Transformers
Pre-trained Language Models

LING572 Advanced Statistical Methods for NLP

March 10, 2020

Announcements

- Thanks for being here!
- Please be active on Zoom chat! That's the only form of interaction; I won't be able to tell what's sticking and what's not without the physical classroom and its visual cues.
- HW7: excellent. 94 avg, no major comments.
- HW9: will post this afternoon. Deep Averaging Network for text classification; you will implement: linear layer, L2 regularization, early stopping.
- Office hours today: https://washington.zoom.us/my/shanest

Outline

- Transformer Architecture
- Transfer learning and pre-training
  - History / main idea
  - In NLP: ELMo, BERT, …

Transformer Architecture

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

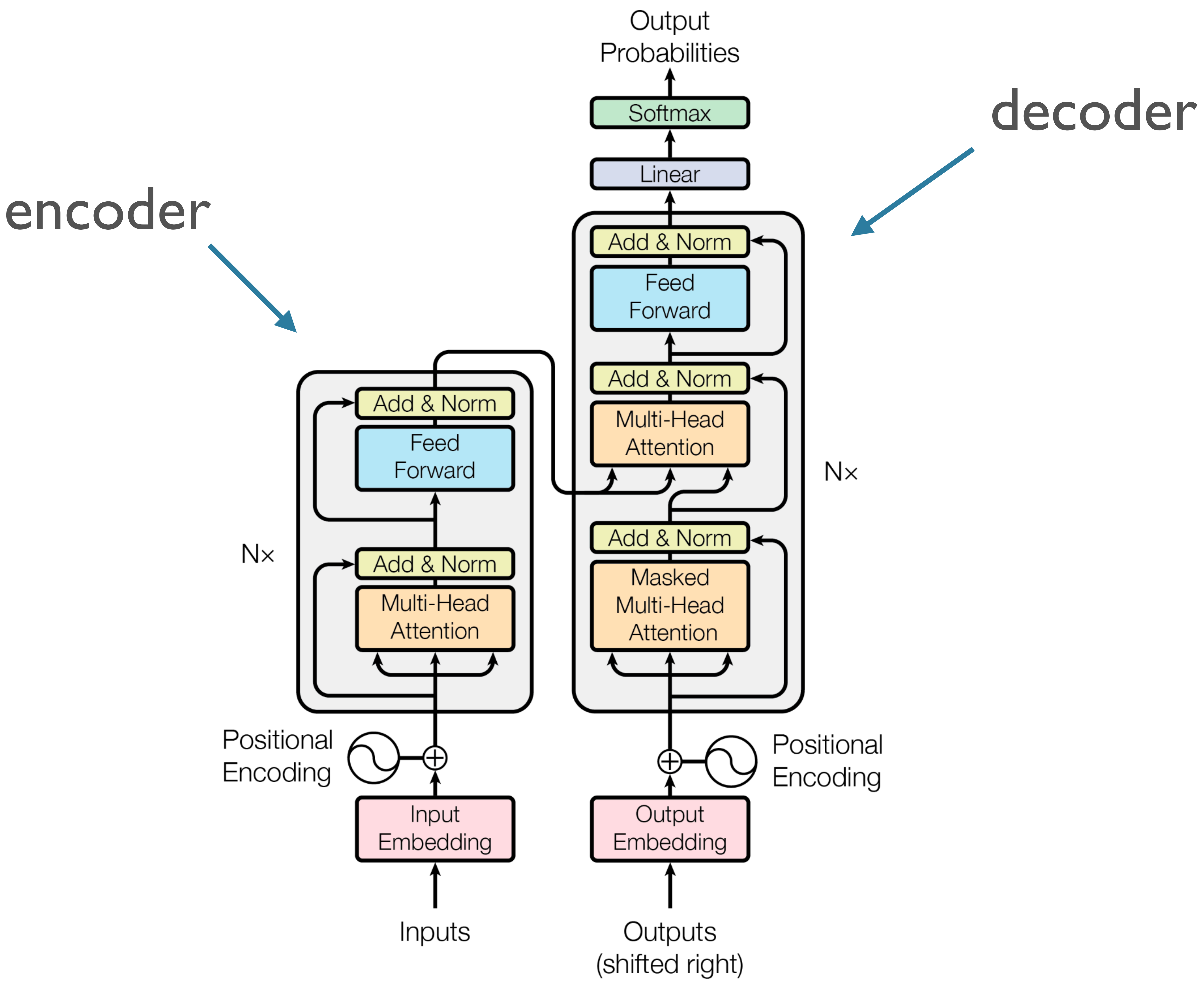
Abstract

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 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Paper link

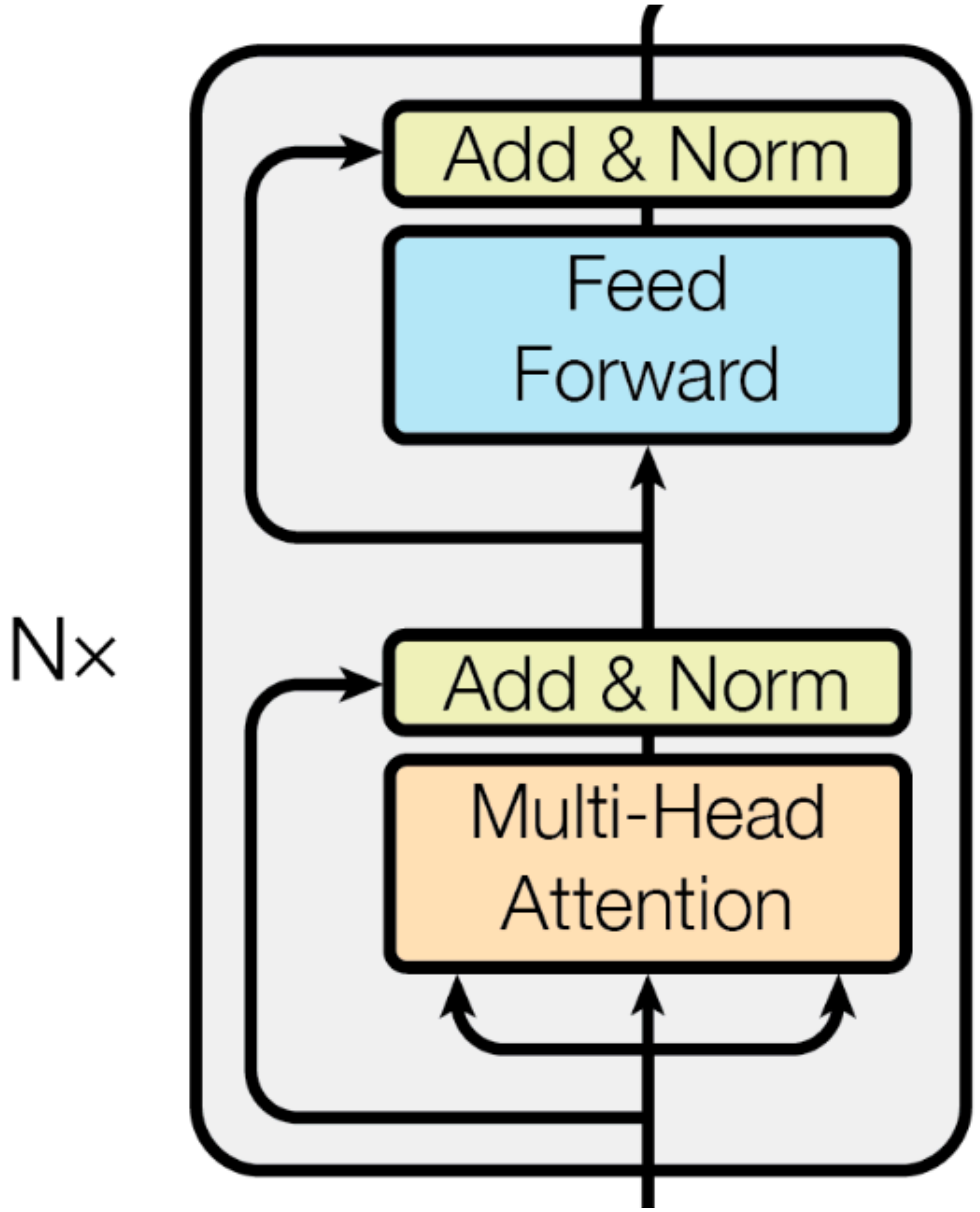
(but see Annotated and Illustrated Transformer)

Full Model

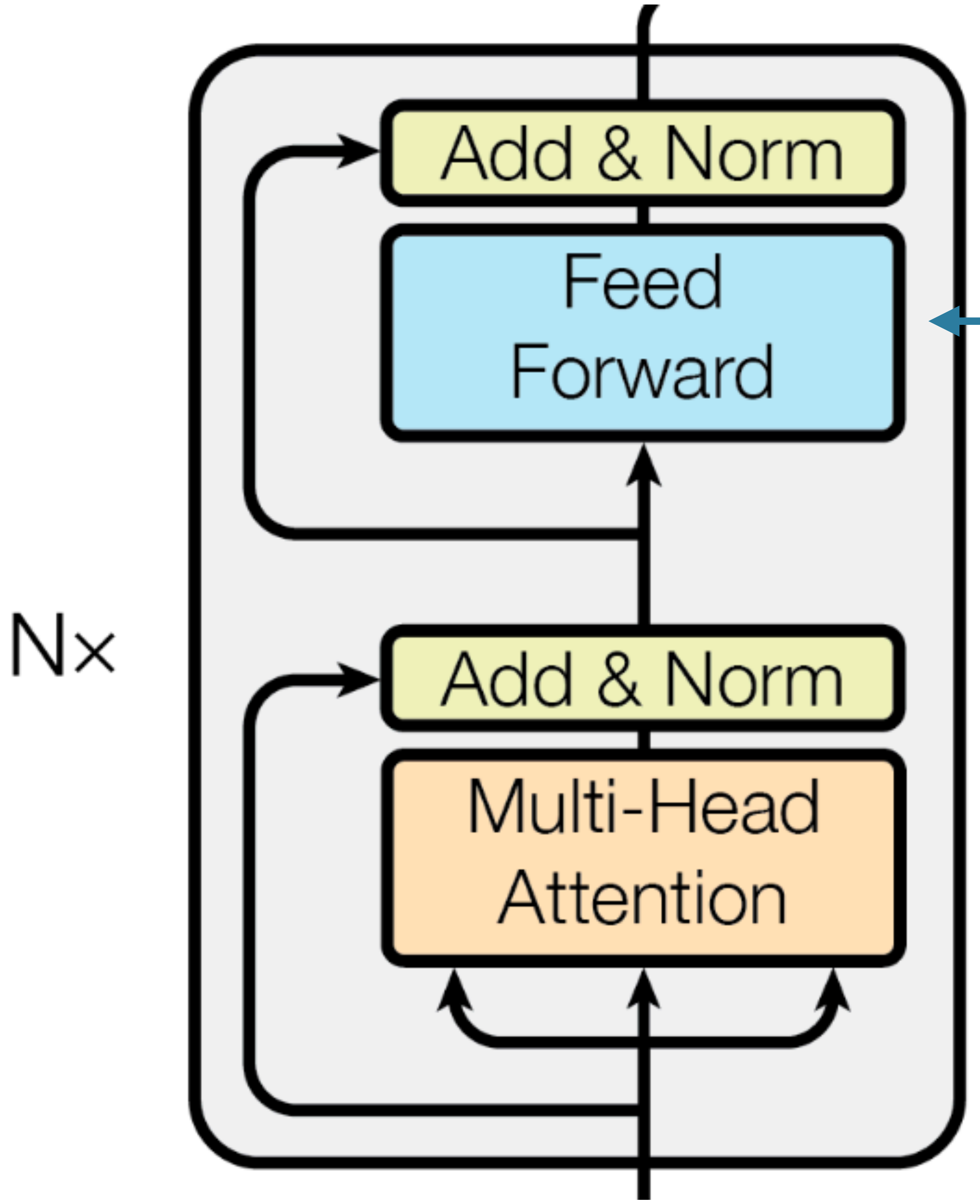


encoder
decoder
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

Transformer Block

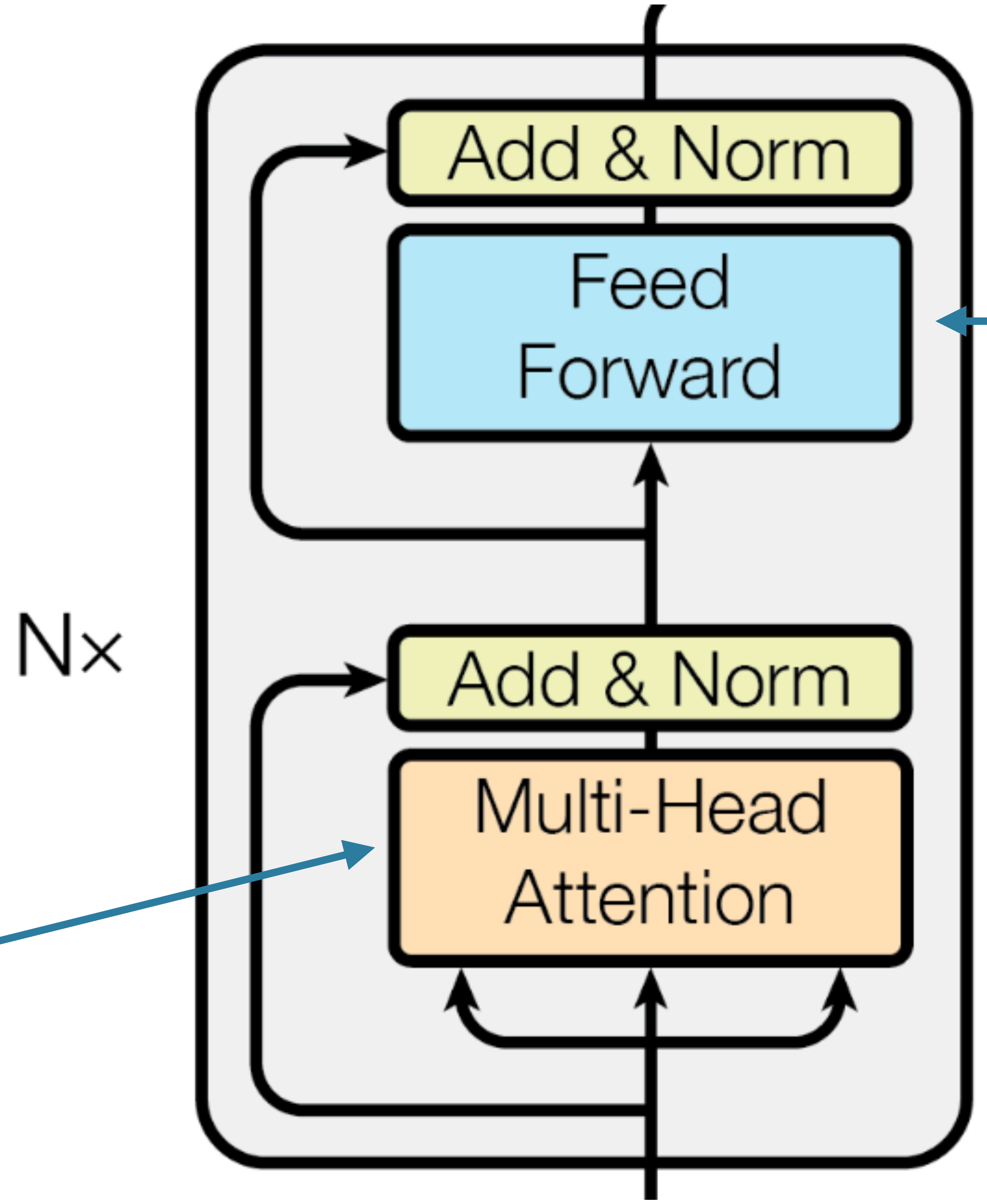


Transformer Block



Single layer, applied to each position

Transformer Block



Single layer, applied to each position

What's this?

Scaled Dot-Product Attention

- Recall:

- Putting it together: (keys/values in matrices)

$$\text{Attention}(q, K, V) = \sum_j \frac{e^{q \cdot k_j}}{\sum_i e^{q \cdot k_i}} v_j$$

- Stacking *multiple* queries: (and scaling)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Scaled Dot-Product Attention

- Recall:

$$\alpha_j = q \cdot k_j$$

$$e_j = e^{\alpha_j} / \Sigma_j e^{\alpha_j}$$

$$c = \Sigma_j e_j v_j$$

- Putting it together:
  (keys/values in matrices)

$$\text{Attention}(q, K, V) = \sum_j \frac{e^{q \cdot k_j}}{\sum_i e^{q \cdot k_i}} v_j$$

- Stacking *multiple* queries:
  (and scaling)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Scaled Dot-Product Attention

- Recall:

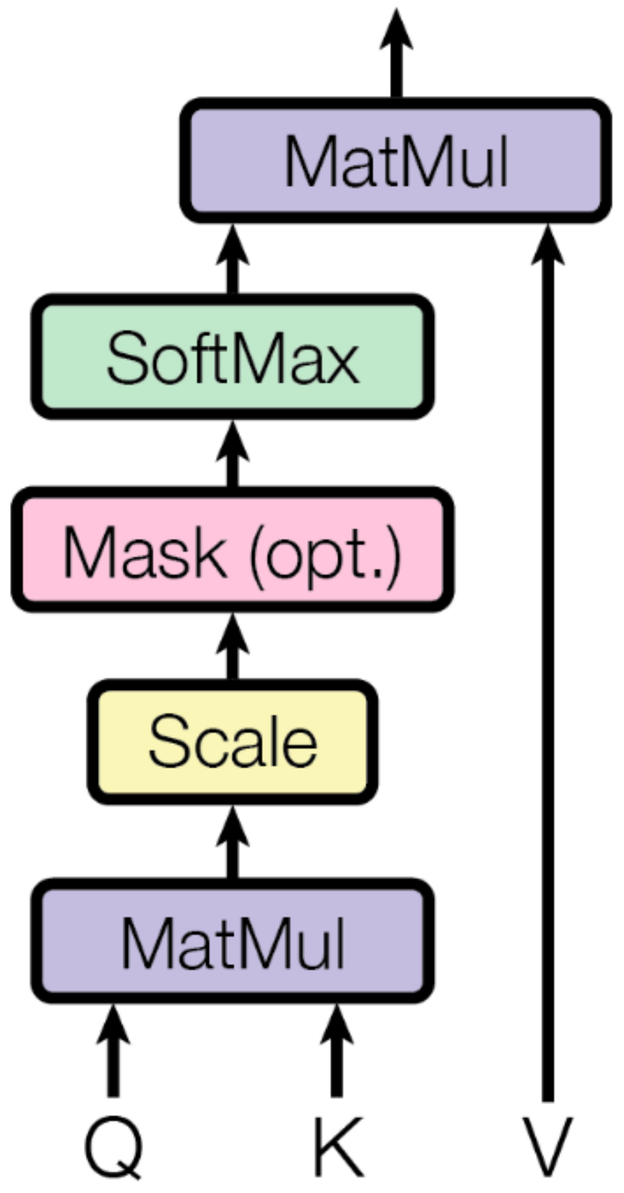
$$\alpha_j = q \cdot k_j$$
$$e_j = e^{\alpha_j} / \Sigma_j e^{\alpha_j}$$
$$c = \Sigma_j e_j v_j$$

- Putting it together:
  (keys/values in matrices)

$$\text{Attention}(q, K, V) = \sum_j \frac{e^{q \cdot k_j}}{\sum_i e^{q \cdot k_i}} v_j$$

- Stacking *multiple* queries:
  (and scaling)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$



Why multiple queries?

Why multiple queries?

- seq2seq: single decoder token attends to *all* encoder states

Why multiple queries?

- seq2seq: single decoder token attends to *all* encoder states
- Transformer: *self*-attention
  - Every (token) position attends to every other position [including self!]
  - Caveat: in the encoder, and only by default
    - Mask in decoder to attend only to previous positions
    - Masking technique applied in some Transformer-based LMs
  - So vector at each position is a query

Multi-headed Attention

- So far: a *single* attention mechanism.
- Could be a bottleneck: need to pay attention to different vectors *for different reasons*
- Multi-headed: several attention mechanisms in parallel

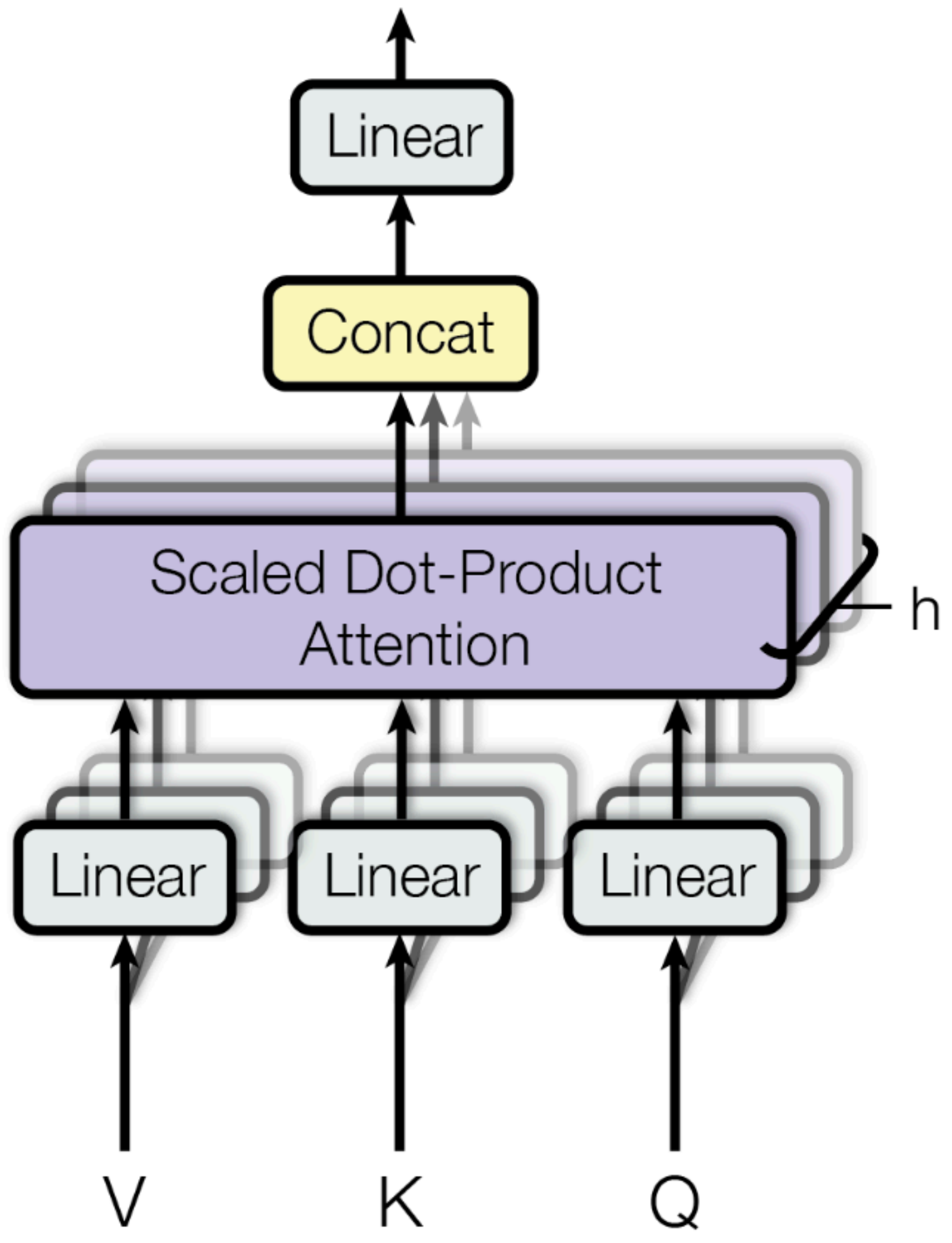
Multi-headed Attention

- So far: a *single* attention mechanism.
- Could be a bottleneck: need to pay attention to different vectors *for different reasons*
- Multi-headed: several attention mechanisms in parallel

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Multi-headed Attention

- So far: a *single* attention mechanism.
- Could be a bottleneck: need to pay attention to different vectors *for different reasons*
- Multi-headed: several attention mechanisms in parallel



$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

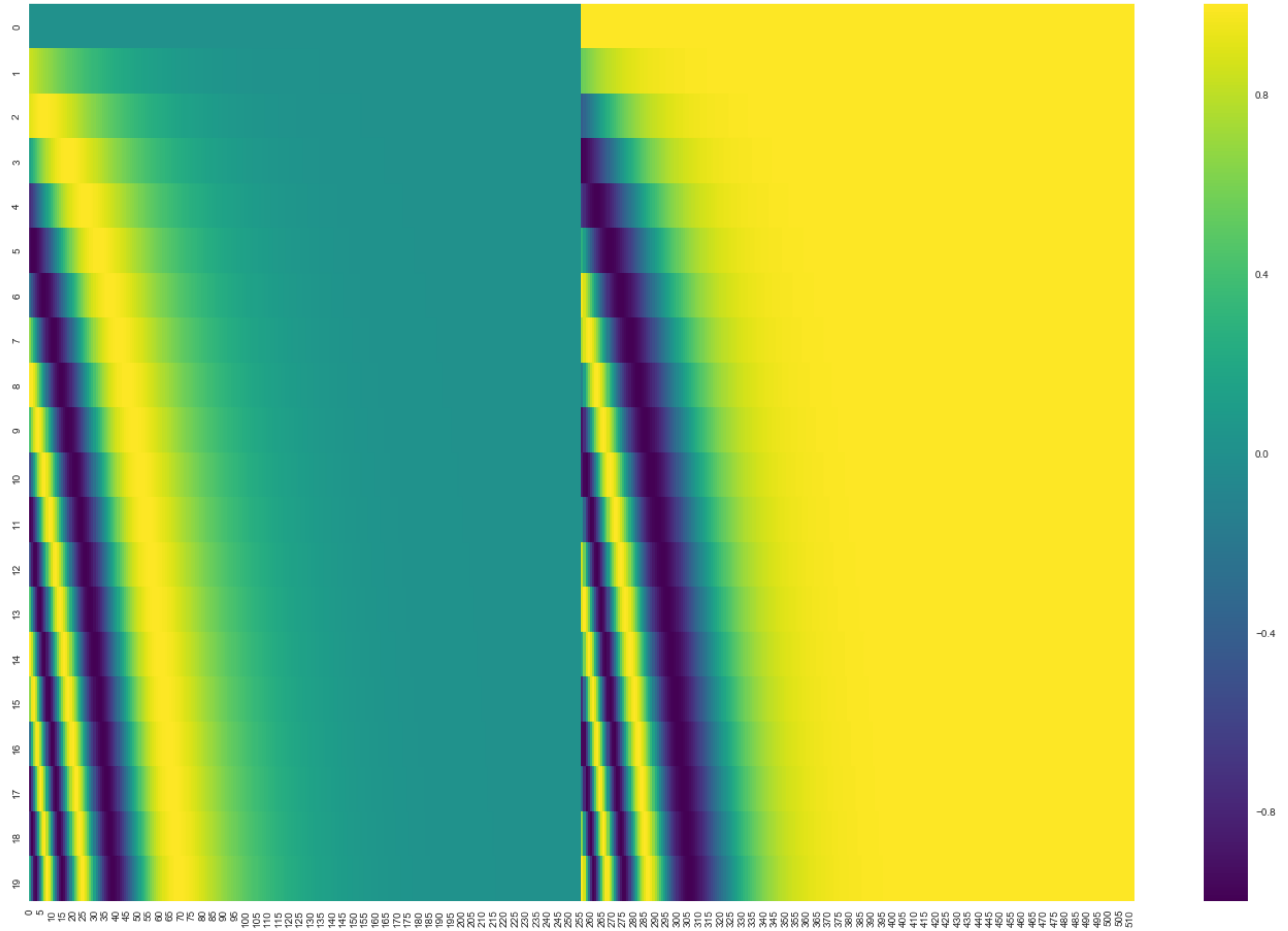
Representing Order

Representing Order

- No notion of order in Transformer. Represented via *positional* encodings.

Representing Order

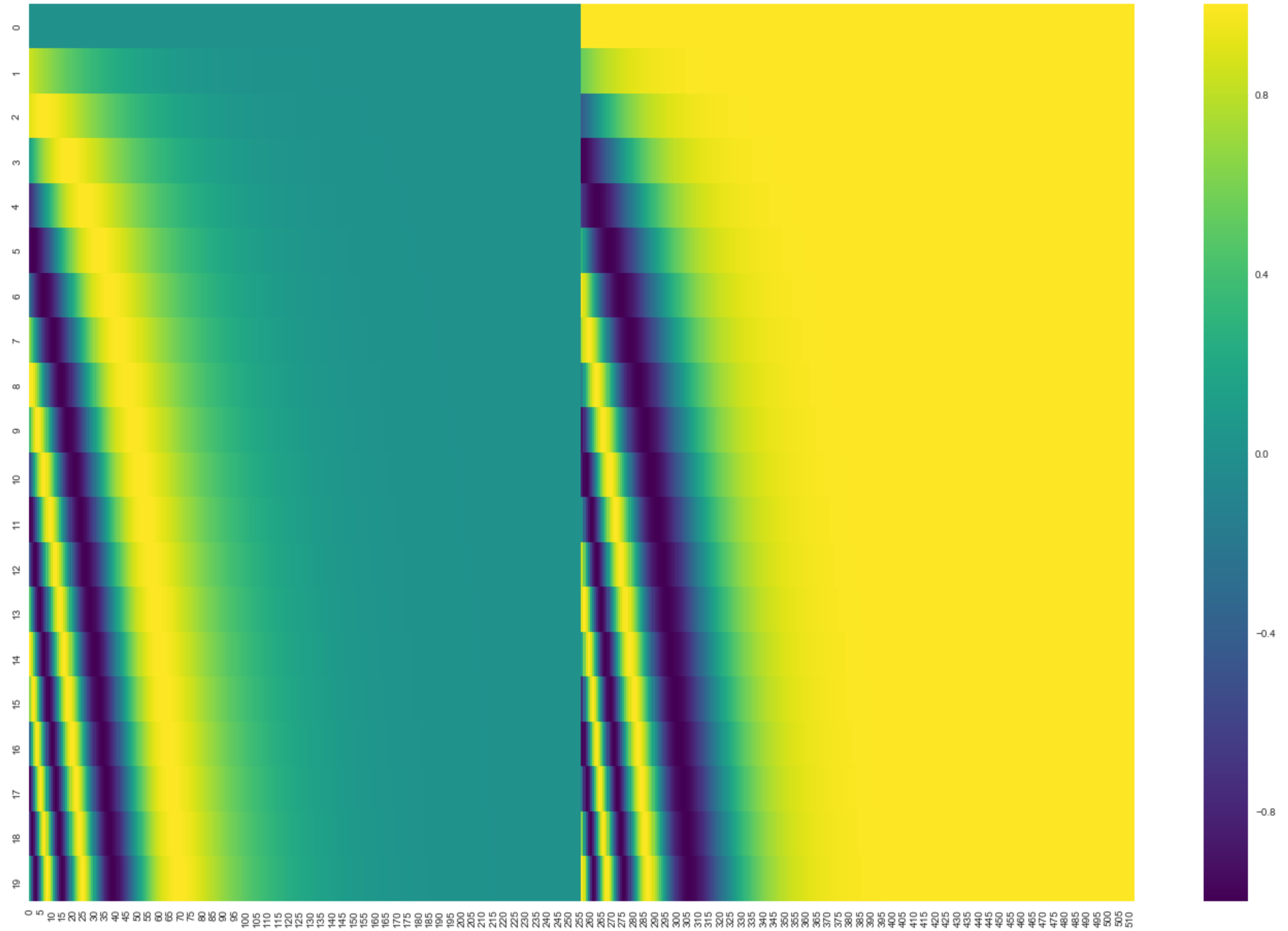
- No notion of order in Transformer. Represented via *positional* encodings.



[source]

Representing Order

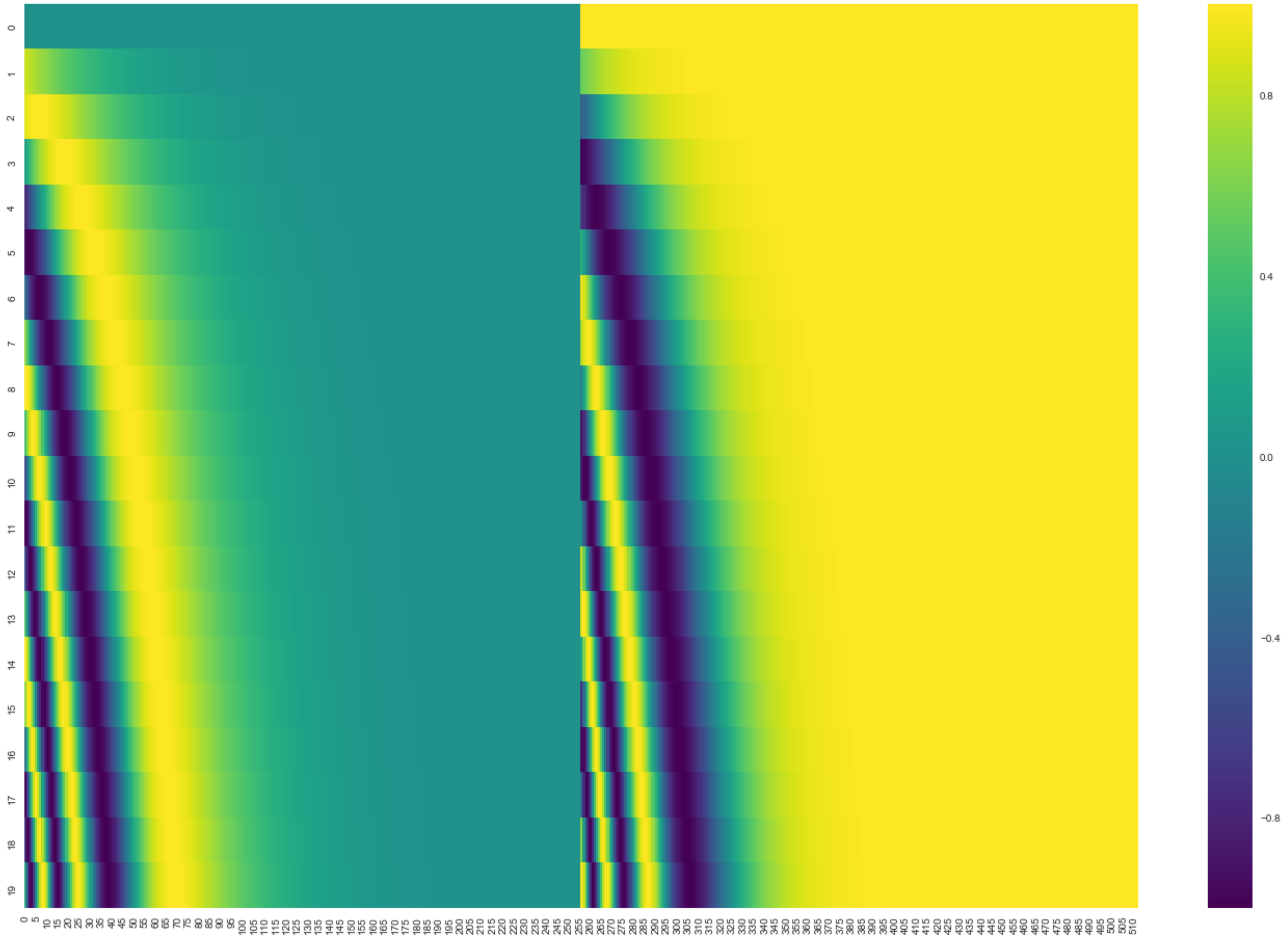
- No notion of order in Transformer. Represented via *positional* encodings.
- Usually fixed, though can be learned.



source

Representing Order

- No notion of order in Transformer. Represented via *positional* encodings.
- Usually fixed, though can be learned.
  - No significant improvement; less generalization.



source

Initial WMT Results

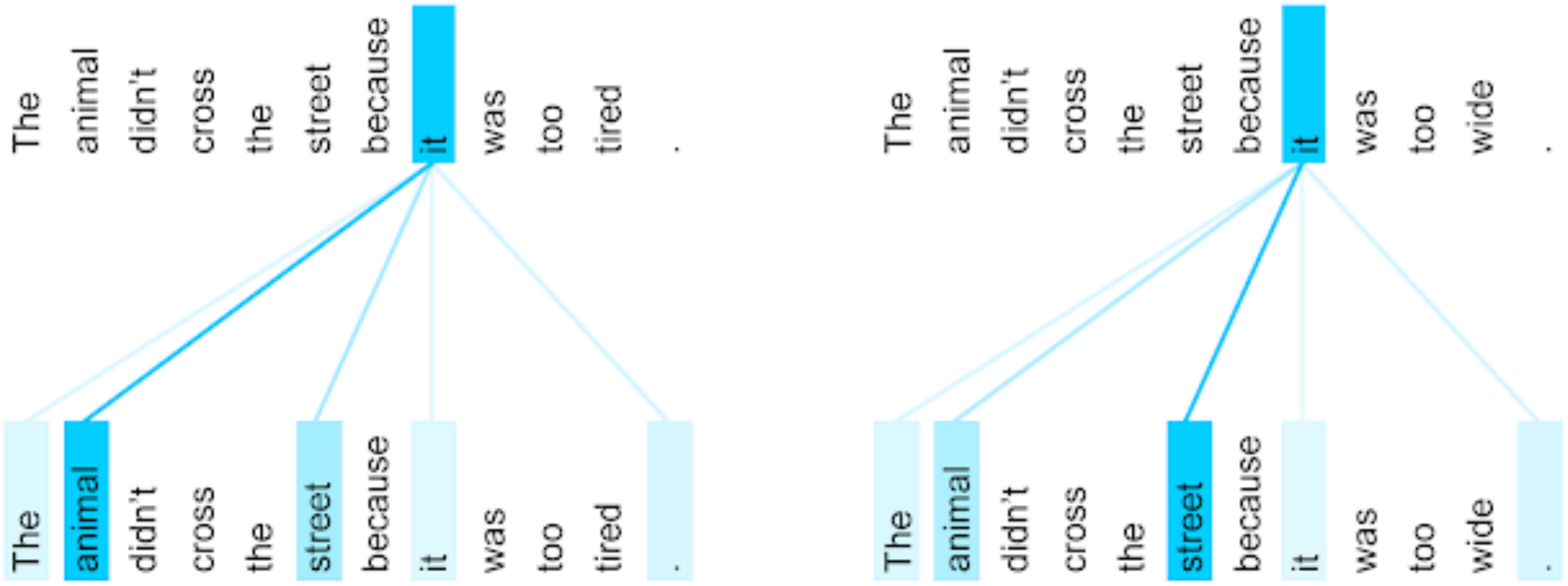
Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

Initial WMT Results

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

More on why important later

Attention Visualization: Coreference?



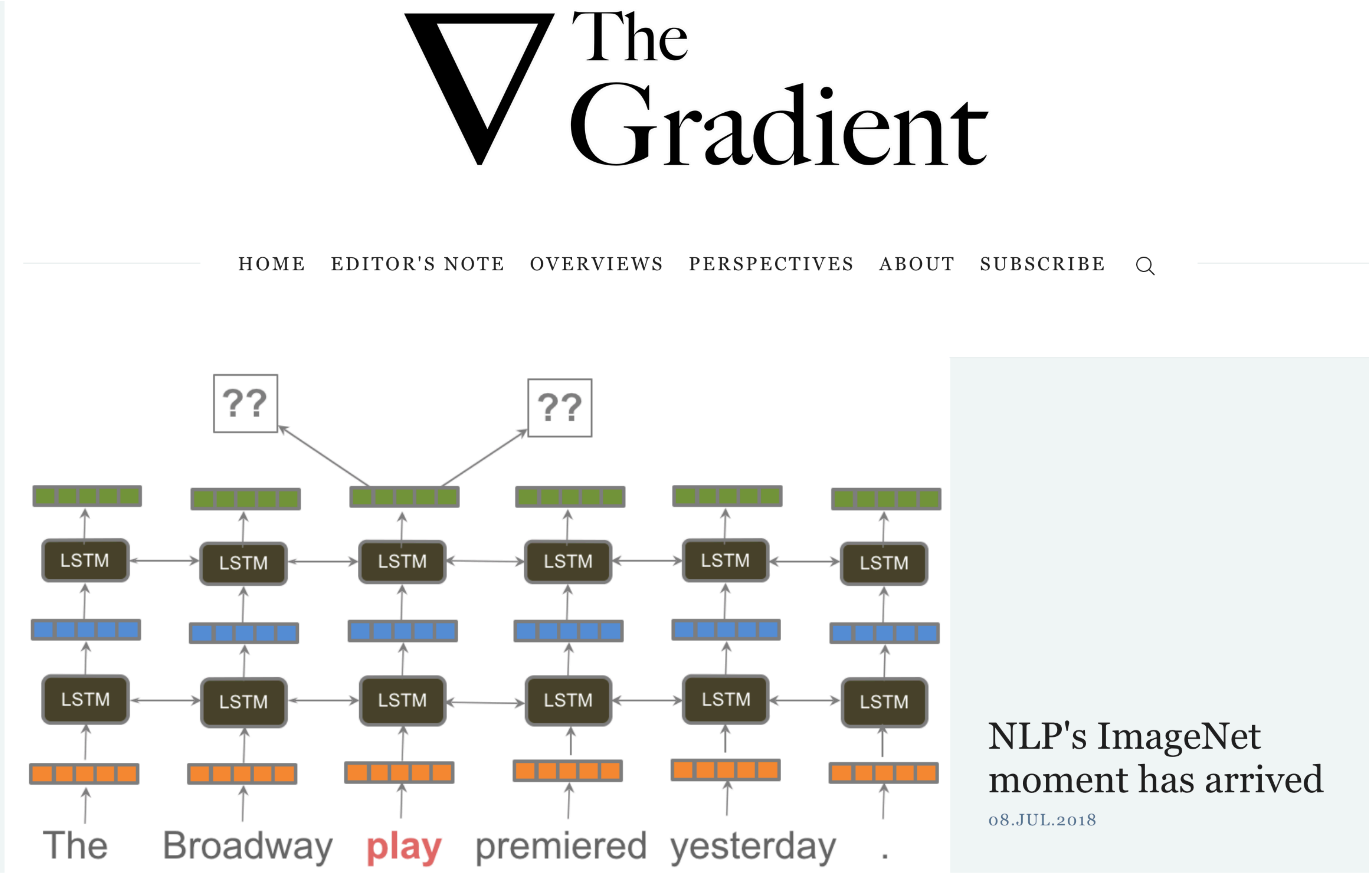
source

Transformer: Summary

- *Entirely* feed-forward
  - Therefore massively parallelizable
  - RNNs are inherently sequential, a parallelization bottleneck
- (Self-)attention everywhere
- Long-term dependencies:
  - LSTM: has to maintain representation of early item
  - Transformer: very short “path-lengths”

Transfer Learning and Pre-training

NLP's "ImageNet Moment"



link

What is ImageNet?

ImageNet: A Large-Scale Hierarchical Image Database

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei
Dept. of Computer Science, Princeton University, USA
{jiadeng, wdong, rsocher, jial, li, feifeili}@cs.princeton.edu

Abstract

The explosion of image data on the Internet has the potential to foster more sophisticated and robust models and algorithms to index, retrieve, organize and interact with images and multimedia data. But exactly how such data can be harnessed and organized remains a critical problem. We *content-based image search and image understanding algorithms, as well as for providing critical training and benchmarking data for such algorithms.*

ImageNet uses the hierarchical structure of WordNet [9]. Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a "synonym set" or "synset". There are around $80,000$ noun synsets

CVPR '09

Why is ImageNet Important?

ATURED QUARTZ EMAI

IT'S NOT ABOUT THE ALGORITHM

The data that transformed AI research—and possibly the world

By Dave Gershgorn • July 26, 2017

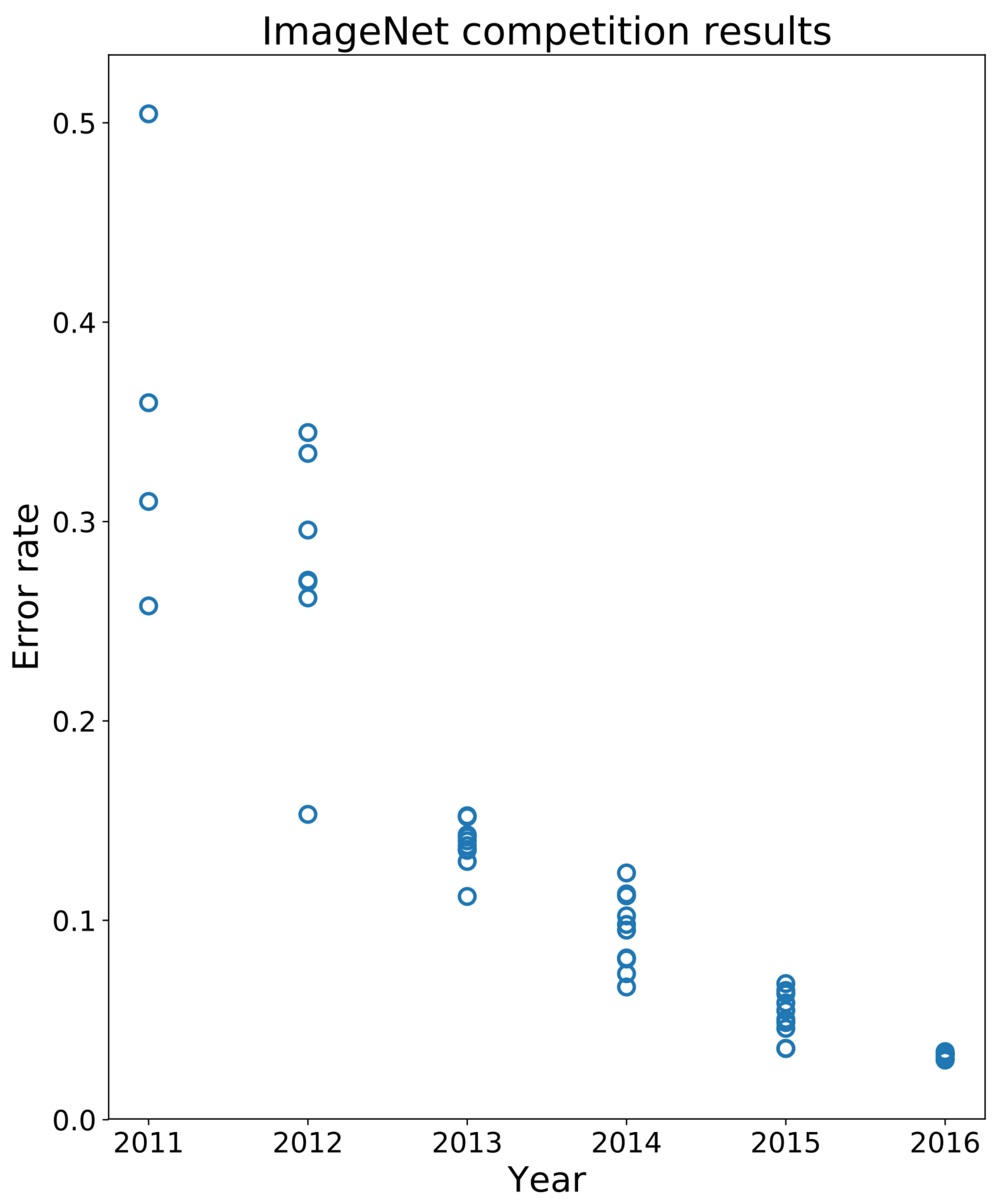
[link]

Why is ImageNet Important?

link

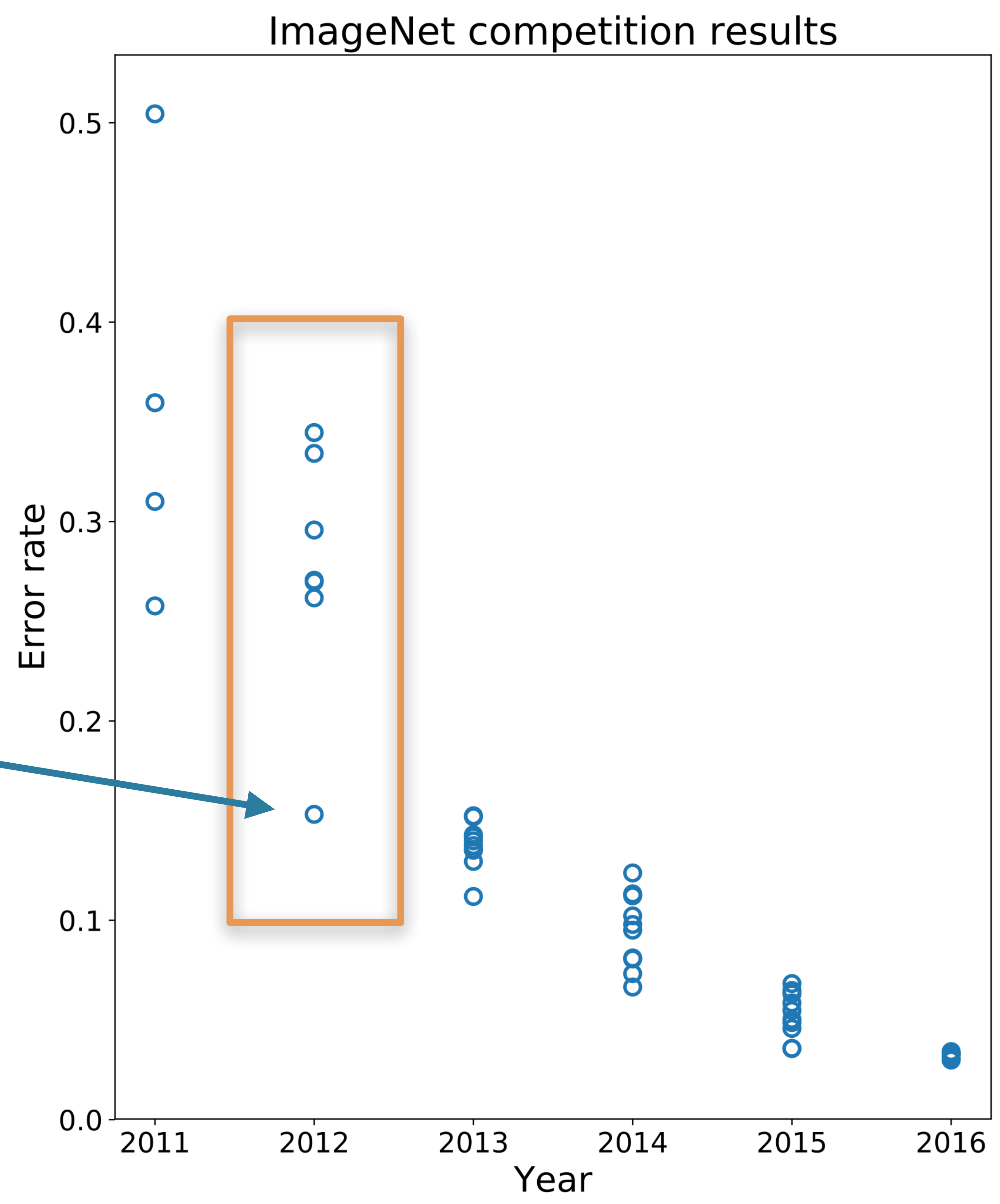
1. Deep learning
2. Transfer learning

ILSVRC results



source

ILSVRC results



AlexNet (CNN)

source

Transfer Learning

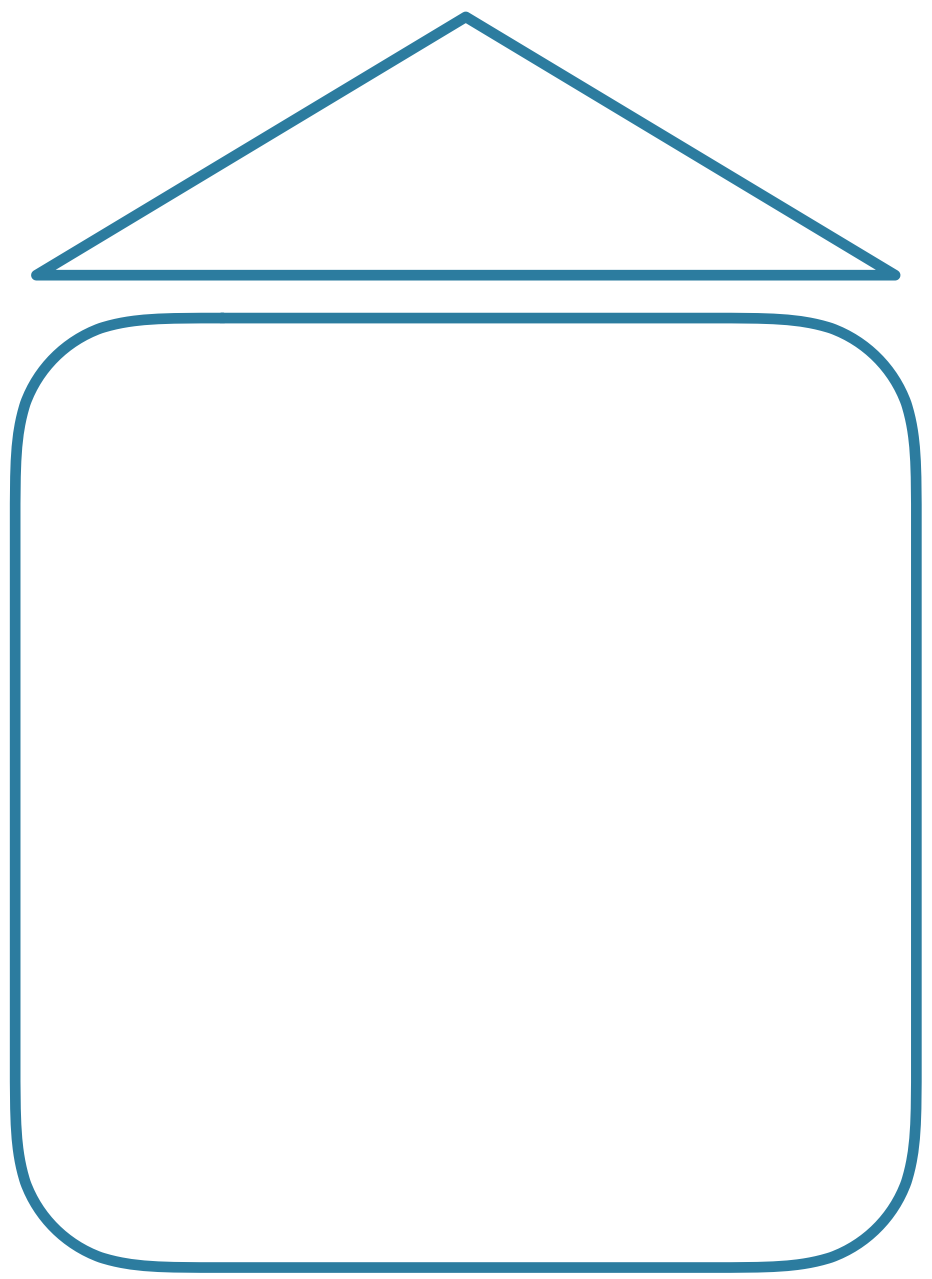
CNN Features off-the-shelf: an Astounding Baseline for Recognition

Ali Sharif Razavian Hossein Azizpour Josephine Sullivan Stefan Carlsson
CVAP, KTH (Royal Institute of Technology)
Stockholm, Sweden
{razavian,azizpour,sullivan,stefanc}@csc.kth.se

"We use features extracted from the `OverFeat` network as a generic image representation to tackle the diverse range of recognition tasks of object image classification, scene recognition, fine grained recognition, attribute detection and image retrieval applied to a diverse set of datasets. We selected these tasks and datasets as they grad-ually move further away from the original task and data the `OverFeat` network was trained to solve [cf. ImageNet]. Astonishingly, we report consistent superior results compared to the highly tuned state-of-the-art systems in all the visual classification tasks on various datasets"

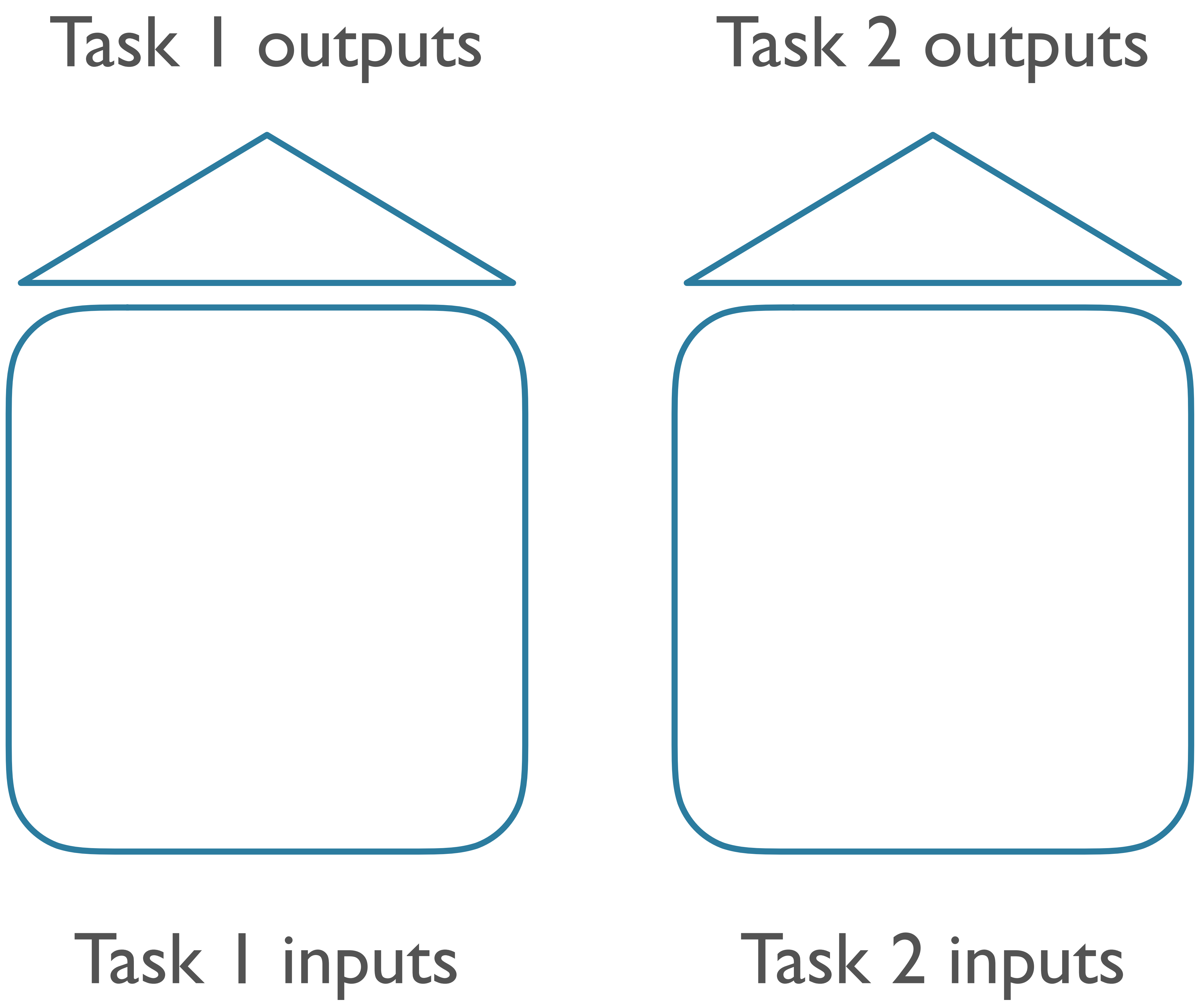
Standard Supervised Learning

Task 1 outputs

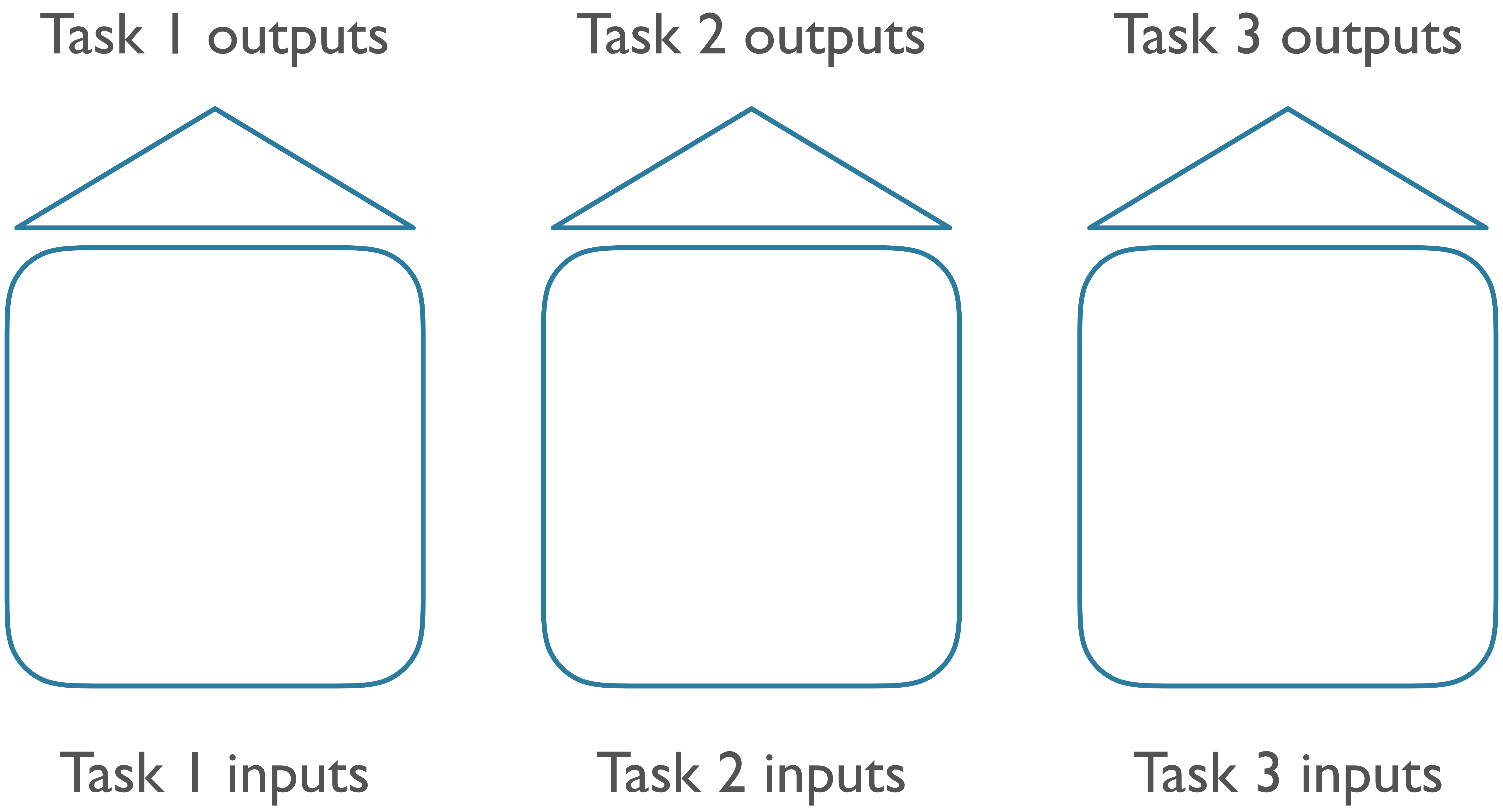


Task 1 inputs

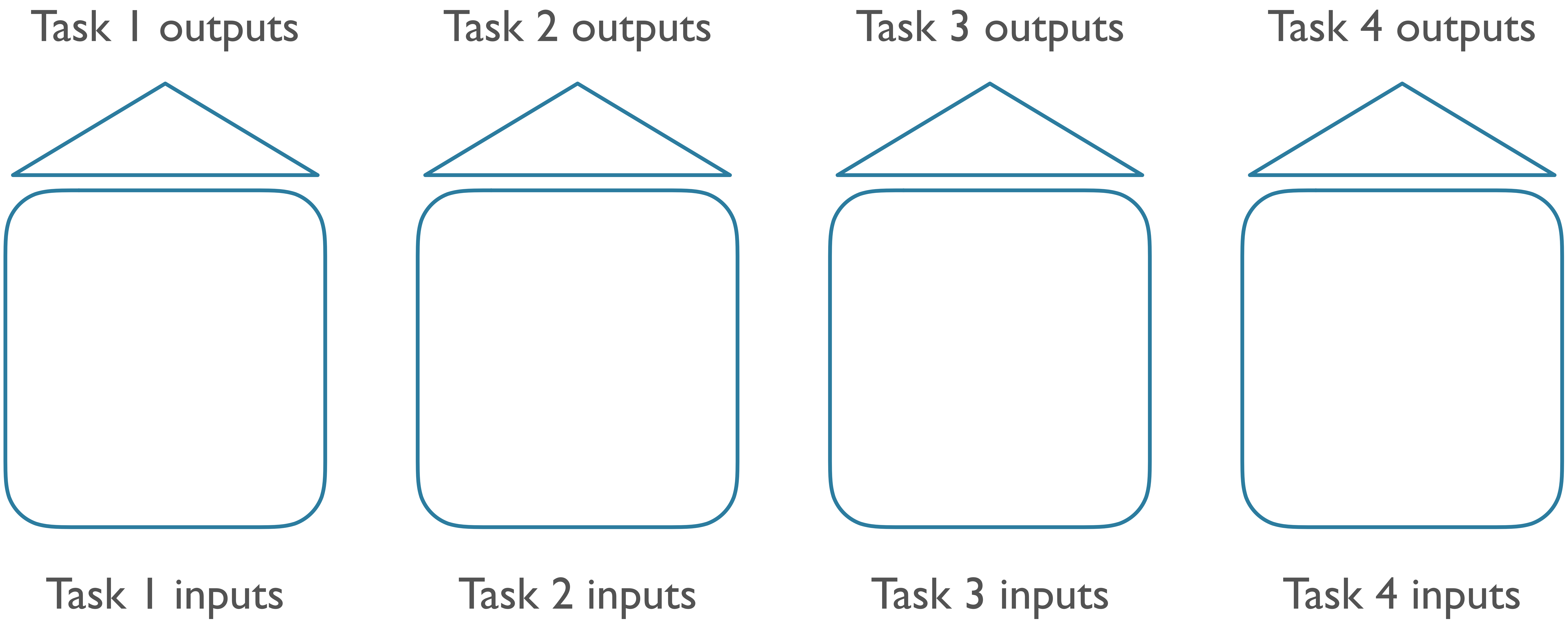
Standard Supervised Learning



Standard Supervised Learning



Standard Supervised Learning

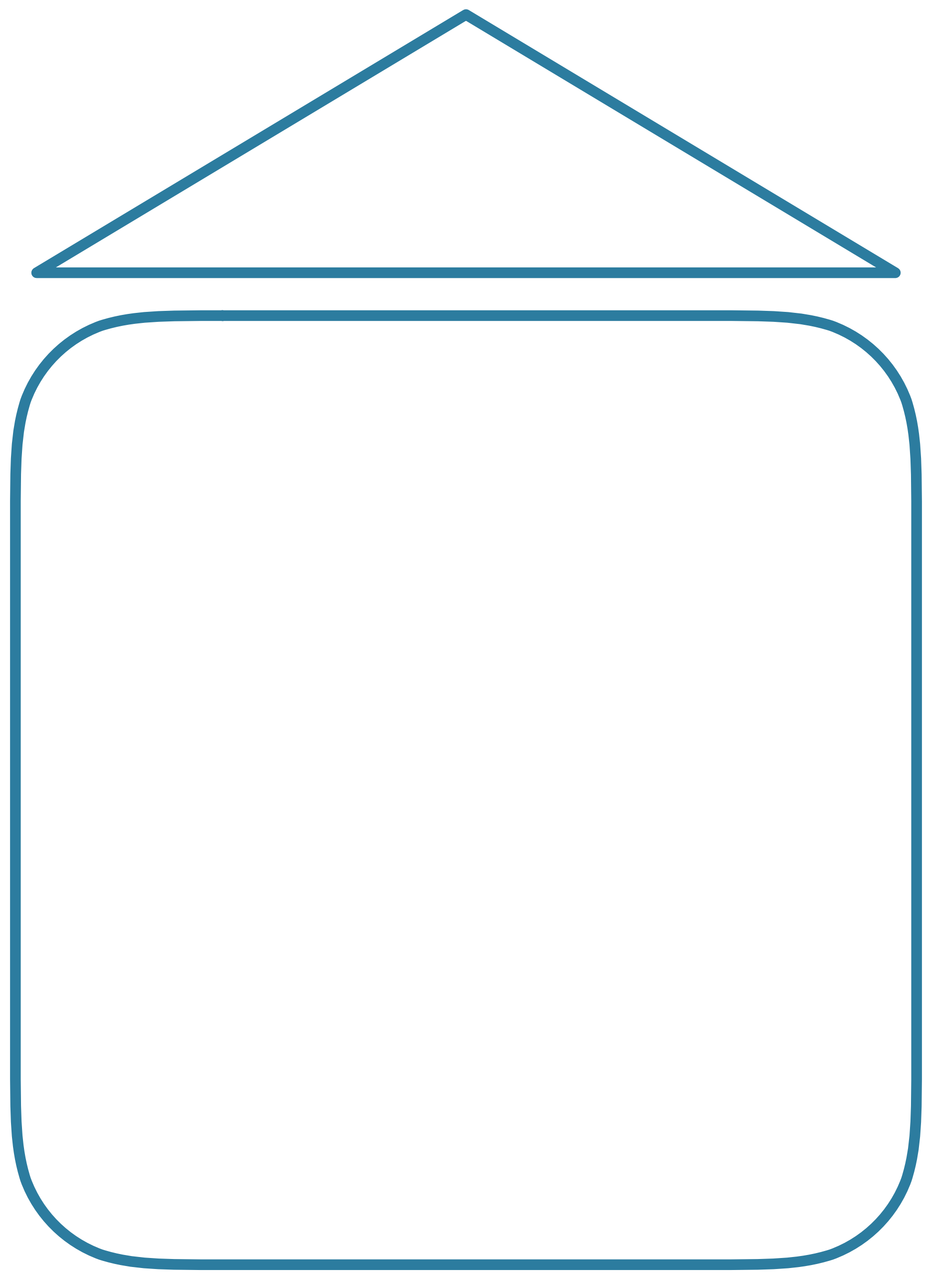


Standard Learning

- New task = new model
- Expensive!
  - Training time
  - Storage space
  - Data availability
    - Can be impossible in low-data regimes

Transfer Learning

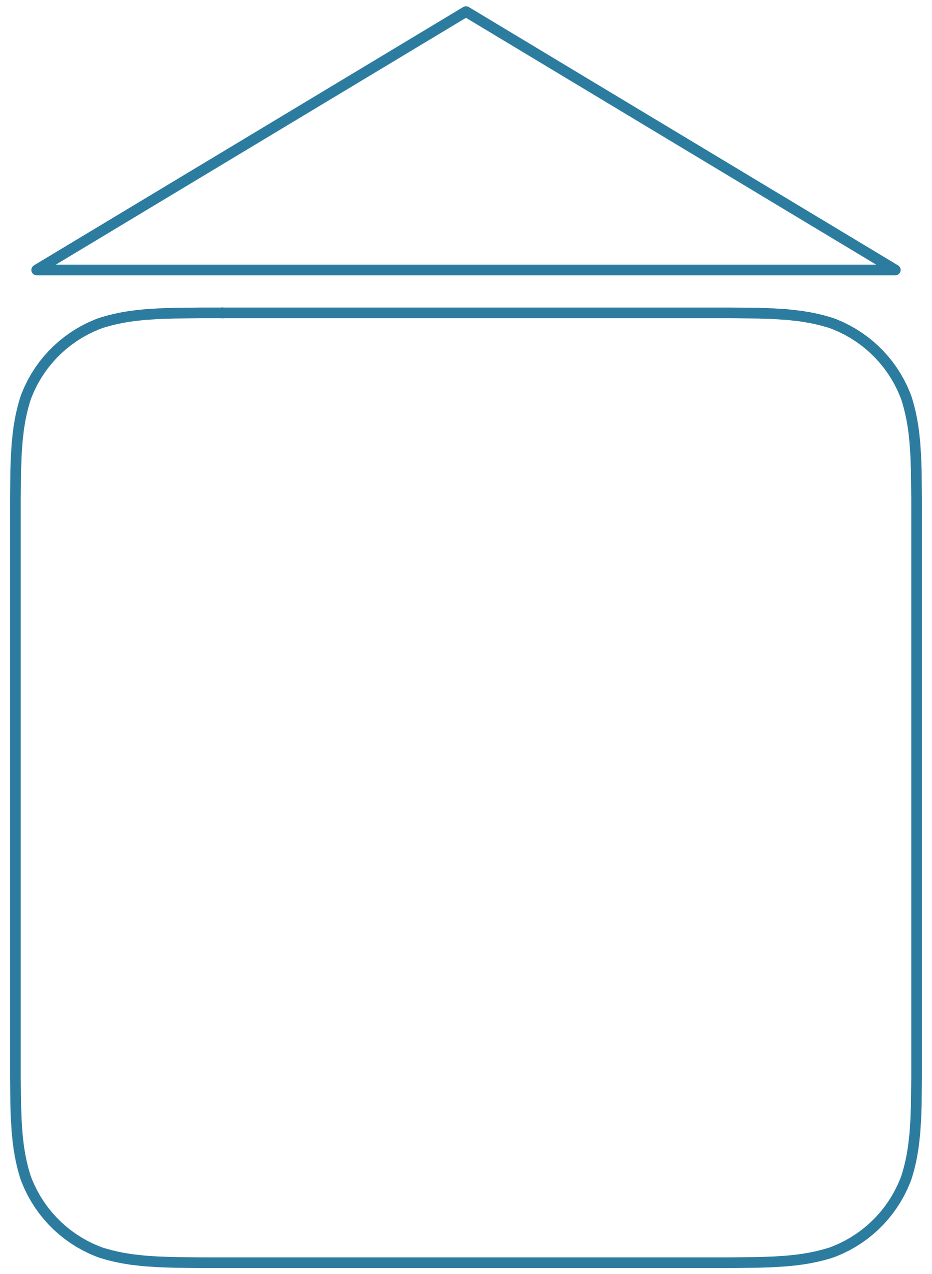
“pre-training” task outputs



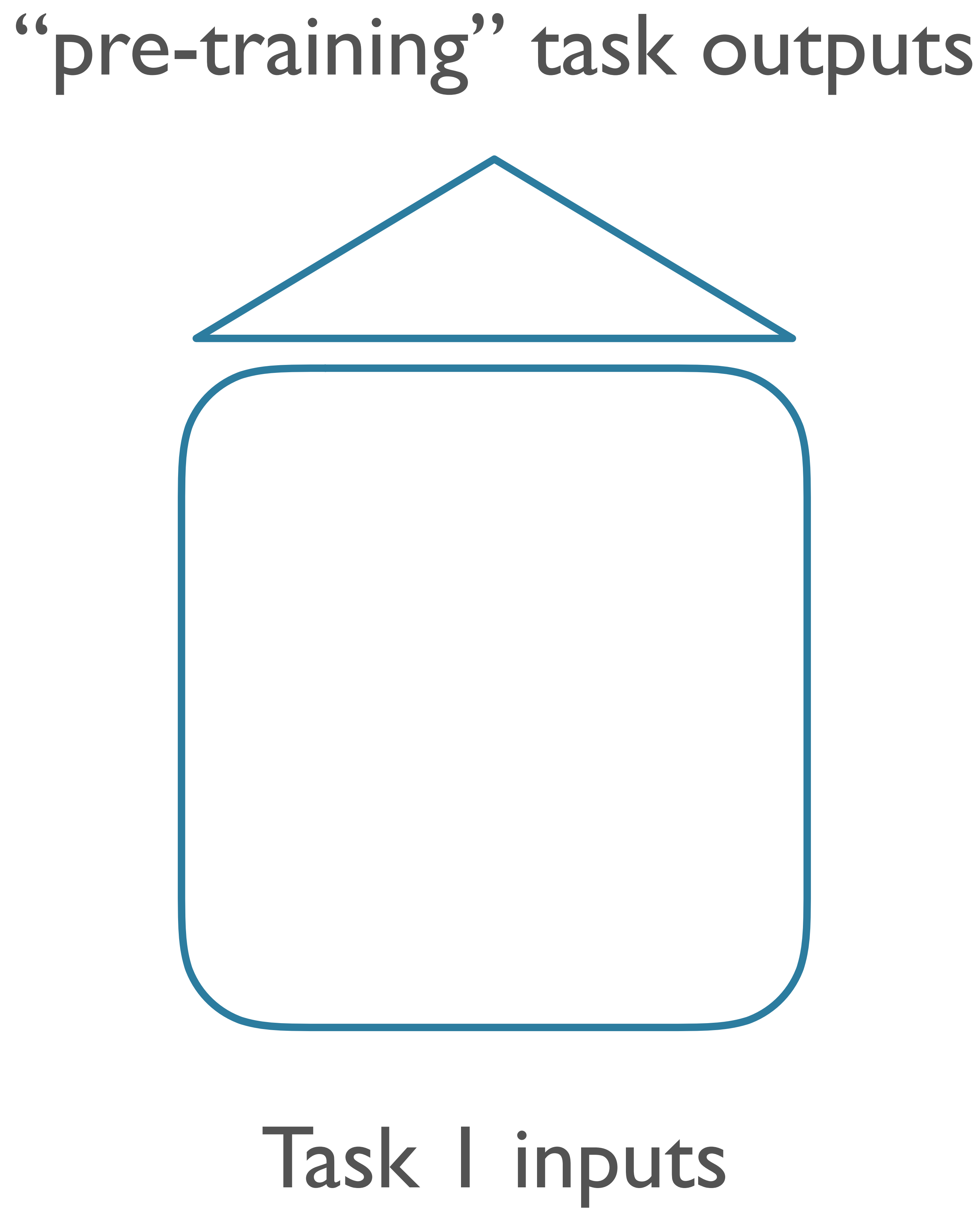
“pre-training” task inputs

Transfer Learning

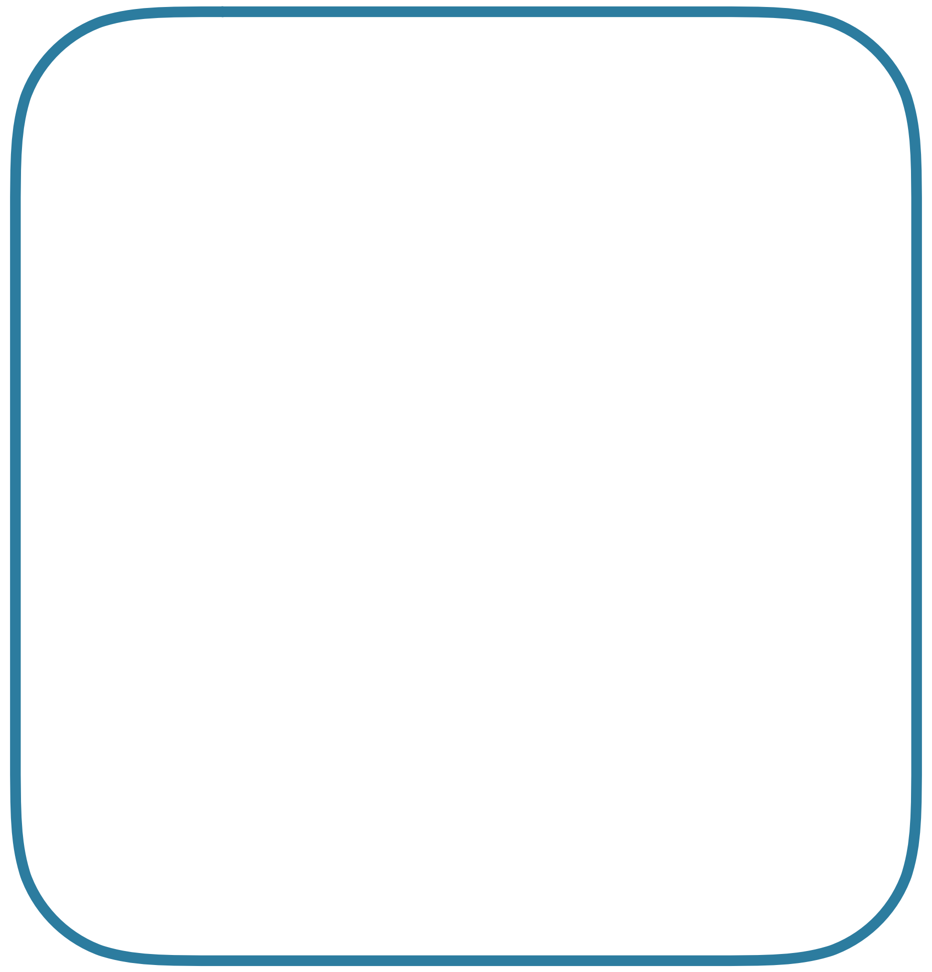
"pre-training" task outputs



Transfer Learning

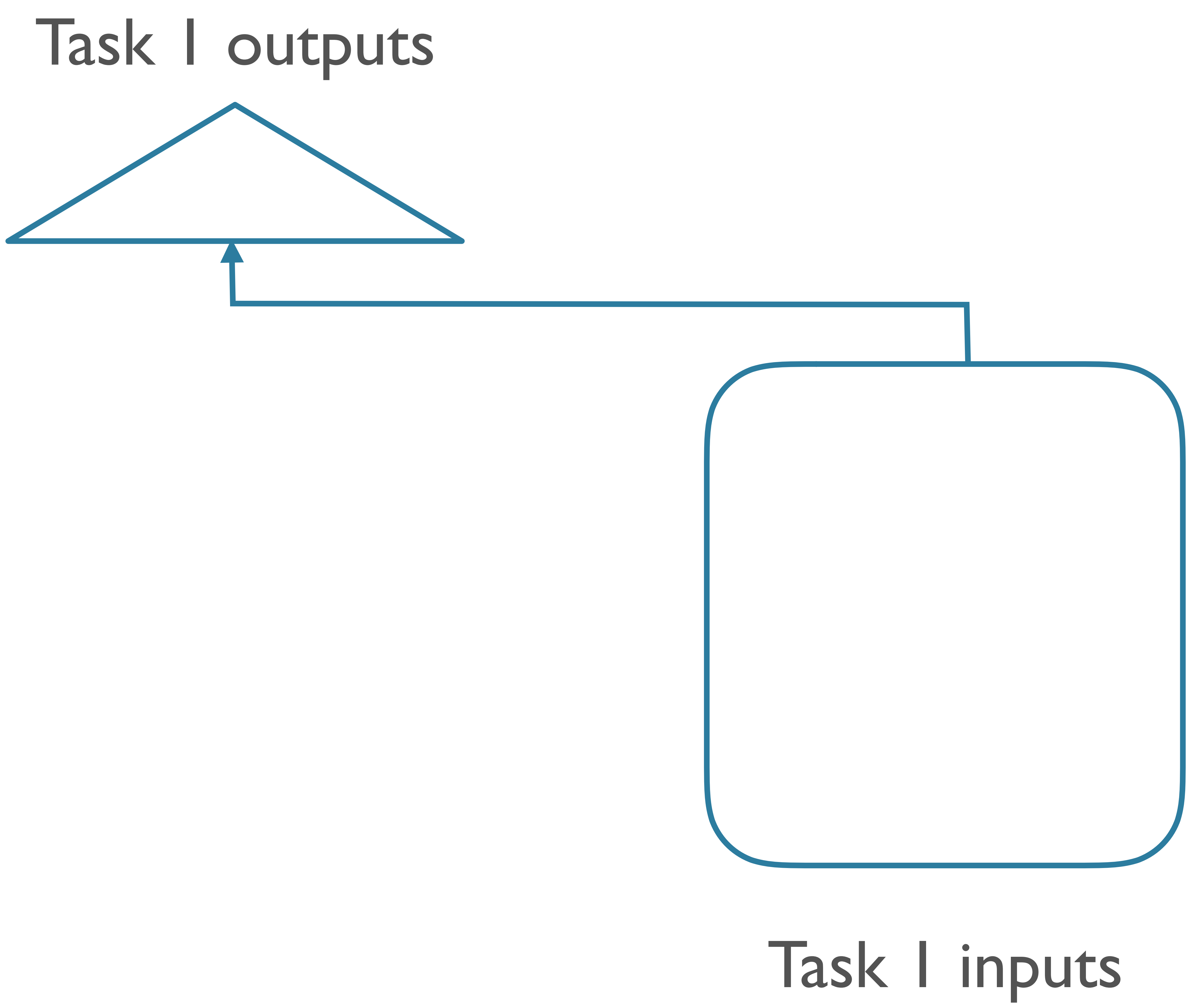


Transfer Learning

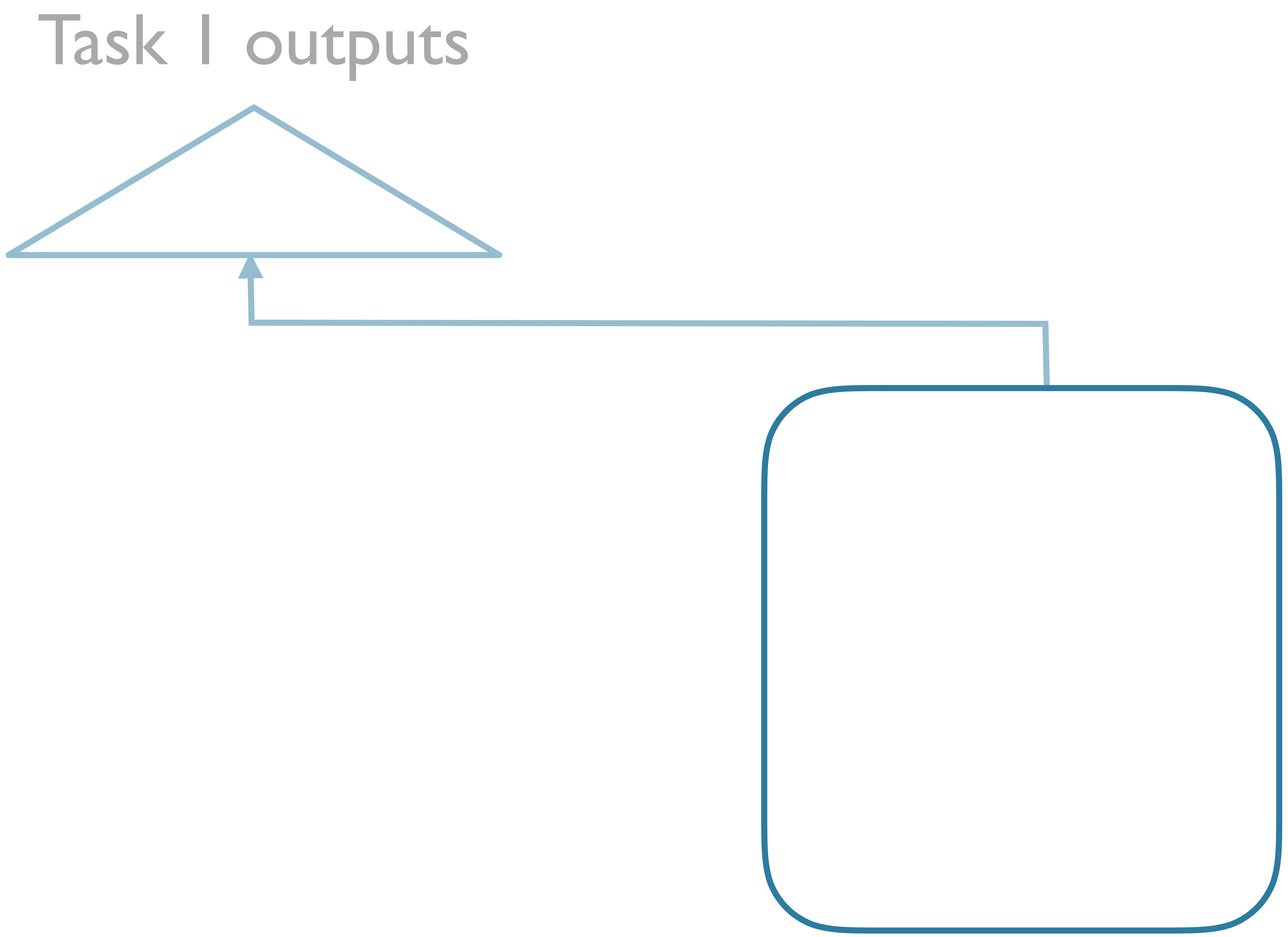


Task 1 inputs

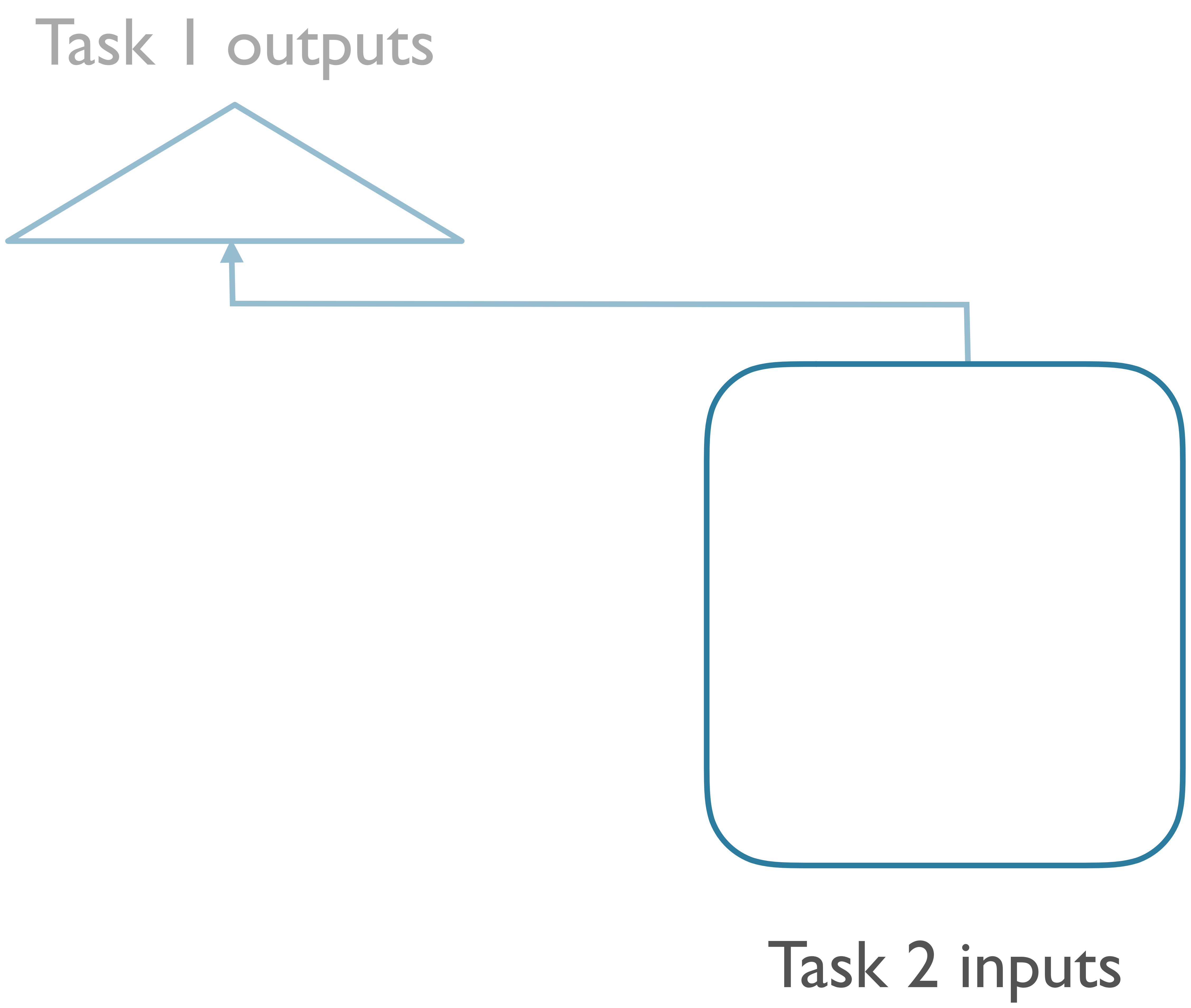
Transfer Learning



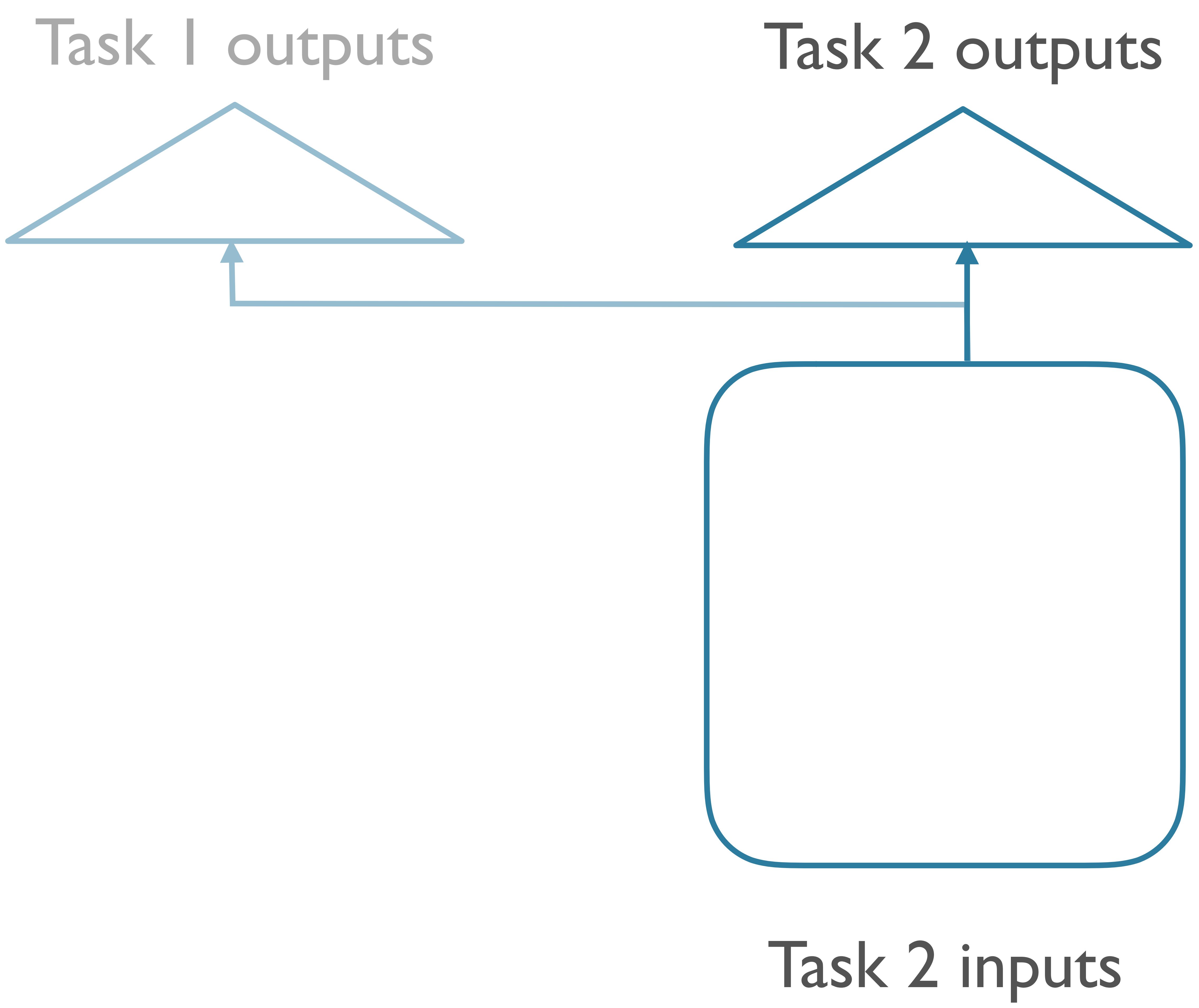
Transfer Learning



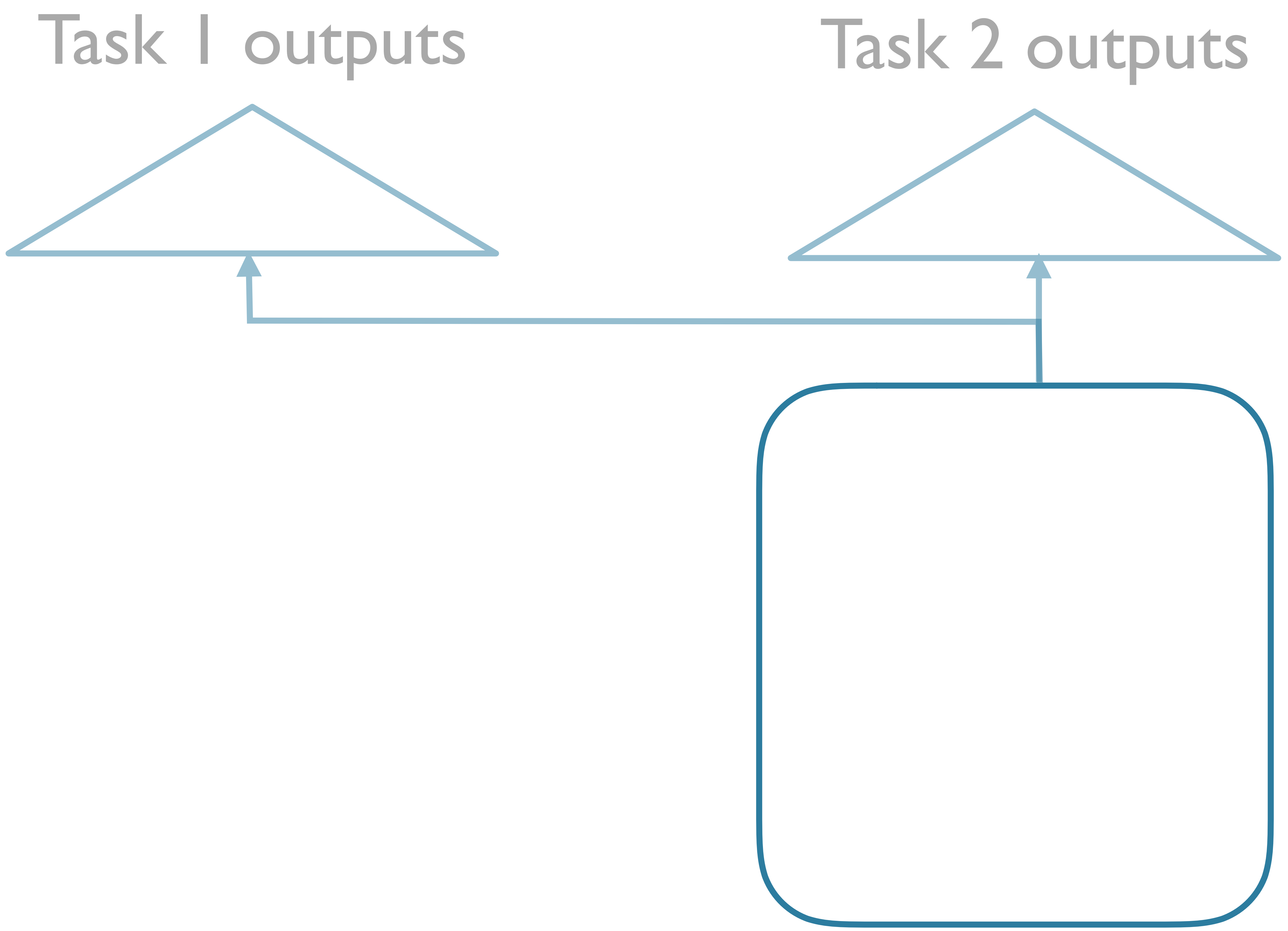
Transfer Learning



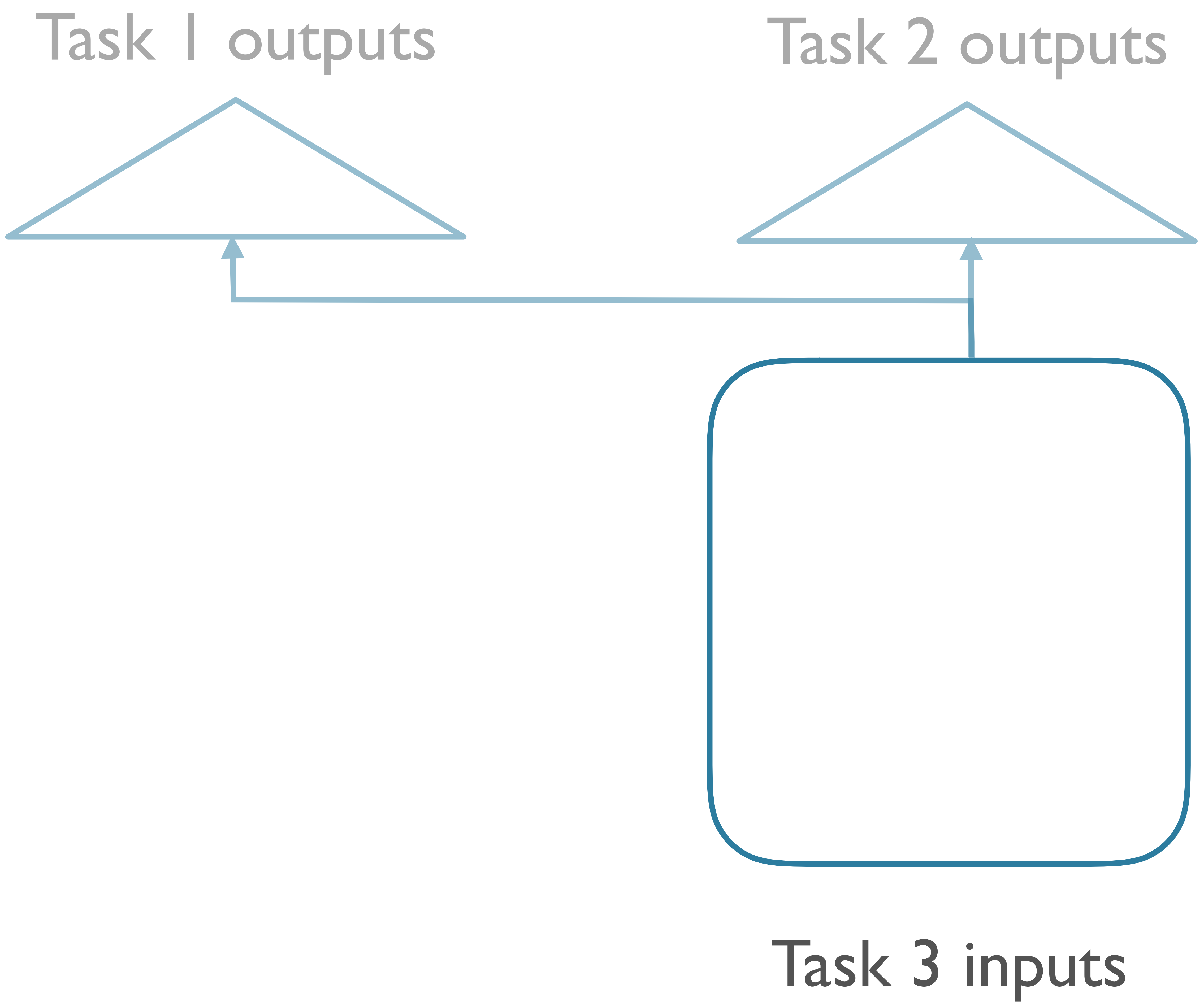
Transfer Learning



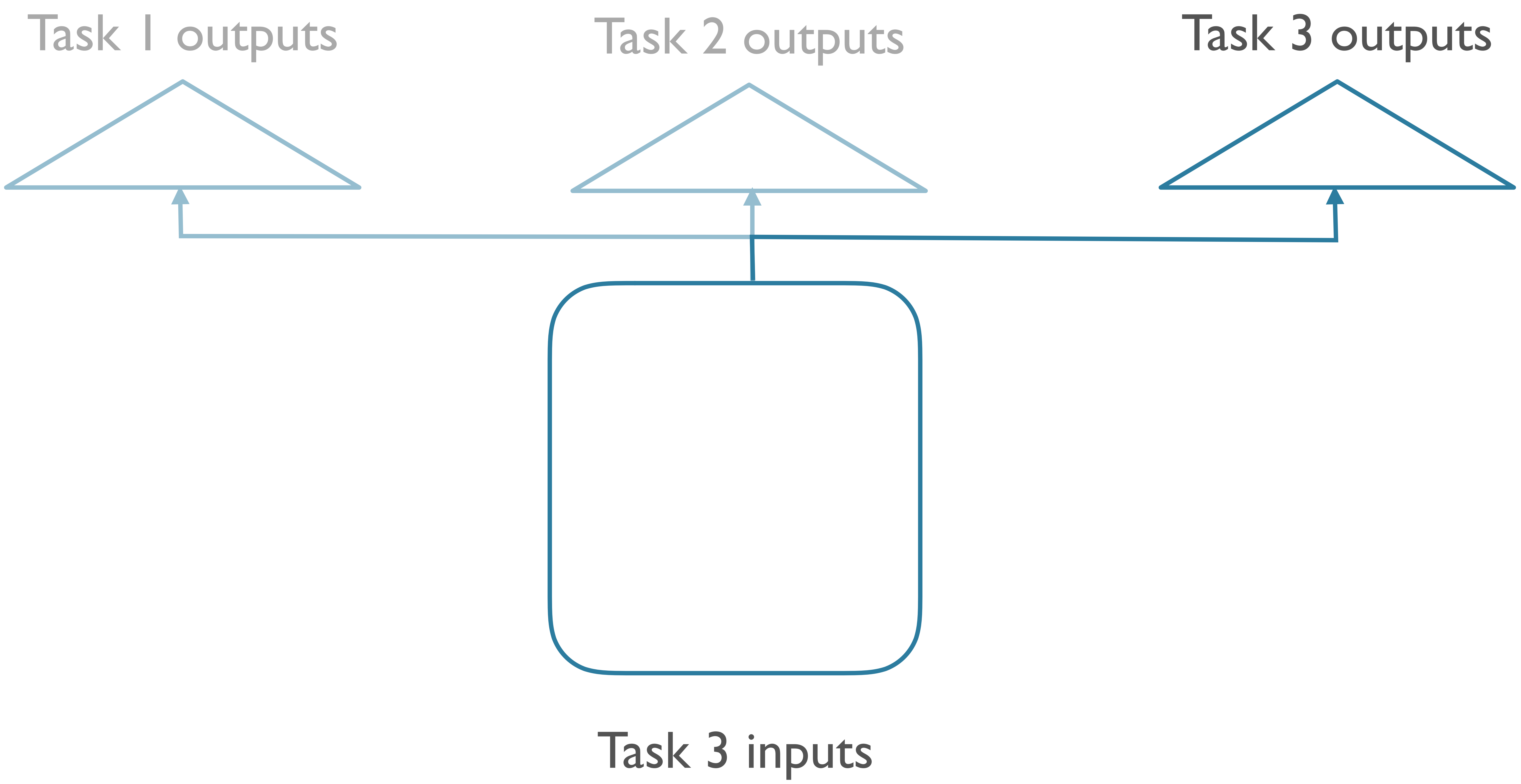
Transfer Learning



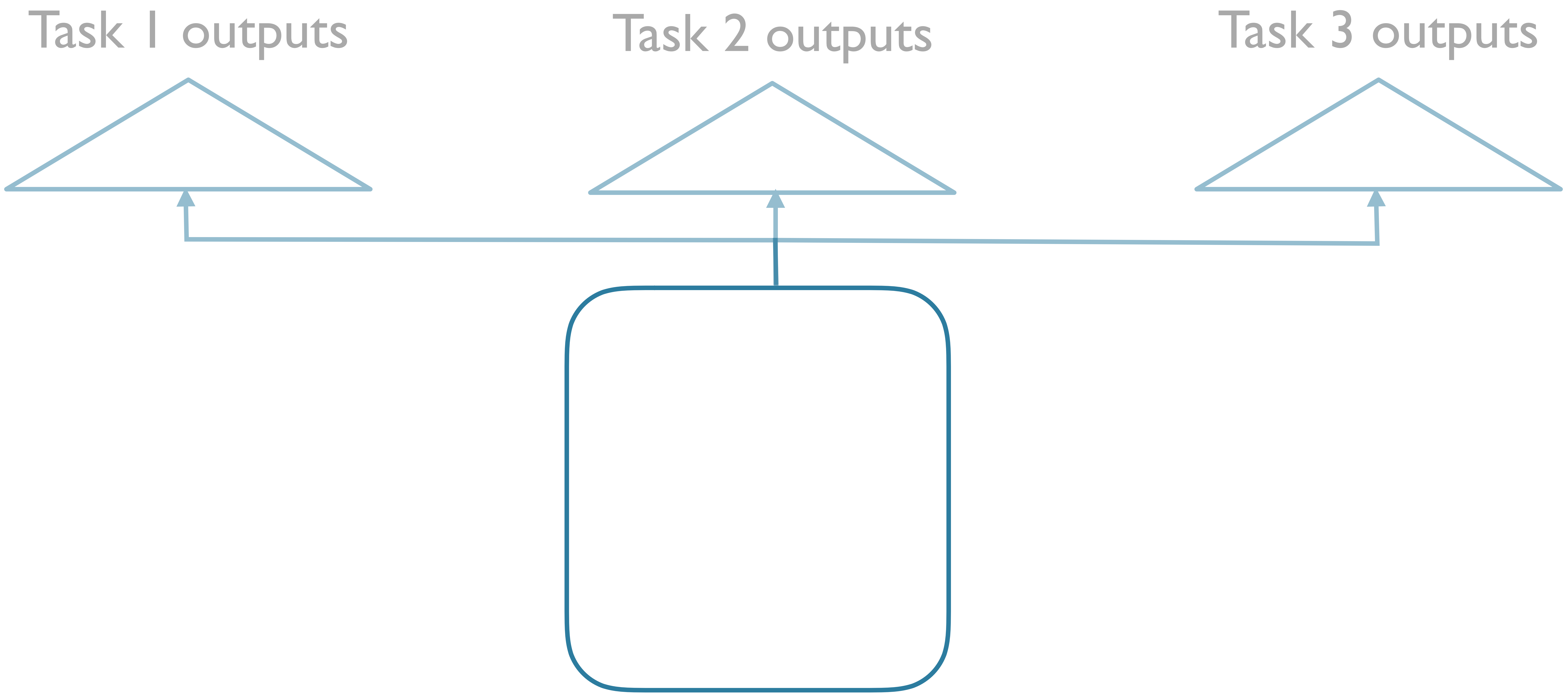
Transfer Learning



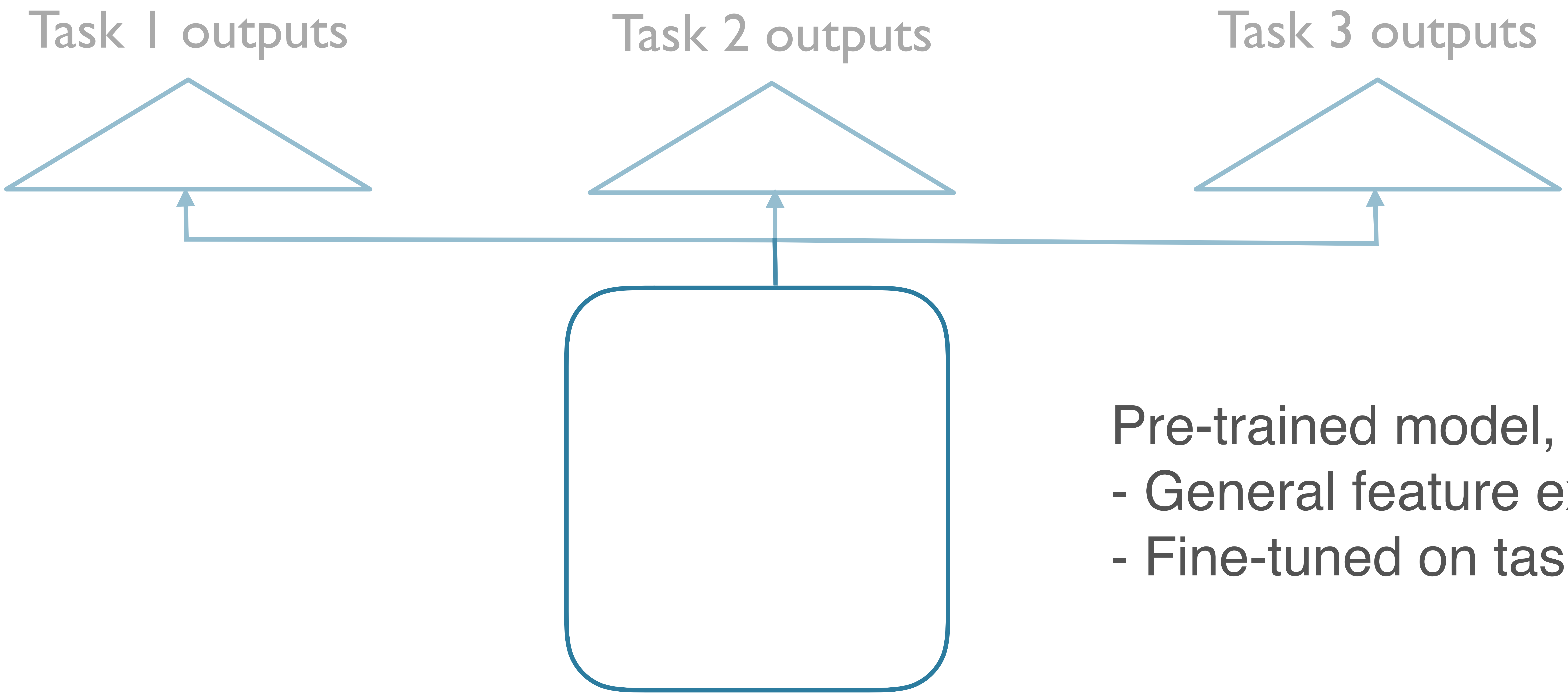
Transfer Learning



Transfer Learning



Transfer Learning

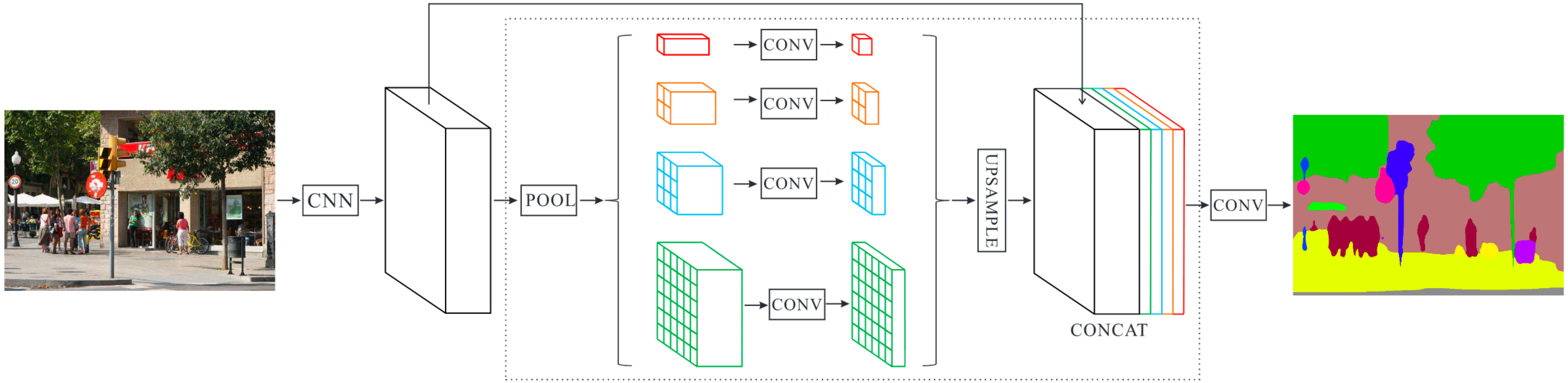


Pre-trained model, either:
- General feature extractor
- Fine-tuned on tasks

Example: Scene Parsing

(a) Image (b) Ground Truth

Example: Scene Parsing

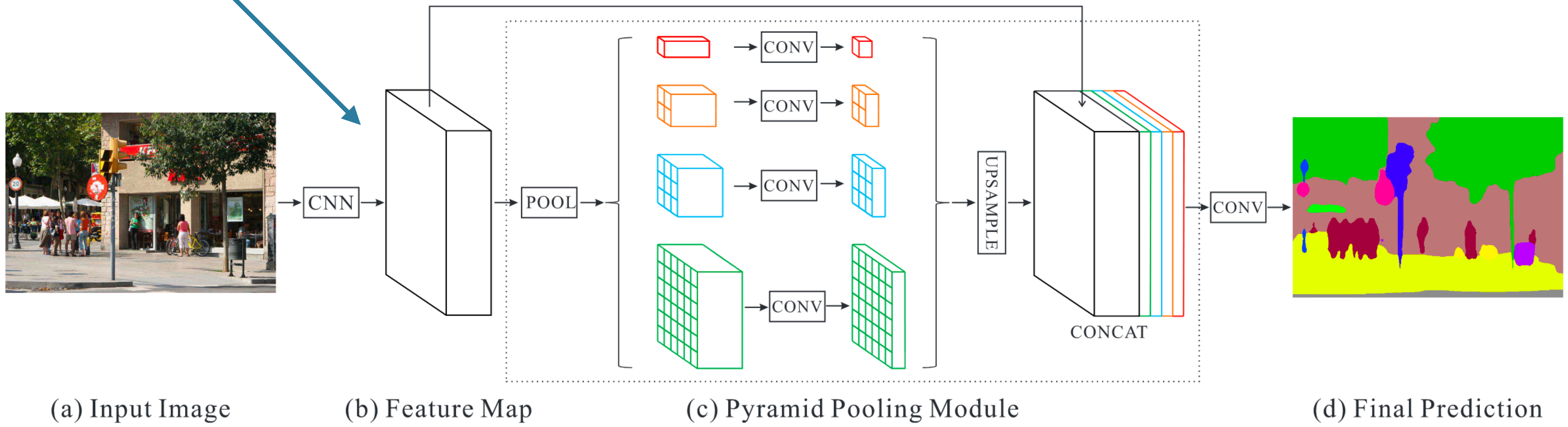


(a) Input Image (b) Feature Map (c) Pyramid Pooling Module (d) Final Prediction

[CVPR '17 paper]

Example: Scene Parsing

Pre-trained ResNet



[CVPR '17 paper]

Transfer Learning in NLP

Where to transfer *from*?

Where to transfer *from*?

- Goal: find a linguistic task that will build general-purpose / *transferable* representations

Where to transfer *from*?

- Goal: find a linguistic task that will build general-purpose / *transferable* representations
- Possibilities:

Where to transfer *from*?

- Goal: find a linguistic task that will build general-purpose / *transferable* representations
- Possibilities:
  - Constituency or dependency parsing

Where to transfer *from*?

- Goal: find a linguistic task that will build general-purpose / *transferable* representations
- Possibilities:
  - Constituency or dependency parsing
  - Semantic parsing

Where to transfer *from*?

- Goal: find a linguistic task that will build general-purpose / *transferable* representations
- Possibilities:
  - Constituency or dependency parsing
  - Semantic parsing
  - Machine translation

Where to transfer *from*?

- Goal: find a linguistic task that will build general-purpose / *transferable* representations
- Possibilities:
  - Constituency or dependency parsing
  - Semantic parsing
  - Machine translation
  - QA

Where to transfer *from*?

- Goal: find a linguistic task that will build general-purpose / *transferable* representations
- Possibilities:
  - Constituency or dependency parsing
  - Semantic parsing
  - Machine translation
  - QA
  - …

Where to transfer *from*?

- Goal: find a linguistic task that will build general-purpose / *transferable* representations
- Possibilities:
  - Constituency or dependency parsing
  - Semantic parsing
  - Machine translation
  - QA
  - …
- Scalability issue: all require expensive annotation

Language Modeling

Language Modeling

- Recent innovation: use *language modeling* (a.k.a. next word prediction)
  - And variants thereof

Language Modeling

- Recent innovation: use *language modeling* (a.k.a. next word prediction)
  - And variants thereof
- Linguistic knowledge:
  - The students were happy because ____ …
  - The student was happy because ____ …

Language Modeling

- Recent innovation: use *language modeling* (a.k.a. next word prediction)
  - And variants thereof
- Linguistic knowledge:
  - The students were happy because ____ …
  - The student was happy because ____ …
- World knowledge:
  - The POTUS gave a speech after missiles were fired by _____
  - The Seattle Sounders are so-named because Seattle lies on the Puget _____

Language Modeling is “Unsupervised”

- An example of “unsupervised” or “semi-supervised” learning
  - NB: I think that “un-annotated” is a better term. Formally, the learning is supervised. But the labels come directly from the data, not an annotator.
- E.g.: “Today is the first day of 575.”
  - (<s>, Today)
  - (<s> Today, is)
  - (<s> Today is, the)
  - (<s> Today is the, first)
  - …

Data for LM is cheap

Data for LM is cheap

Data for LM is cheap

Text

Text is abundant

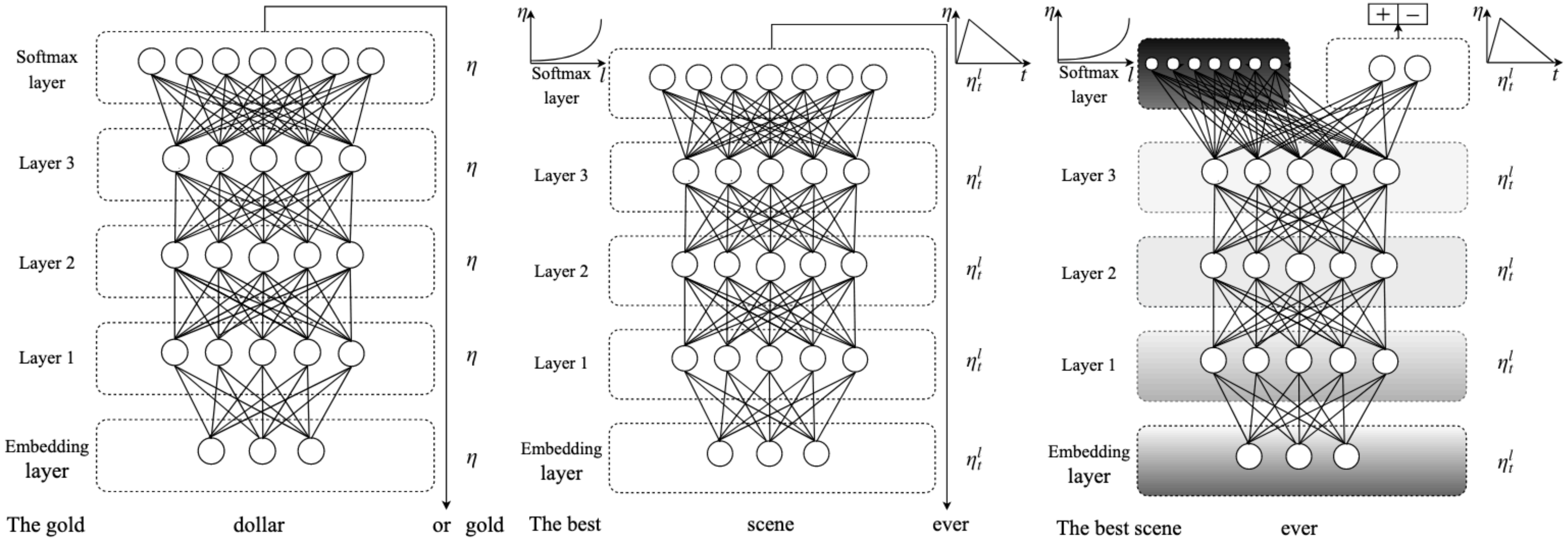
- News sites (e.g. Google 1B)
- Wikipedia (e.g. WikiText103)
- Reddit
- ….
- General web crawling:
  - https://commoncrawl.org/

The Revolution will not be [Annotated]

Yann LeCun

https://twitter.com/rgblong/status/916062474545319938?lang=en

ULMFiT



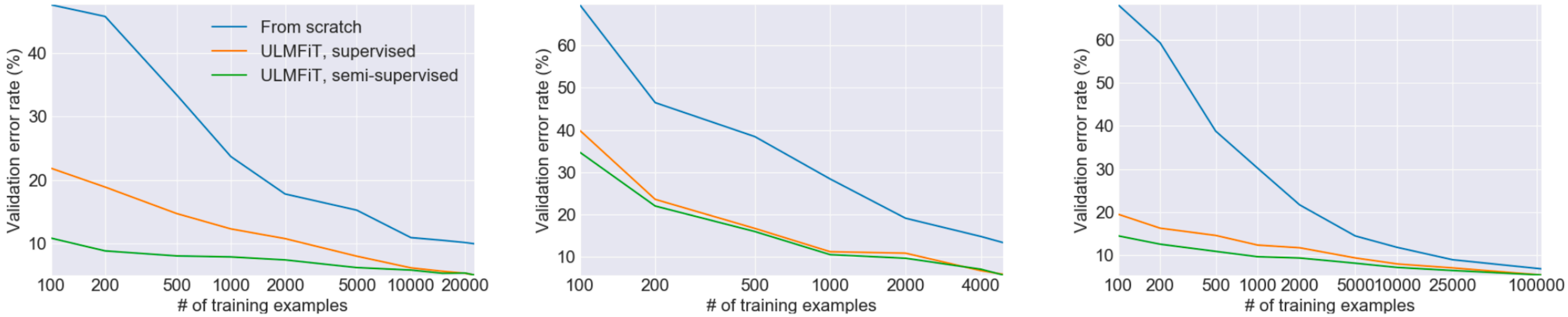
(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

[Universal Language Model Fine-tuning for Text Classification (ACL '18)]

ULMFiT

	Model	Test		Model	Test
IMDb	CoVe (McCann et al., 2017)	8.2	TREC-6	CoVe (McCann et al., 2017)	4.2
	oh-LSTM (Johnson and Zhang, 2016)	5.9		TBCNN (Mou et al., 2015)	4.0
	Virtual (Miyato et al., 2016)	5.9		LSTM-CNN (Zhou et al., 2016)	3.9
	ULMFiT (ours)	**4.6**		ULMFiT (ours)	**3.6**

ULMFiT


From scratch
ULMFiT, supervised
ULMFiT, semi-supervised
Validation error rate (%)
# of training examples
100
200
500
1000
2000
5000
10000
20000
4000
25000
100000
10
20
30
40
50
60

Deep Contextualized Word Representations

Peters et. al (2018)

Deep Contextualized Word Representations

Peters et. al (2018)

- NAACL 2018 Best Paper Award

Deep Contextualized Word Representations

[Peters et. al (2018)]

- NAACL 2018 Best Paper Award
- **E**mbeddings from **L**anguage **Mo**dels (ELMo)
  - [aka the OG NLP Muppet]

Deep Contextualized Word Representations

Peters et. al (2018)

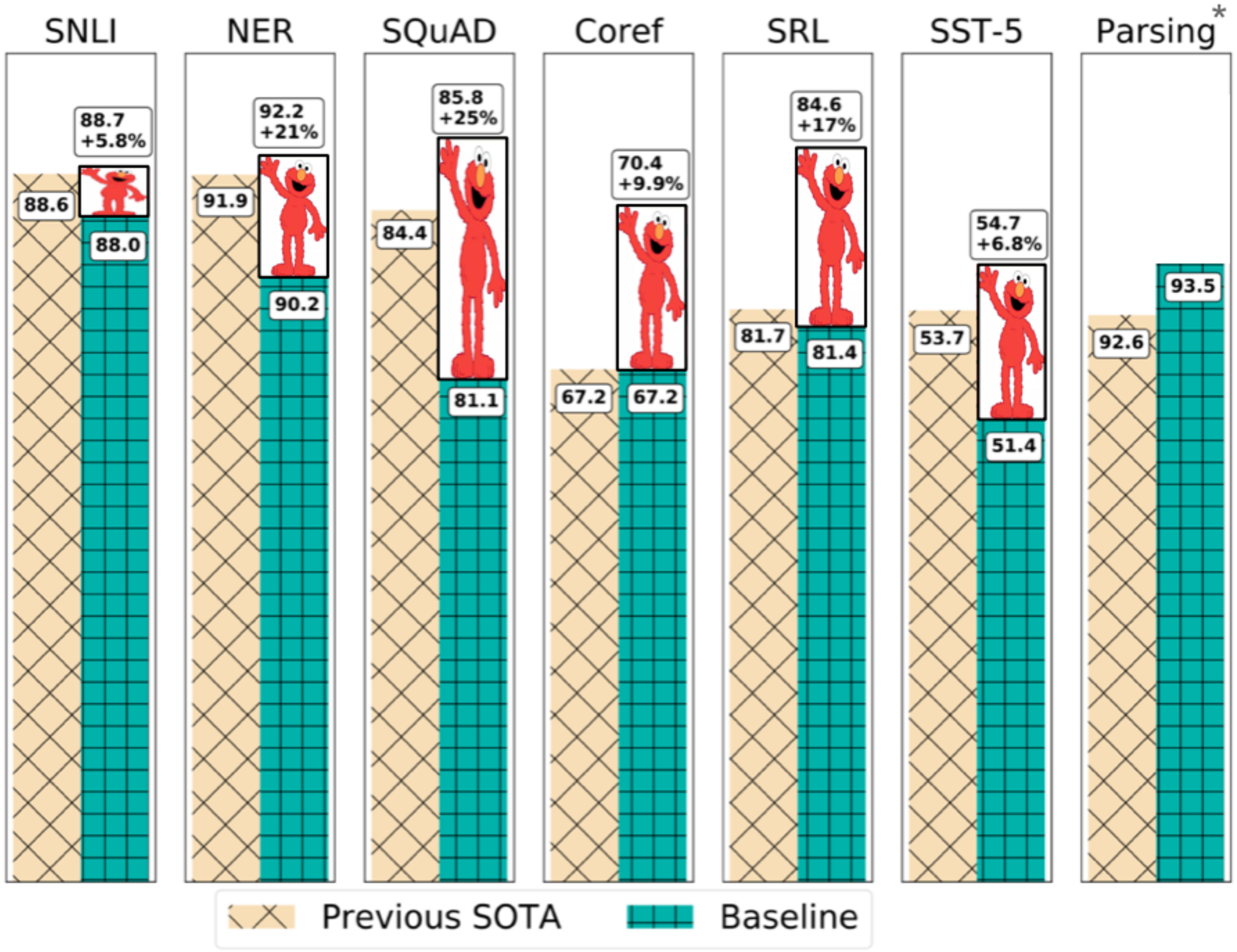
- Comparison to GloVe:

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular **play** on Alusik's grounder…	Kieffer, the only junior in the group, was commended for his ability to hit in the clutch, as well as his all-round excellent **play**.
	Olivia De Havilland signed to do a Broadway **play** for Garson…	…they were actors who had been handed fat roles in a successful **play**, and had talent enough to fill the roles competently, with nice understatement.

Deep Contextualized Word Representations

[Peters et. al (2018)](#)

- Used in place of other embeddings on multiple tasks:



*Kitaev and Klein, ACL 2018 (see also Joshi et al., ACL 2018)

SQuAD = Stanford Question Answering Dataset
SNLI = Stanford Natural Language Inference Corpus
SST-5 = Stanford Sentiment Treebank

figure: Matthew Peters

BERT: Bidirectional Encoder Representations from Transformers

Devlin et al NAACL 2019

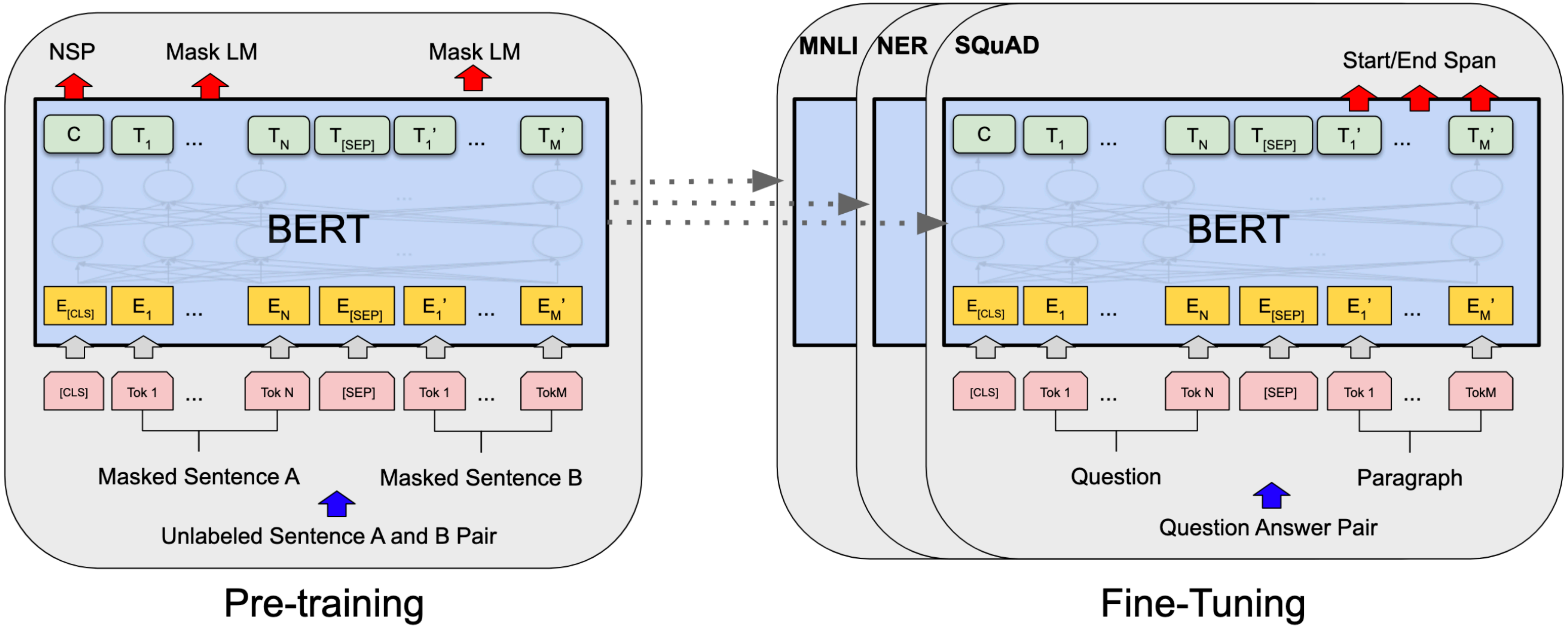
Overview

- Encoder Representations from Transformers: ✓
- Bidirectional: ………?
  - BiLSTM (ELMo): left-to-right and right-to-left
  - Self-attention: every token can see every other
- How do you treat the encoder as an LM (as computing $P(w_t \mid w_{t-1}, w_{t-2}, \ldots, w_1)$)?
  - Don't: modify the task

Masked Language Modeling

- Language modeling: next word prediction
- *Masked* Language Modeling (a.k.a. cloze task): fill-in-the-blank
  - Nancy Pelosi sent the articles of ____ to the Senate.
  - Seattle ____ some snow, so UW was delayed due to ____ roads.
- I.e. $P(w_t \mid w_{t+k}, w_{t+(k-1)}, \ldots, w_{t+1}, w_{t-1}, \ldots, w_{t-(m+1)}, w_{t-m})$
  - (very similar to CBOW: continuous bag of words from word2vec)
- Auxiliary training task: next sentence prediction.
  - Given sentences A and B, binary classification: did B follow A in the corpus or not?

Schematically



Some details

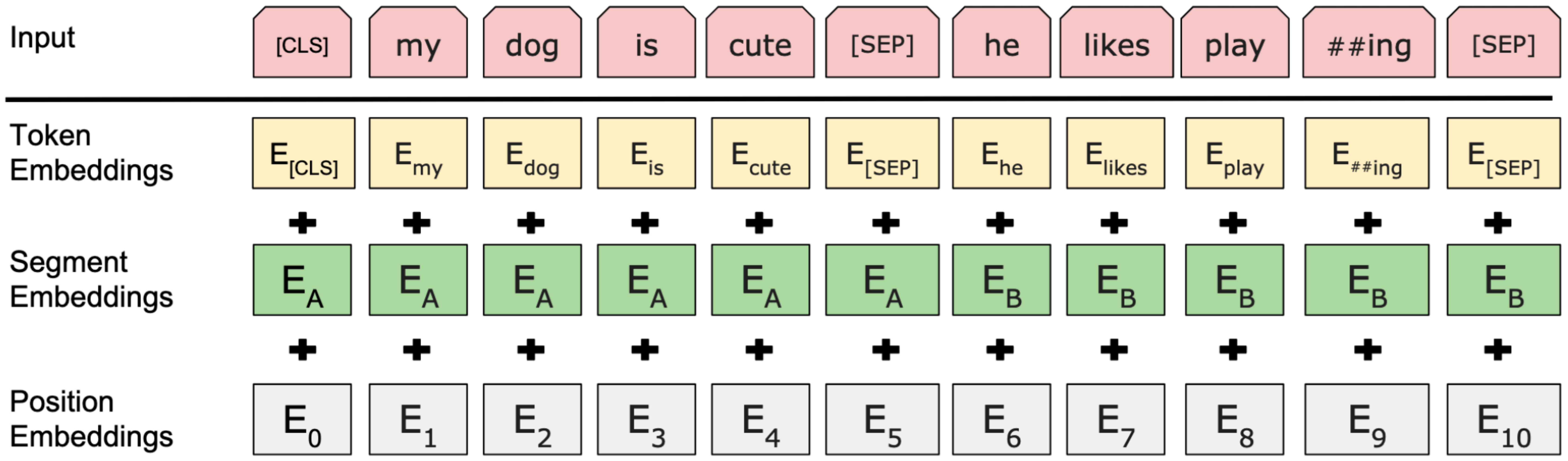
Some details

- BASE model:
  - 12 Transformer Blocks
  - Hidden vector size: 768
  - Attention heads / layer: 12
  - Total parameters: 110M

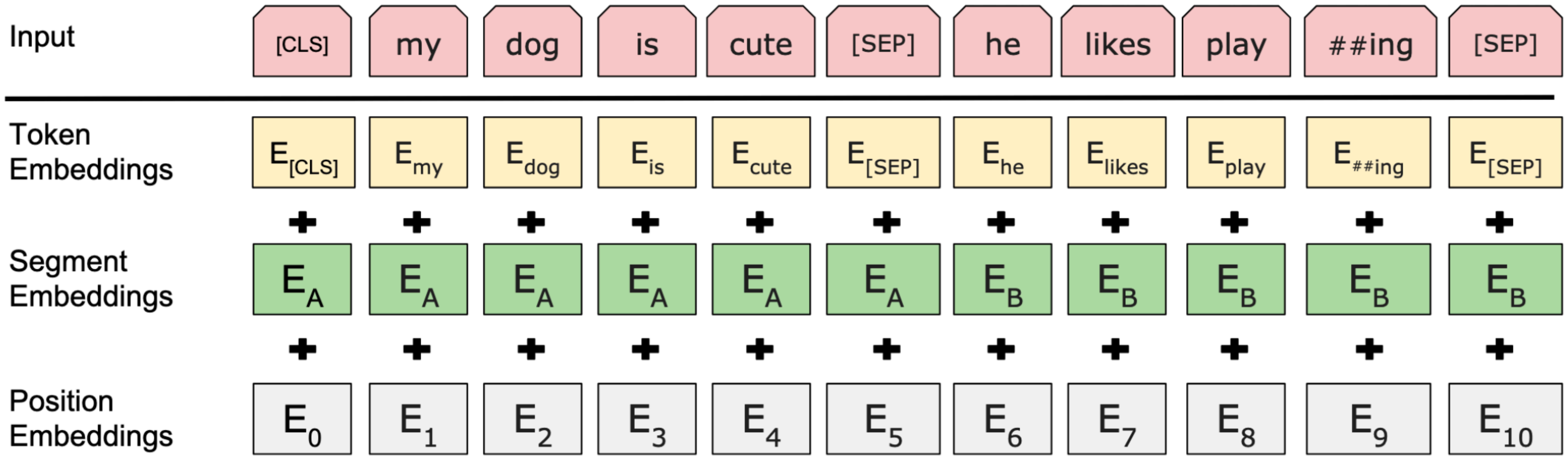
Some details

- BASE model:
  - 12 Transformer Blocks
  - Hidden vector size: 768
  - Attention heads / layer: 12
  - Total parameters: 110M
- LARGE model:
  - 24 Transformer Blocks
  - Hidden vector size: 1024
  - Attention heads / layer: 16
  - Total parameters: 340M

Input Representation

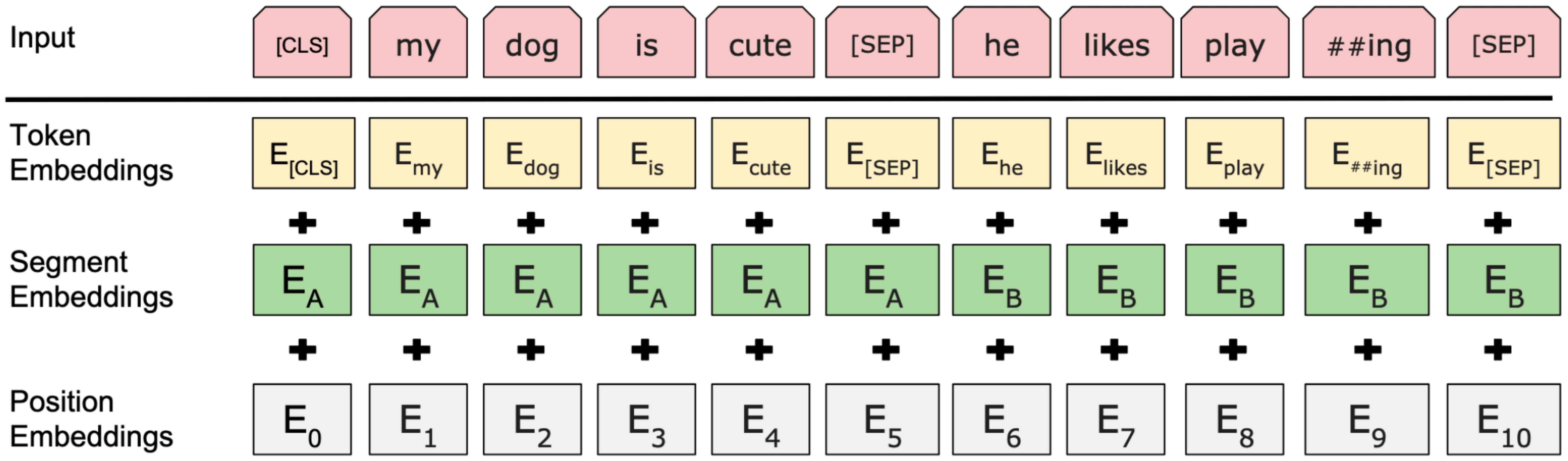


Input Representation



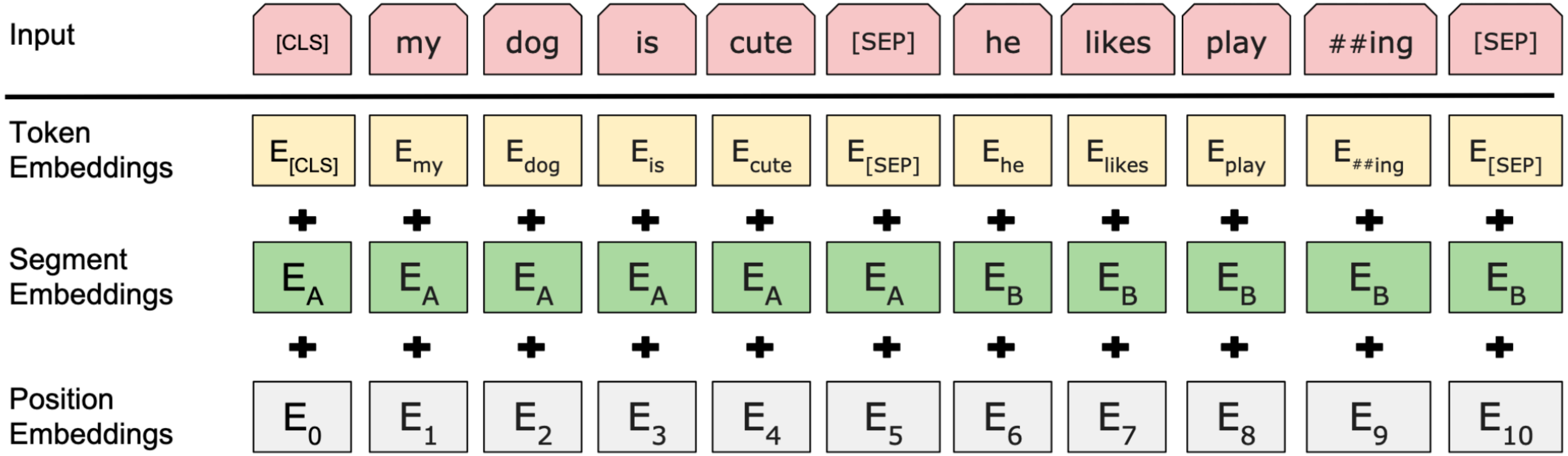
- [CLS], [SEP]: special tokens

Input Representation



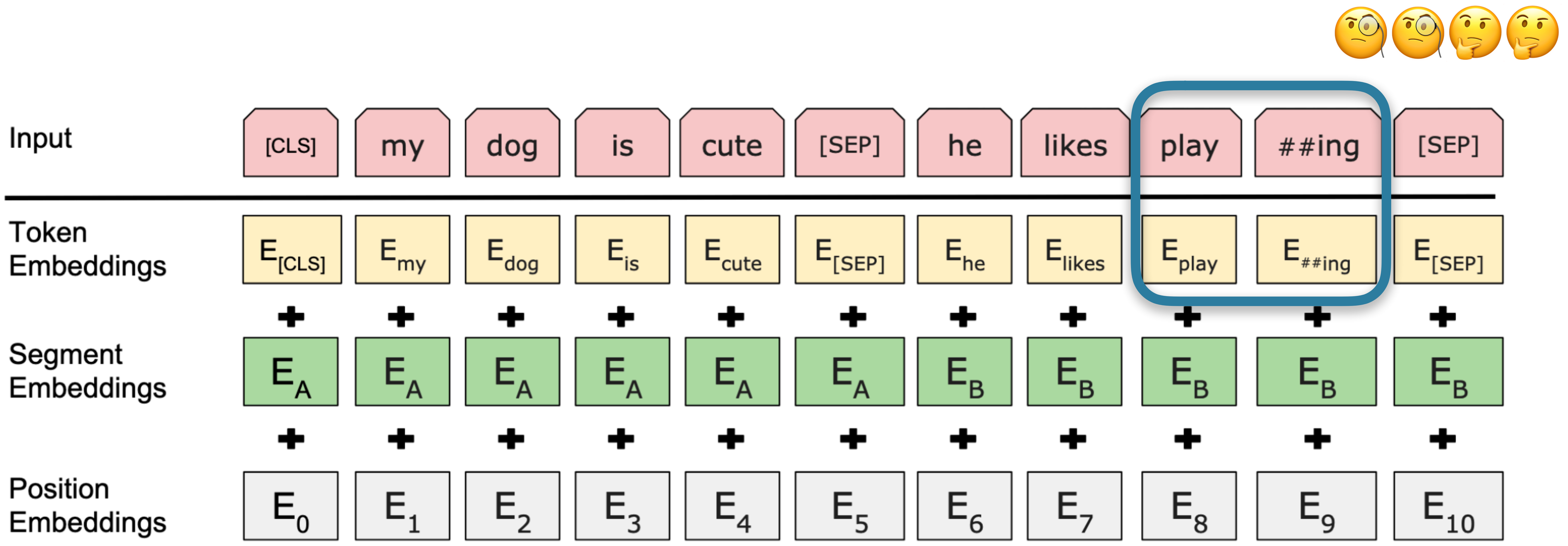
- [CLS], [SEP]: special tokens
- Segment: is this a token from sentence A or B?

Input Representation



- [CLS], [SEP]: special tokens
- Segment: is this a token from sentence A or B?
- Position embeddings: provide position in sequence (learned, not fixed, in this case)

Input Representation



- [CLS], [SEP]: special tokens
- Segment: is this a token from sentence A or B?
- Position embeddings: provide position in sequence (learned, not fixed, in this case)

WordPiece Embeddings

- Another solution to OOV problem, from NMT context (see Wu et al 2016)
- Main idea:
  - Fix vocabulary size |V| in advance [for BERT: 30k]
  - Choose |V| wordpieces (subwords) such that total number of wordpieces in the corpus is minimized
- Frequent words aren’t split, but rarer ones are
- NB: this is a small issue when you transfer to / evaluate on pre-existing tagging datasets with their own vocabularies.

Training Details

- BooksCorpus (800M words) + Wikipedia (2.5B)
- Masking the input text. 15% of all tokens are chosen. Then:
  - 80% of the time: replaced by designated '[MASK]' token
  - 10% of the time: replaced by random token
  - 10% of the time: unchanged
- Loss is cross-entropy of the prediction at the masked positions.
- Max seq length: 512 tokens (final 10%; 128 for first 90%)
- 1M training steps, batch size 256 = 4 days on 4 or 16 TPUs

Initial Results

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Ablations

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

- Not a given (depth doesn't help ELMo); possibly a difference between fine-tuning vs. feature extraction

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

- Many more variations to explore

Major Application

Google

The Keyword   Latest Stories   Product Updates   Company News

SEARCH

Understanding searches better than ever before

Pandu Nayak
Google Fellow and Vice President, Search

Published Oct 25, 2019

If there's one thing I've learned over the 15 years working on Google Search, it's that people's curiosity is endless. We see billions of searches every day, and 15 percent of those queries are ones we haven't seen before--so we've built ways to return results for queries we can't anticipate.

https://www.blog.google/products/search/search-language-understanding-bert/

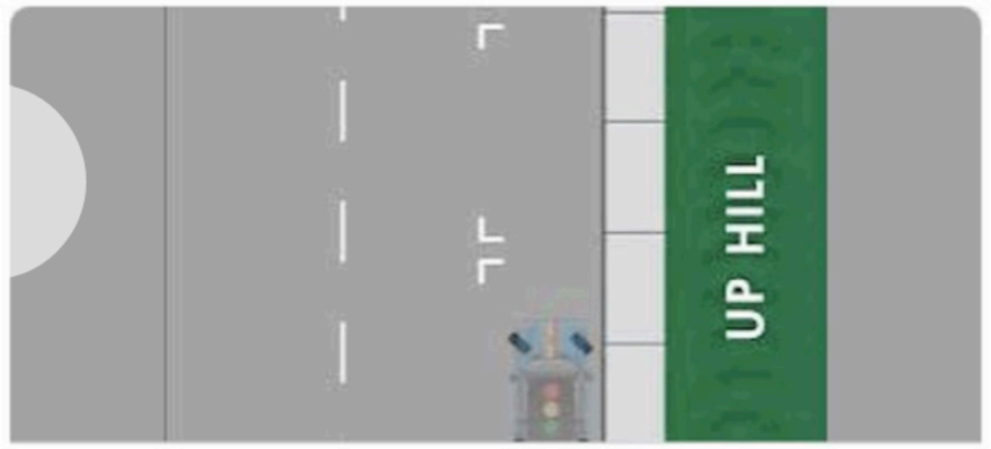
Major Application

parking on a hill with no curb

BEFORE

9:00

google.com



Parking on a Hill. Uphill: When headed uphill at a **curb**, turn the front wheels away from the **curb** and let your vehicle roll backwards slowly until the rear part of the front wheel rests against the **curb** using it as a block. Downhill: When you stop your car headed downhill, turn your front wheels

AFTER

9:00

google.com

For either uphill or downhill **parking**, if there is no **curb**, turn the wheels toward the side of the road so the car will roll away from the center of the road if the brakes fail. When you park on a sloping driveway, turn the wheels so that the car will not roll into the

Pre-trained Neural Models Everywhere

GLUE SuperGLUE Paper </> Code Tasks Leaderboard FAQ Diagnostics Submit

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	ERNIE Team - Baidu	ERNIE		90.2	72.2	97.5	93.0/90.7	92.9/92.5	75.2/90.8	91.2	90.6	98.0	90.9	94.5	49.4
2	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		90.1	73.2	97.1	93.9/91.9	93.0/92.5	74.8/91.0	90.8	90.6	99.2	87.4	94.5	48.7
3	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
4	T5 Team - Google	T5		89.7	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4	92.0	91.7	96.7	92.5	93.2	53.1
5	XLNet Team	XLNet (ensemble)		89.5	70.2	97.1	92.9/90.5	93.0/92.6	74.7/90.4	90.9	90.9	99.0	88.5	92.5	48.4
6	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3	91.0	99.2	89.2	91.8	50.2
7	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.8	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3	91.1	90.7	98.8	88.7	89.0	50.1
8	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	98.9	88.2	89.0	48.7
9	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
10	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
11	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-

General Language Understanding Evaluation (GLUE) / SuperGLUE

Note on the costs of LMs

Note on the costs of LMs

- Currently something of an ‘arms race’ between e.g. Google, Facebook, OpenAI, MS, Baidu

Note on the costs of LMs

- Currently something of an ‘arms race’ between e.g. Google, Facebook, OpenAI, MS, Baidu
- Hugely expensive
  - Carbon emissions
  - Monetarily
    - Inequitable access

Note on the costs of LMs

- Currently something of an ‘arms race’ between e.g. Google, Facebook, OpenAI, MS, Baidu
- Hugely expensive
  - Carbon emissions
  - Monetarily
    - Inequitable access

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Abstract

Recent progress in hardware and methodology for training neural networks has ushered in a new generation of large networks trained on abundant data. These models have obtained notable gains in accuracy across many NLP tasks. However, these accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result these models are costly to train and develop, both financially, due to the cost of hardware and electricity or cloud compute time, and environmentally, due to the carbon footprint required to fuel modern tensor

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Note on the costs of LMs

- Currently something of an ‘arms race’ between e.g. Google, Facebook, OpenAI, MS, Baidu
- Hugely expensive
  - Carbon emissions
  - Monetarily
    - Inequitable access

Note on the costs of LMs

- Currently something of an ‘arms race’ between e.g. Google, Facebook, OpenAI, MS, Baidu
- Hugely expensive
  - Carbon emissions
  - Monetarily
    - Inequitable access

Green AI

Roy Schwartz[*♢] Jesse Dodge[*♢♣] Noah A. Smith[♢♡] Oren Etzioni[♢]

[♢]Allen Institute for AI, Seattle, Washington, USA
[♣] Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[♡] University of Washington, Seattle, Washington, USA

July 2019

Abstract

The computations required for deep learning research have been doubling every few months, resulting in an estimated 300,000x increase from 2012 to 2018 [2]. These computations have a surprisingly large carbon footprint [40]. Ironically, deep learning was inspired by the human brain, which is remarkably energy efficient. Moreover, the financial cost of the computations can make it difficult for academics, students, and researchers, in particular those from emerging economies, to engage in deep learning research.

This position paper advocates a practical solution by making **efficiency** an evaluation criterion for research alongside accuracy and related measures. In addition, we propose reporting the financial cost or “price tag” of developing, training, and running models to provide baselines for the investigation of increasingly efficient methods. Our goal is to make AI both greener and more inclusive—enabling any inspired undergraduate with a laptop to write high-quality research papers. Green AI is an emerging focus at the Allen Institute for AI.

Note on the costs of LMs

- Currently something of an ‘arms race’ between e.g. Google, Facebook, OpenAI, MS, Baidu
- Hugely expensive
  - Carbon emissions
  - Monetarily
    - Inequitable access
- A role for interpretability/analysis:
  - Bigger is better, but:
  - Which factors really matter

Green AI

Roy Schwartz[*♢] Jesse Dodge[*♢♣] Noah A. Smith[♢♡] Oren Etzioni[♢]

[♢]Allen Institute for AI, Seattle, Washington, USA
[♣] Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[♡] University of Washington, Seattle, Washington, USA

July 2019

Abstract

The computations required for deep learning research have been doubling every few months, resulting in an estimated 300,000x increase from 2012 to 2018 [2]. These computations have a surprisingly large carbon footprint [40]. Ironically, deep learning was inspired by the human brain, which is remarkably energy efficient. Moreover, the financial cost of the computations can make it difficult for academics, students, and researchers, in particular those from emerging economies, to engage in deep learning research.

This position paper advocates a practical solution by making **efficiency** an evaluation criterion for research alongside accuracy and related measures. In addition, we propose reporting the financial cost or “price tag” of developing, training, and running models to provide baselines for the investigation of increasingly efficient methods. Our goal is to make AI both greener and more inclusive—enabling any inspired undergraduate with a laptop to write high-quality research papers. Green AI is an emerging focus at the Allen Institute for AI.

Sidebar: Word Embeddings

Sidebar: Word Embeddings

- Aren't word embeddings like word2vec and GloVe examples of transfer learning?
  - Yes: get linguistic representations from raw text to use in downstream tasks
  - No: not to be used as *general-purpose* representations

Sidebar: Word Embeddings

Sidebar: Word Embeddings

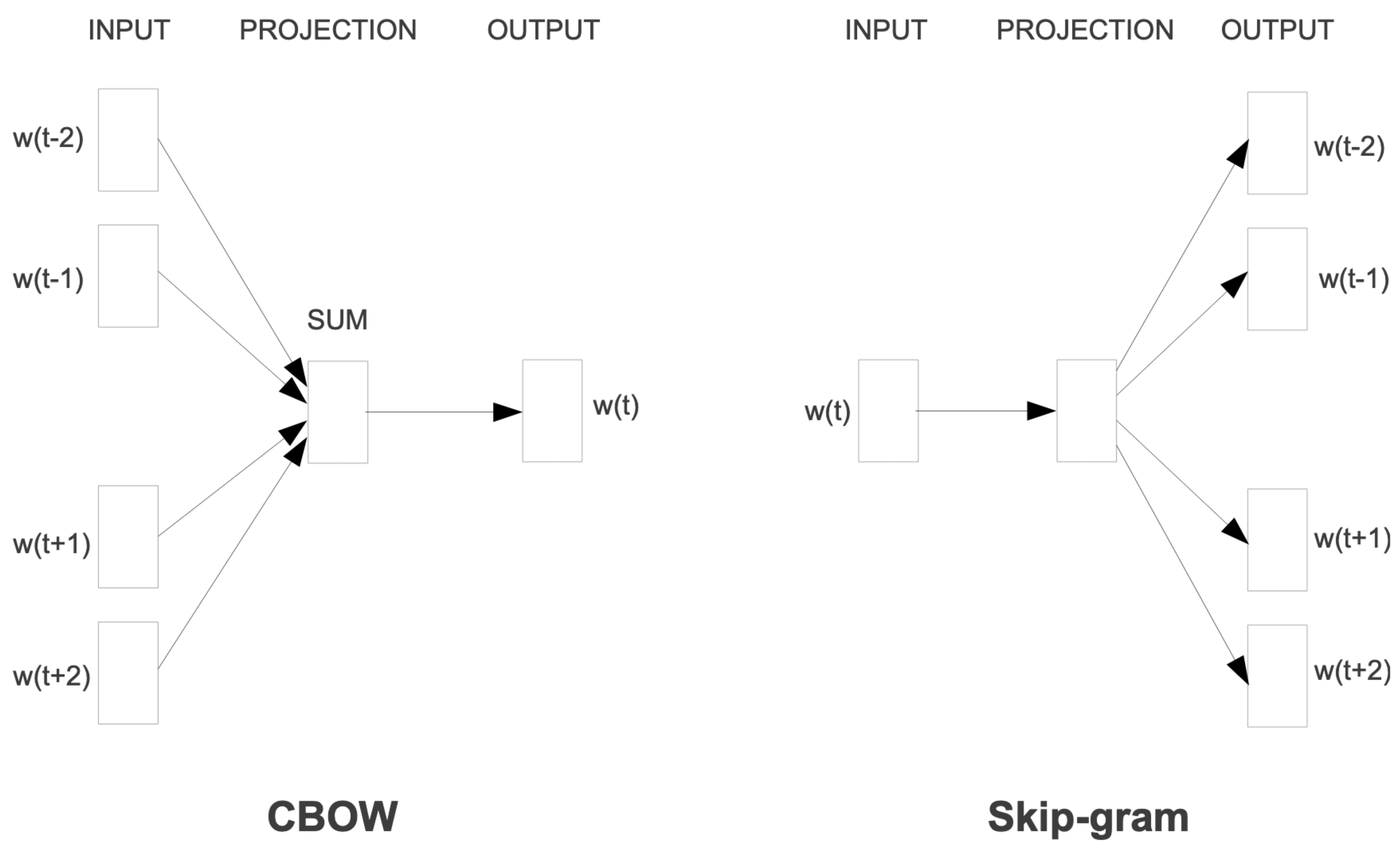
- One distinction:
  - *Global* representations:
    - word2vec, GloVe: *one* vector for each word *type* (e.g. ‘play’)
  - *Contextual* representations (from LMs):
    - Representation of word in context, not independently

Sidebar: Word Embeddings

- One distinction:
  - *Global* representations:
    - word2vec, GloVe: *one* vector for each word *type* (e.g. ‘play’)
  - *Contextual* representations (from LMs):
    - Representation of word in context, not independently
- Another:
  - *Shallow* (global) vs. *Deep* (contextual) pre-training

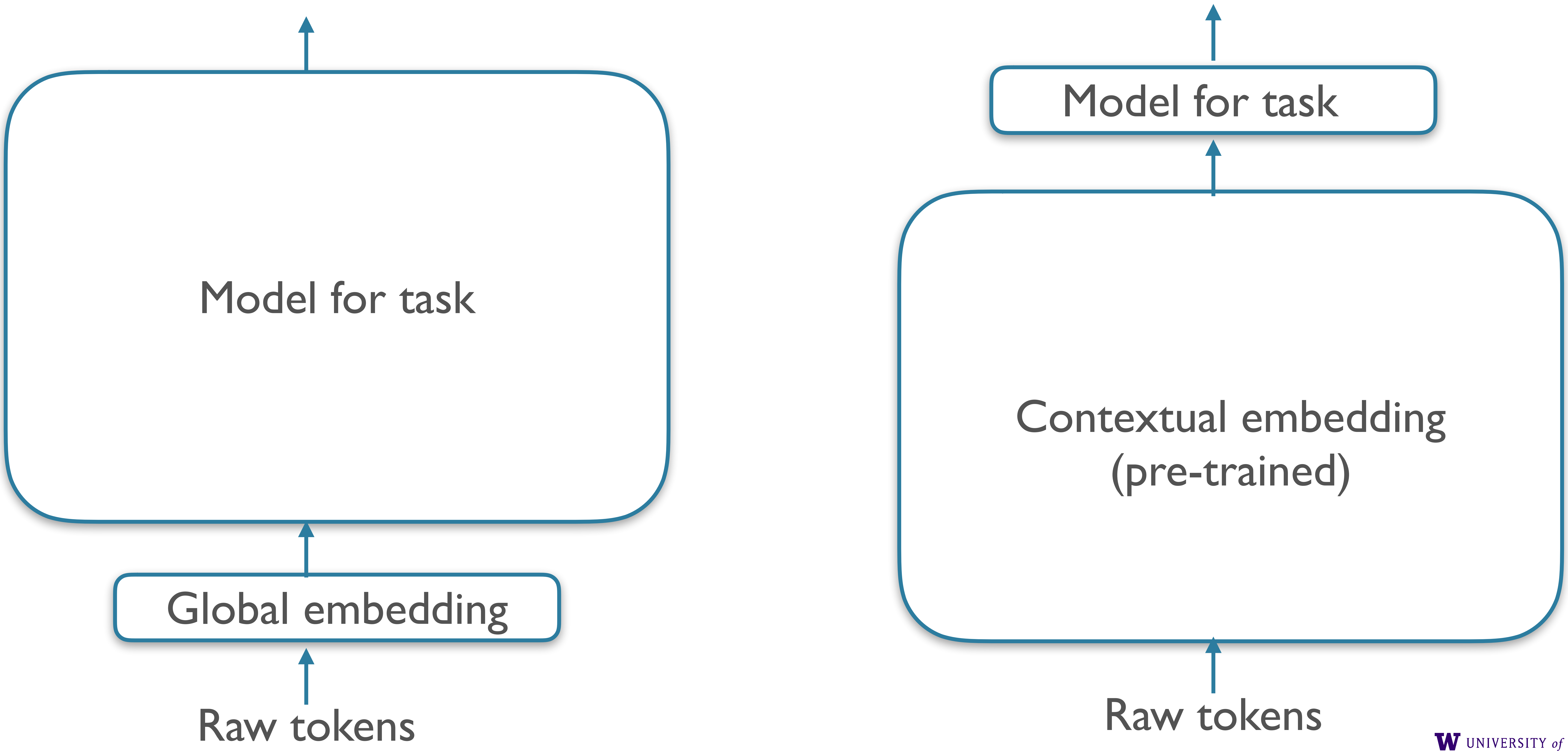
Global Embeddings: Models

Global Embeddings: Models



[Mikolov et al 2013a]() (the OG `word2vec` paper)

Shallow vs Deep Pre-training



State of the Field

- Manning 2017: “The BiLSTM Hegemony”
- Right now: “The pre-trained Transformer Hegemony”
  - By default: fine-tune a large pre-trained Transformer on the task you care about
  - Will often yield the best results
  - Beware: often not significantly better than *very simple* baselines (SVM, etc)
- Very useful library to quickly use these models: HuggingFace Transformers
  - https://huggingface.co/transformers/
- Variants of BERT differ on: hyper-parameters, architectural choices, pre-training tasks, ….