TensorFlow

NLP using Transformer Architectures

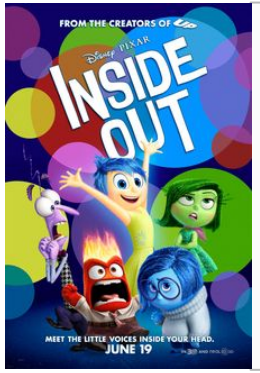
TF World 2019

Aurélien Géron
ML Consultant
@aureliengeron

- NLP Tasks and Datasets
- The Transformer Architecture
- Recent Language Models

Natural Language Processing

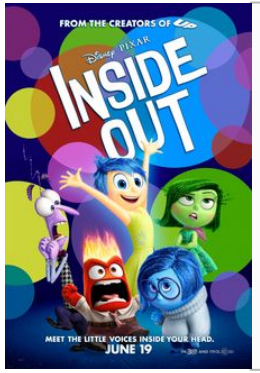
Sentiment Analysis



An artistic triumph
21 September 2015

For some reason, I couldn't quite catch this movie in theaters and I managed to watch it on an international flight. And boy, am I glad I did!

Sentiment Analysis

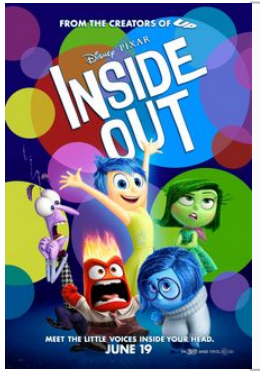


An artistic triumph
21 September 2015

For some reason, I couldn't quite catch this movie in theaters and I managed to watch it on an international flight. And boy, am I glad I did!

POSITIVE ?

Sentiment Analysis



★ 9/10 → **LABEL = POSITIVE**

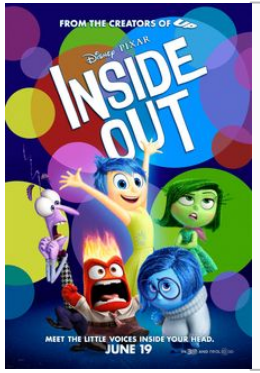
An artistic triumph
21 September 2015

For some reason, I couldn't quite catch this movie in theaters and I managed to watch it on an international flight. And boy, am I glad I did!

↓

POSITIVE ?

Sentiment Analysis



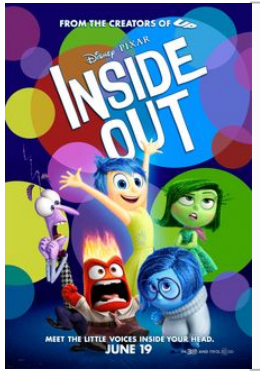
★ 9/10

An artistic triumph
21 September 2015

For some reason, I couldn't quite catch this movie in theaters and I managed to watch it on an international flight. And boy, am I glad I did!

- Recommender Systems
- Market Sentiment

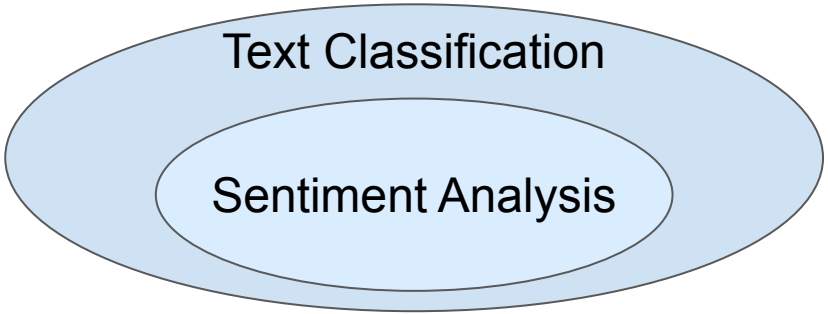
Sentiment Analysis



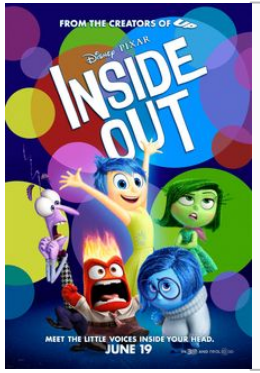
★ 9/10

An artistic triumph
21 September 2015

For some reason, I couldn't quite catch this movie in theaters and I managed to watch it on an international flight. And boy, am I glad I did!



Sentiment Analysis



★ 9/10

An artistic triumph
21 September 2015

For some reason, I couldn't quite catch this movie in theaters and I managed to watch it on an international flight. And boy, am I glad I did!

```
!pip install tensorflow-datasets # or tfds-nightly

import tensorflow_datasets as tfds

datasets = tfds.load("imdb_reviews")
```

TensorFlow Datasets

```
datasets = tfds.load("imdb_reviews")
train_set = datasets["train"] # 25,000 reviews
test_set = datasets["test"]   # 25,000 reviews
```

TensorFlow Datasets

```
train_set, test_set = tfds.load(
    "imdb_reviews",
    split=["train", "test"])
```

TensorFlow Datasets

```
train_set, test_set = tfds.load(
    "imdb_reviews:1.0.0",
    split=["train", "test"])
```

TensorFlow Datasets

```
train_set, test_set = tfds.load(
    "imdb_reviews:1.0.0",
    split=["train", "test[:60%]"])
```

TensorFlow Datasets

```
train_set, test_set, valid_set = tfds.load(
    "imdb_reviews:1.0.0",
    split=["train", "test[:60%]", "test[60%:]"])
```

TensorFlow Datasets

```
train_set, test_set, valid_set = tfds.load(
    "imdb_reviews:1.0.0",
    split=["train", "test[:60%]", "test[60%:]"],
    as_supervised=True)
```

TensorFlow Datasets

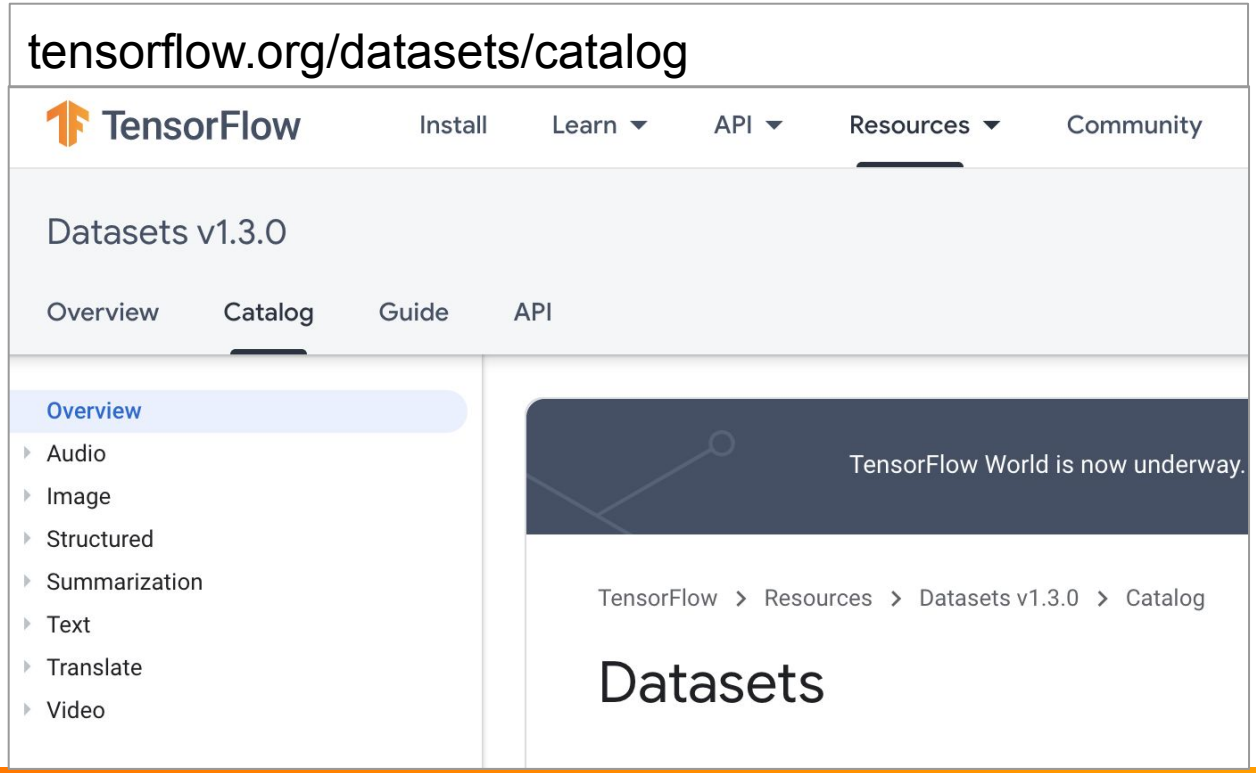
```
for review, label in train_set.take(2):
  print(review.numpy().decode("utf-8"))
  print(label.numpy())
```

TensorFlow Datasets

```
for review, label in train_set.take(2):
  print(review.numpy().decode("utf-8"))
  print(label.numpy())
```

```
This was an absolutely terrible movie. Don't be lured in by
Christopher Walken or Michael Ironside. Both are great actors, but
this must simply be their worst role in history.[...]
0
This is the kind of film for a snowy Sunday afternoon when the
rest of the world can go ahead with its own [...] A family film in
every sense and one that deserves the praise it received.
1
```

TensorFlow Datasets



Translation

"Das Leben ist wunderbar."

"Life is wonderful."

Translation

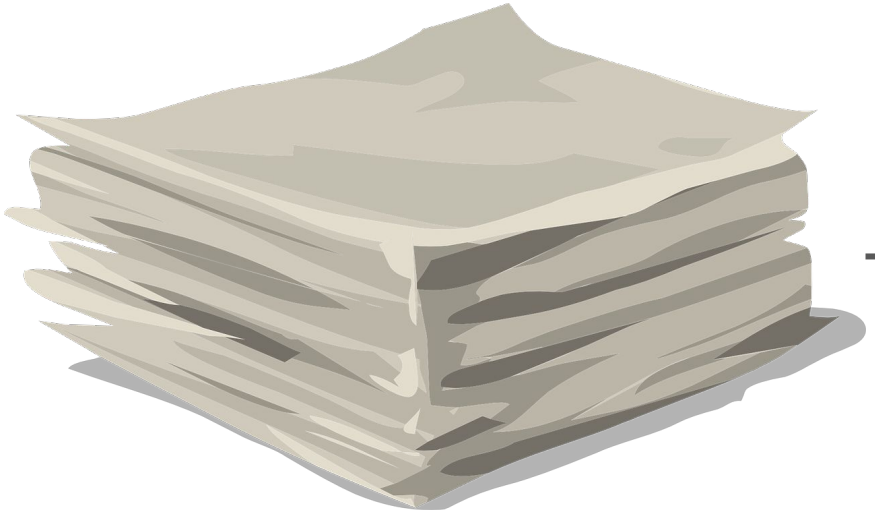
"Das Leben ist wunderbar."

"Life is wonderful."

Workshop on Machine Translation (WMT)

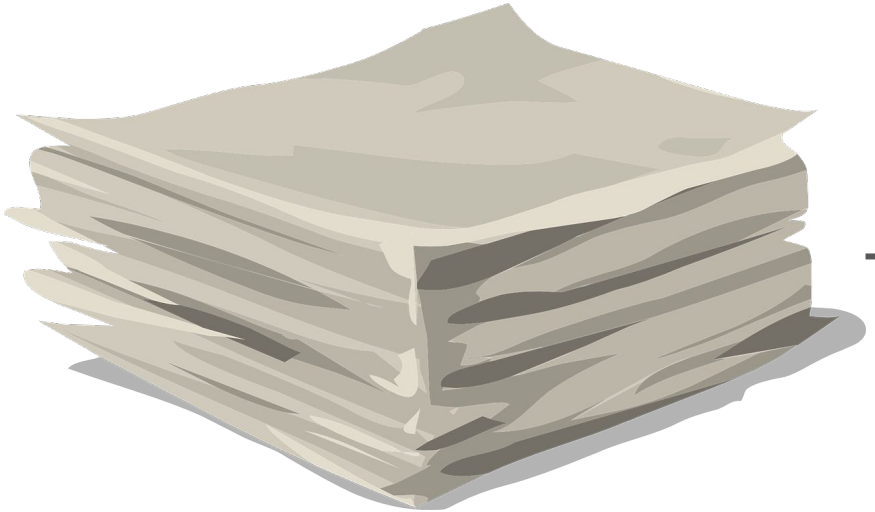
```
datasets = tfds.load("wmt19_translate/de-en") # 10 GB!
```

Summarization



"Popular YouTubers raise 10 million dollars in 5 days to plant 10 million trees, in an effort to fight global warming."

Summarization



"Popular YouTubers raise 10 million dollars in 5 days to plant 10 million trees, in an effort to fight global warming."

Predict abstract from news articles

```
datasets = tfds.load("multi_news")
```

Question Answering

Context

Southern California, often abbreviated SoCal, is a geographic and cultural region that generally comprises California's southernmost 10 counties. The region is traditionally described as "eight counties", based on demographics and [...]

Question

What is Southern California often abbreviated as?

Answer

SoCal

Question Answering

Context

> Southern California, often abbreviated SoCal, is a geographic and cultural region that generally comprises California's southernmost 10 counties. The region is traditionally described as "eight counties", based on demographics and [...]

Question

> What is Southern California often abbreviated as?

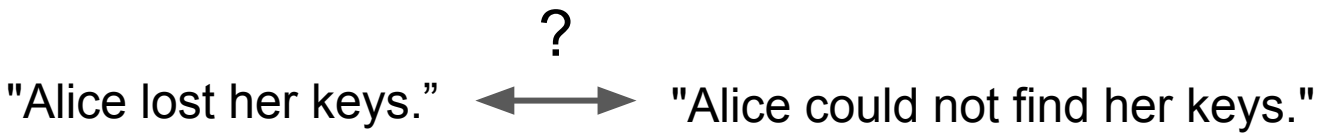
Answer

> SoCal

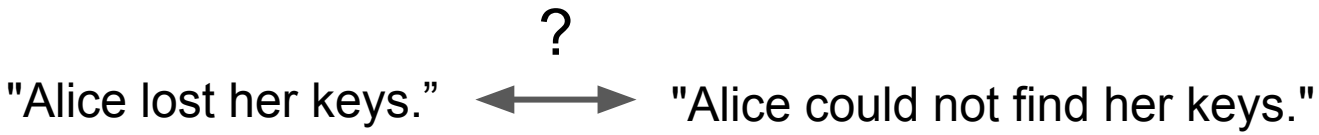
Stanford Question Answering Dataset (SQuAD)

```
datasets = tfds.load("squad")
```

Semantic Equivalence



Semantic Equivalence



Microsoft Research Paraphrase Corpus

```
datasets = tfds.load("glue/mrpc")
```

Entailment

"Alice lost her keys, but Betty found them and returned them to her.”

"Alice got her keys back."

Entailment

"Alice lost her keys, but Betty found them and returned them to her."

Entails

"Alice got her keys back."

Entailment

"Alice lost her keys, but Betty found them and returned them to her."

Contradicts

"Alice lost her keys forever."

Entailment

"Alice lost her keys, but Betty found them and returned them to her."

"Alice and Betty are sisters."

Neutral

Entailment

"Alice lost her keys, but Betty found them and returned them to her."

Neutral

"Alice and Betty are sisters."

Stanford Natural Language Inference Dataset

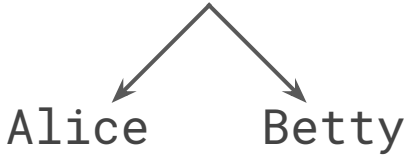
```
datasets = tfds.load("snli/plain_text")
```

Coreference Resolution

"Alice lost her keys, but Betty found them and returned them to **her**."

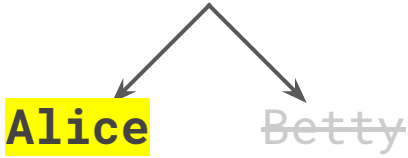
Coreference Resolution

"Alice lost her keys, but Betty found them and returned them to **her**."



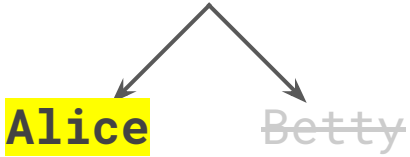
Coreference Resolution

"Alice lost her keys, but Betty found them and returned them to **her**."



Coreference Resolution

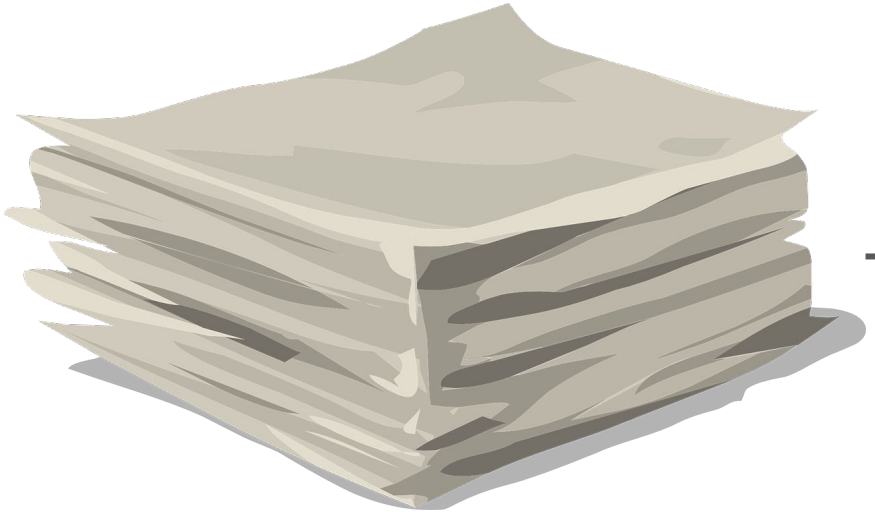
"Alice lost her keys, but Betty found them and returned them to **her**."

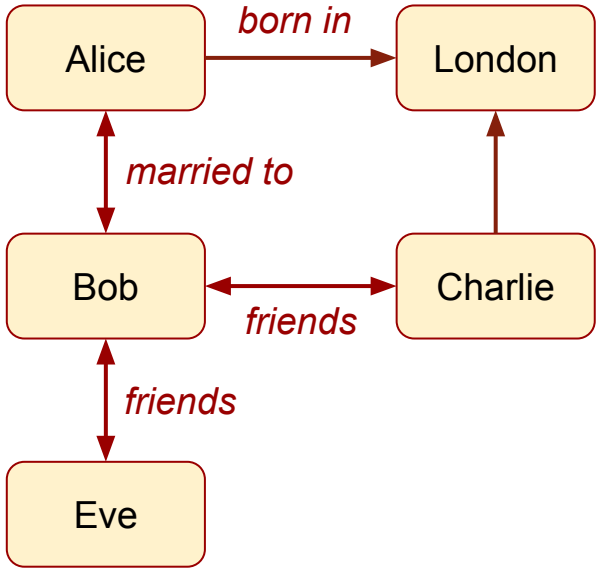


Stanford Natural Language Inference Dataset

```
datasets = tfds.load("gap")
```

Information Extraction

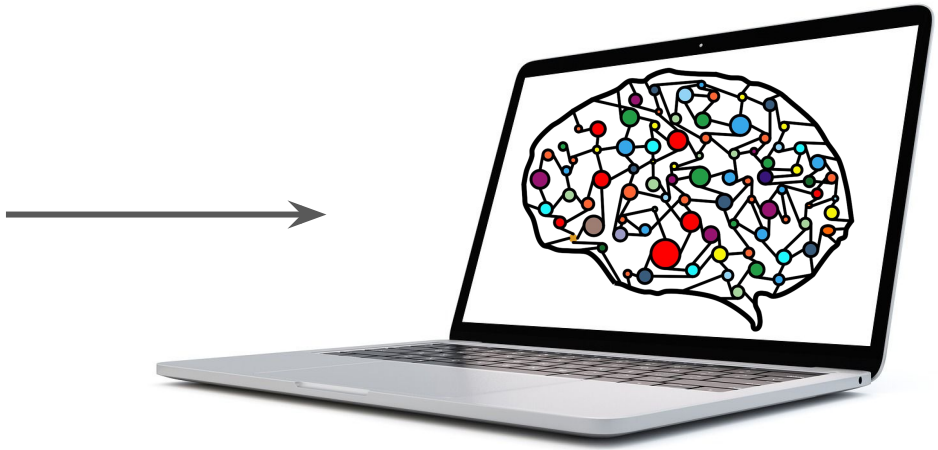




Speech to Text

Text to Speech

Optical Character Recognition



Metrics

Dataset	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**
Wikipedia	16GB	81.85	**19.31**	81.29	68.01	**26.94**	39.69	**27.67**
Wikipedia + TBC	20GB	83.65	**19.28**	**82.08**	**73.24**	**26.77**	39.63	**27.57**

Table 8: Performance resulting from pre-training on different datasets. The first four variants are based on our new C4 dataset.

Source: "*Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*" by Raffel, Shazeer, Roberts, Lee et al., https://arxiv.org/abs/1910.10683

TensorFlow

Transformer Architecture

Attention Is All You Need

Attention Is All You Need

https://arxiv.org/abs/1706.03762

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

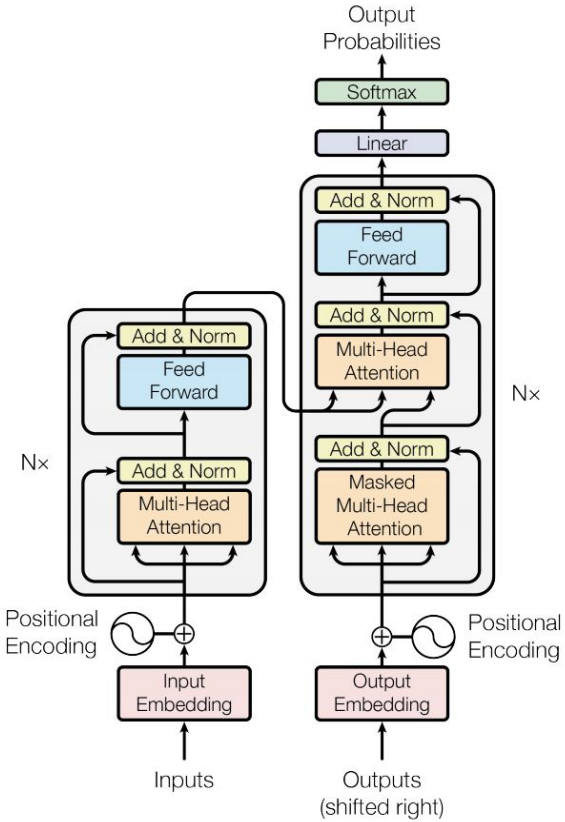
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

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

Attention Is All You Need



Source: Figure 1 from the paper "Attention Is All You Need" by Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser and Polosukhin https://arxiv.org/abs/1706.03762

BERT
Oct 2018

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language repr
tion model called **BERT**, which star
Bidirectional **E**ncoder **R**epresentation
Transformers. Unlike recent language
sentation models (Peters et al., 2018; R
et al., 2018), BERT is designed to pr
deep bidirectional representations by
conditioning on both left and right cor
all layers. As a result, the pre-trained
representations can be fine-tuned with j
additional output layer to create state-
art models for a wide range of tasks
as question answering and language inf
without substantial task-specific archi
modifications.

BERT is conceptually simple and emp
powerful. It obtains new state-of-the-
sults on eleven natural language proc

GPT-2
Feb 2019

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] **Jeffrey Wu** [*1] **Rewon Child** [1] **David Luan** [1] **Dario Amodei** [**1] **Ilya Sutskever** [**1]

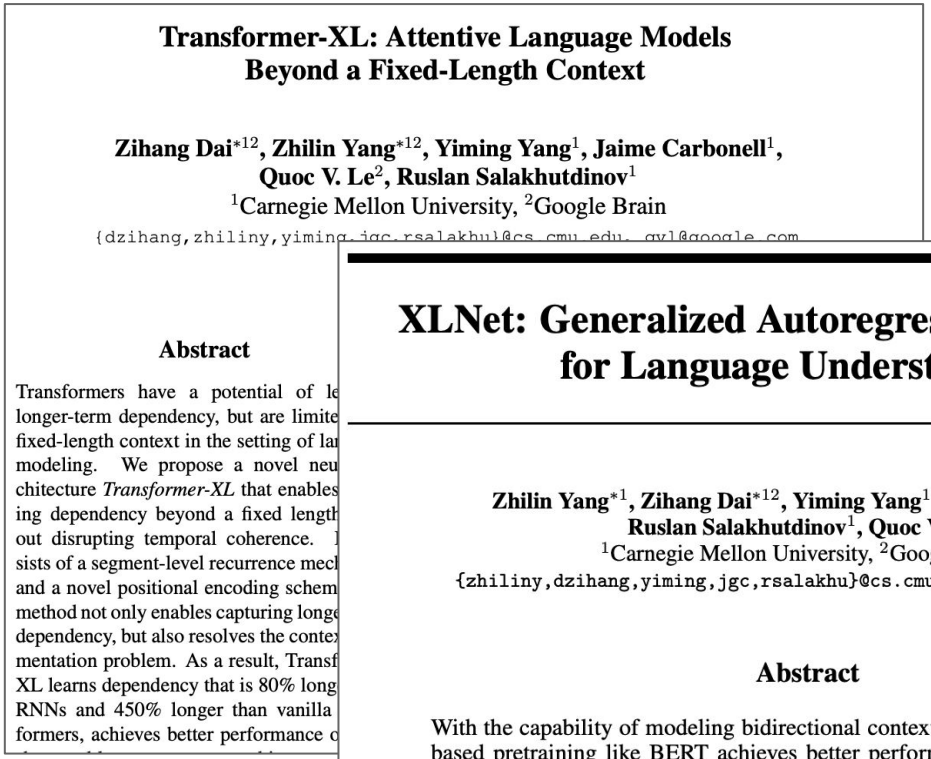
Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest

competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al., 2017), reading comprehension systems (Jia & Liang, 2017), and image classifiers (Alcorn et al., 2018) on the diversity and variety of possible inputs highlights some of the shortcomings of this approach.

Our suspicion is that the prevalence of single task training on single domain datasets is a major contributor to the lack of generalization observed in current systems. Progress towards robust systems with current architectures is likely to require training and measuring performance on a wide range of domains and tasks. Recently, several benchmarks have been proposed such as GLUE (Wang et al., 2018) and decaNLP (McCann et al., 2018) to begin studying this.



Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context

Zihang Dai[*12], **Zhilin Yang**[*12], **Yiming Yang**[1], **Jaime Carbonell**[1], **Quoc V. Le**[2], **Ruslan Salakhutdinov**[1]
[1]Carnegie Mellon University, [2]Google Brain
{dzihang,zhiliny,yiming,jgc,rsalakhu}@cs.cmu.edu, qvl@google.com

Abstract

Transformers have a potential of le
longer-term dependency, but are limite
fixed-length context in the setting of la
modeling. We propose a novel neu
chitecture *Transformer-XL* that enables
ing dependency beyond a fixed length
out disrupting temporal coherence.
sists of a segment-level recurrence mech
and a novel positional encoding schem
method not only enables capturing longe
dependency, but also resolves the contex
mentation problem. As a result, Transf
XL learns dependency that is 80% long
RNNs and 450% longer than vanilla
formers, achieves better performance o

Transformer-XL
Jan 2019

XLNet: Generalized Autoregressive Pretraining for Language Understanding

Zhilin Yang[*1], **Zihang Dai**[*12], **Yiming Yang**[1], **Jaime Carbonell**[1], **Ruslan Salakhutdinov**[1], **Quoc V. Le**[2]
[1]Carnegie Mellon University, [2]Google Brain
{zhiliny,dzihang,yiming,jgc,rsalakhu}@cs.cmu.edu, qvl@google.com

Abstract

With the capability of modeling bidirectional contexts, denoising autoencoding based pretraining like BERT achieves better performance than pretraining approaches based on autoregressive language modeling. However, relying on corrupting the input with masks, BERT neglects dependency between the masked positions and suffers from a pretrain-finetune discrepancy. In light of these pros and cons, we propose XLNet, a generalized autoregressive pretraining method that (1) enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order and (2) overcomes the limitations of BERT thanks to its autoregressive formulation. Furthermore, XLNet integrates ideas from Transformer-XL, the state-of-the-art autoregressive model, into pretraining. Empirically, XLNet outperforms BERT on 20 tasks, often by a large margin, and achieves state-of-the-art results on 18 tasks including question answering, natural

XLNet
Jun 2019

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu[*§] Myle Ott[*§] Naman Goyal[*§] Jingfei Du[*§] Mandar Joshi[†]
Danqi Chen[§] Omer Levy[§] Mike Lewis[§] Luke Zettlemoyer[†§] Veselin Stoyanov[§]

[†] Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA
{mandar90,lsz}@cs.washington.edu

{yinhanli
danqi,om

Abstract

Language model pretraining has le
nificant performance gains but care
parison between different approache
lenging. Training is computationall
sive, often done on private datasets of
sizes, and, as we will show, hyperp
choices have significant impact on th
sults. We present a replication study
pretraining (Devlin et al., 2019) that
measures the impact of many key hyp
eters and training data size. We find th
was significantly undertrained, and c
or exceed the performance of ever
published after it. Our best model
state-of-the-art results on GLUE, R
SQuAD. These results highlight th
tance of previously overlooked design

RoBERTa
Jul 2019

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel[*] Noam Shazeer[*] Adam Roberts[*] Katherine Lee[*]
Sharan Narang Michael Matena Yanqi Zhou Wei Li Peter J. Liu

Google

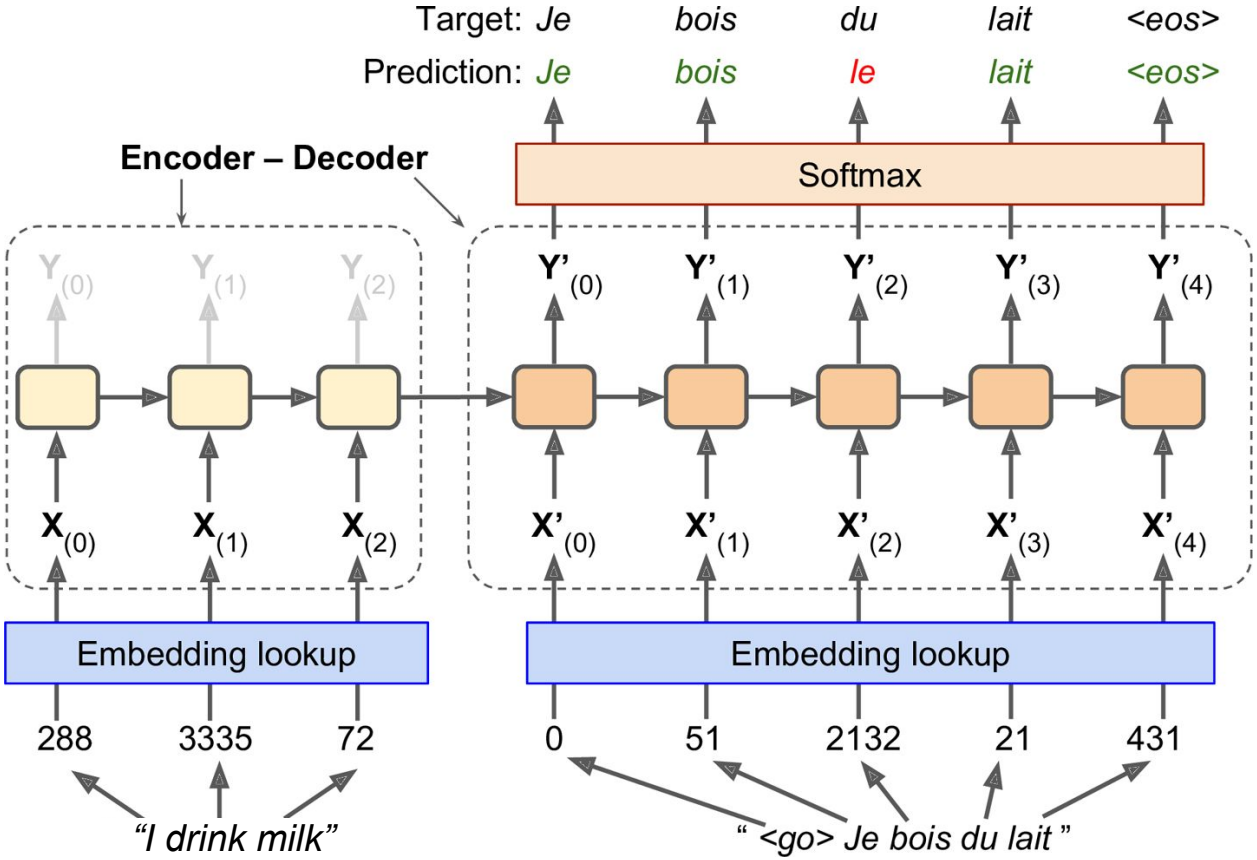
Abstract

Transfer learning, where a model is first pre-trained on a data-rich task before being fine-tuned on a downstream task, has emerged as a powerful technique in natural language processing (NLP). The effectiveness of transfer learning has given rise to a diversity of approaches, methodology, and practice. In this paper, we explore the landscape of transfer learning techniques for NLP by introducing a unified framework that converts every language problem into a text-to-text format. Our systematic study compares pre-training objectives, architectures, unlabeled datasets, transfer approaches, and other factors on dozens of language understanding tasks. By combining the insights from our exploration with scale and our new "Colossal Clean Crawled Corpus", we achieve state-of-the-art results on many benchmarks covering summarization, question answering, text classification, and more. To facilitate future work on transfer learning for NLP, we release our dataset, pre-trained models, and code.[1]

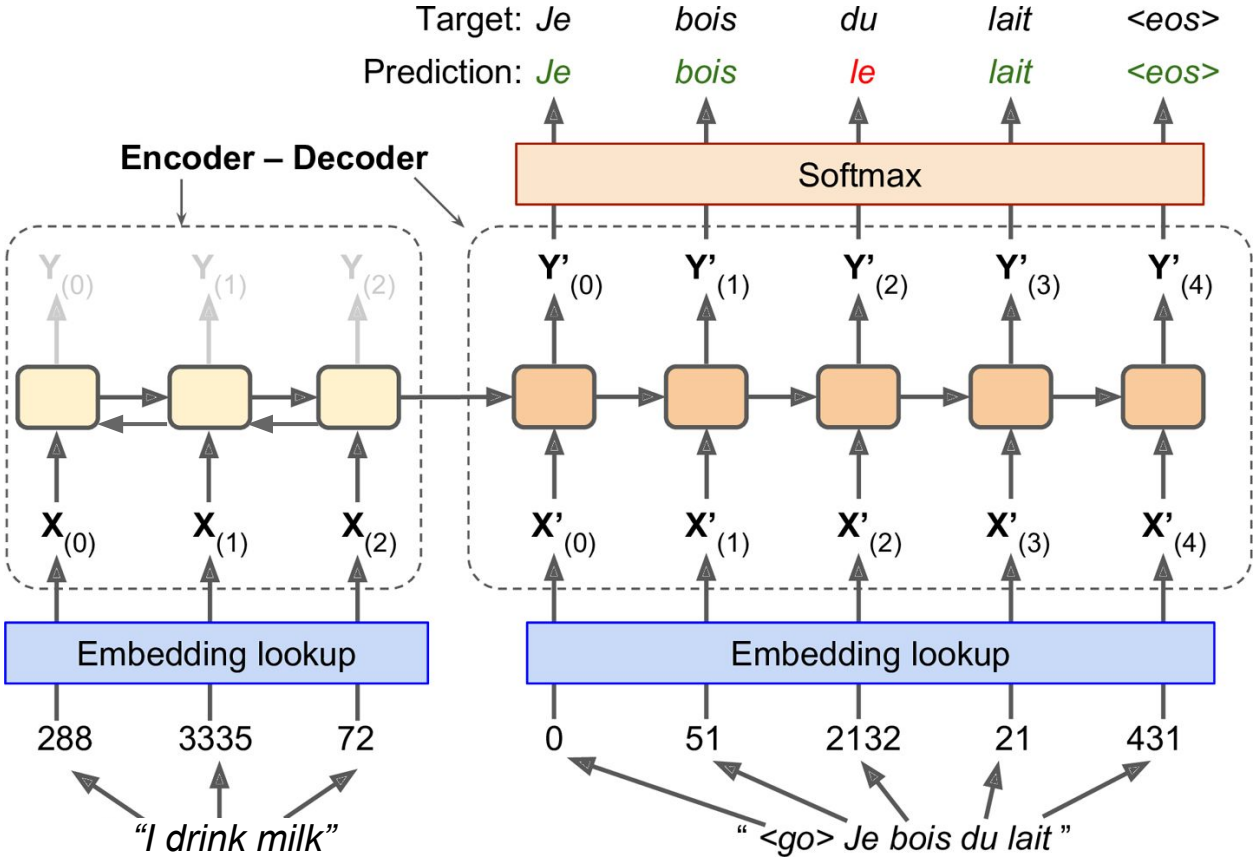
1 Introduction

T5
Oct 2019

LSTM Encoder/Decoder



BiLSTM Encoder/Decoder



Attention Mechanisms

Sep 2014

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

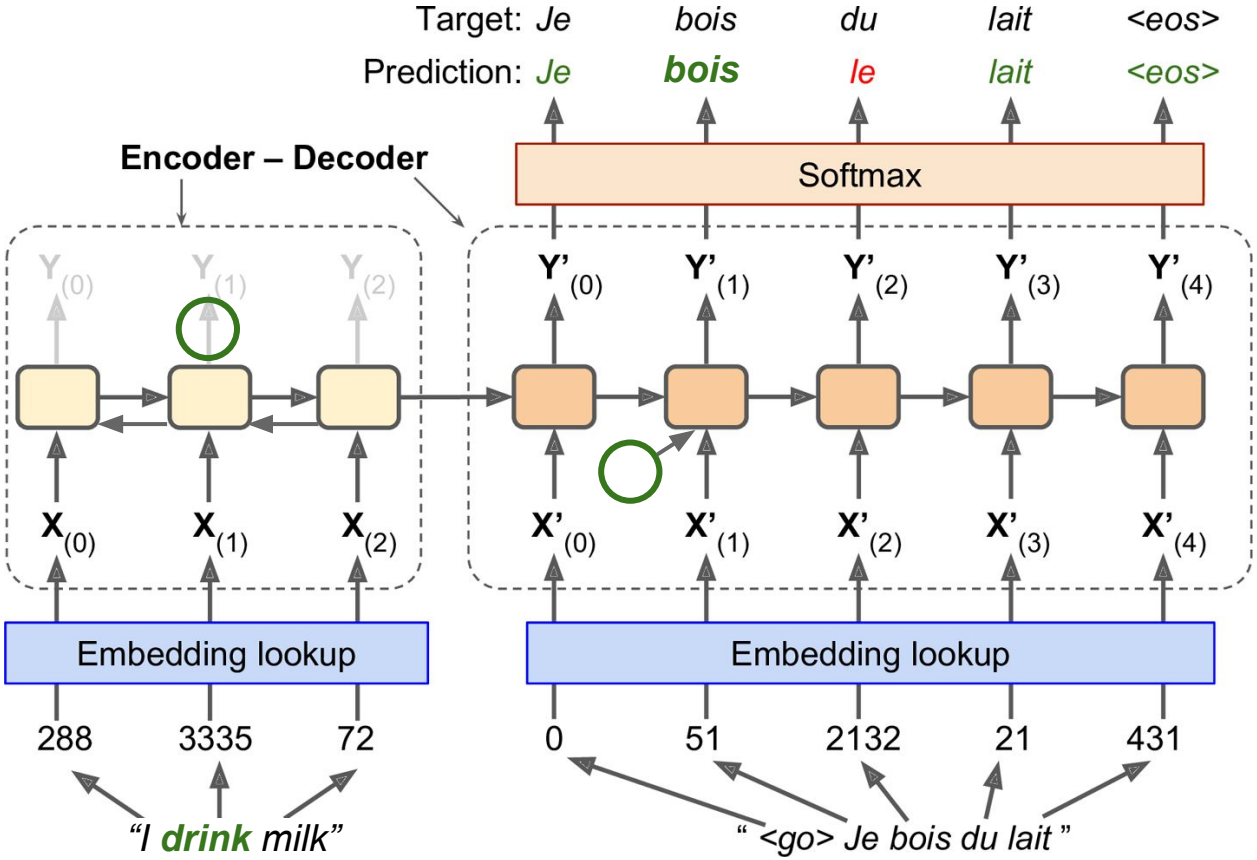
Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

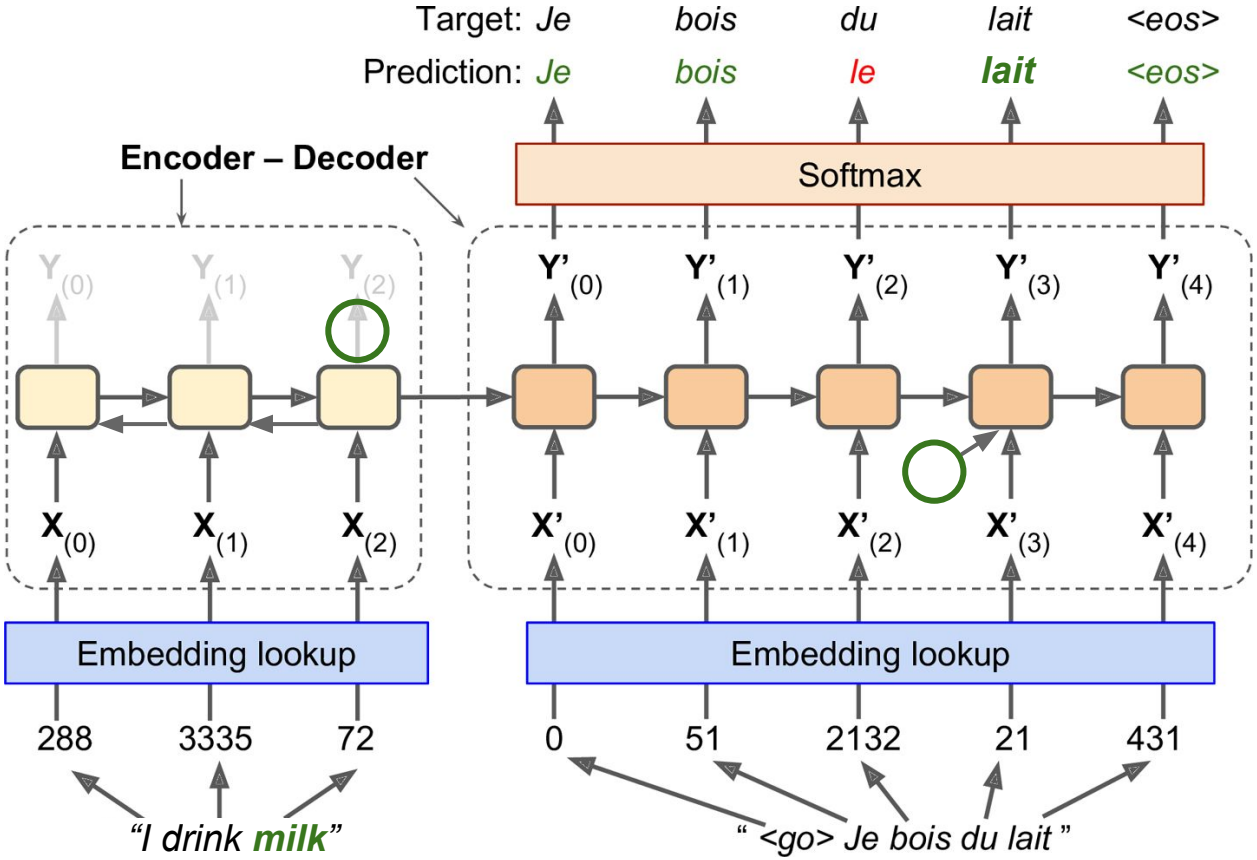
ABSTRACT

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.

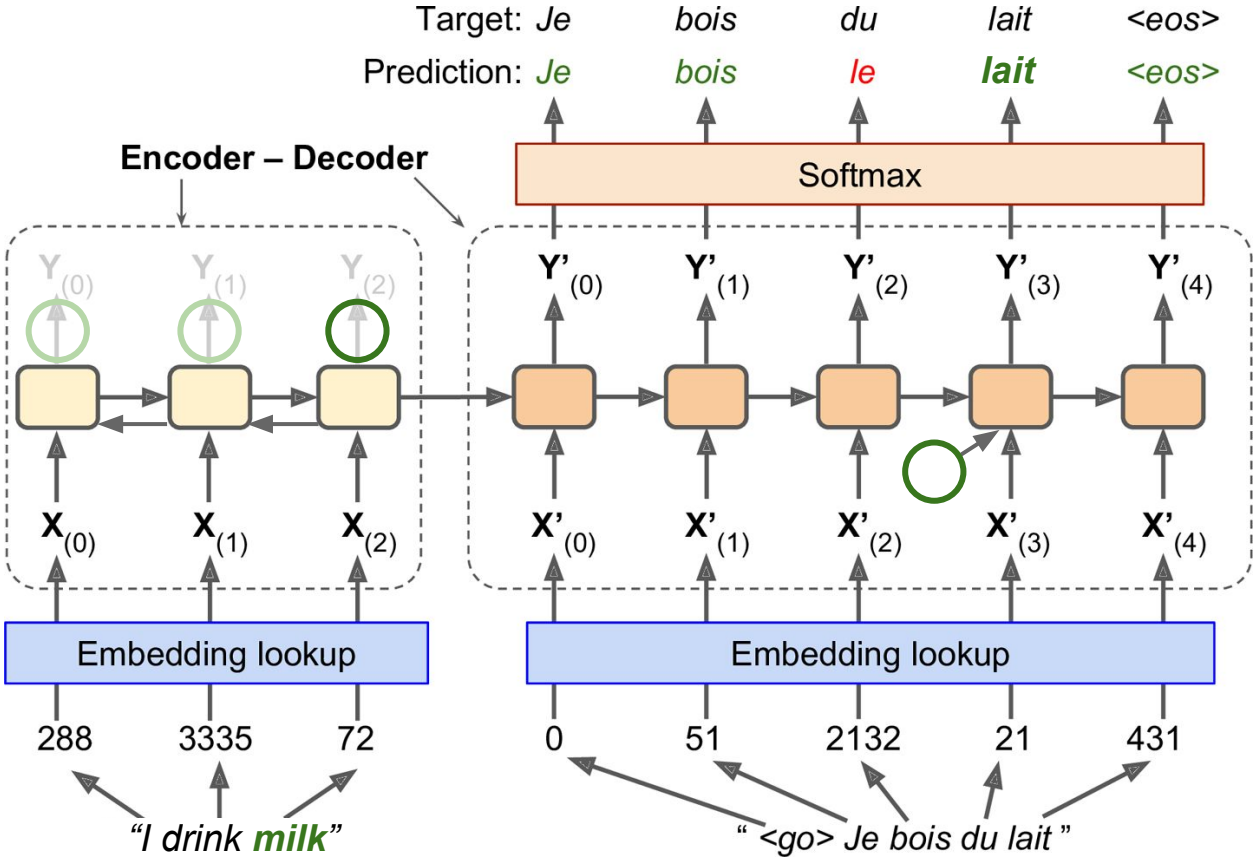
Encoder/Decoder with Attention



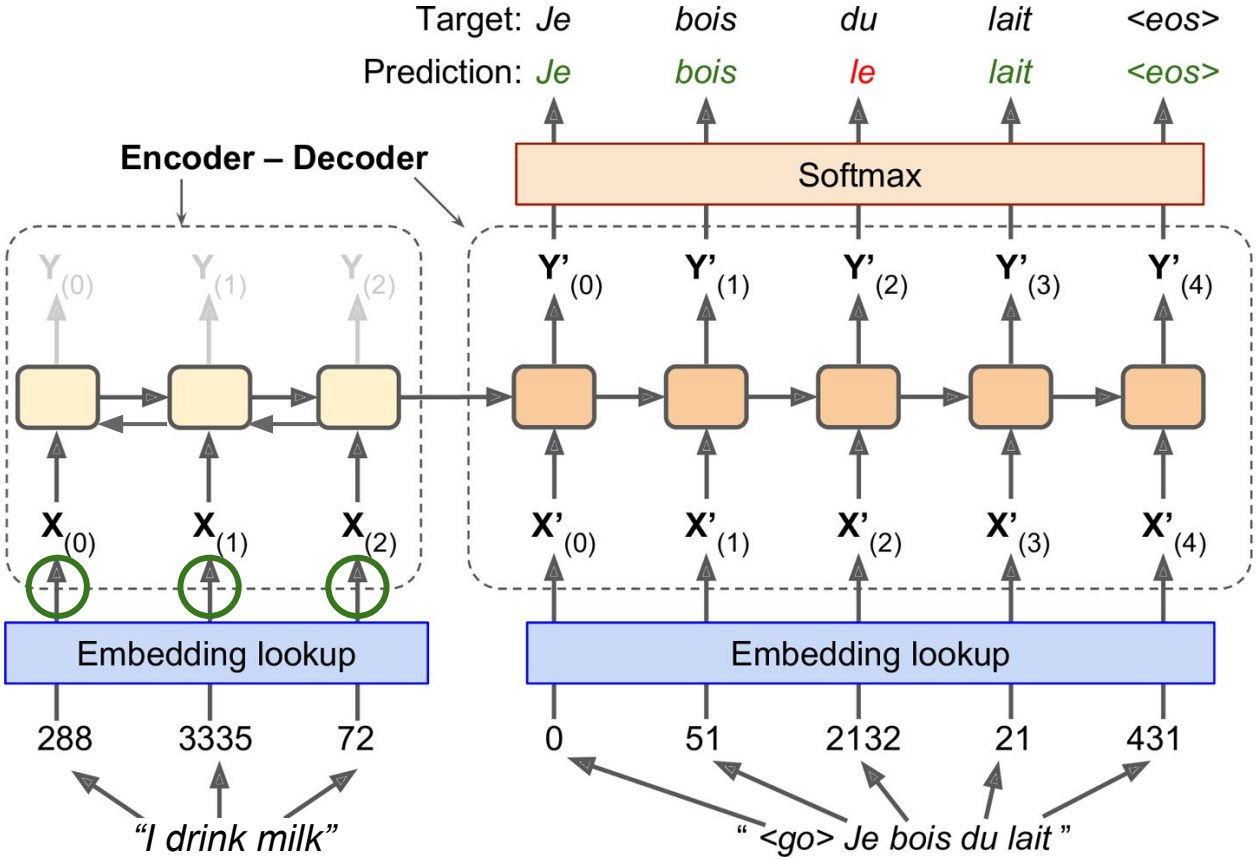
Encoder/Decoder with Attention



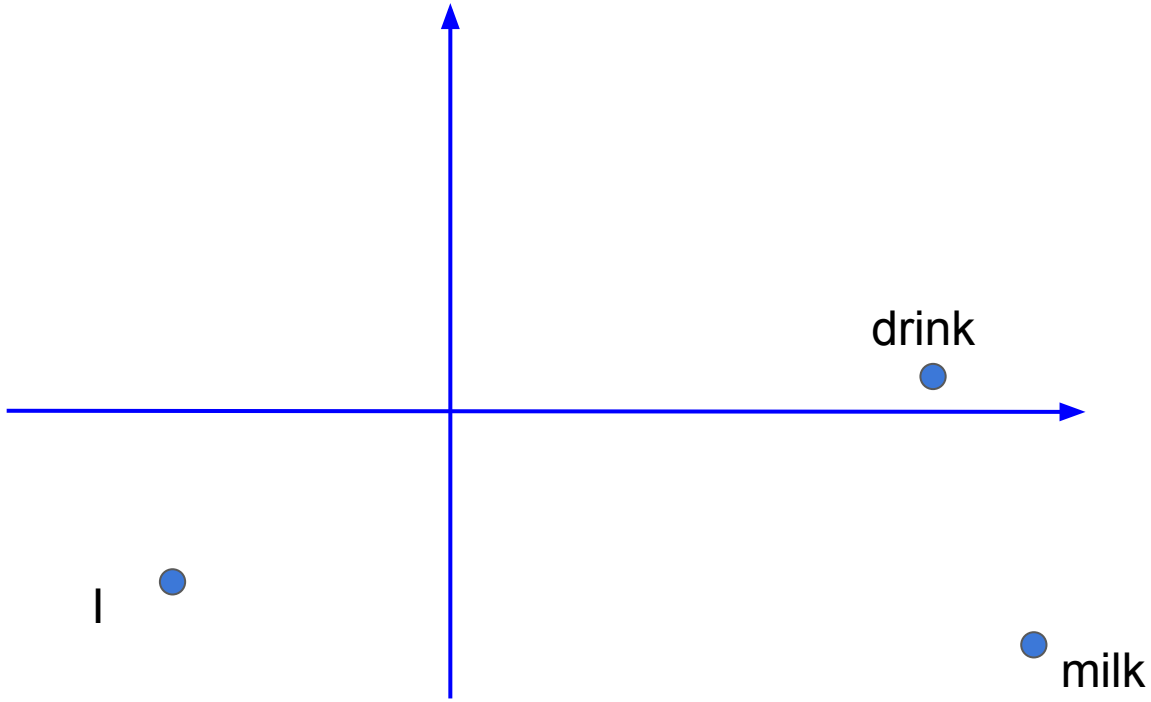
Encoder/Decoder with Attention



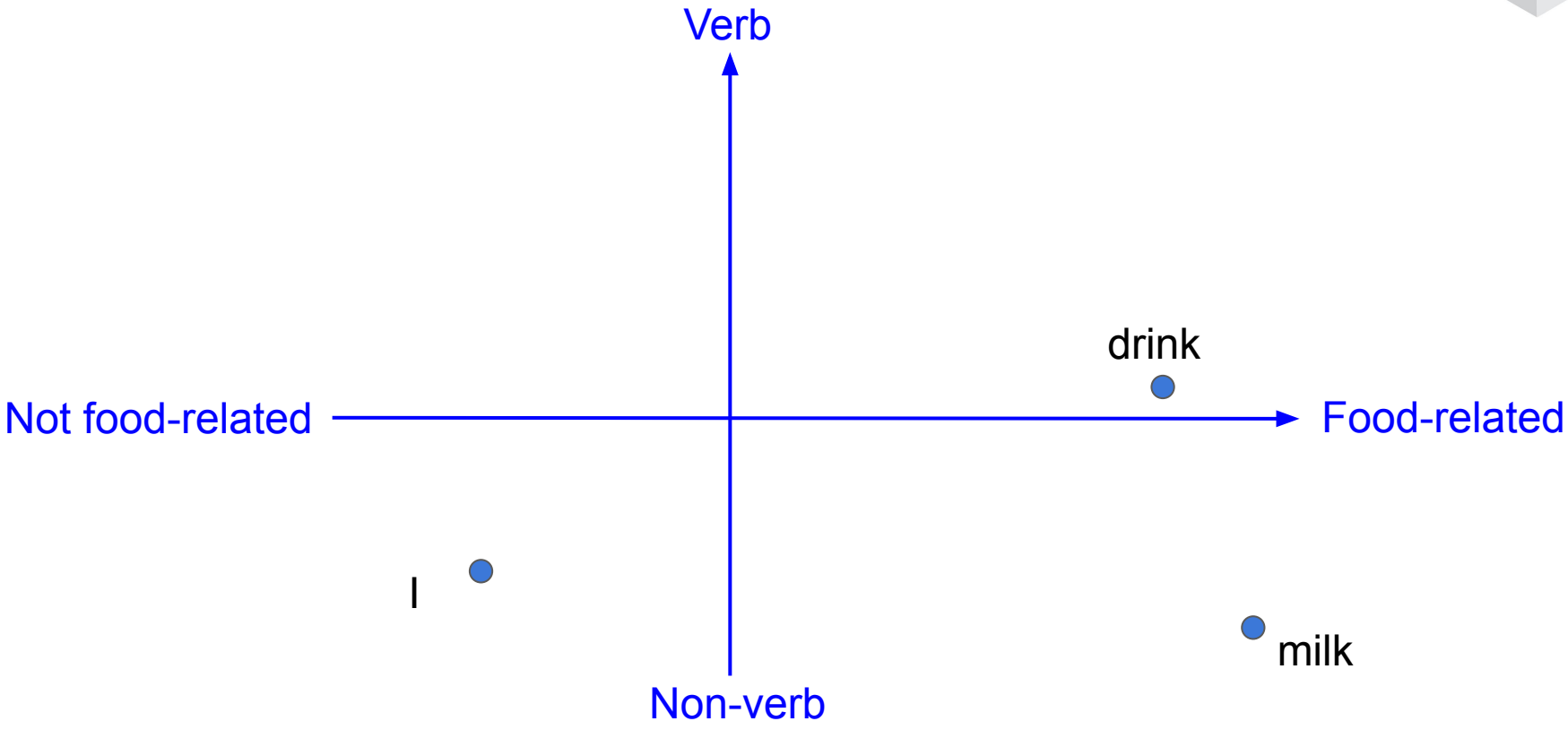
Encoder/Decoder with Attention



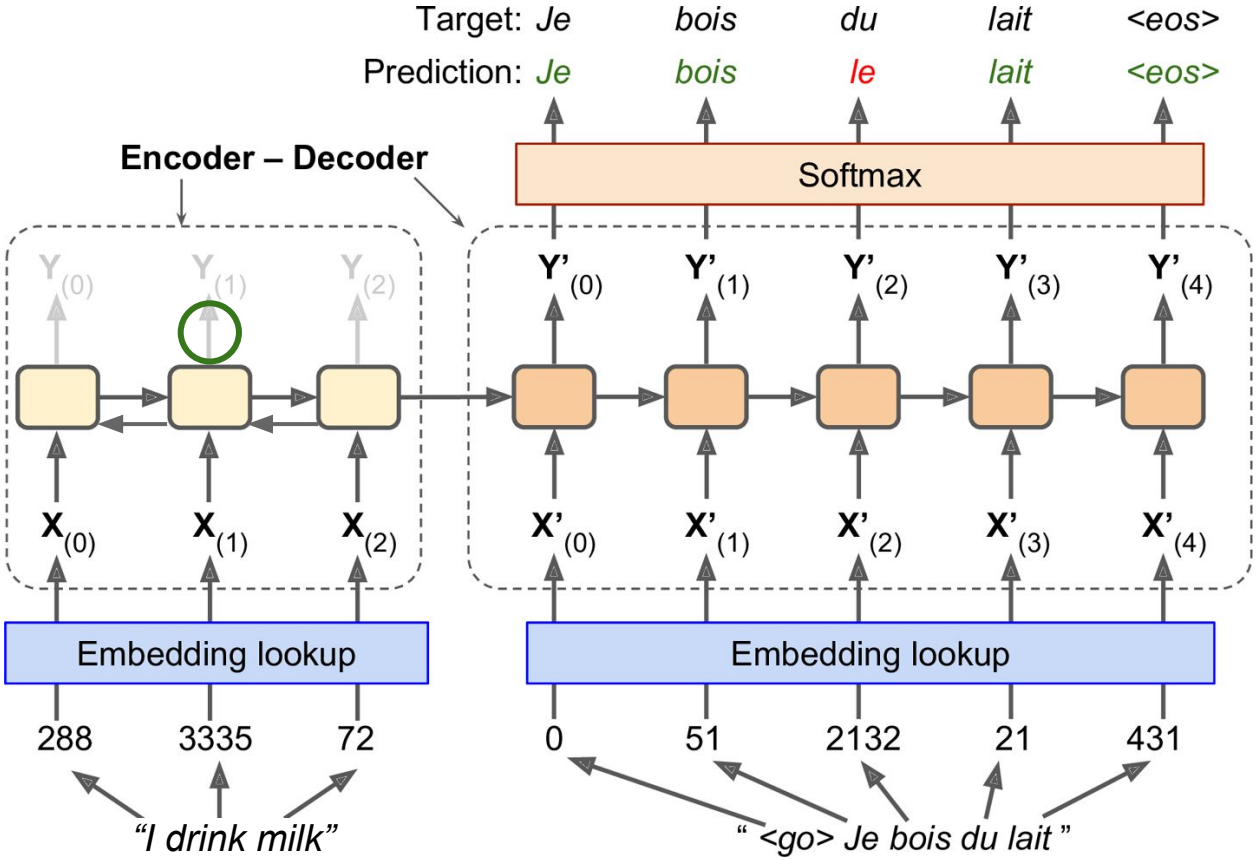
Attention Mechanism



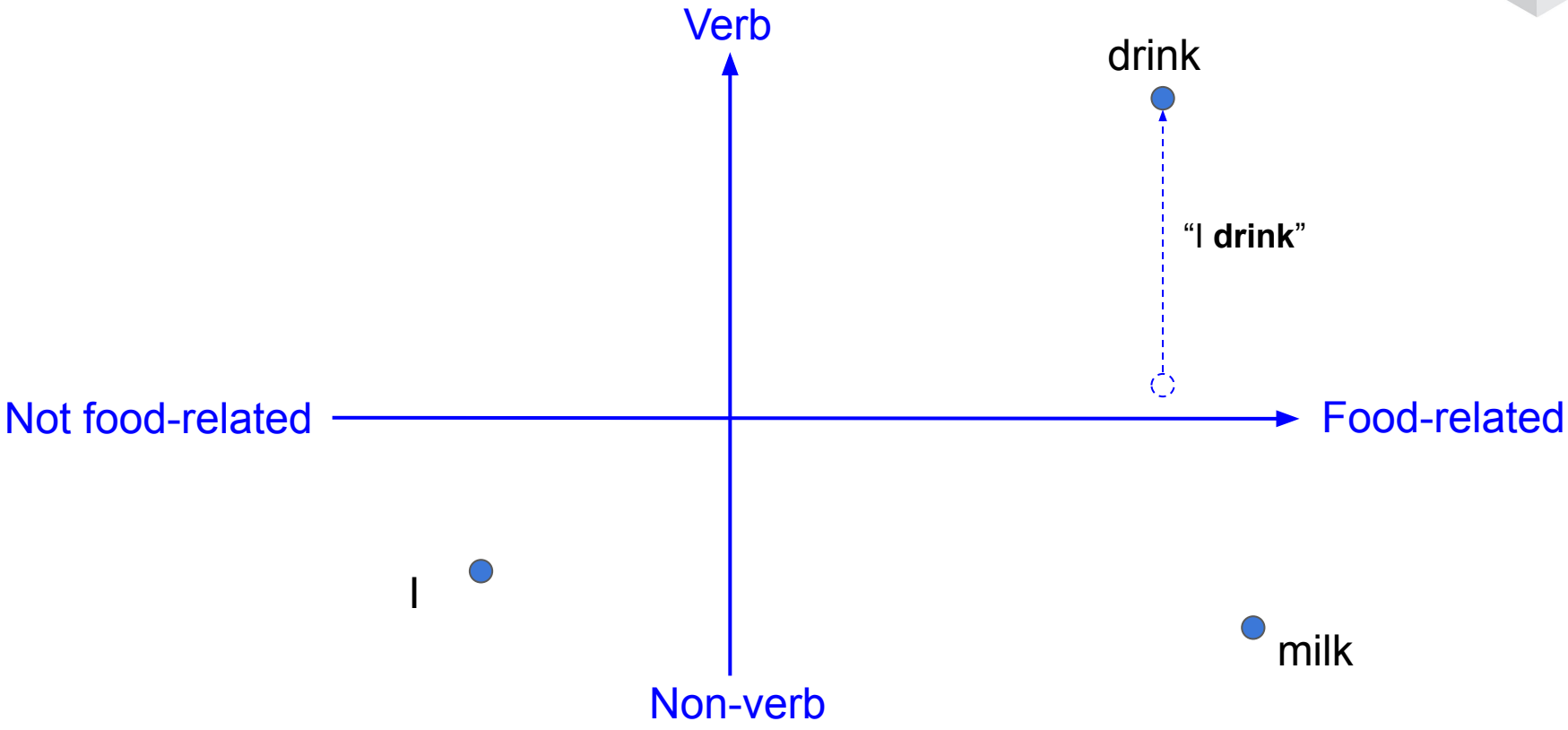
Attention Mechanism



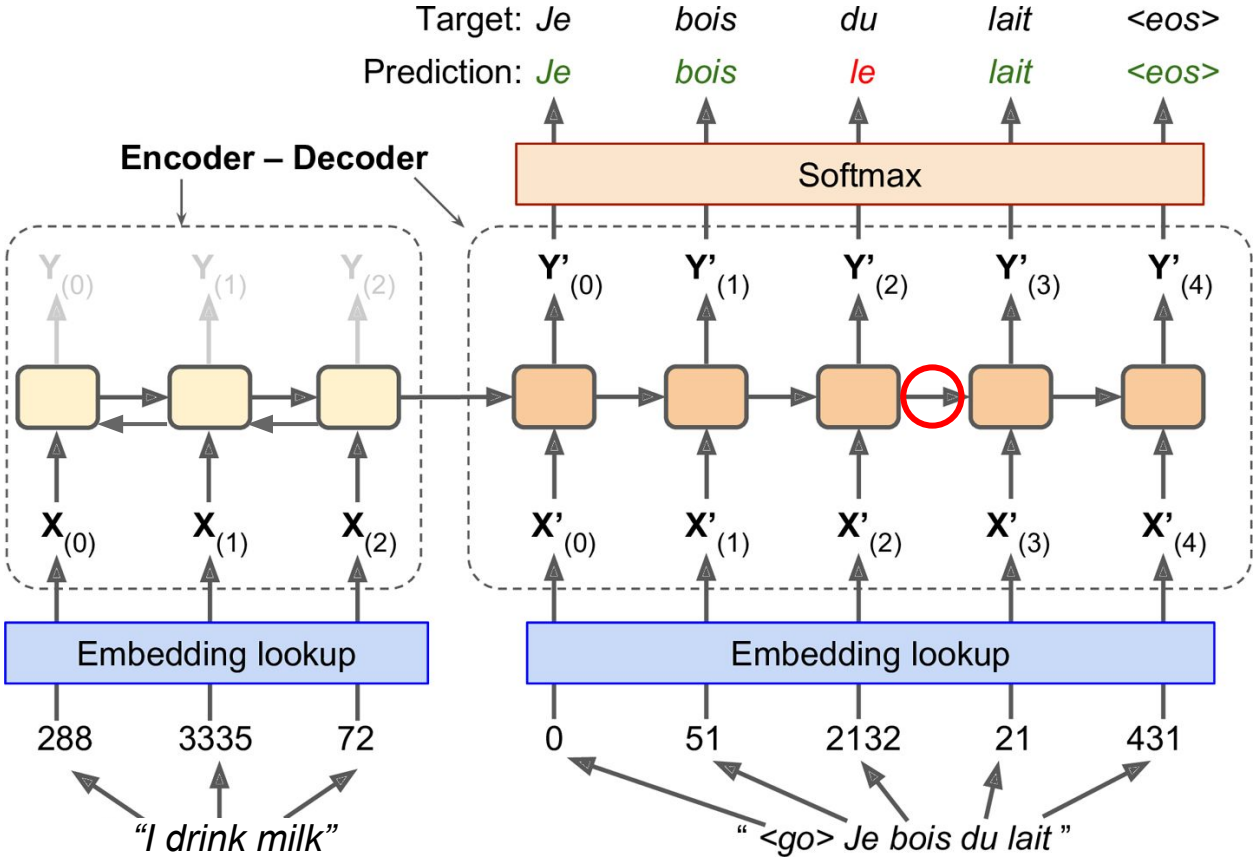
Encoder/Decoder with Attention



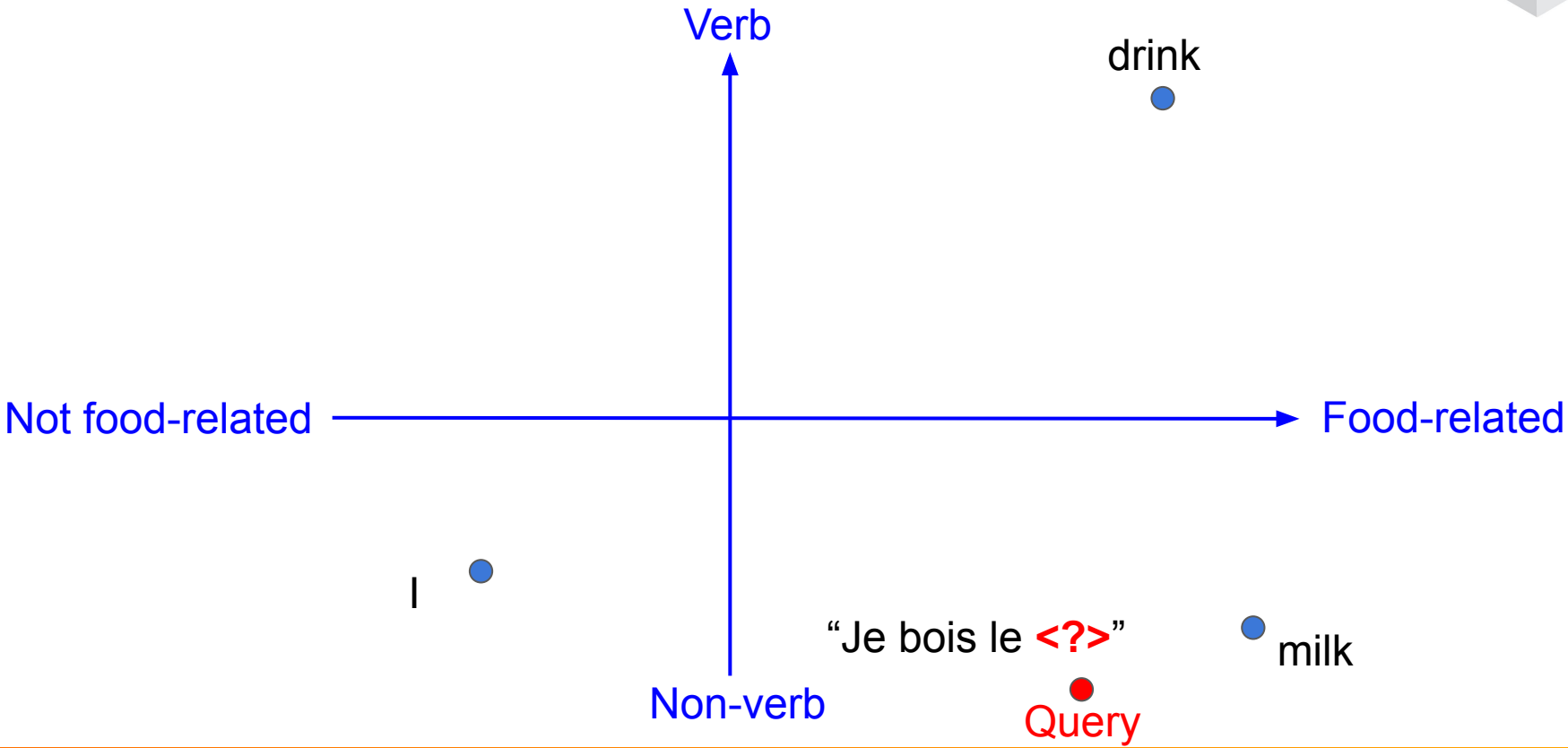
Attention Mechanism



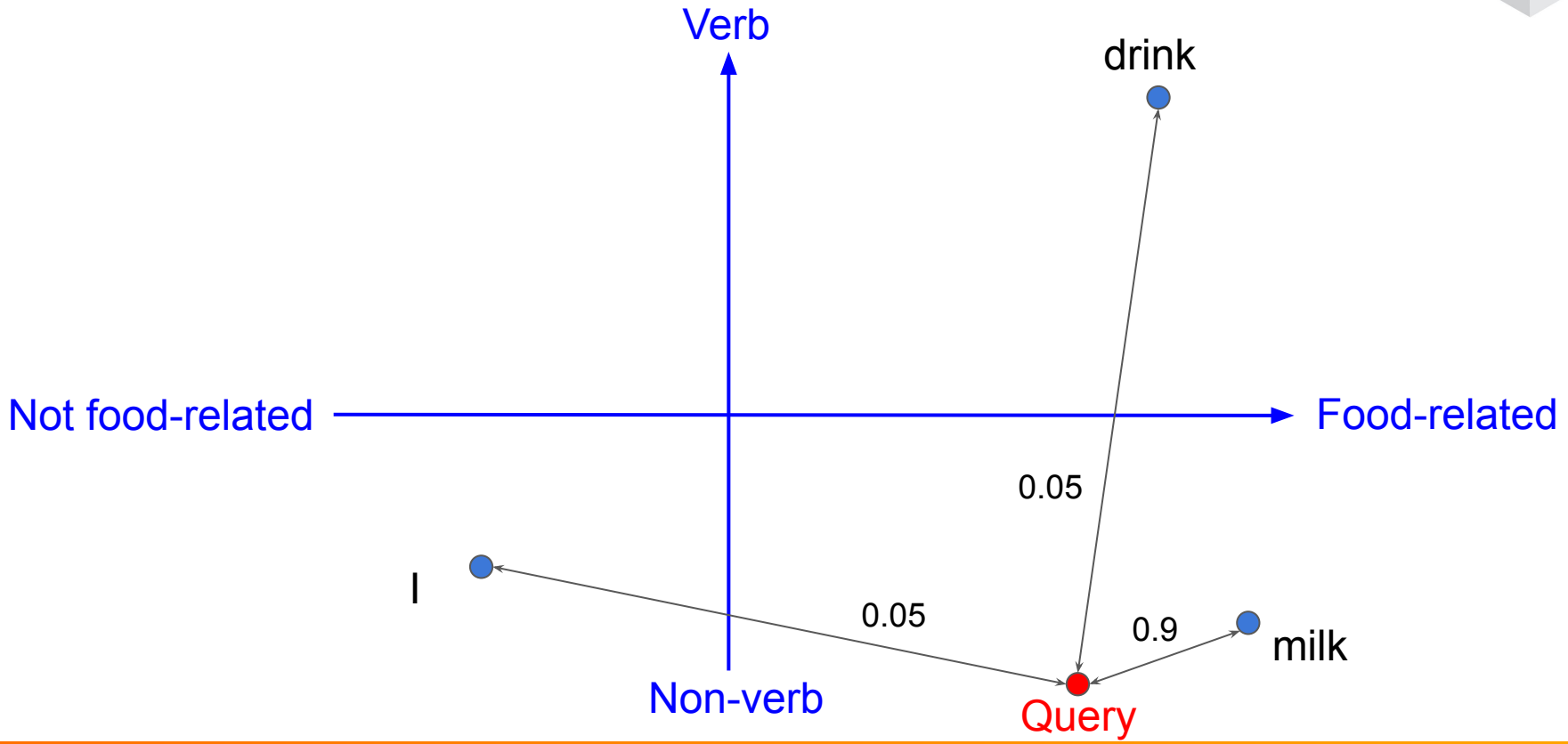
Encoder/Decoder with Attention



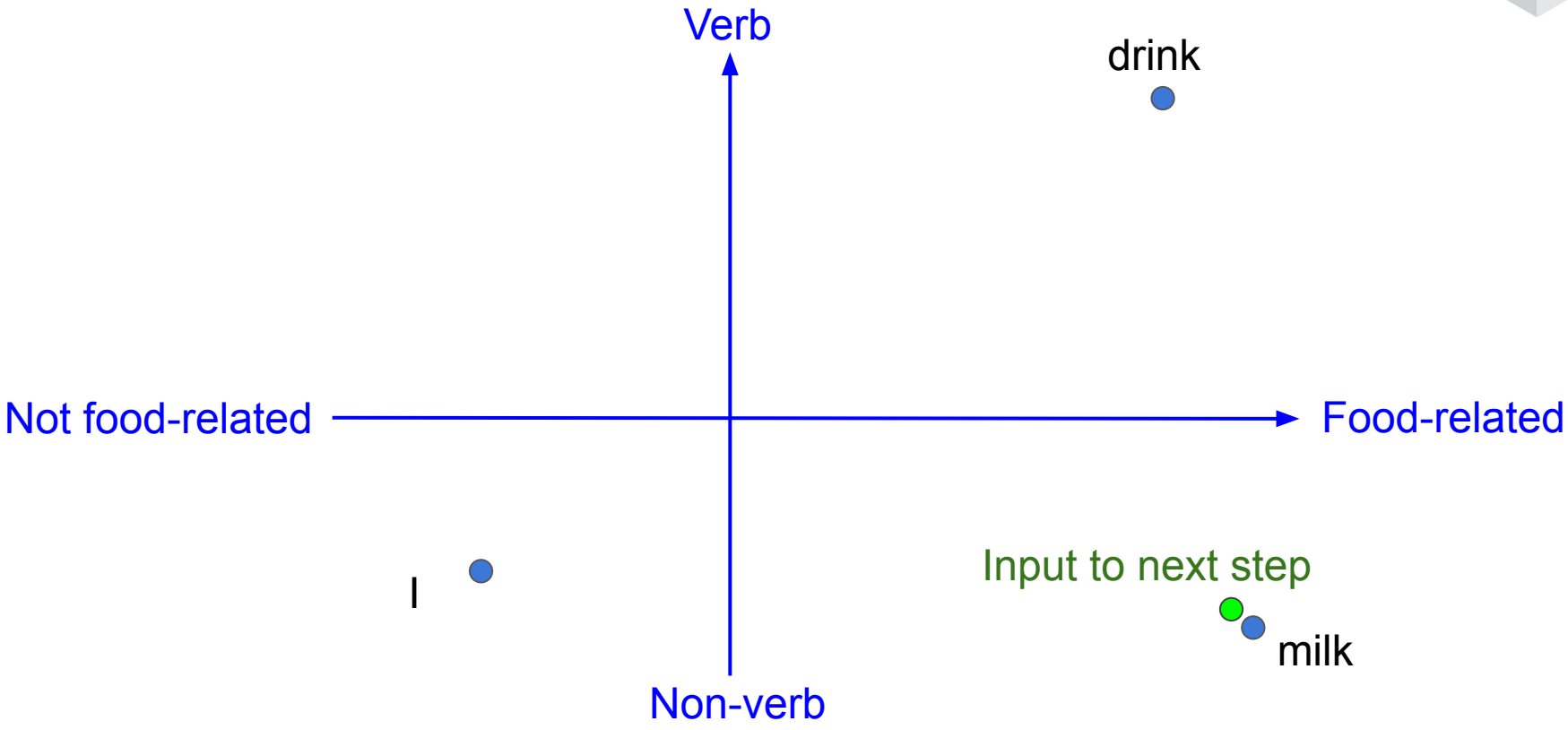
Attention Mechanism



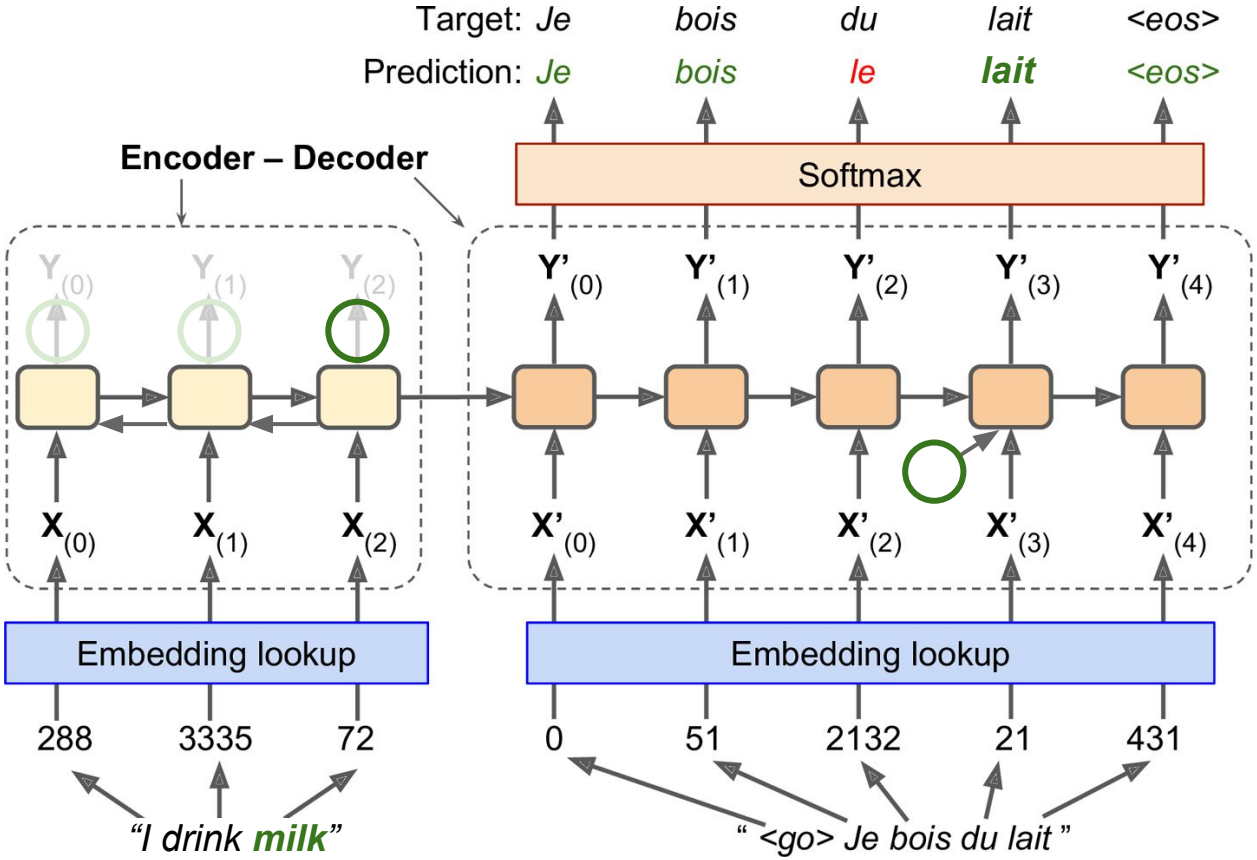
Attention Mechanism



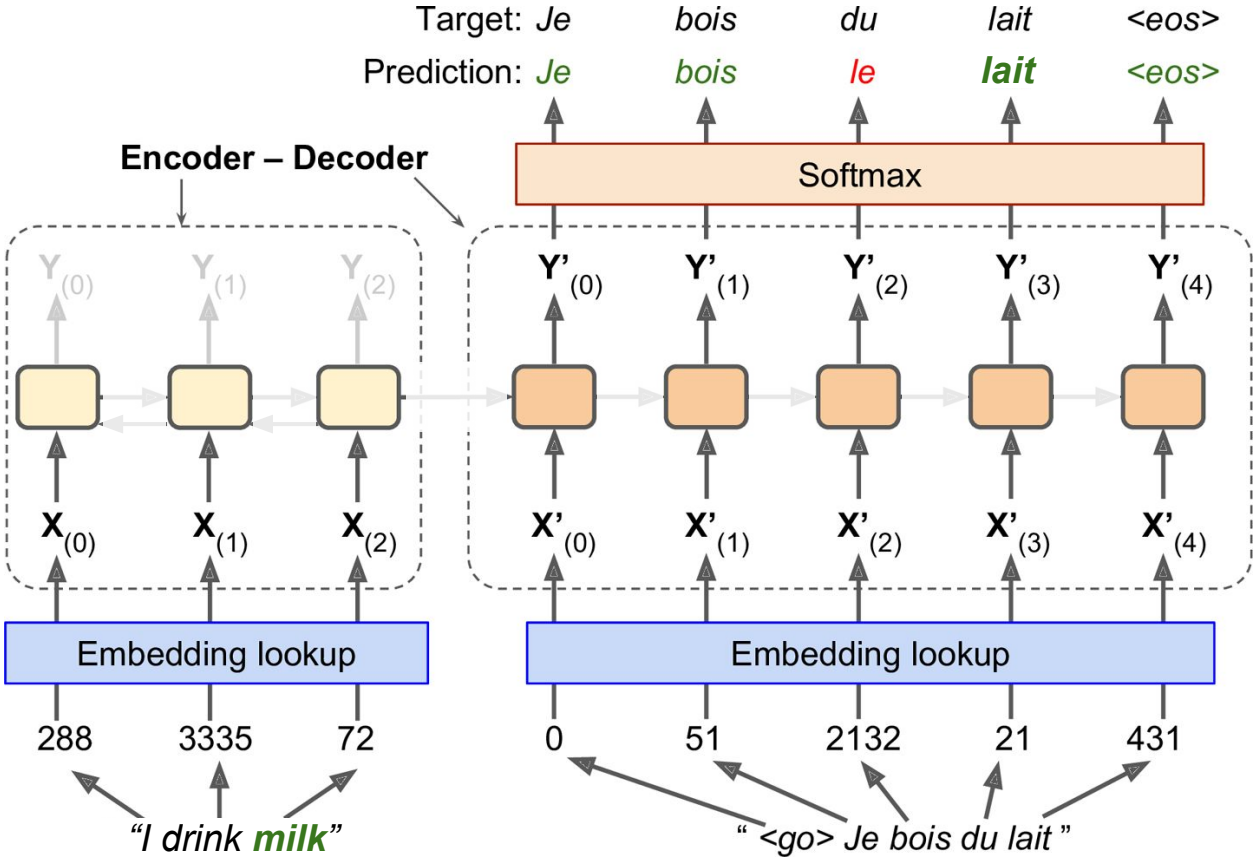
Attention Mechanism



Encoder/Decoder with Attention



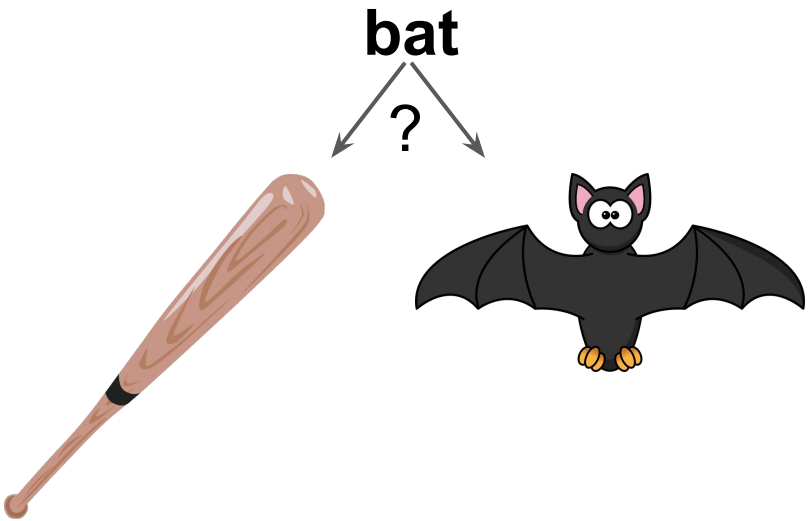
Attention Is All You Need



Attention Is All You Need

bat

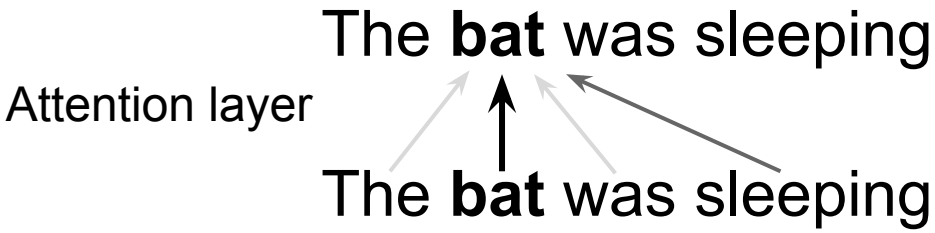
Attention Is All You Need



Attention Is All You Need

The **bat** was sleeping

Attention Is All You Need



Attention Is All You Need

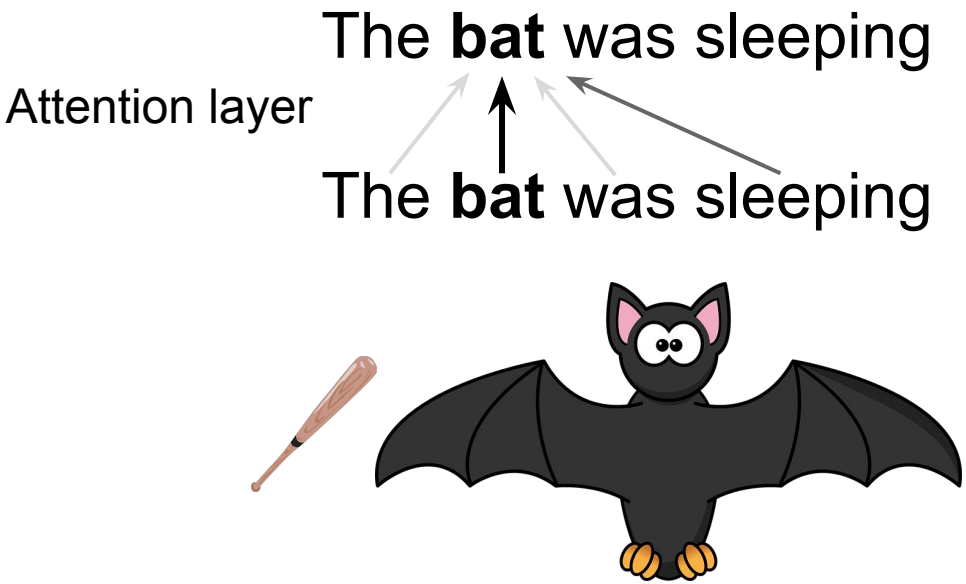


Figure 1 from the paper

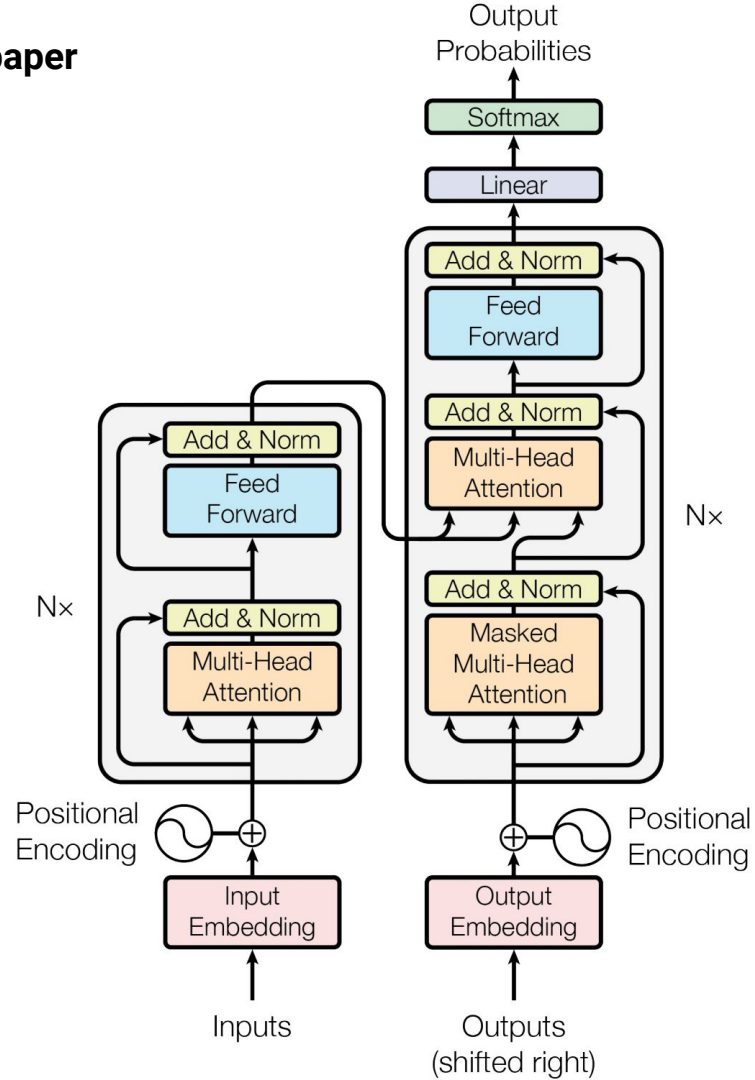


Figure 1 from the paper
(simplified)

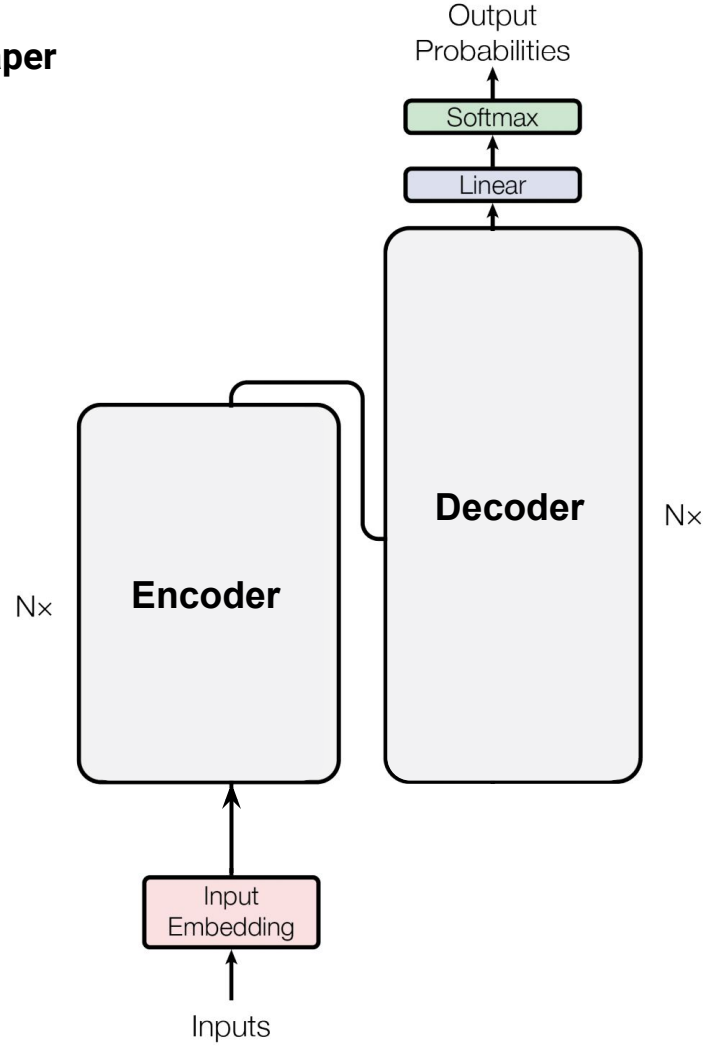


Figure 1 from the paper
(simplified)

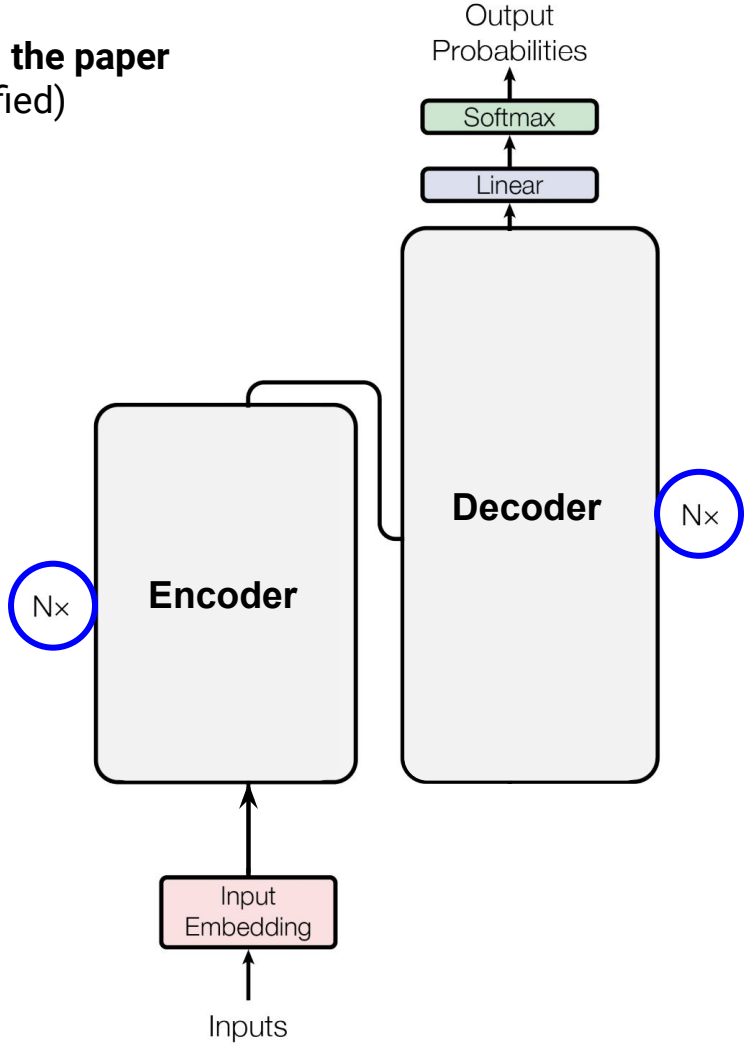


Figure 1 from the paper
(simplified)

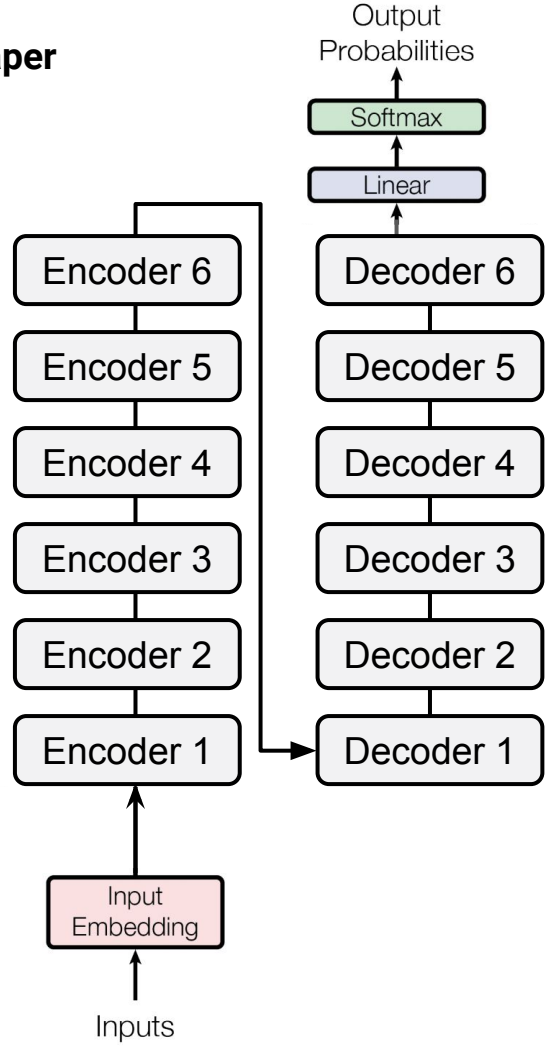
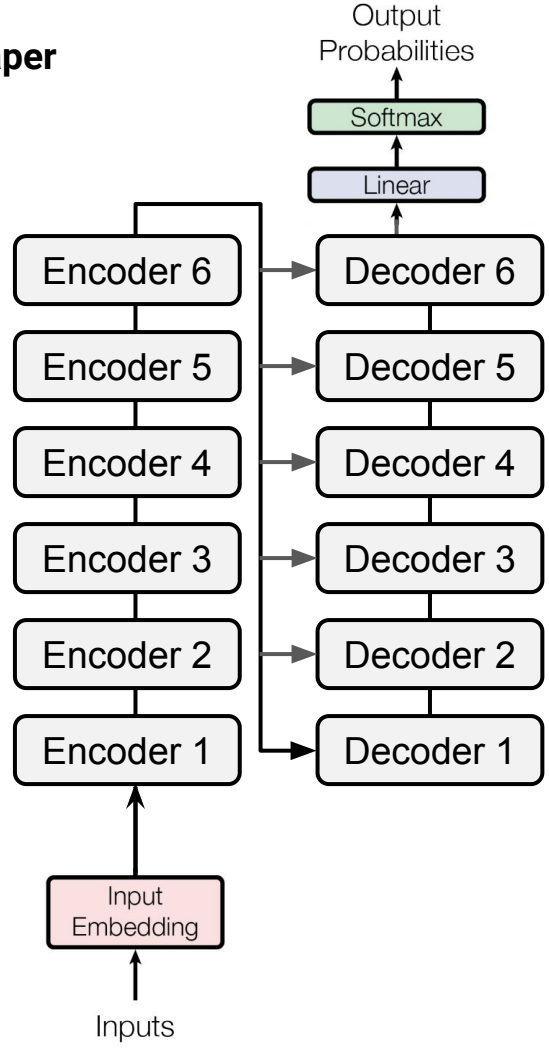


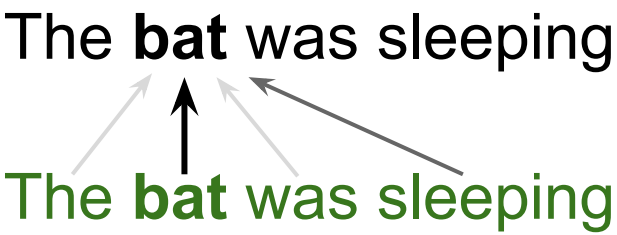
Figure 1 from the paper
(simplified)



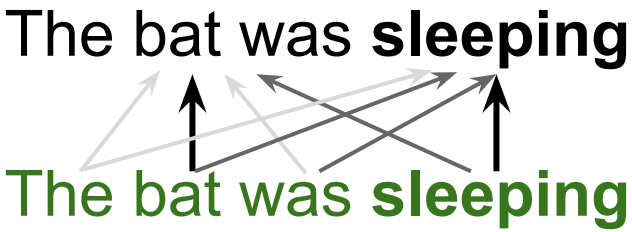
Attention Is All You Need

The bat was sleeping

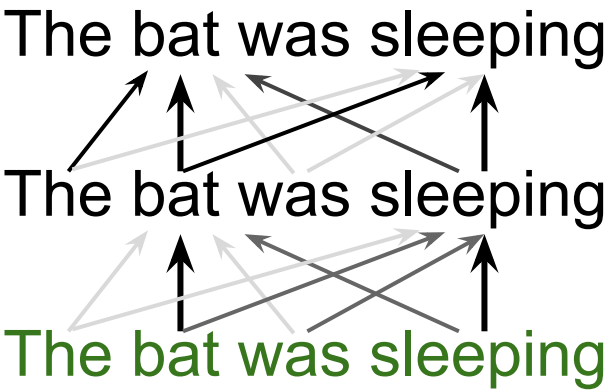
Attention Is All You Need



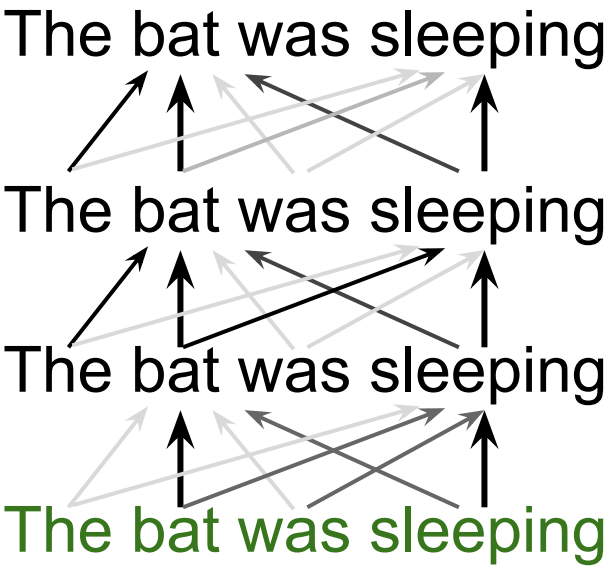
Attention Is All You Need



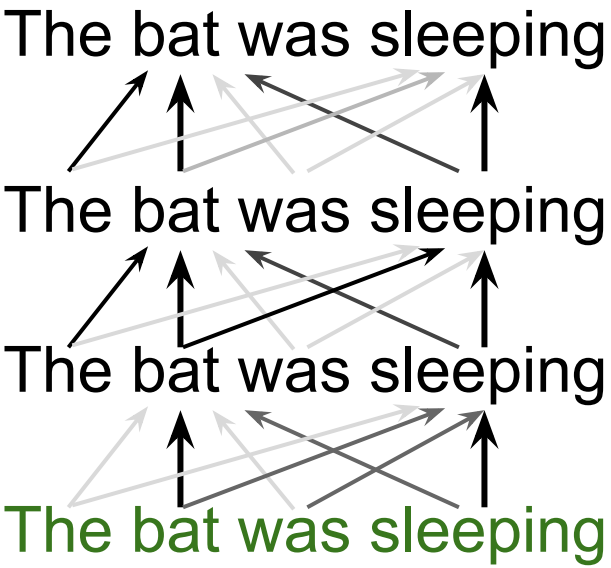
Attention Is All You Need



Attention Is All You Need

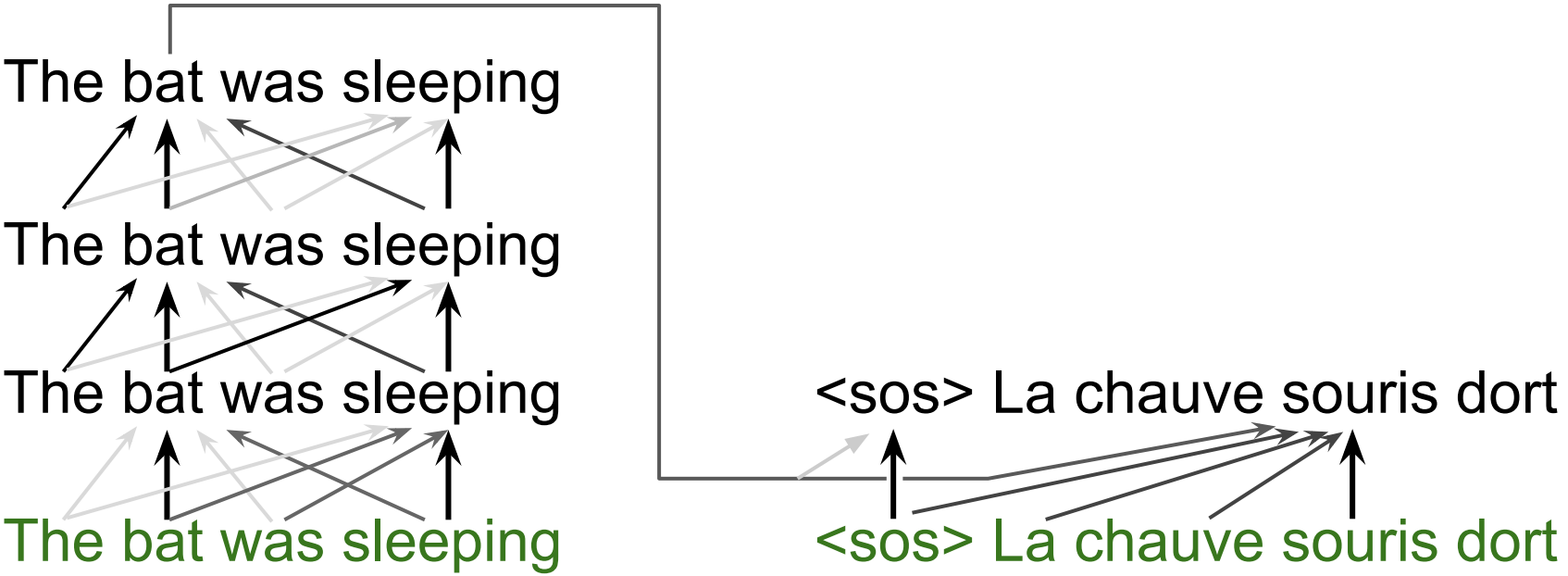


Attention Is All You Need

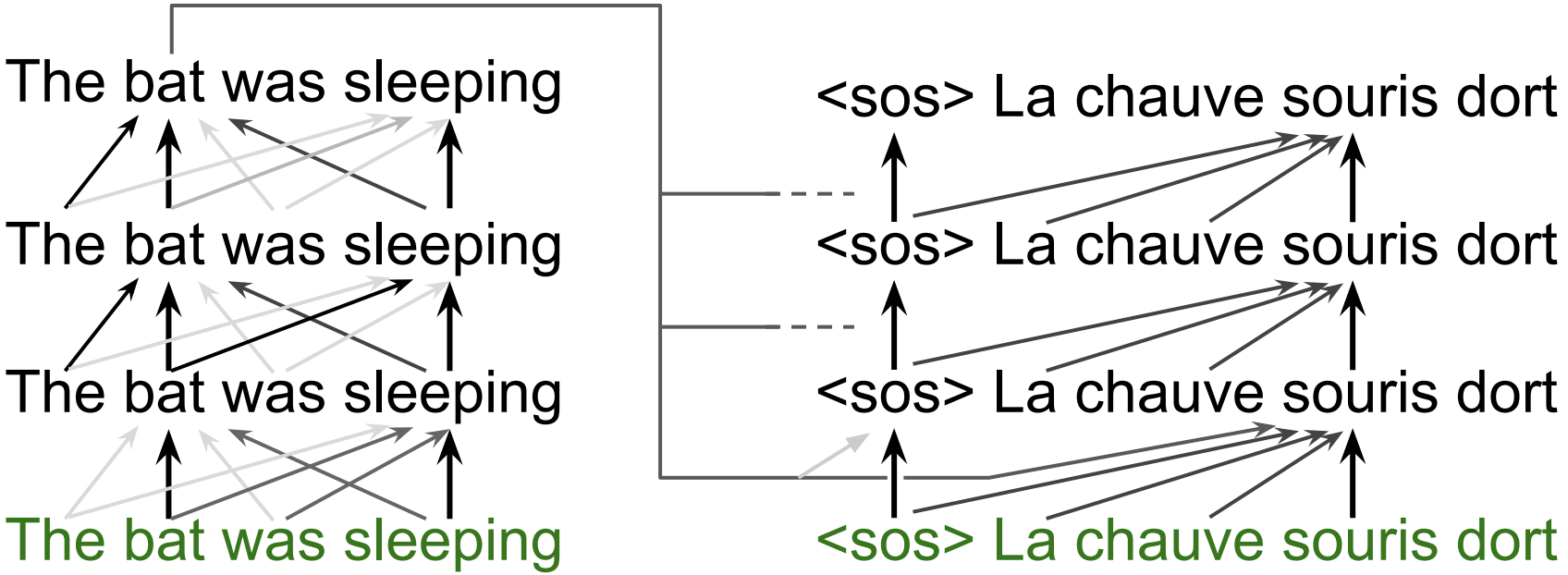


<sos> La chauve souris dort

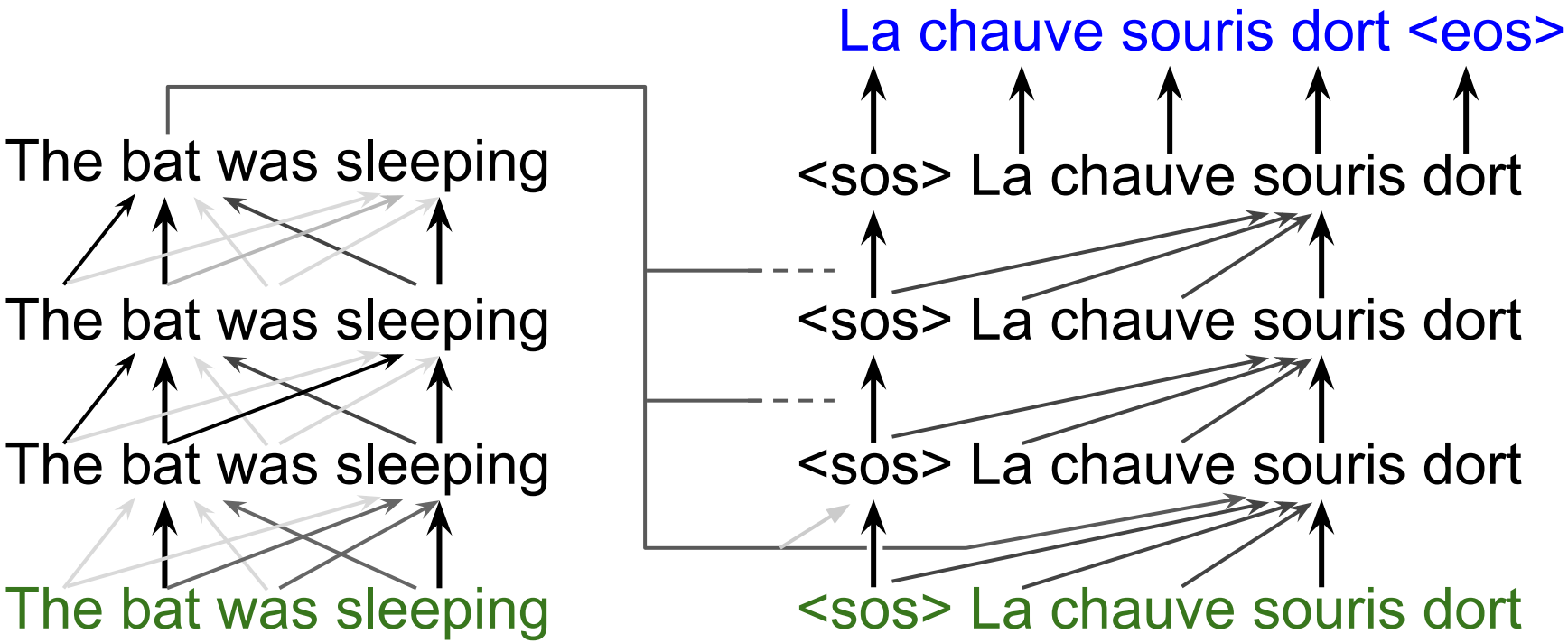
Attention Is All You Need



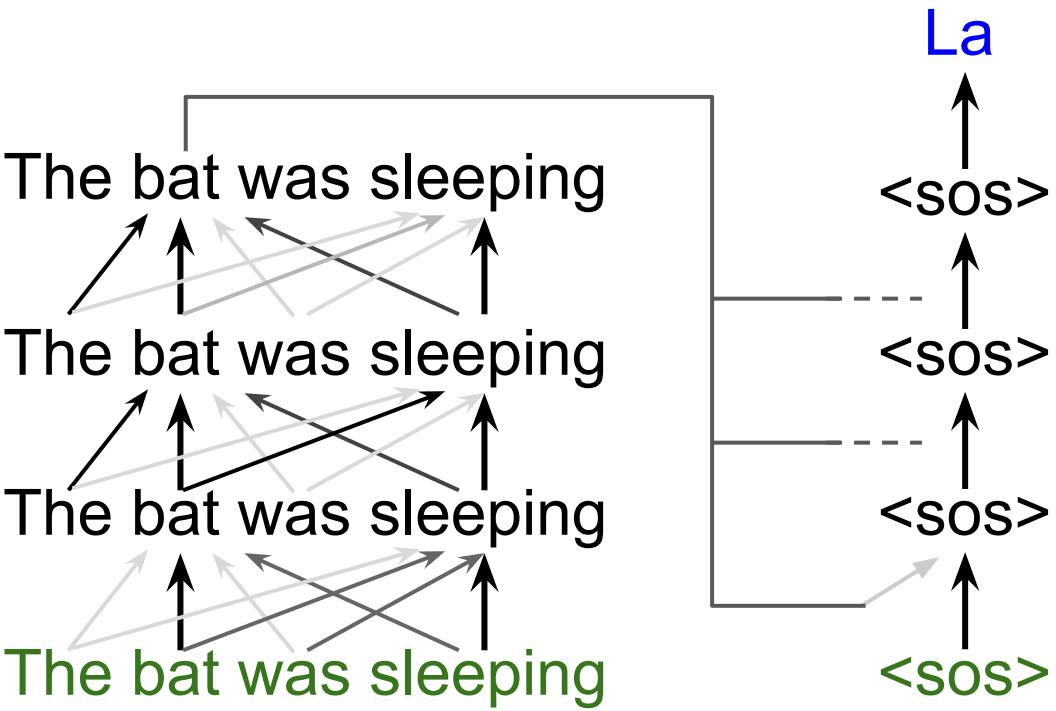
Attention Is All You Need



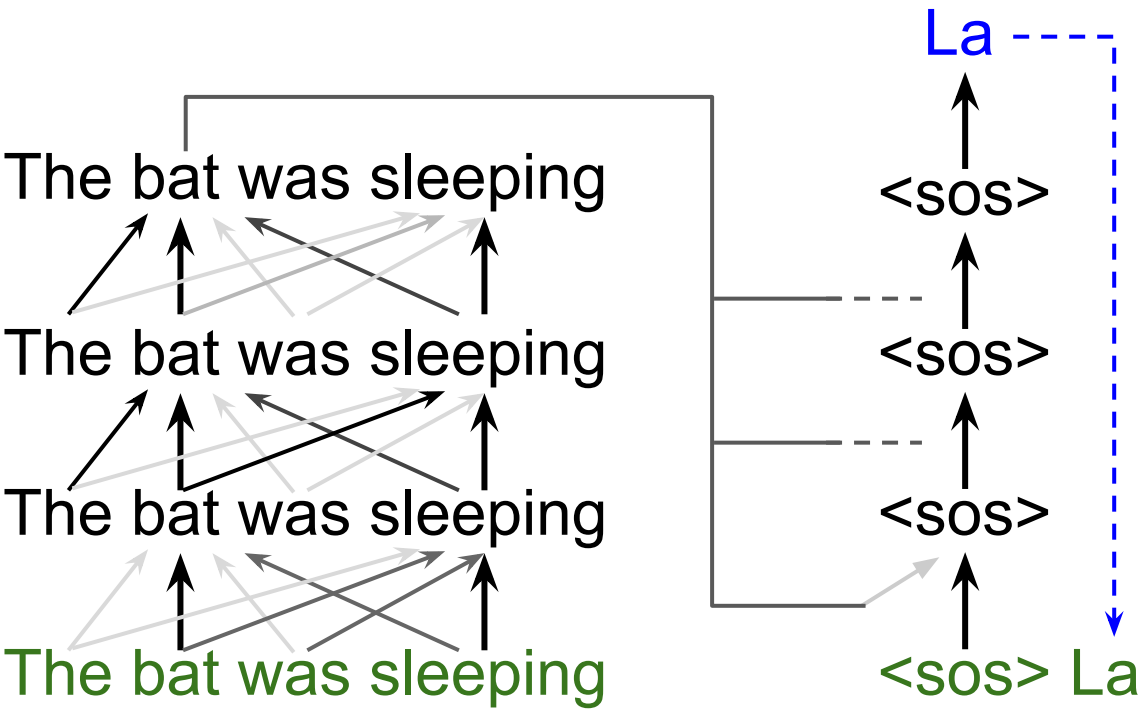
Attention Is All You Need



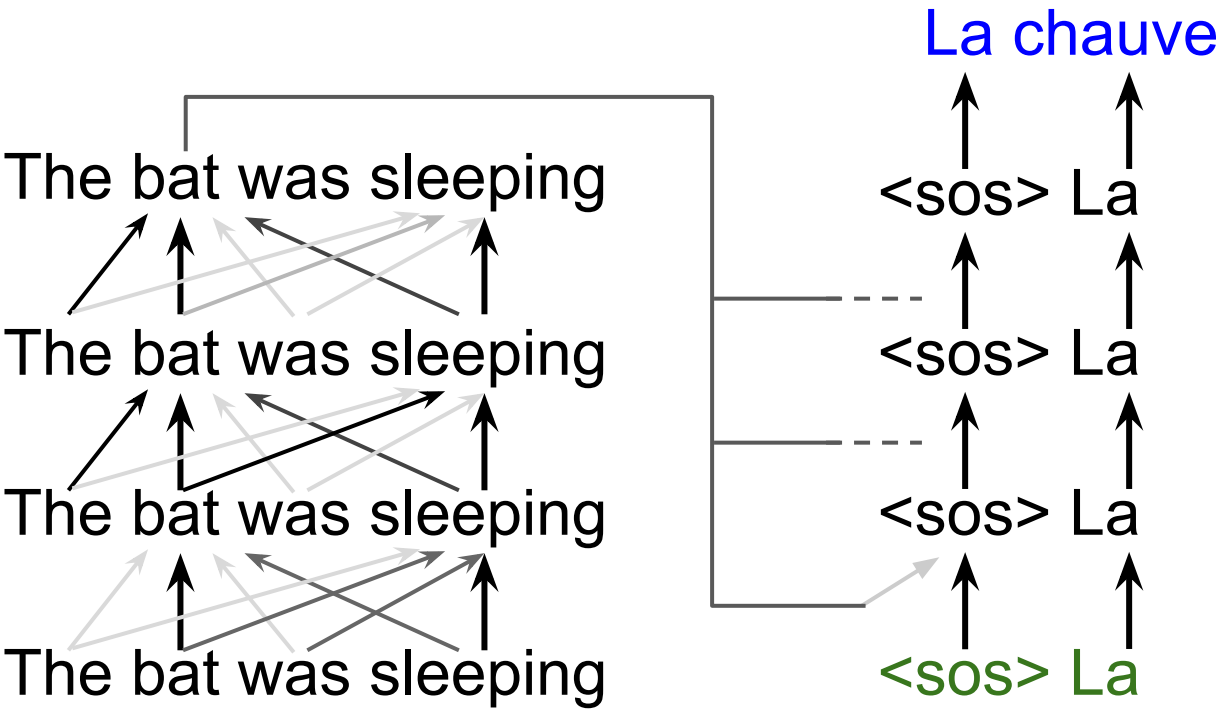
Attention Is All You Need



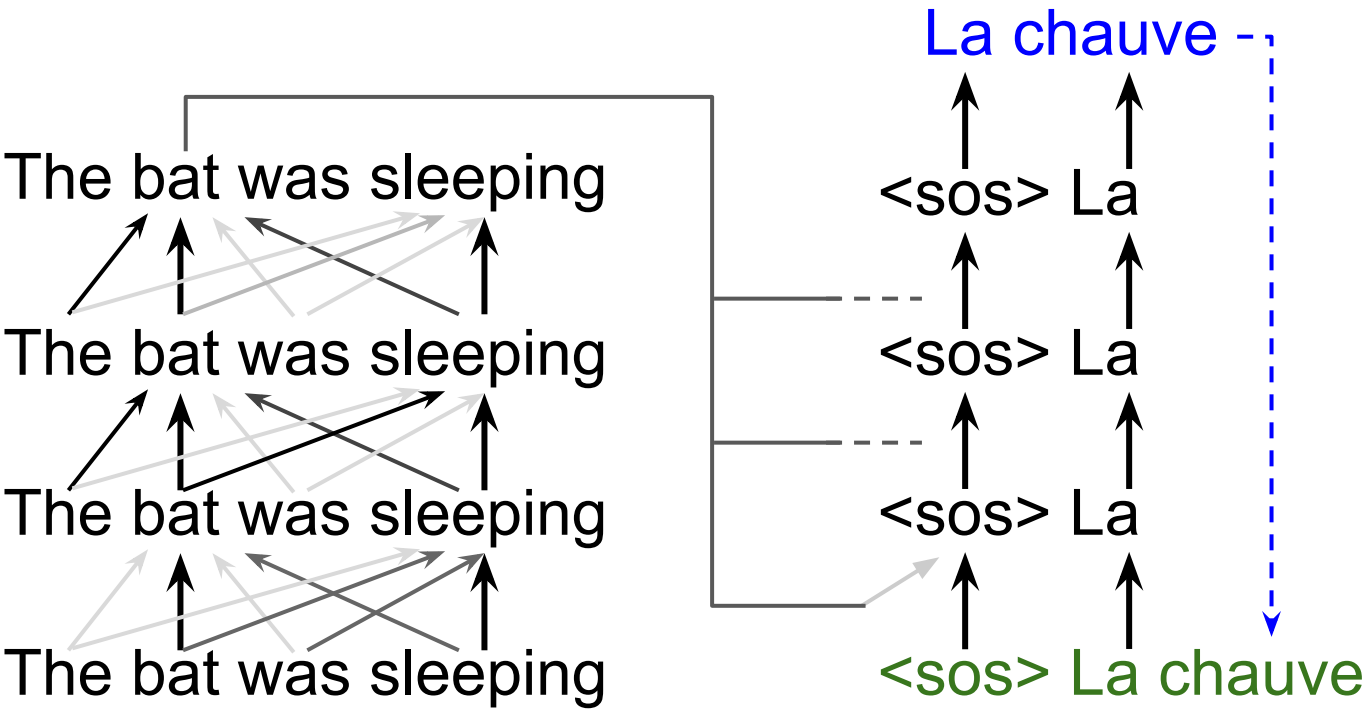
Attention Is All You Need



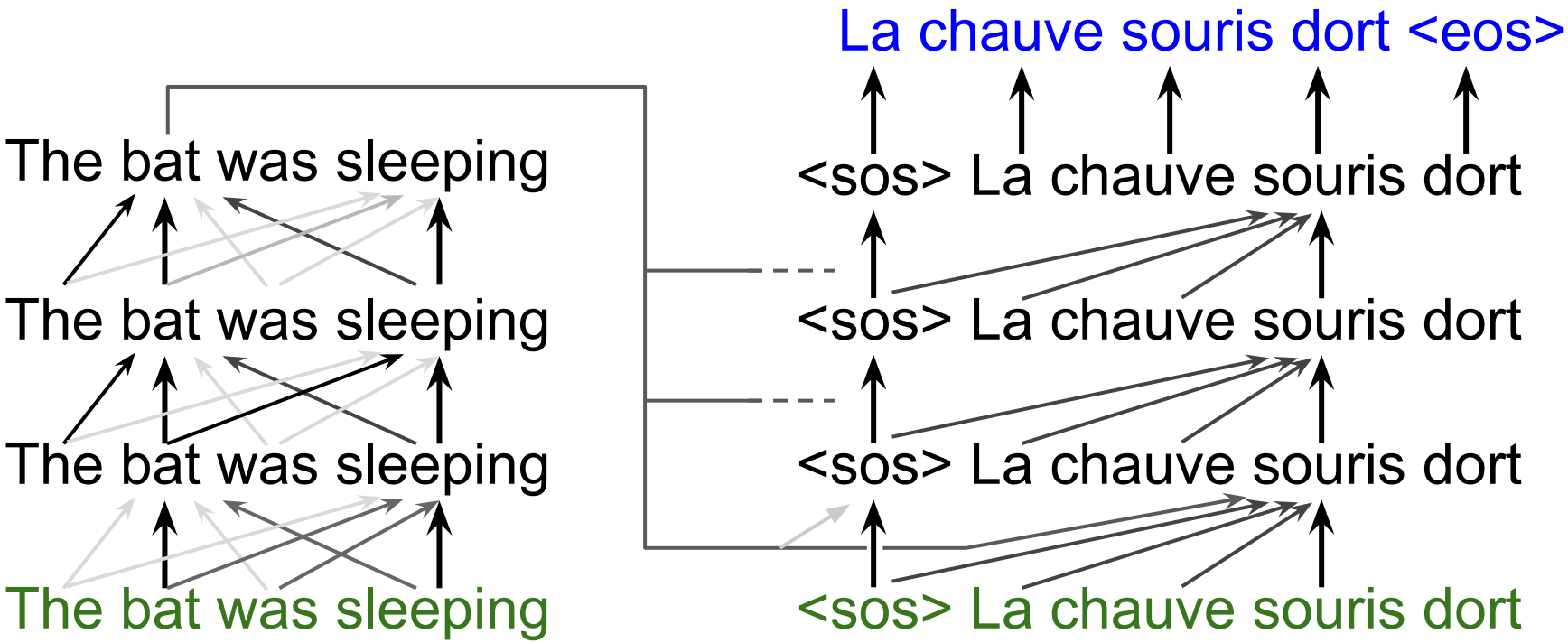
Attention Is All You Need



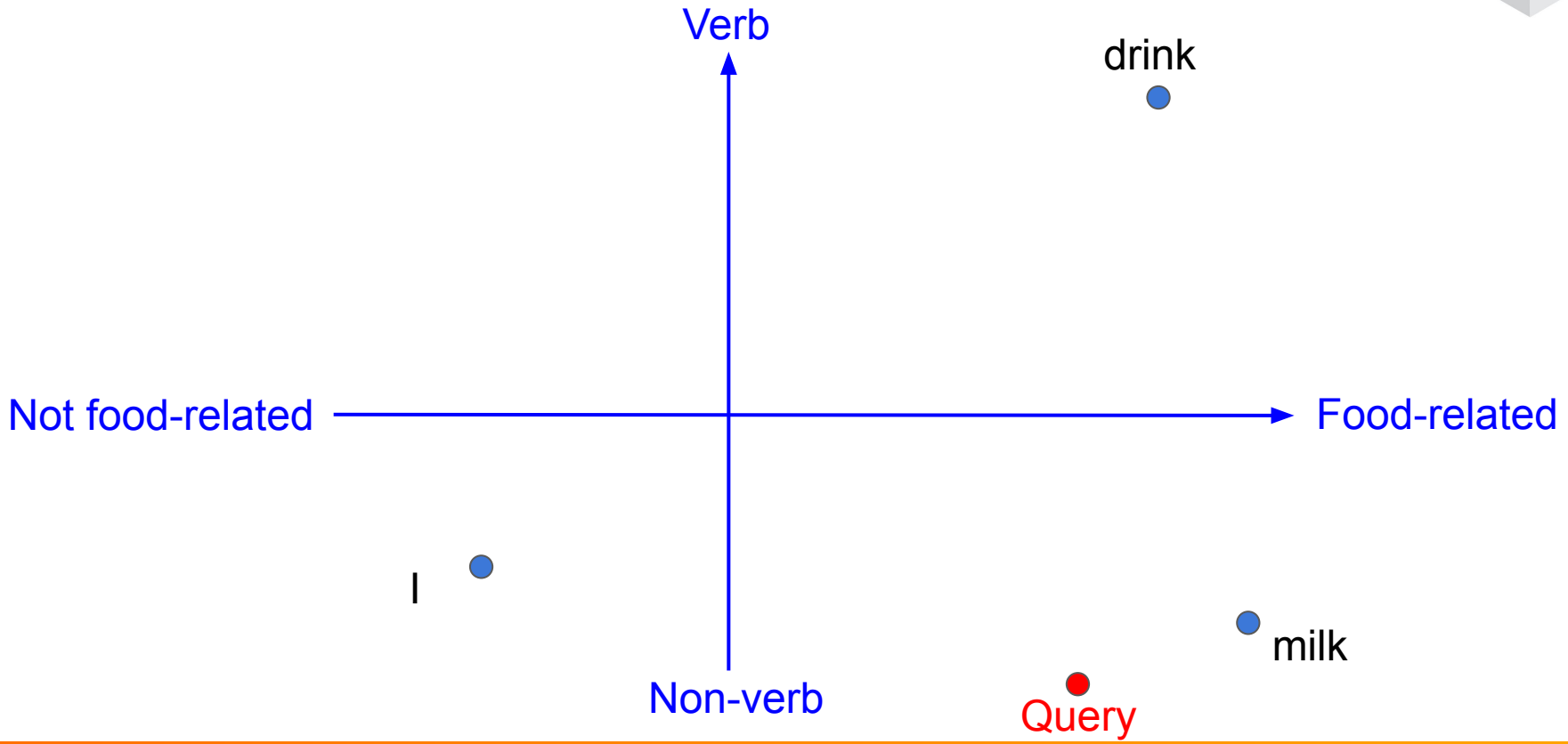
Attention Is All You Need



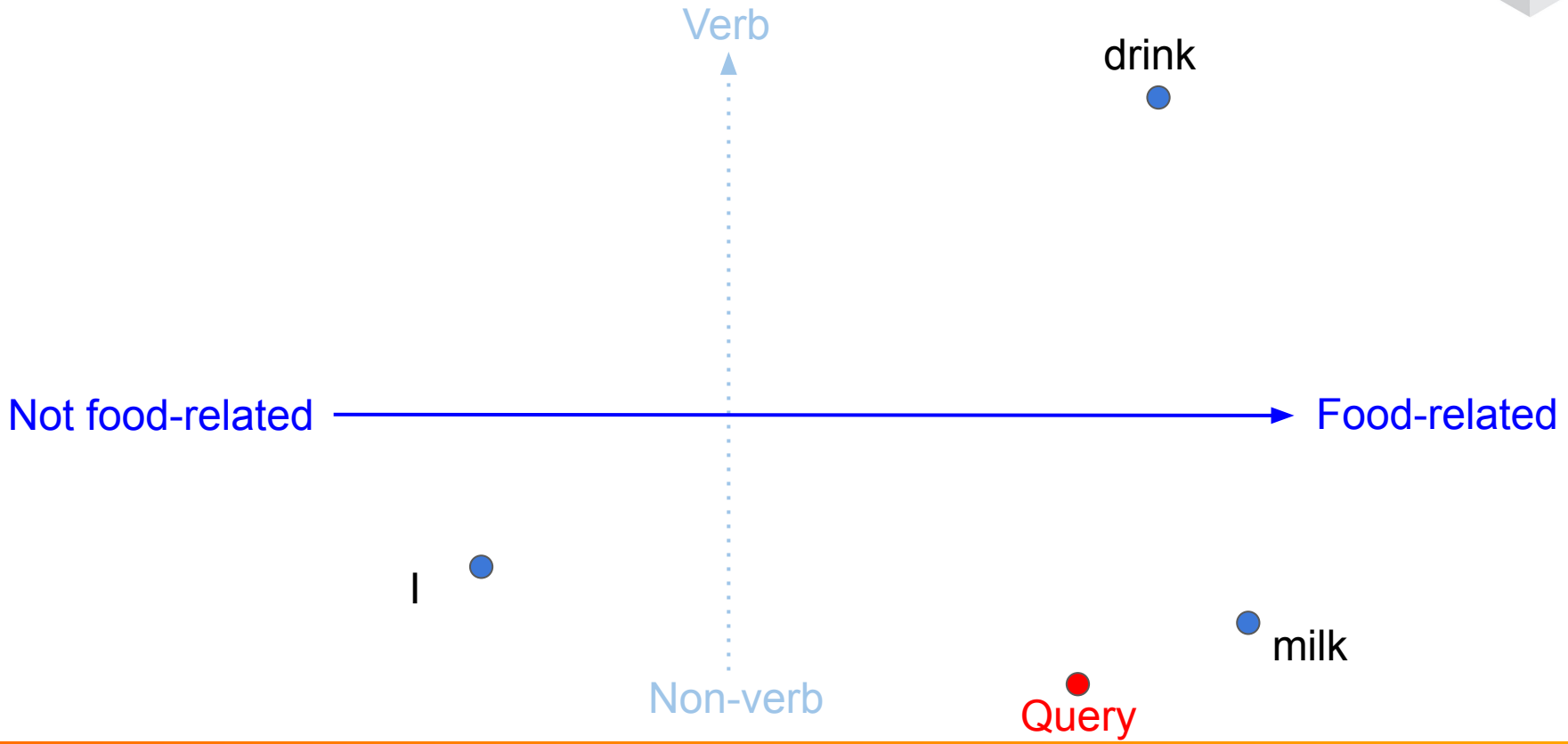
Attention Is All You Need



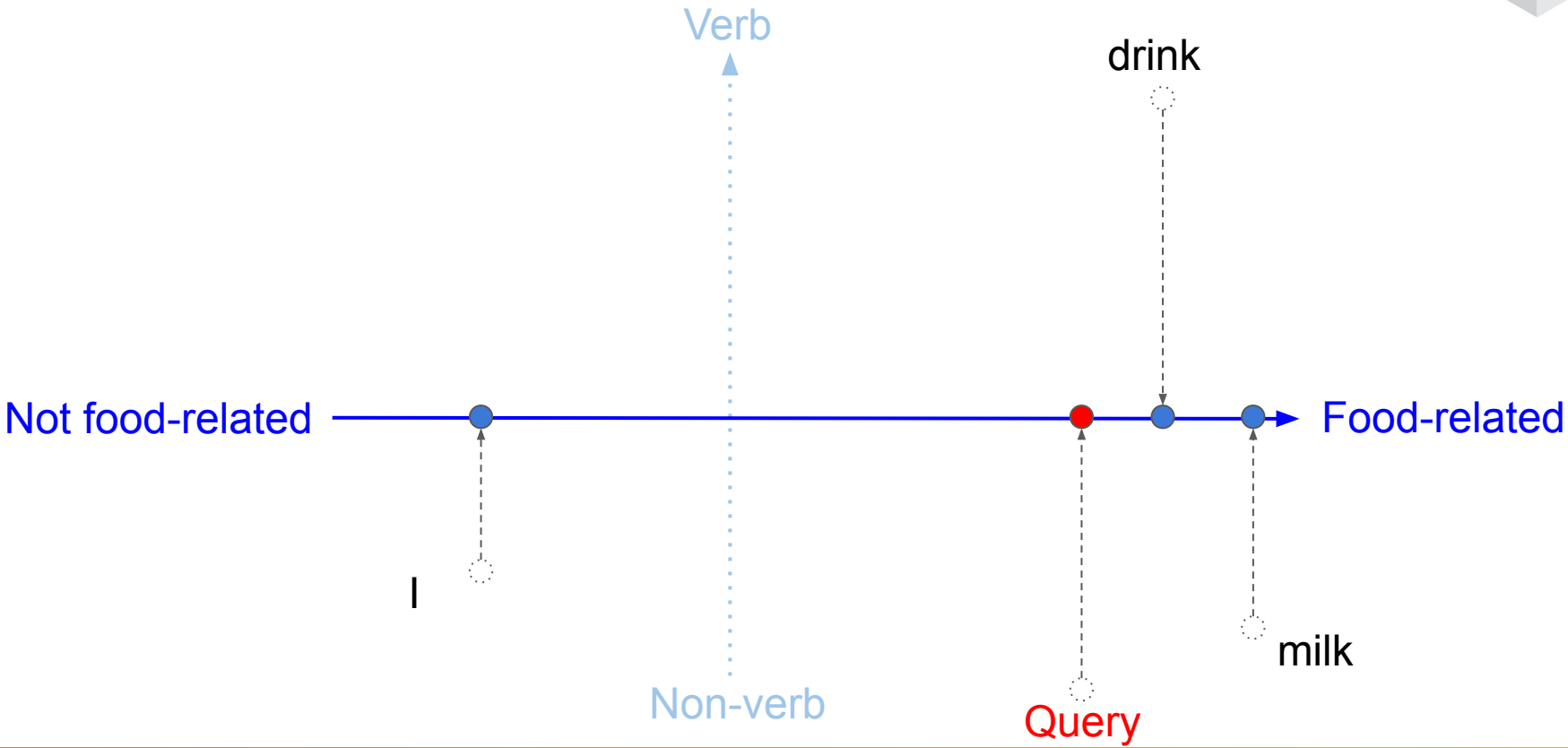
Multi-Head Attention



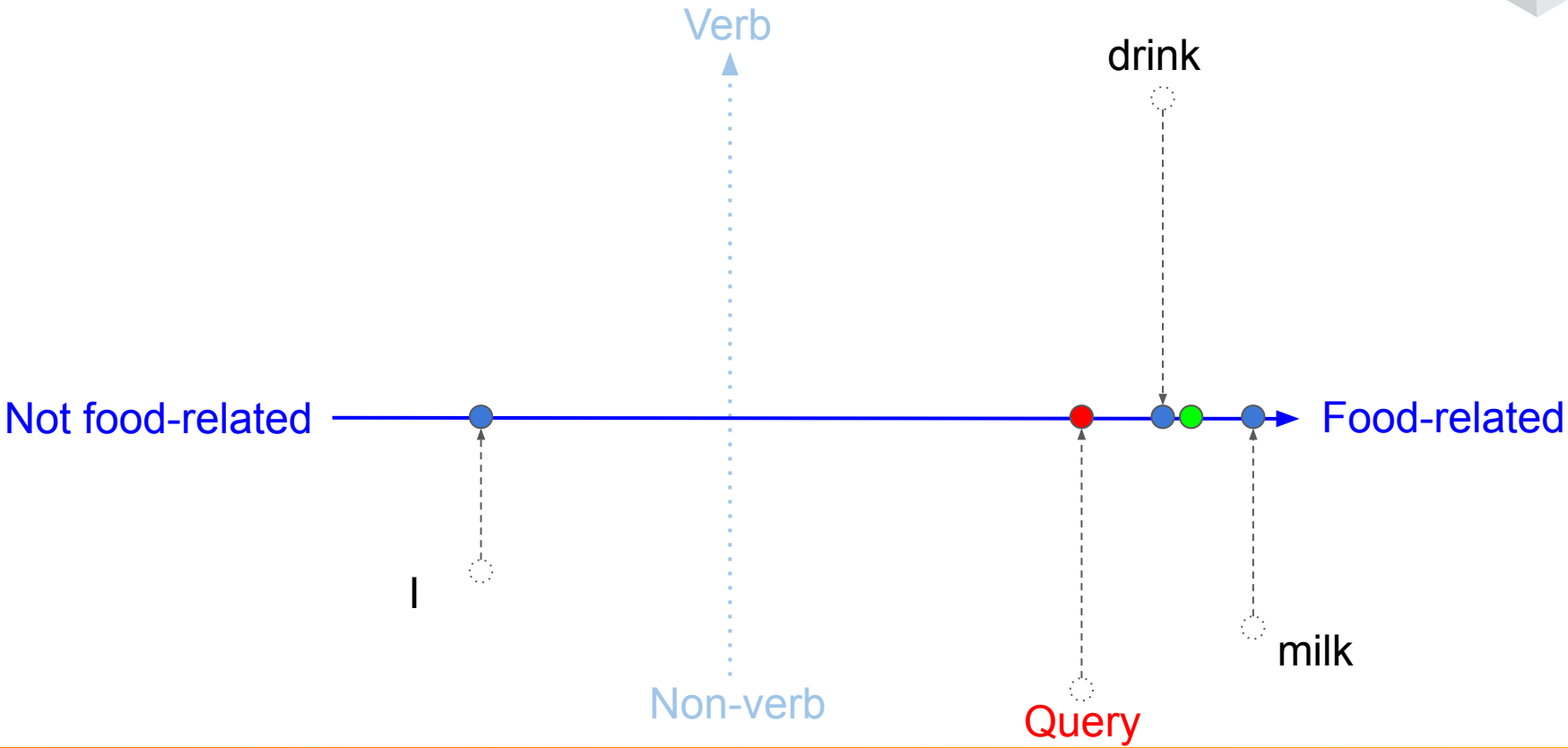
Multi-Head Attention



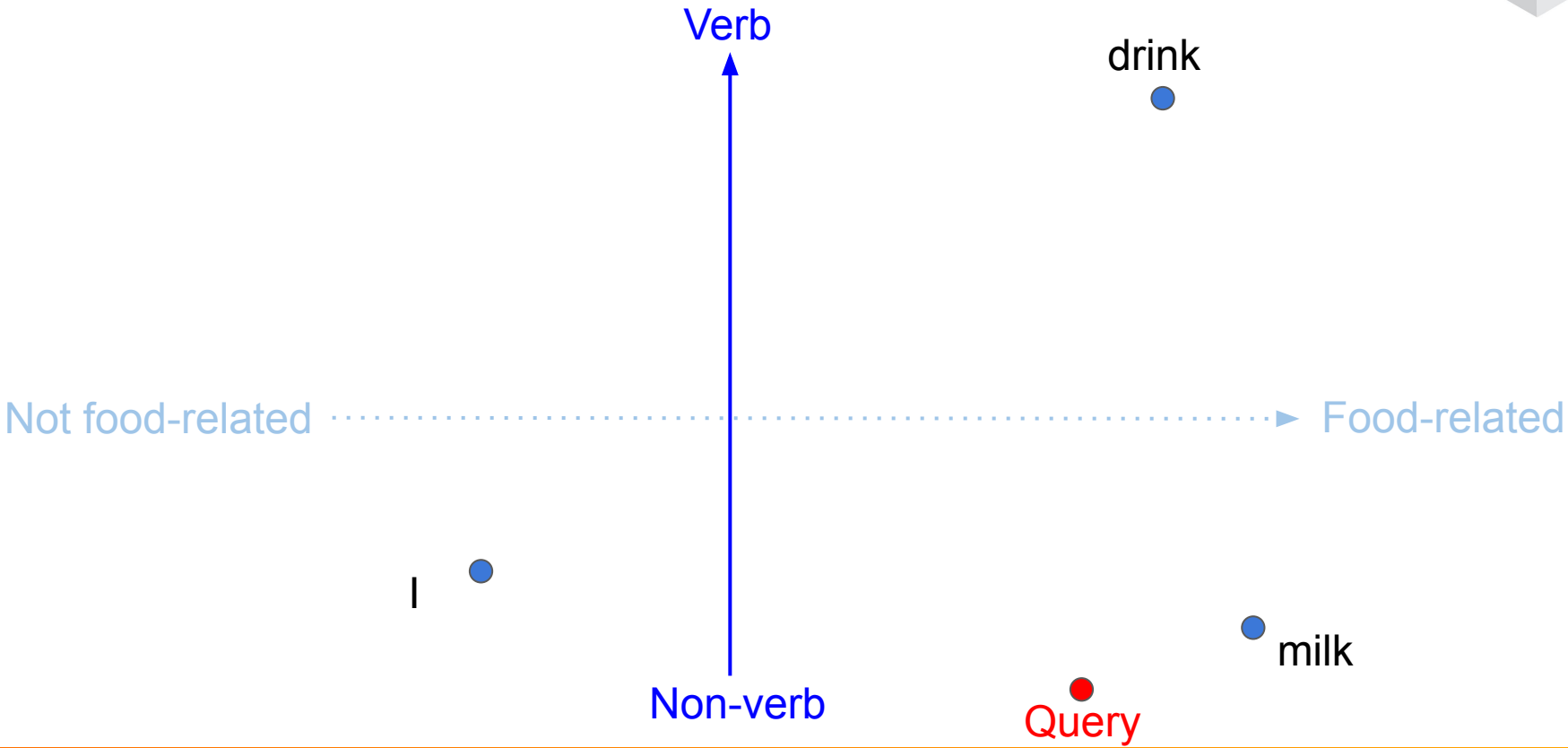
Multi-Head Attention



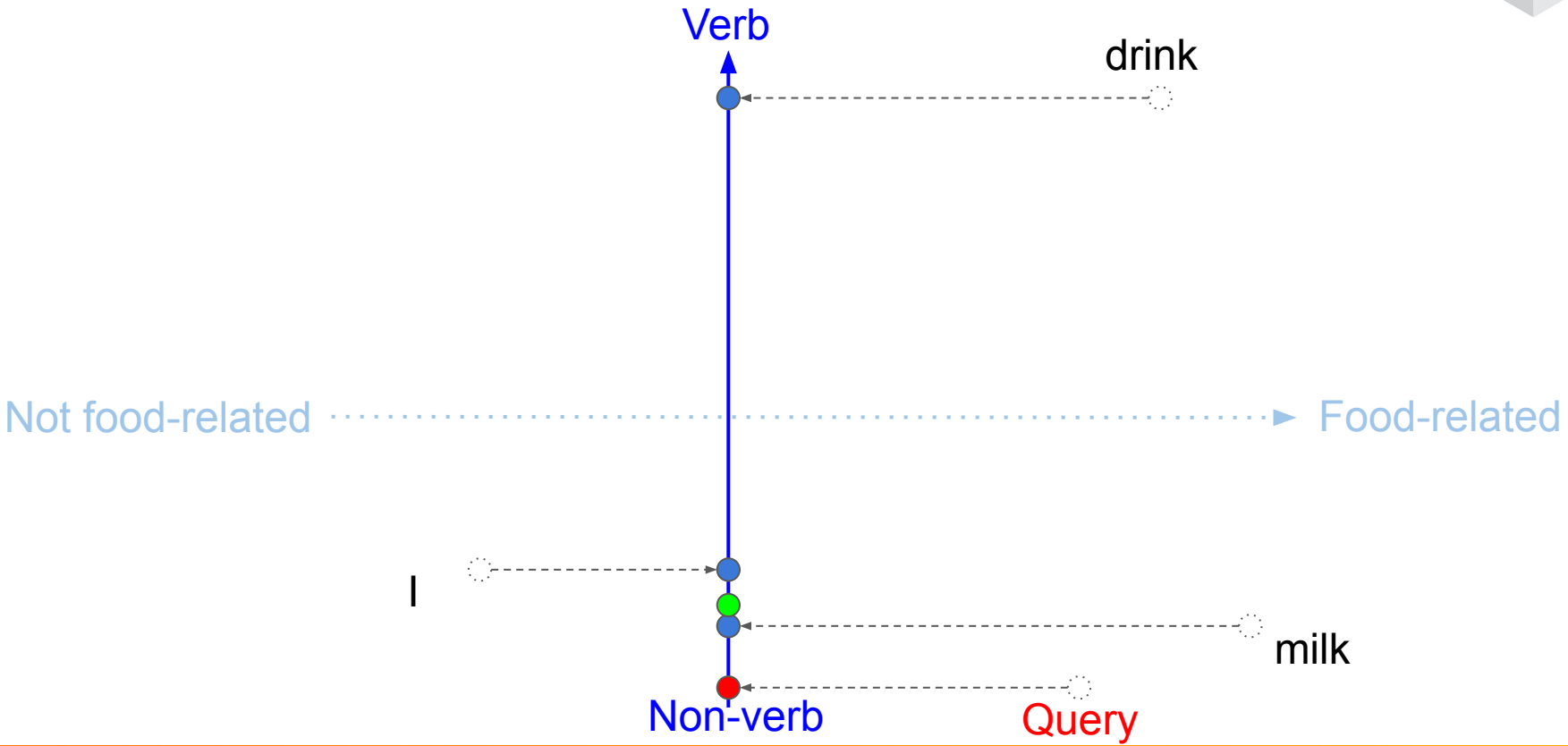
Multi-Head Attention



Multi-Head Attention



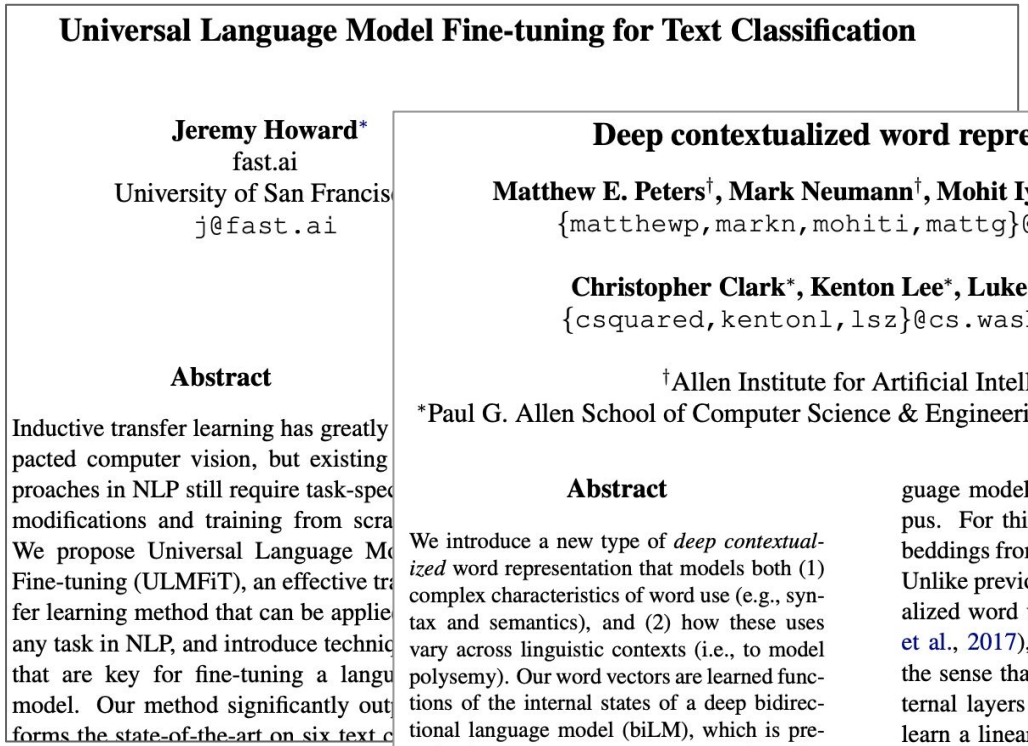
Multi-Head Attention



TensorFlow

Language Models

Pretraining + Fine-tuning



Universal Language Model Fine-tuning for Text Classification

Jeremy Howard*
fast.ai
University of San Francis
j@fast.ai

Abstract

Inductive transfer learning has greatly
pacted computer vision, but existing
proaches in NLP still require task-spec
modifications and training from scra
We propose Universal Language Mo
Fine-tuning (ULMFiT), an effective tra
fer learning method that can be applie
any task in NLP, and introduce technic
that are key for fine-tuning a langu
model. Our method significantly out
forms the state-of-the-art on six text c

ULMFit
Jan 2018

Deep contextualized word representations

Matthew E. Peters†, Mark Neumann†, Mohit Iyyer†, Matt Gardner†,
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark*, Kenton Lee*, Luke Zettlemoyer†*
{csquared,kentonl,lsz}@cs.washington.edu

†Allen Institute for Artificial Intelligence
*Paul G. Allen School of Computer Science & Engineering, University of Washington

Abstract

We introduce a new type of *deep contextualized* word representation that models both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy). Our word vectors are learned functions of the internal states of a deep bidirectional language model (biLM), which is pretrained on a large text corpus. We show that these representations can be easily added to existing models and significantly improve the

guage model (LM) objective on a large text corpus. For this reason, we call them ELMo (Embeddings from Language Models) representations. Unlike previous approaches for learning contextualized word vectors (Peters et al., 2017; McCann et al., 2017), ELMo representations are deep, in the sense that they are a function of all of the internal layers of the biLM. More specifically, we learn a linear combination of the vectors stacked above each input word for each end task, which markedly improves performance over just using the top LSTM layer.

ELMo
Feb 2018

BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT representations can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, *without* substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing models are required to produce fine-grained output at the token-level.

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018), uses tasks-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning the pre-trained parameters. In previous work, both approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

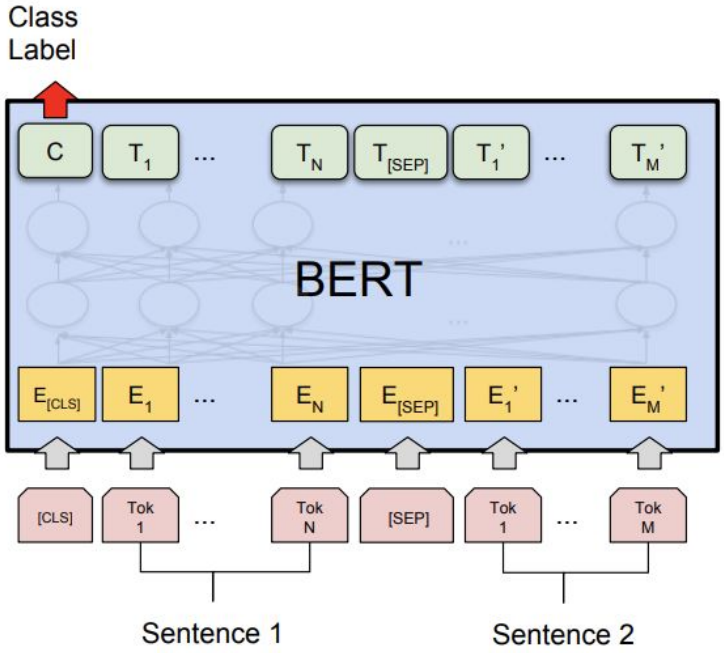
- Large number of parameters
- Pretraining on a huge corpus
- Subword tokenization
- Single architecture for multiple tasks

Subword Tokenization

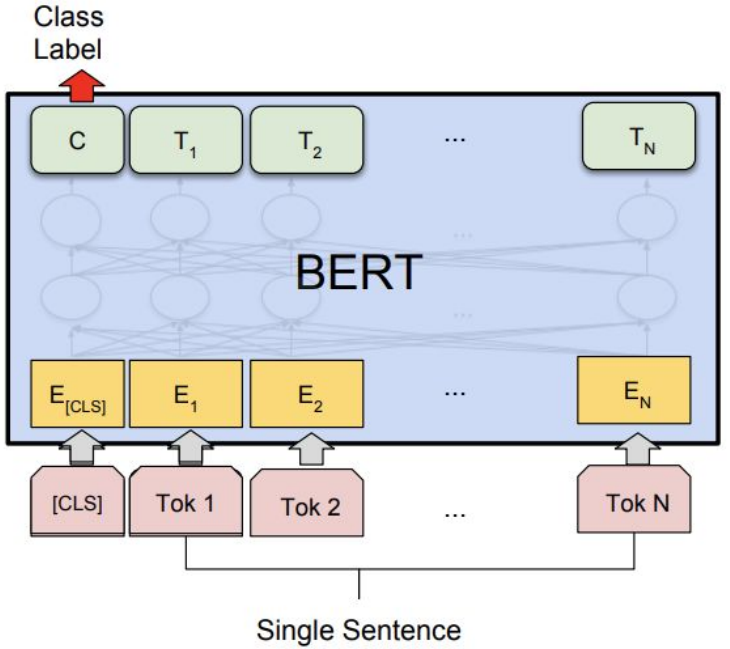
```
Going on computerless vacation
```

```
Going on a computer ##less vacation
```

BERT — Figure 4 From Paper

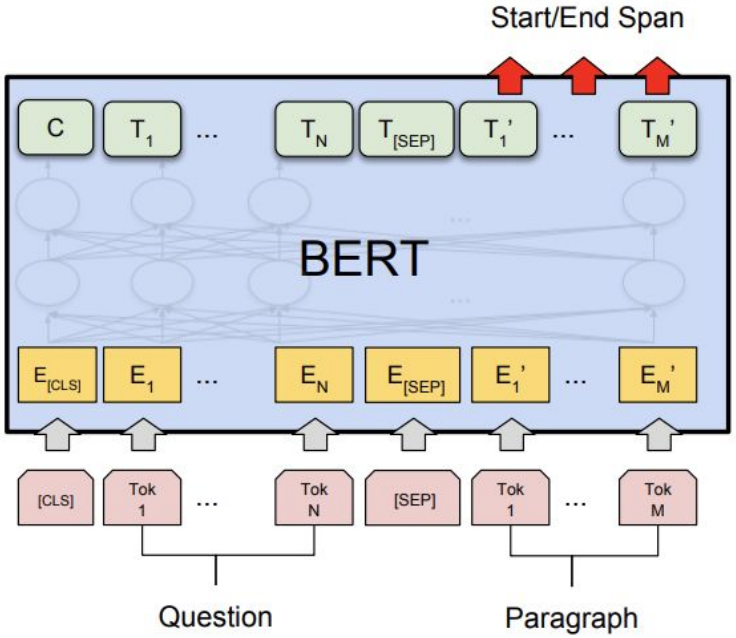


(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

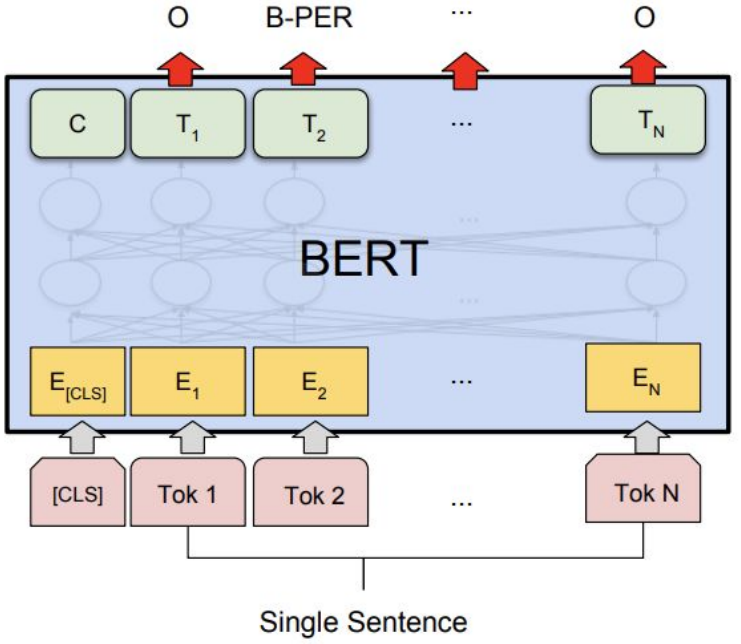


(b) Single Sentence Classification Tasks: SST-2, CoLA

BERT — Figure 4 From Paper



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

GPT-2

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] **Jeffrey Wu** [*1] **Rewon Child** [1] **David Luan** [1] **Dario Amodei** [**1] **Ilya Sutskever** [**1]

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al., 2017), reading comprehension systems (Jia & Liang, 2017), and image classifiers (Alcorn et al., 2018) on the diversity and variety of possible inputs highlights some of the shortcomings of this approach.

Our suspicion is that the prevalence of single task training on single domain datasets is a major contributor to the lack of generalization observed in current systems. Progress towards robust systems with current architectures is likely to require training and measuring performance on a wide range of domains and tasks. Recently, several benchmarks have been proposed such as GLUE (Wang et al., 2018) and decaNLP (McCann et al., 2018) to begin studying this.

Transformer-XL

Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context

Zihang Dai[*12], Zhilin Yang[*12], Yiming Yang[1], Jaime Carbonell[1], Quoc V. Le[2], Ruslan Salakhutdinov[1]
[1]Carnegie Mellon University, [2]Google Brain
{dzihang,zhiliny,yiming,jgc,rsalakhu}@cs.cmu.edu, qvl@google.com

Abstract

Transformers have a potential of learning longer-term dependency, but are limited by a fixed-length context in the setting of language modeling. We propose a novel neural architecture *Transformer-XL* that enables learning dependency beyond a fixed length without disrupting temporal coherence. It consists of a segment-level recurrence mechanism and a novel positional encoding scheme. Our method not only enables capturing longer-term dependency, but also resolves the context fragmentation problem. As a result, Transformer-XL learns dependency that is 80% longer than RNNs and 450% longer than vanilla Transformers, achieves better performance on both

Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), have been a standard solution to language modeling and obtained strong results on multiple benchmarks. Despite the wide adaption, RNNs are difficult to optimize due to gradient vanishing and explosion (Hochreiter et al., 2001), and the introduction of gating in LSTMs and the gradient clipping technique (Graves, 2013) might not be sufficient to fully address this issue. Empirically, previous work has found that LSTM language models use 200 context words on average (Khandelwal et al., 2018), indicating room for further improvement.

On the other hand, the direct connections between long-distance word pairs baked in atten-

XLNet

XLNet: Generalized Autoregressive Pretraining for Language Understanding

Zhilin Yang[*1], **Zihang Dai**[*12], **Yiming Yang**[1], **Jaime Carbonell**[1],
Ruslan Salakhutdinov[1], **Quoc V. Le**[2]
[1]Carnegie Mellon University, [2]Google Brain
{zhiliny,dzihang,yiming,jgc,rsalakhu}@cs.cmu.edu, qvl@google.com

Abstract

With the capability of modeling bidirectional contexts, denoising autoencoding based pretraining like BERT achieves better performance than pretraining approaches based on autoregressive language modeling. However, relying on corrupting the input with masks, BERT neglects dependency between the masked positions and suffers from a pretrain-finetune discrepancy. In light of these pros and cons, we propose XLNet, a generalized autoregressive pretraining method that (1) enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order and (2) overcomes the limitations of BERT thanks to its autoregressive formulation. Furthermore, XLNet integrates ideas from Transformer-XL, the state-of-the-art autoregressive model, into pretraining. Empirically, XLNet outperforms BERT on 20 tasks, often by a large margin, and achieves state-of-the-art results on 18 tasks including question answering, natural

RoBERTa

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu[*§] **Myle Ott**[*§] **Naman Goyal**[*§] **Jingfei Du**[*§] **Mandar Joshi**[†]
Danqi Chen[§] **Omer Levy**[§] **Mike Lewis**[§] **Luke Zettlemoyer**[†§] **Veselin Stoyanov**[§]

[†] Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA
{mandar90,lsz}@cs.washington.edu

[§] Facebook AI
{yinhanliu,myleott,naman,jingfeidu,
danqi,omerlevy,mikelewis,lsz,ves}@fb.com

Abstract

Language model pretraining has led to significant performance gains but careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes, and, as we will show, hyperparameter choices have significant impact on the final results. We present a replication study of BERT pretraining (Devlin et al., 2019) that carefully measures the impact of many key hyperparameters and training data size. We find that BERT was significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE and SQuAD. These results highlight the importance of previously overlooked design choices,

We present a replication study of BERT pretraining (Devlin et al., 2019), which includes a careful evaluation of the effects of hyperparmeter tuning and training set size. We find that BERT was significantly undertrained and propose an improved recipe for training BERT models, which we call RoBERTa, that can match or exceed the performance of all of the post-BERT methods. Our modifications are simple, they include: (1) training the model longer, with bigger batches, over more data; (2) removing the next sentence prediction objective; (3) training on longer sequences; and (4) dynamically changing the masking pattern applied to the training data. We also collect a large new dataset (CC-NEWS) of comparable size to other privately used datasets, to better control for training set size effects.

T5

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel* **Noam Shazeer*** **Adam Roberts*** **Katherine Lee***
Sharan Narang **Michael Matena** **Yanqi Zhou** **Wei Li** **Peter J. Liu**

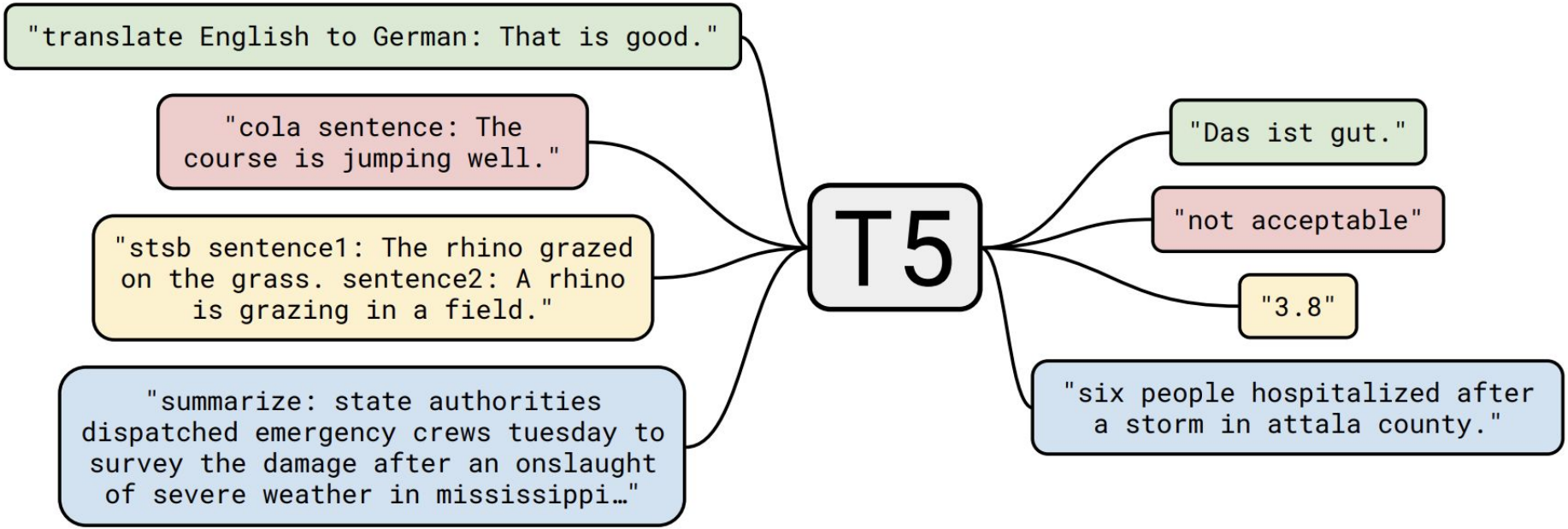
Google

Abstract

Transfer learning, where a model is first pre-trained on a data-rich task before being fine-tuned on a downstream task, has emerged as a powerful technique in natural language processing (NLP). The effectiveness of transfer learning has given rise to a diversity of approaches, methodology, and practice. In this paper, we explore the landscape of transfer learning techniques for NLP by introducing a unified framework that converts every language problem into a text-to-text format. Our systematic study compares pre-training objectives, architectures, unlabeled datasets, transfer approaches, and other factors on dozens of language understanding tasks. By combining the insights from our exploration with scale and our new "Colossal Clean Crawled Corpus", we achieve state-of-the-art results on many benchmarks covering summarization, question answering, text classification, and more. To facilitate future work on transfer learning for NLP, we release our dataset, pre-trained models, and code.[1]

1 Introduction

T5 — Figure 1 From Paper



Get Coding!

×

https://github.com/huggingface/transformers

https://medium.com/@lysandrejik

Thank you!