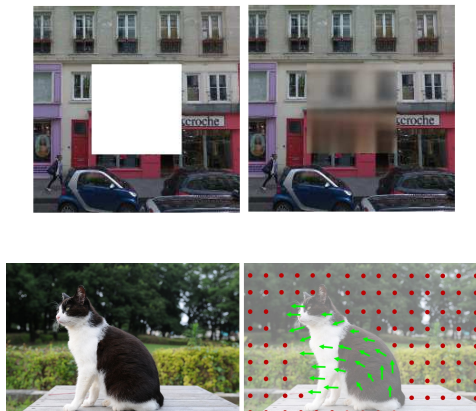
Building a multi-sensory understanding of the world

Language

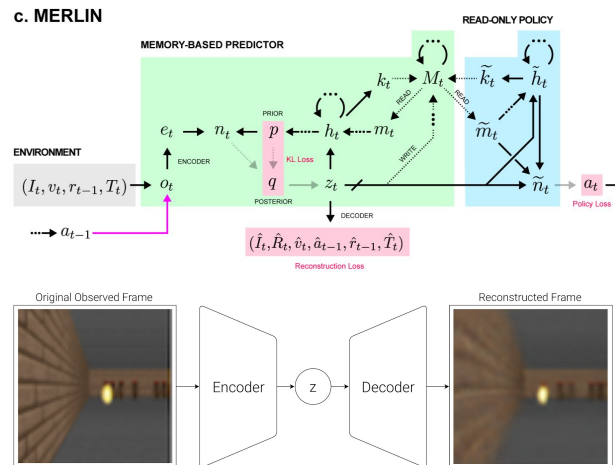


<http://jalammar.github.io/illustrated-bert>

Vision



Actions



Peters, Matthew E., et al. "Deep contextualized word representations." arXiv:1802.05365 (2018).

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv:1810.04805 (2018).

Pathak, Deepak, et al. "Context encoders: Feature learning by inpainting." CVPR 2016.

Pathak, Deepak, et al. "Learning features by watching objects move." CVPR 2017.

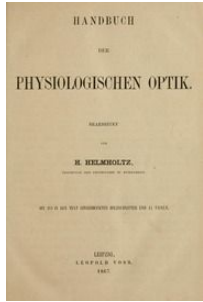
Wayne, Greg, et al. "Unsupervised predictive memory in a goal-directed agent." arXiv:1803.10760 (2018).

Ha, David, and Jürgen Schmidhuber. "World models." arXiv:1803.10122 (2018).



Knowledge aggregation from prediction

Helmholtz

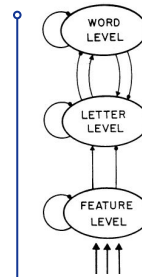


William James

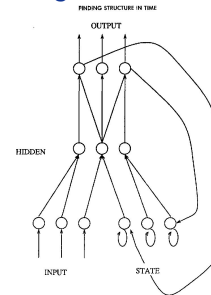
H. Barlow
Cortex as a model builder.

U. Neisser
Analysis by synthesis

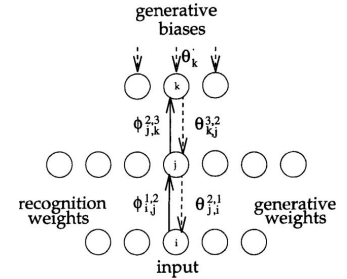
McClelland and Rumelhart
Interactive activation model



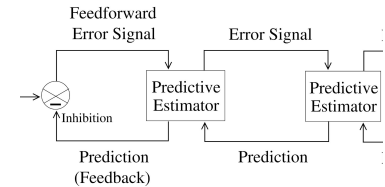
J. Elman
Finding structure in time



Dayan and Hinton
Helmholtz machine



Rao and Ballard



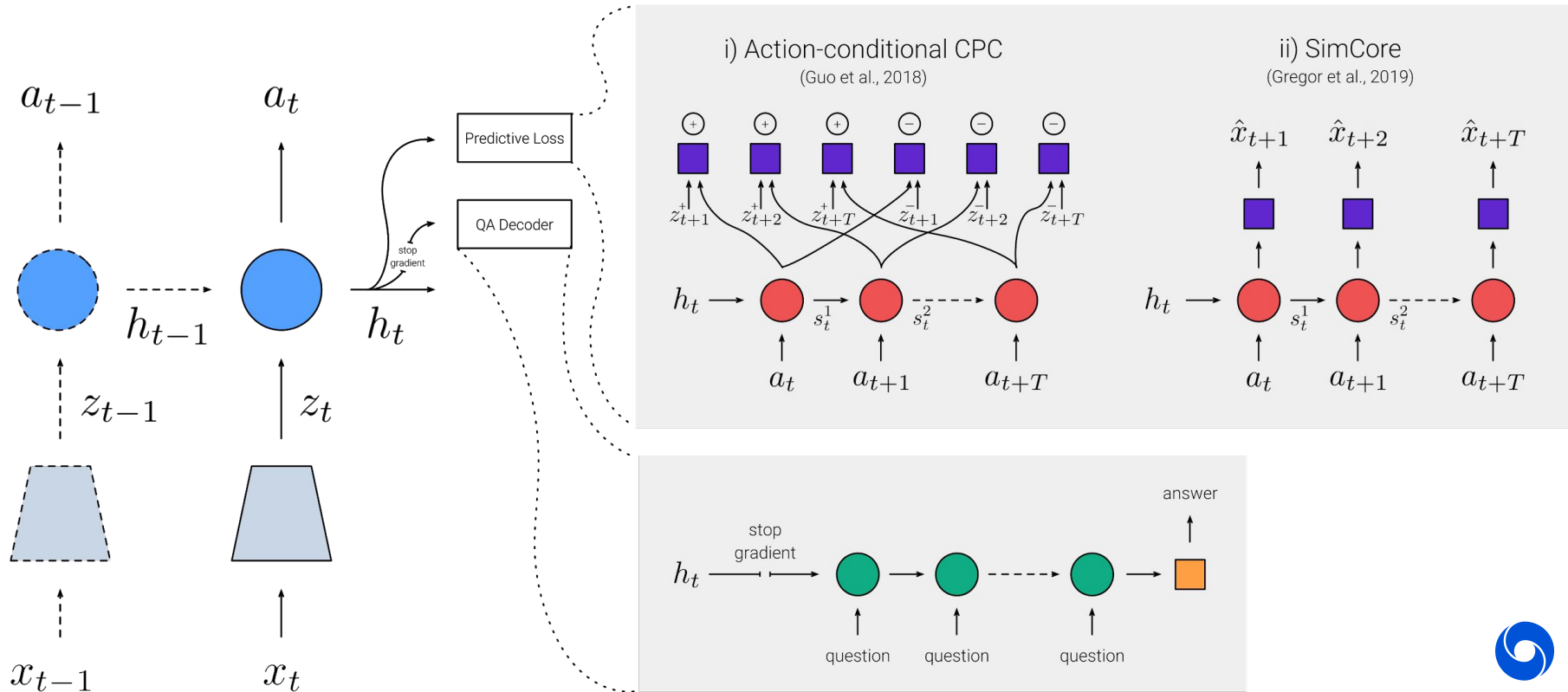


Questions to diagnose knowledge acquisition

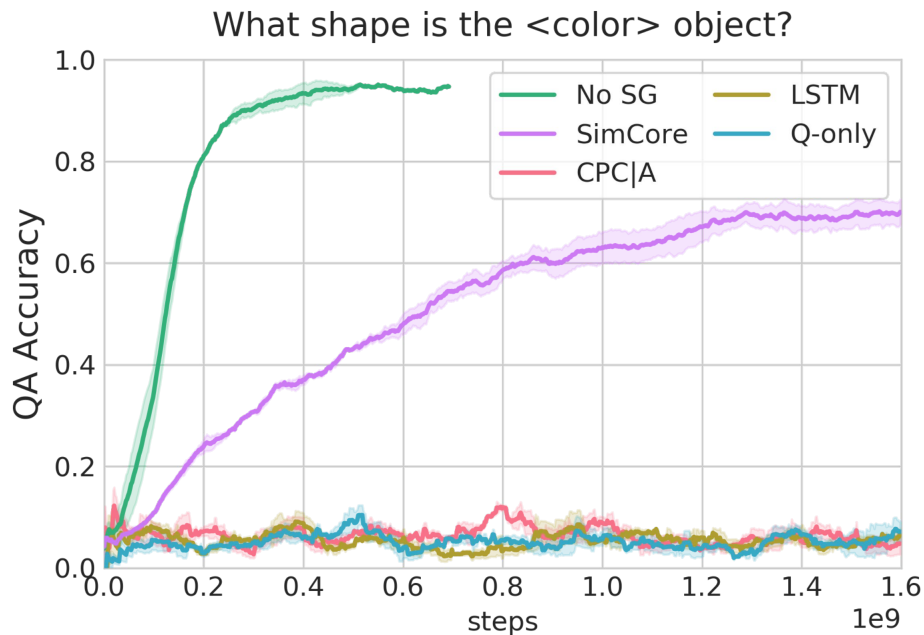
Question type	Template	# QA pairs
Attribute	What is the color of the <shape>?	500
	What shape is the <color> object?	500
Count	How many <shape> are there?	200
	How many <color> objects are there?	40
Exist	Is there a <shape>?	100
Compare + Count	Are there the same number of <color1> objects as <color2> objects?	180
	Are there the same number of <shape1> as <shape2>?	4900
Relation + Attribute	What is the color of the <shape1> near the <shape2>?	24500
	What is the <color> object near the <shape>?	25000



Predictive agents



Results



Oracle:

- Without stop-gradient

Baselines:

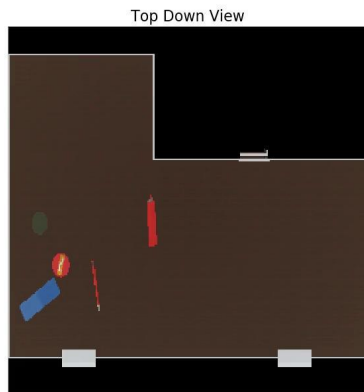
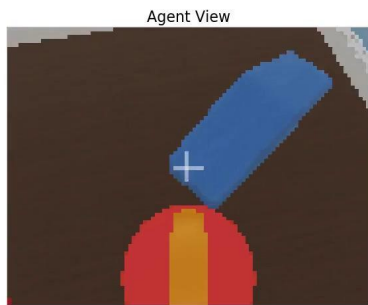
- Question only
- LSTM

Predictive Models:

- SimCore
- CPC|A



Results



Language

Question: What is the color of the pencil ?

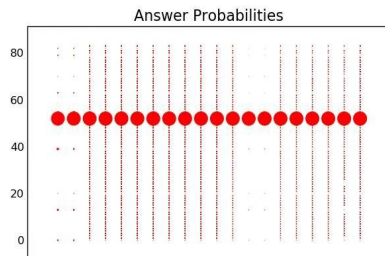
True answer: red

Predicted Answers

red 0.9987

green 0.0008

comb OR purple 0.0002



DeepMind

To conclude



1. Words have many related senses
2. Disambiguation depends on context
3. Relevant context can be non-local **and non linguistic**
4. 'Composition' depends on what words mean
5. Understanding requires background knowledge....
not always from language



Transformers

1. Words have many related senses
2. Disambiguation depends on context
3. Relevant context can be non-local **and non linguistic**
4. 'Composition' depends on what words mean
5. Understanding requires background knowledge....
not always from language



Transformers

1. Words have many related senses
2. Disambiguation depends on context
3. Relevant context can be non-local **and non linguistic**
4. 'Composition' depends on what words mean

Self-supervised / unsupervised learning

5. Understanding requires background knowledge....

not always from language



Transformers

1. Words have many related senses

2. Disambiguation depends on context

3. Relevant context can be non-local

4. 'Composition' depends on what words mean

Self-supervised / unsupervised learning

5. Understanding requires background knowledge....

not always from language

Embodied learning

and non linguistic



Transformers

Embodied learning

1. Words have many related senses
2. Disambiguation depends on context
3. Relevant context can be non-local
4. 'Composition' depends on what words mean

and non linguistic

Self-supervised / unsupervised learning

5. Understanding requires background knowledge....

not always from language

Fast-mapping

Goal-directed dialogue

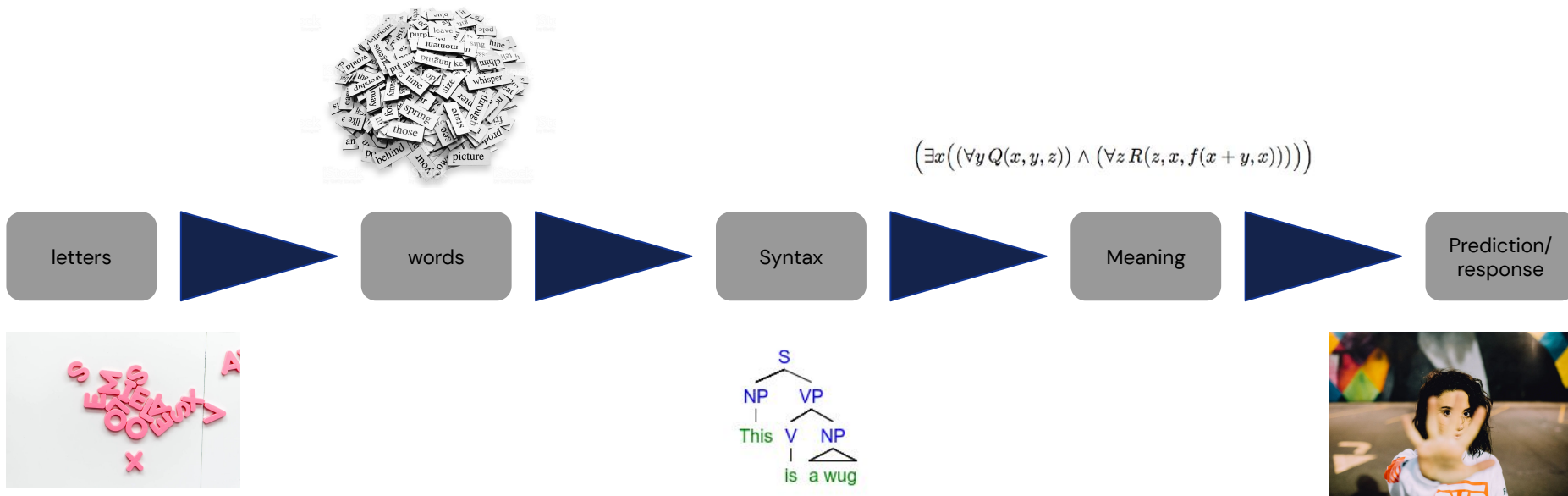
Understanding intentions

Social learning

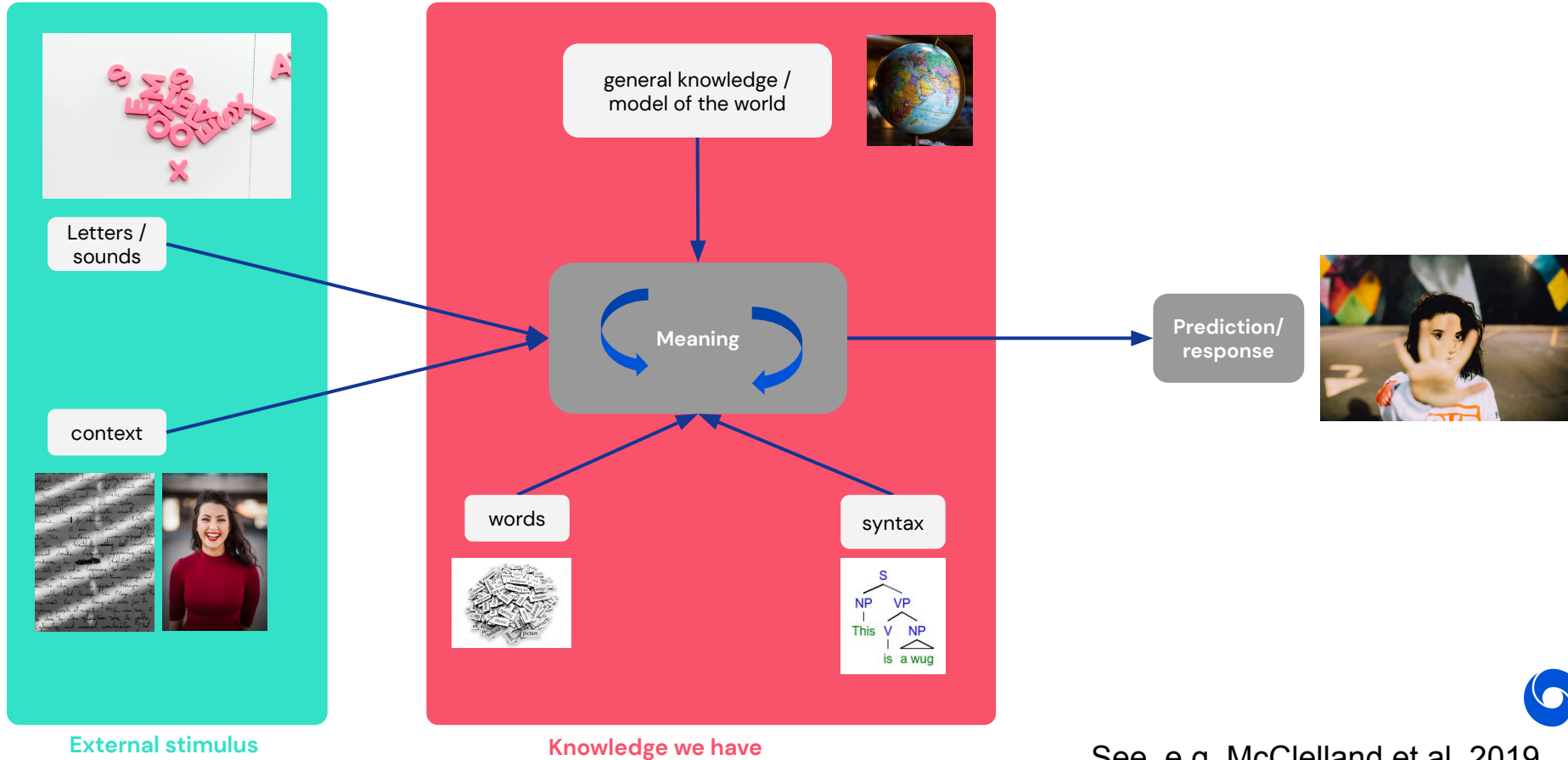
Event cognition



The 'pipeline' view of language processing



An alternative schematic model of language processing



See. e.g. McClelland et al. 2019



Selected references

Early treatment of distributed representations in neural language models

Natural language processing with modular PDP networks and distributed lexicon.
Mikkuilainen, Risto, and Michael G. Dyer. *Cognitive Science* 1991

The transformer architecture

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin:
Attention Is All You Need. Neurips 2017

BERT

Bert: Pre-training of deep bidirectional transformers for language understanding.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. NAACL 2019.

Embodied language learning at DeepMind

Environmental Drivers of Generalization and Systematicity in the Situated Agent
Hill et al. ICLR 2020

Robust Instruction-Following in a Situated Agent via Transfer Learning from Text

Hill et al. Under review

Probing emergent semantic knowledge in predictive agents via question answering

Carnevale et al. Under review



Thank you

