

Analyzing Neural Language Models

Introduction

Shane Steinert-Threlkeld

Jan 9, 2020

Today's Plan

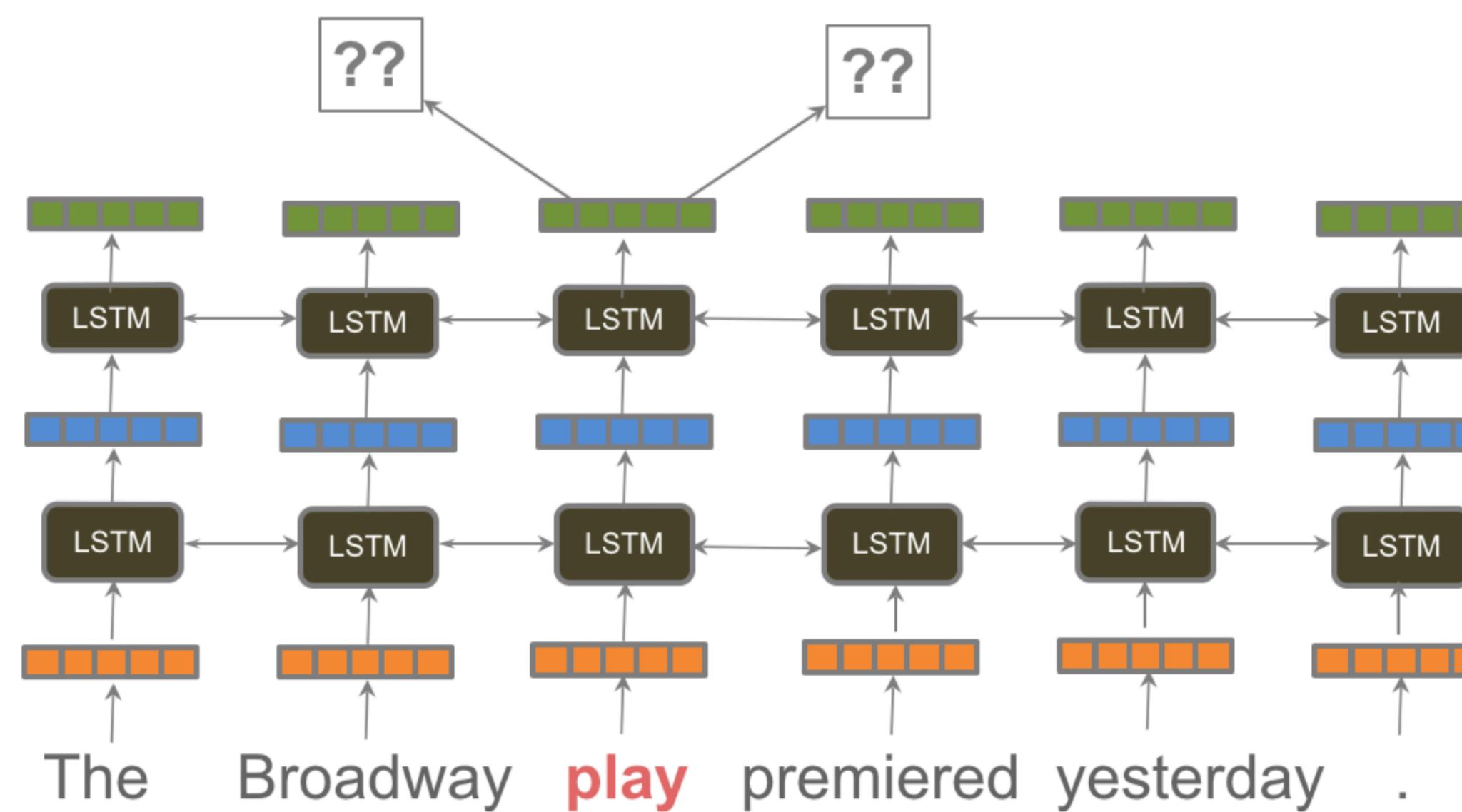
- Motivation / background
 - NLP’s “ImageNet moment”
 - NLP’s “Clever Hans moment”
- 15 minute break
- Course information / logistics

Motivation

NLP's “ImageNet Moment”

∇ The Gradient

HOME EDITOR'S NOTE OVERVIEWS PERSPECTIVES ABOUT SUBSCRIBE Q



NLP's ImageNet
moment has arrived

08.JUL.2018

[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

What is ImageNet?

- Large dataset, v1 in 2009
- Object classification (among others):
 - Input: image
 - Label: synsets from WordNet
- ~14M images currently
- <http://www.image-net.org>

What is ImageNet?

Geological formation, formation
(geology) the geological features of the earth

1808 pictures 86.24% Popularity Percentile Wordnet IDs

Numbers in brackets: (the number of synsets in the subtree).

ImageNet 2011 Fall Release (32326)

- ↳ plant, flora, plant life (4486)
- ↳ geological formation, formation (17)
 - ↳ aquifer (0)
 - ↳ beach (1)
 - ↳ cave (3)
 - ↳ cliff, drop, drop-off (2)
 - ↳ delta (0)
 - ↳ diapir (0)
 - ↳ folium (0)
 - ↳ foreshore (0)
 - ↳ ice mass (10)
 - ↳ lakefront (0)
 - ↳ massif (0)
 - ↳ monocline (0)
 - ↳ mouth (0)
 - ↳ natural depression, depression (0)
 - ↳ natural elevation, elevation (41)
 - ↳ oceanfront (0)
 - ↳ range, mountain range, range of (0)
 - ↳ relict (0)
 - ↳ ridge, ridgeline (2)
 - ↳ ridge (0)
 - ↳ shore (7)
 - ↳ slope, incline, side (17)
 - ↳ spring, fountain, outflow, outpouring (0)
 - ↳ talus, scree (0)
 - ↳ vein, mineral vein (1)
 - ↳ volcanic crater, crater (2)
 - ↳ wall (0)

Treemap Visualization Images of the Synset Downloads

ImageNet 2011 Fall Release > Geological formation, formation

The treemap visualization shows the hierarchical structure of geological formations. The root node 'geological formation, formation' is expanded to show its sub-categories: Natural, Slope, Shore, Ice, Water, Vein, Delta, Foreshore, Massif, Talus, Volcanic, Beach, Mouth, Lakefront, Range, Diapir, Cliff, Wall, Monocline, Oceanfront, Aquifer, Cave, Spring, and Ridge. Each category is represented by a grid of small images illustrating various geological features.

Why is ImageNet Important?

ATURED

QUARTZ

EMAIL

IT'S NOT ABOUT THE ALGORITHM

The data that transformed AI research—and possibly the world

[link](#)

By [Dave Gershgorin](#) • July 26, 2017

Why is ImageNet Important?

ATURED

QUARTZ

EMAIL

IT'S NOT ABOUT THE ALGORITHM

The data that transformed AI research—and possibly the world

[link](#)

By [Dave Gershgorin](#) • July 26, 2017

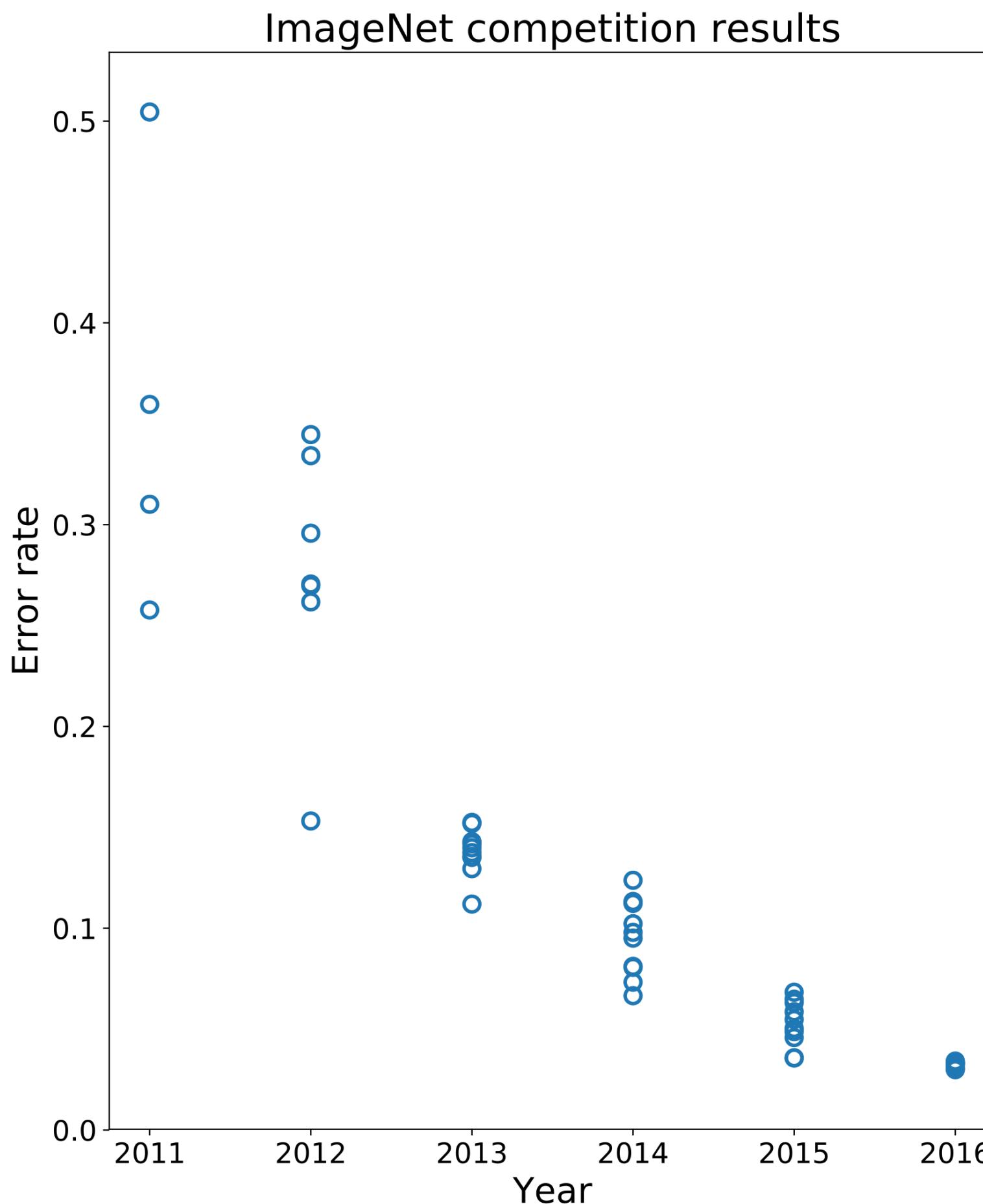
1. Deep learning

2. Transfer learning

ILSVRC

- ImageNet Large Scale Visual Recognition Challenge
- Annual competition on standard benchmark
 - 2010-2017
- ~1.2M training images, 1000 categories
- <http://www.image-net.org/challenges/LSVRC/>

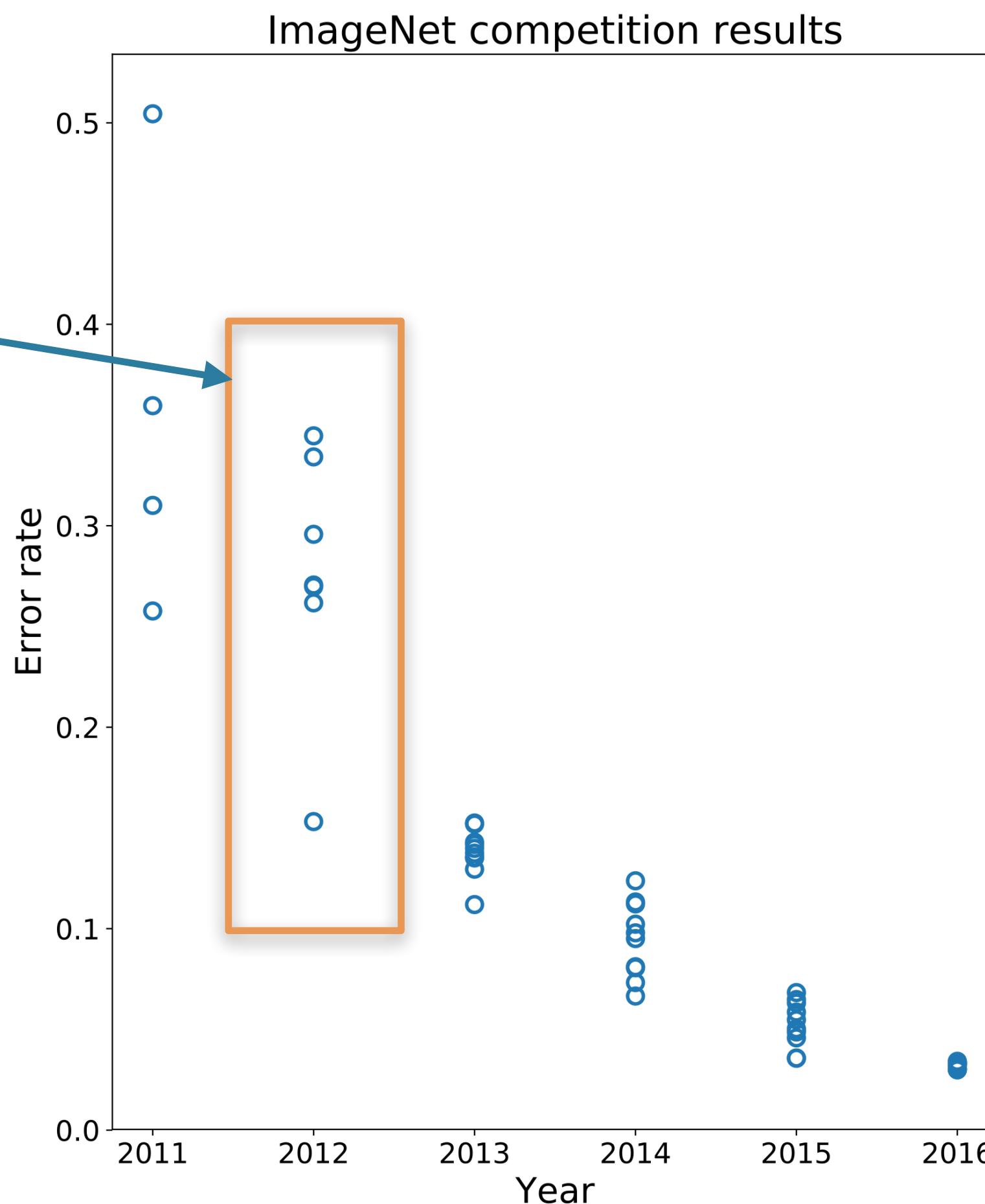
ILSVRC results



[source](#)

ILSVRC results

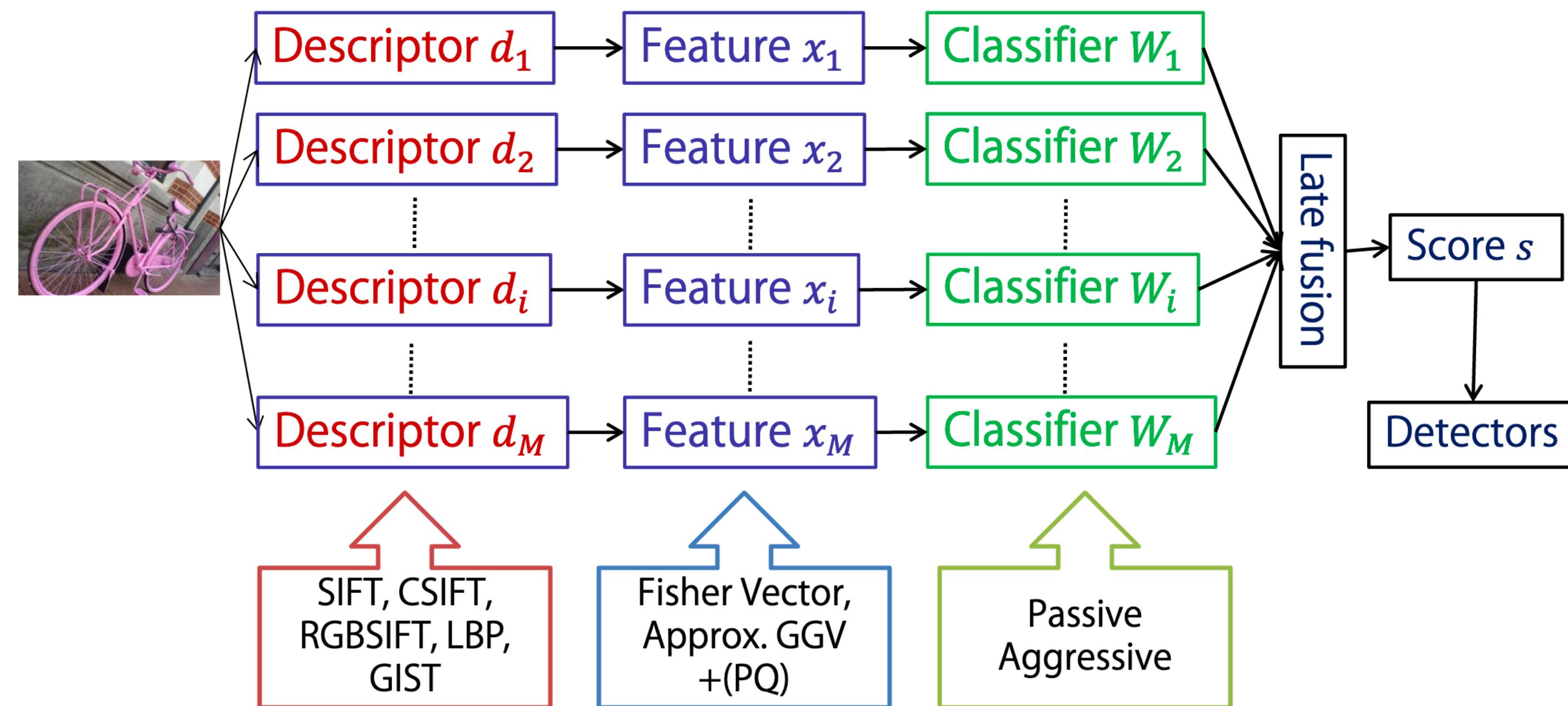
What happened in 2012?



[source](#)

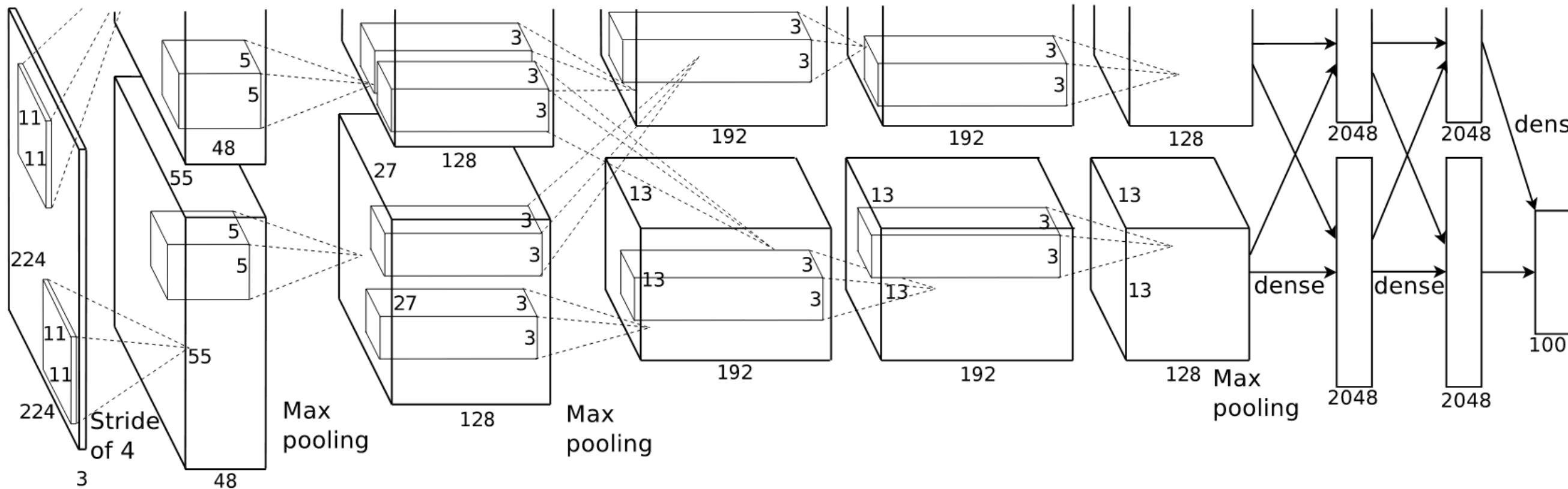
ILSVRC 2012: runner-up

Fisher based features + Multi class linear classifiers



[source](#)

ILSVRC 2012: winner



ImageNet Classification with Deep Convolutional Neural Networks

[NeurIPS 2012 paper](#)

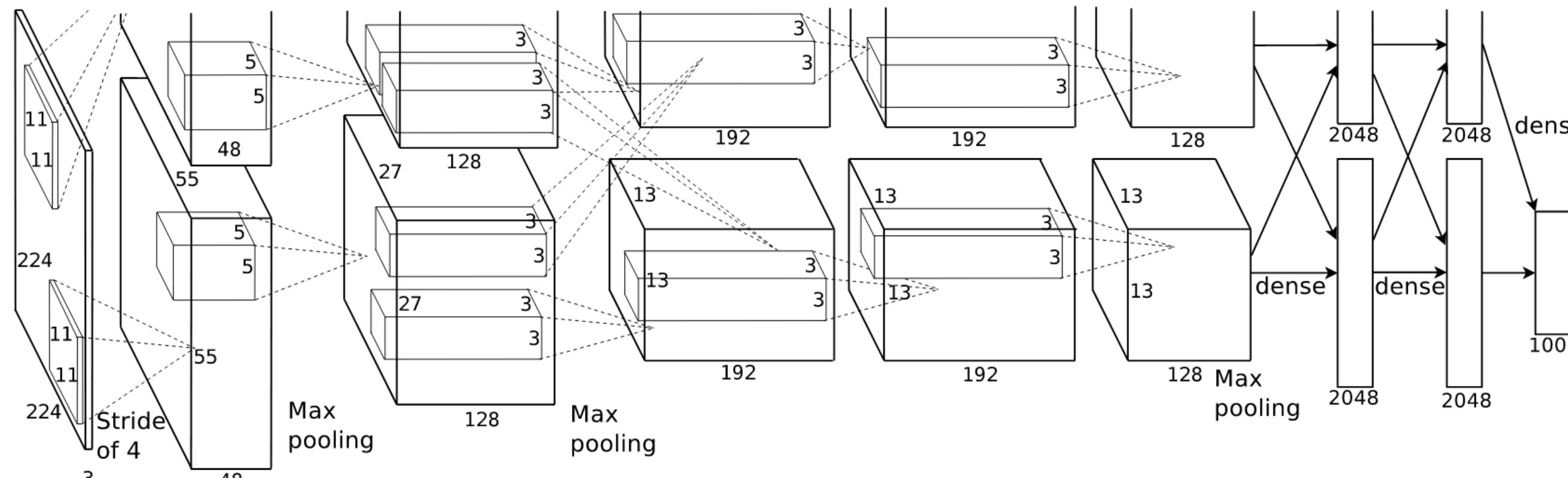
Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

ILSVRC 2012: winner

“AlexNet”



ImageNet Classification with Deep Convolutional Neural Networks

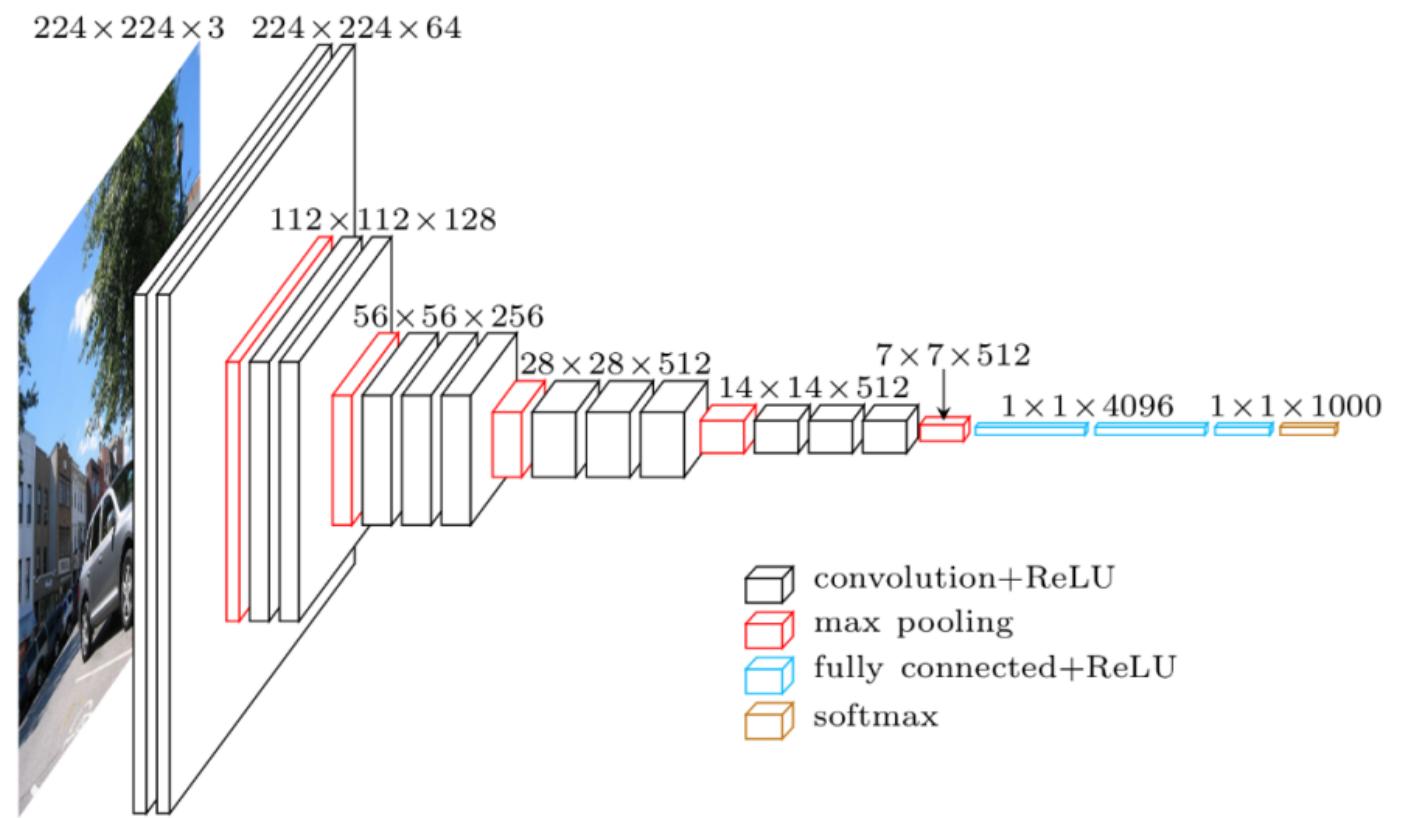
[NeurIPS 2012 paper](#)

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

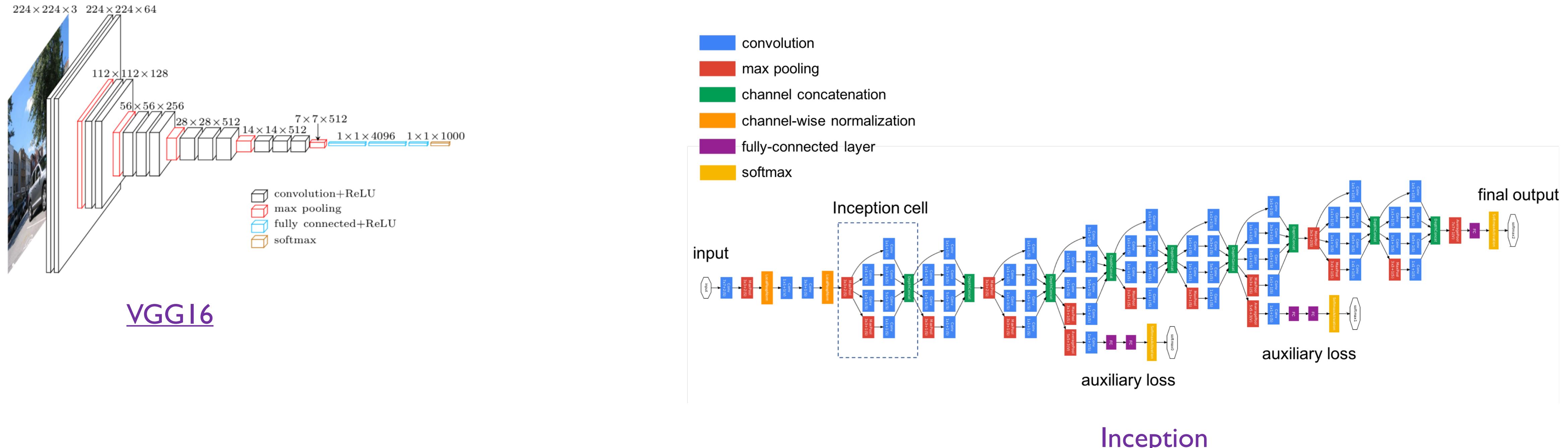
Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Deep Learning Tidal Wave

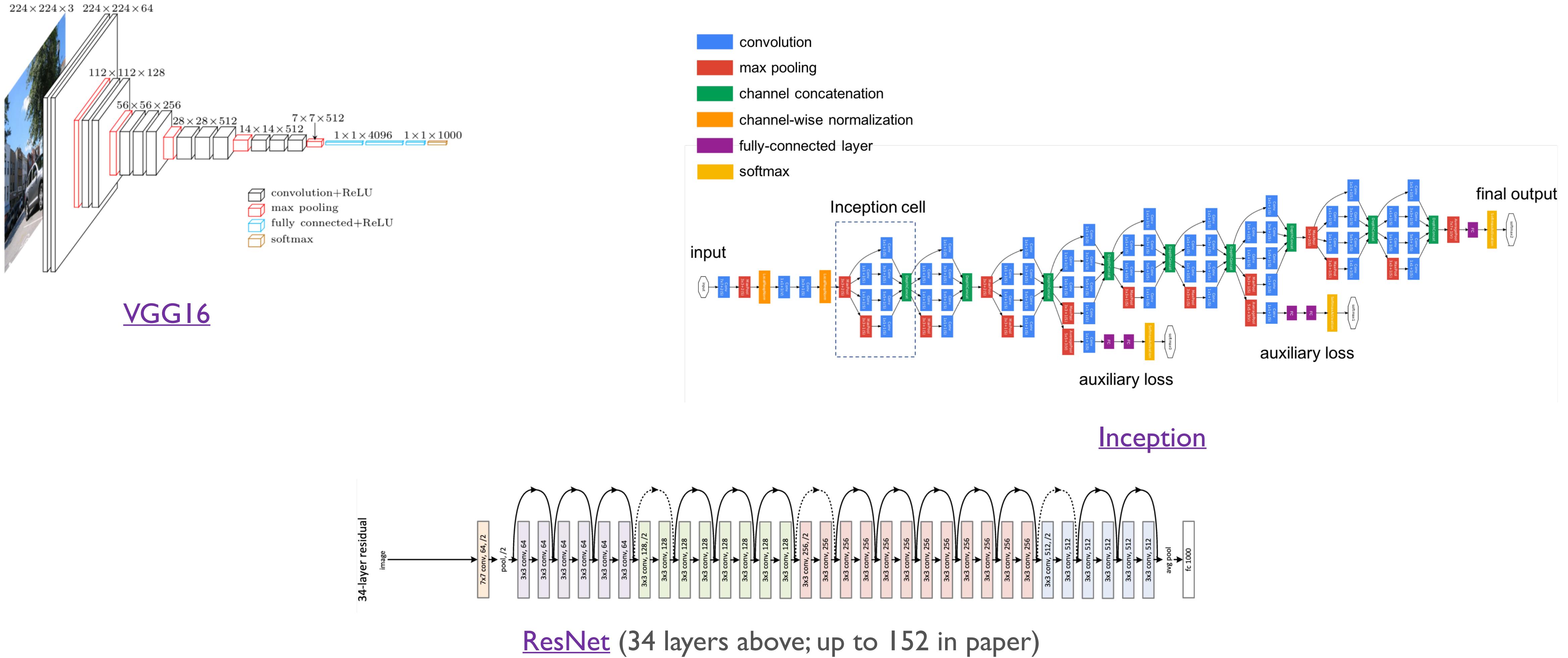


VGG16

Deep Learning Tidal Wave



Deep Learning Tidal Wave



Transfer Learning

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

Ali Sharif Razavian Hossein Azizpour Josephine Sullivan Stefan Carlsson

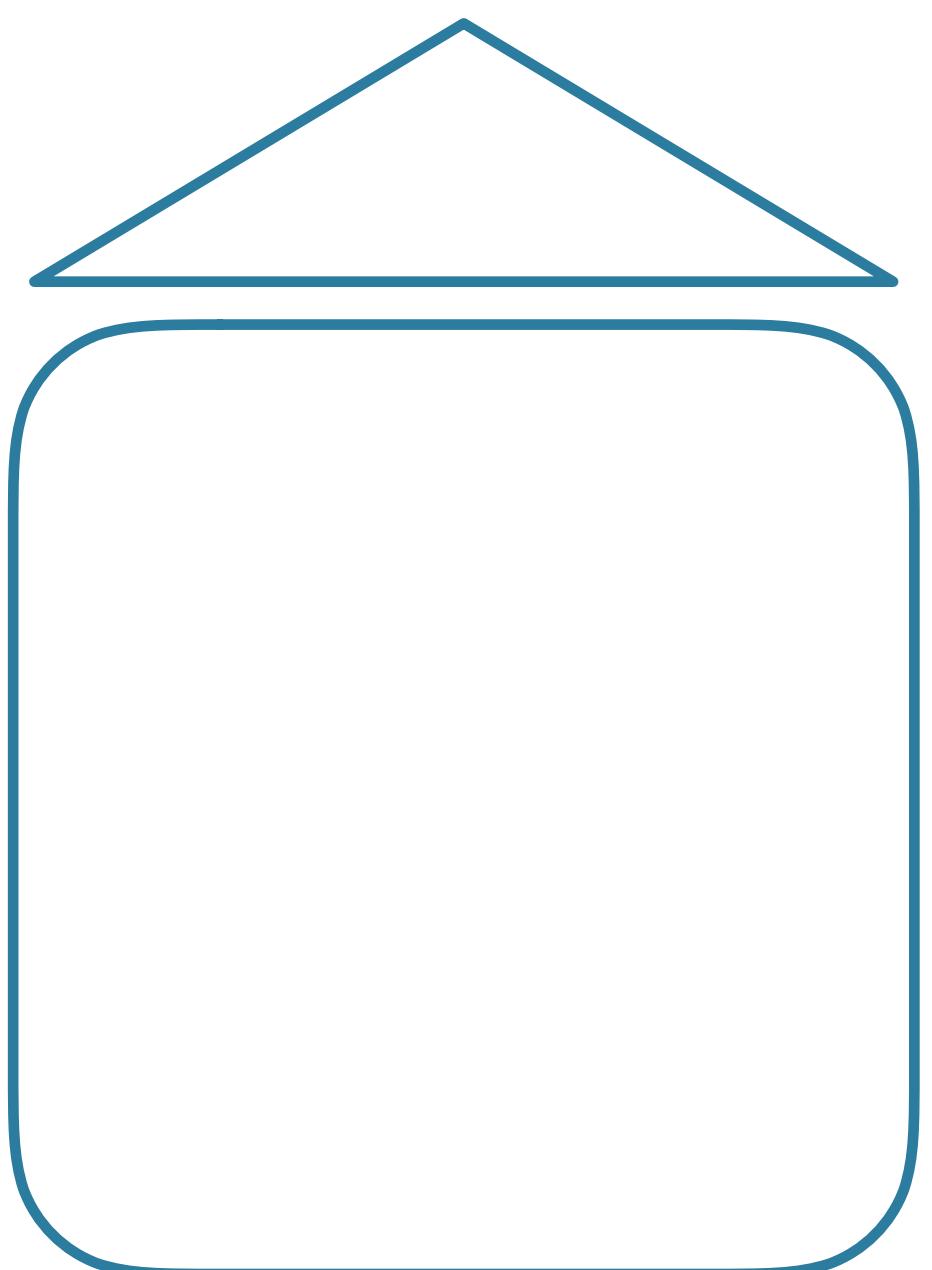
CVAP, KTH (Royal Institute of Technology)
Stockholm, Sweden

{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 gradually 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 Learning

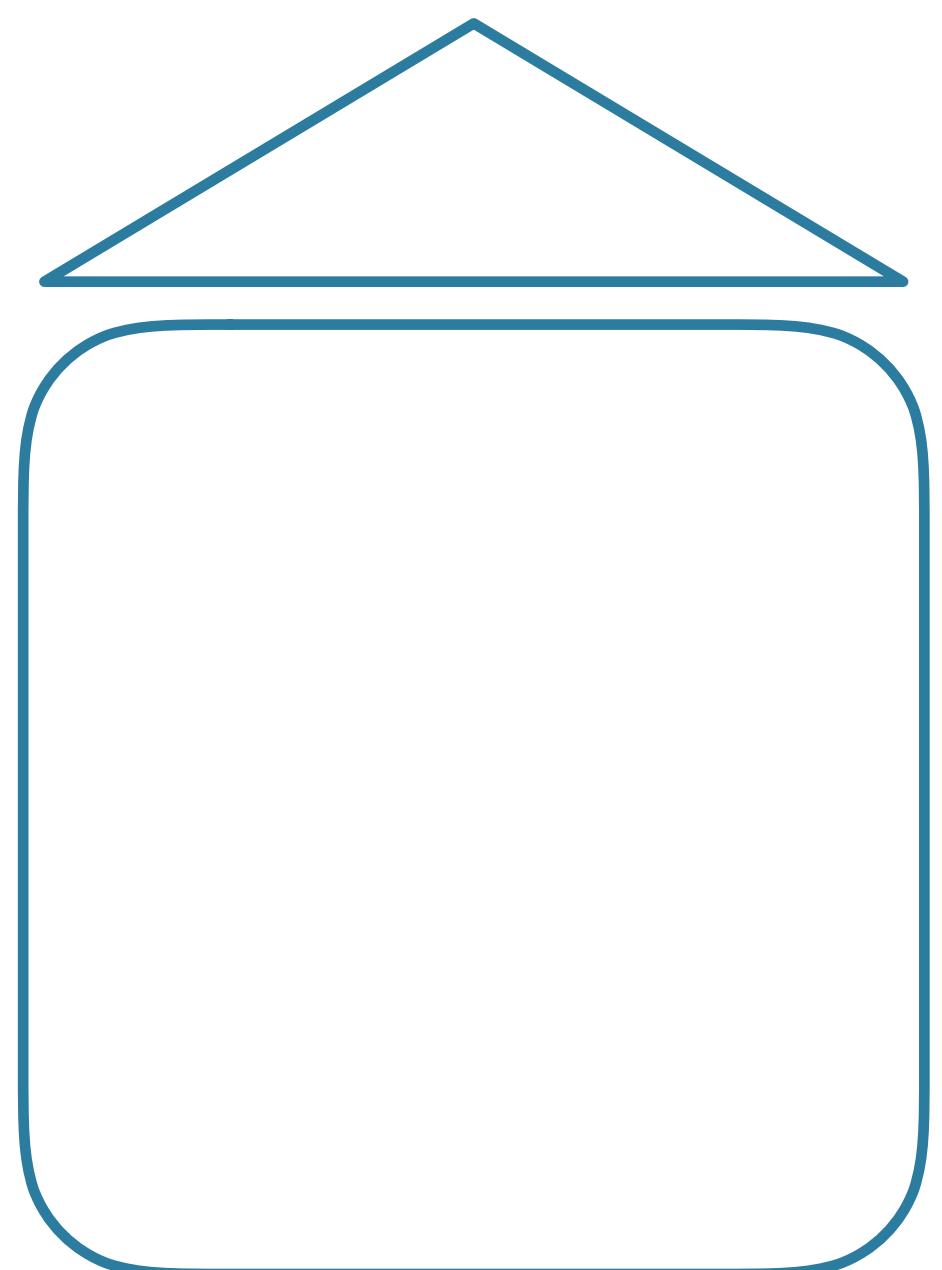
Task I outputs



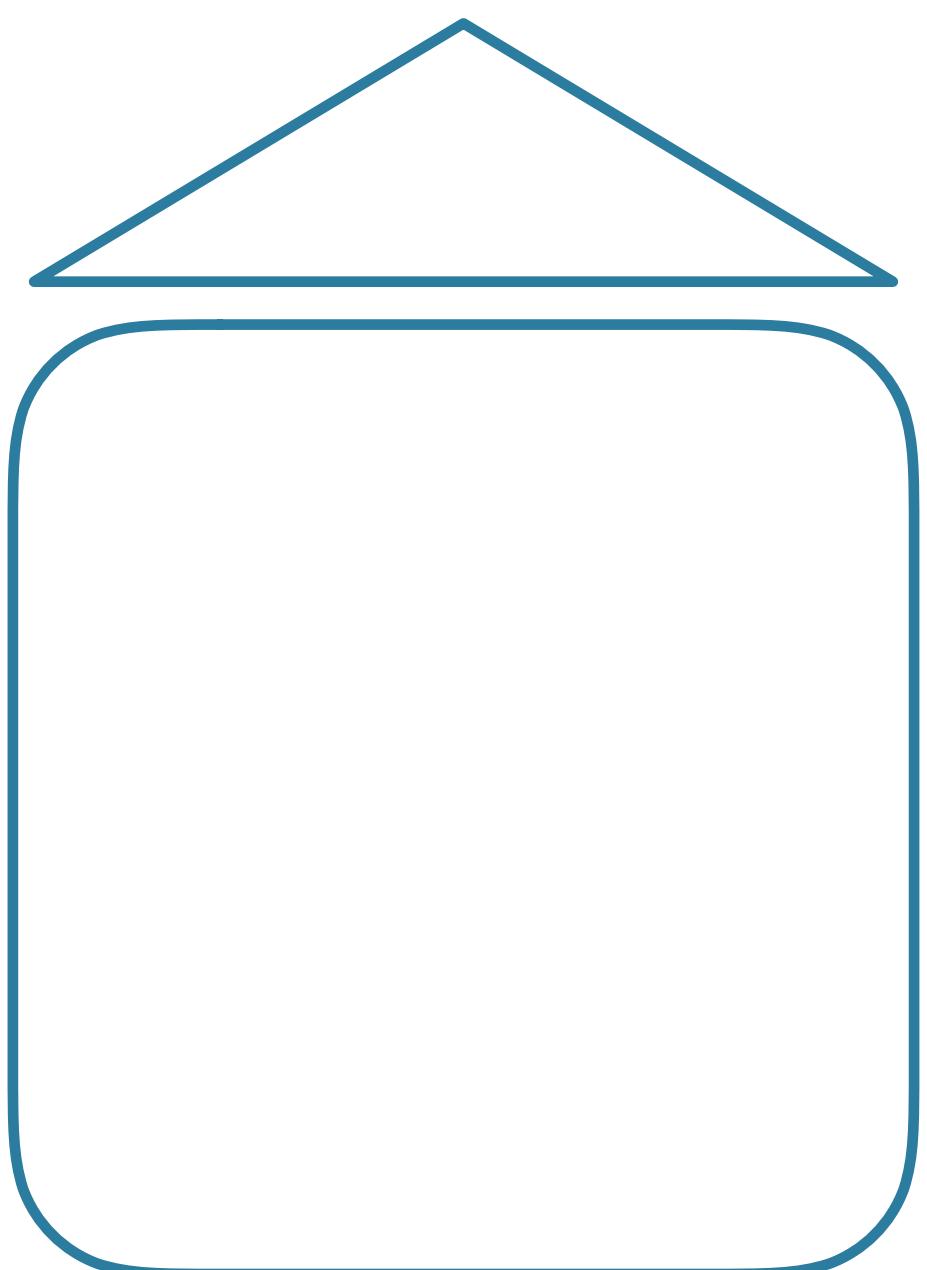
Task I inputs

Standard Learning

Task 1 outputs



Task 2 outputs

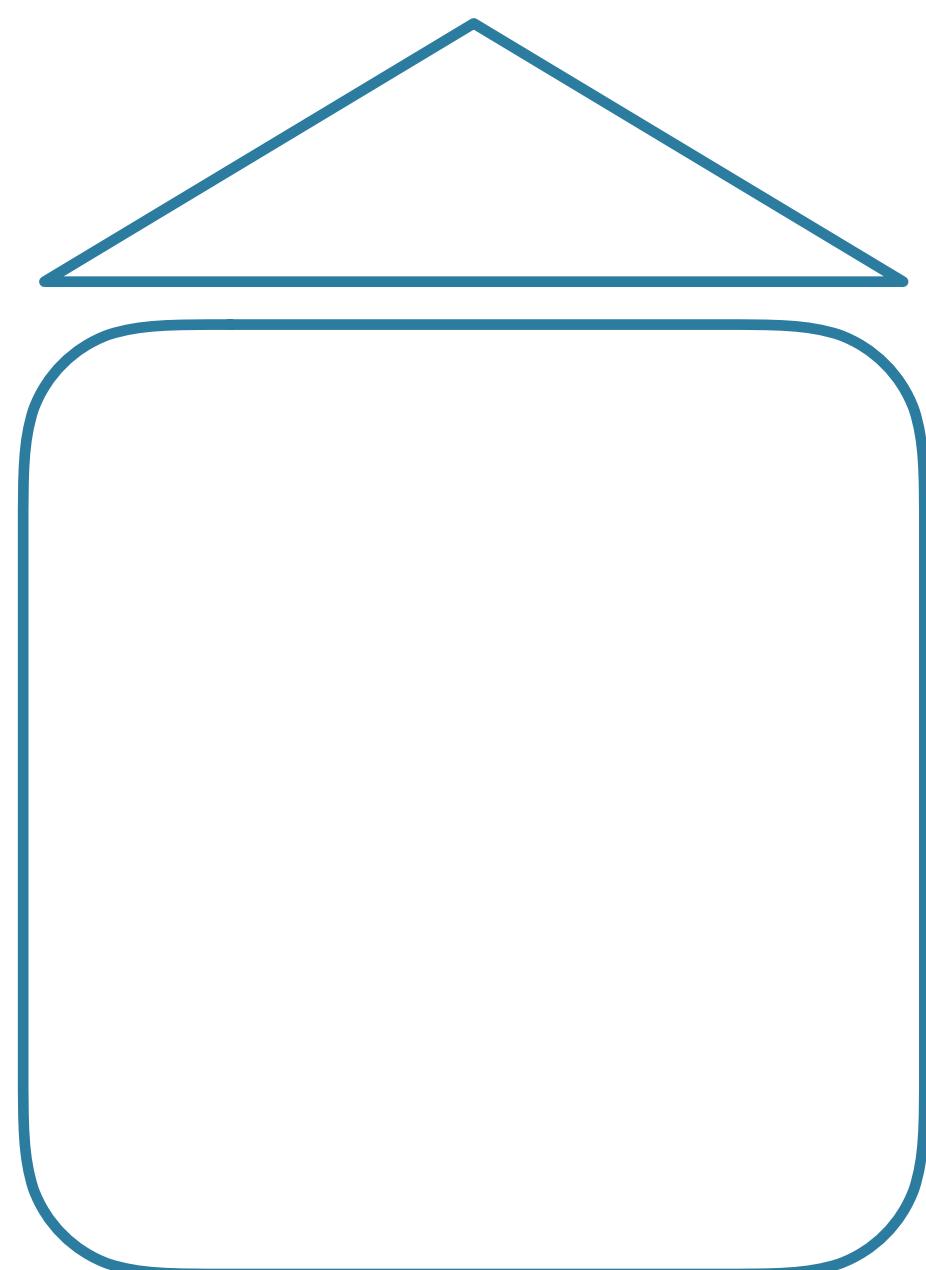


Task 1 inputs

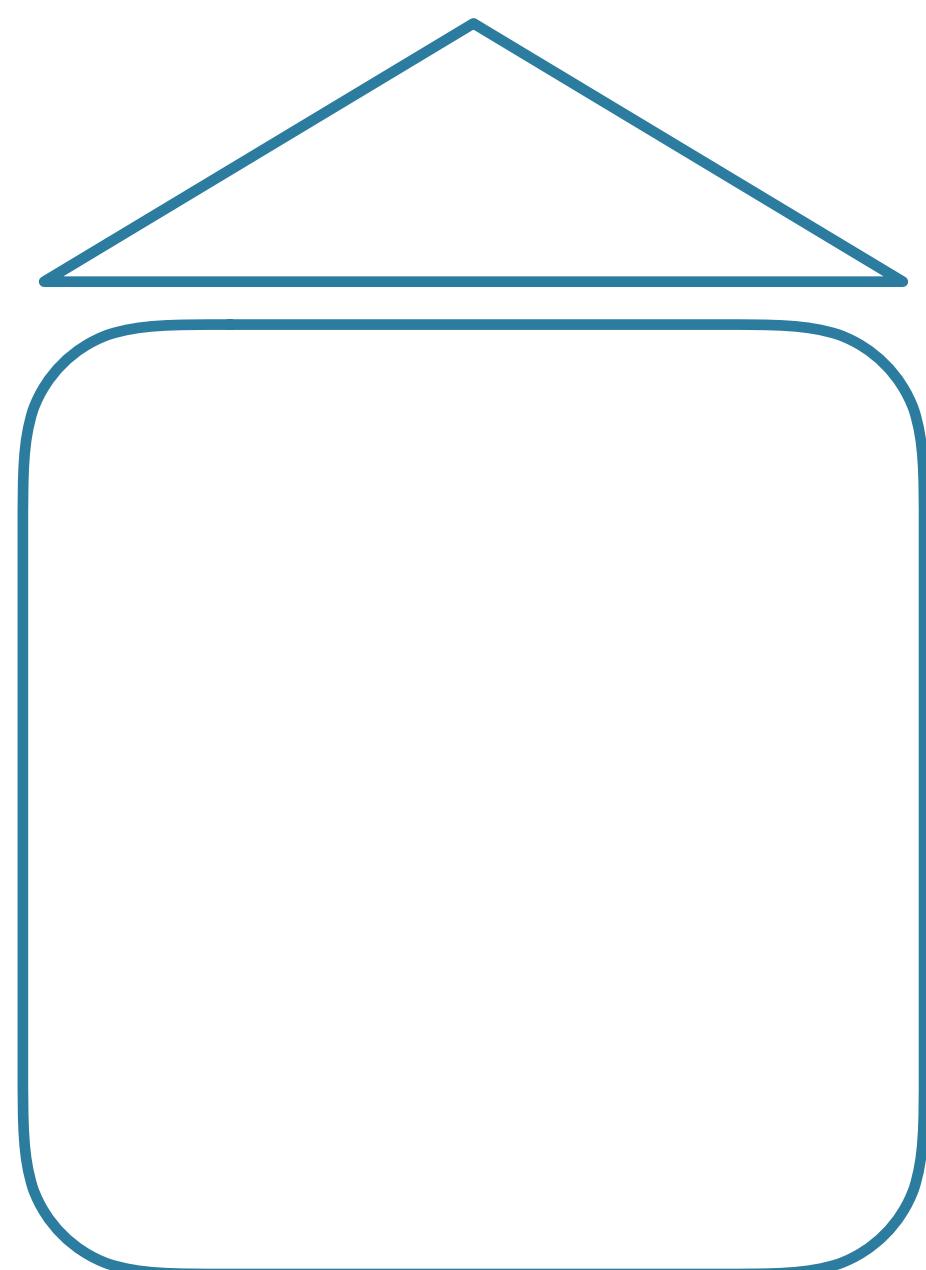
Task 2 inputs

Standard Learning

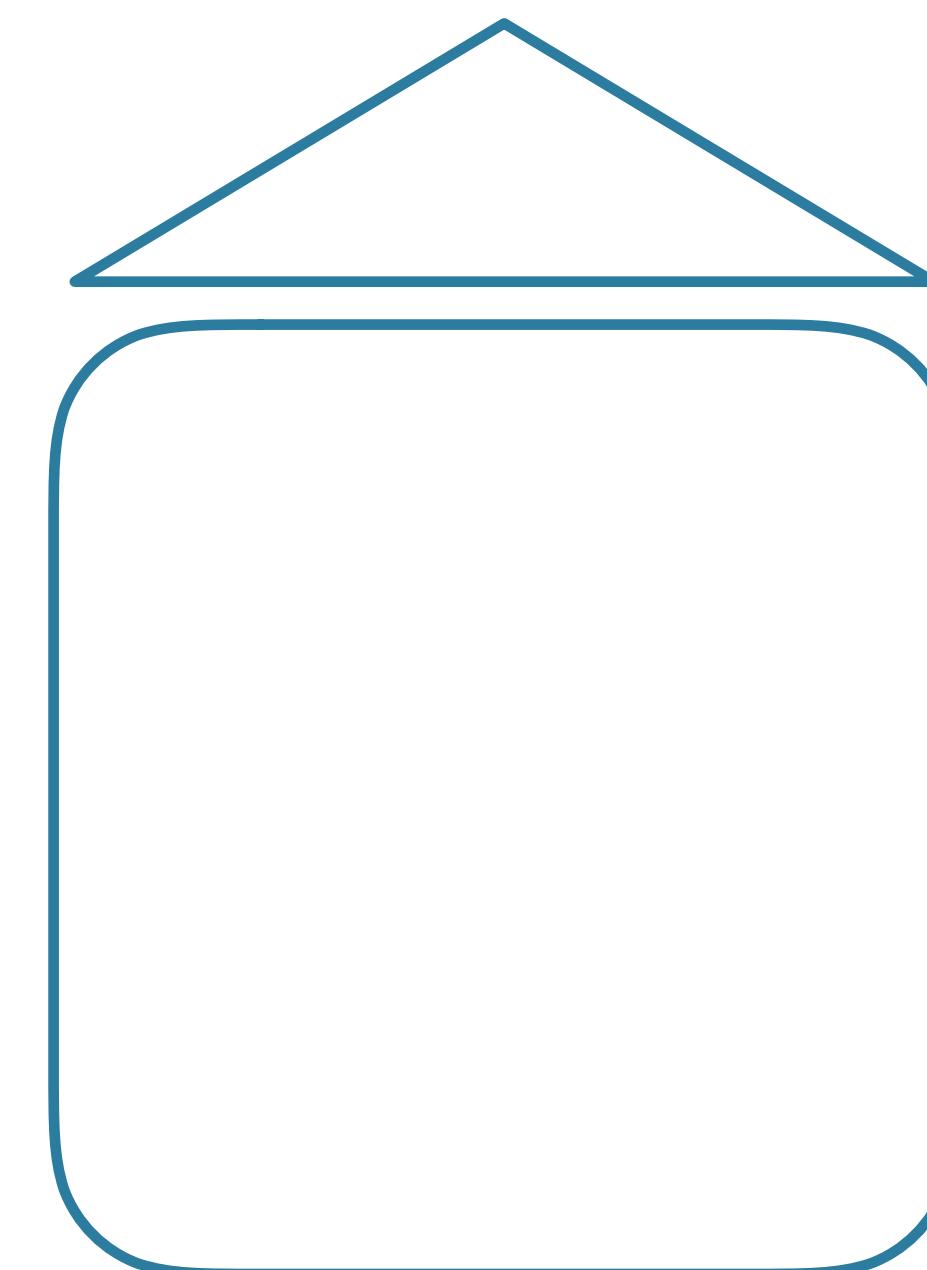
Task 1 outputs



Task 2 outputs



Task 3 outputs



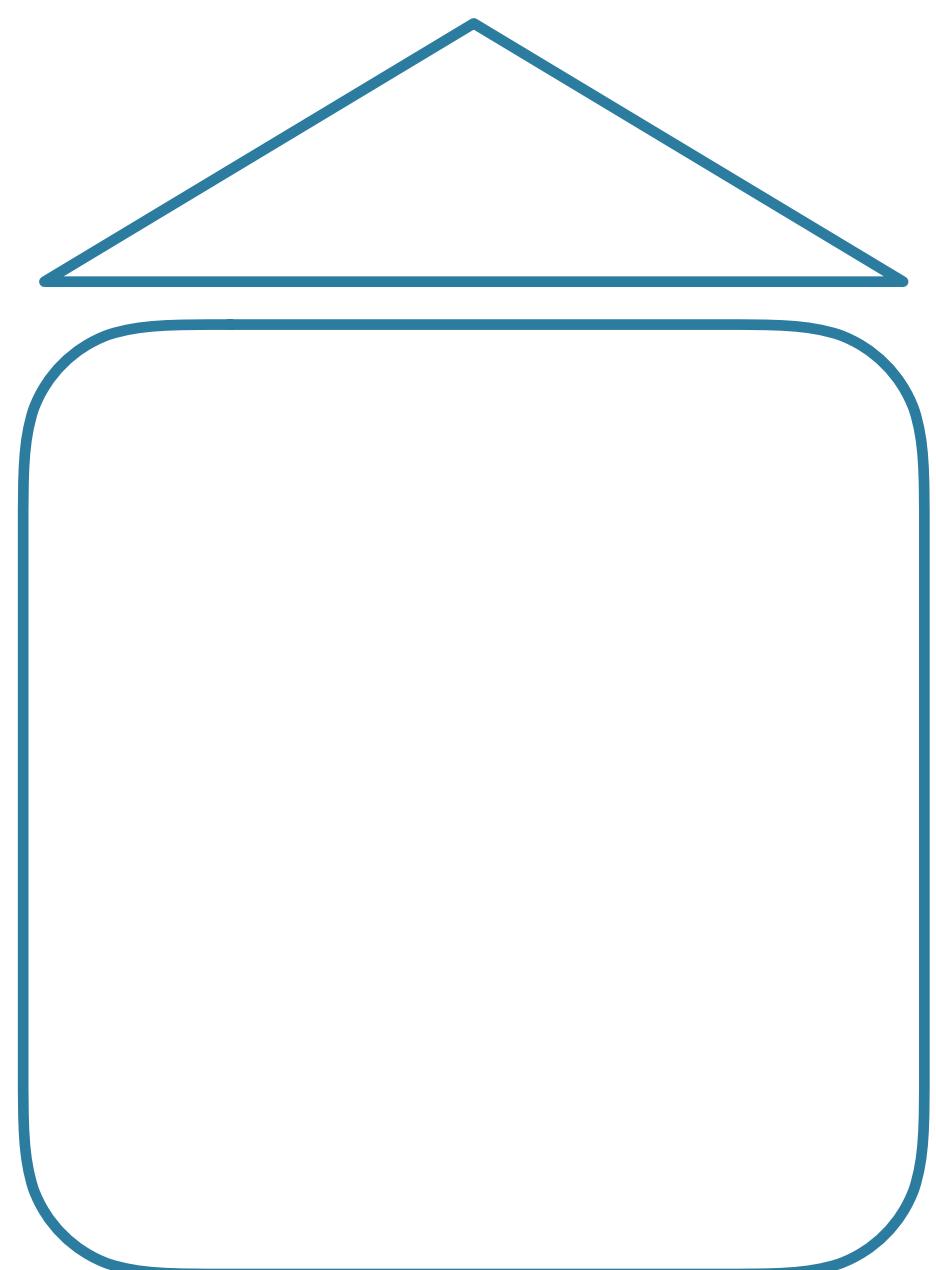
Task 1 inputs

Task 2 inputs

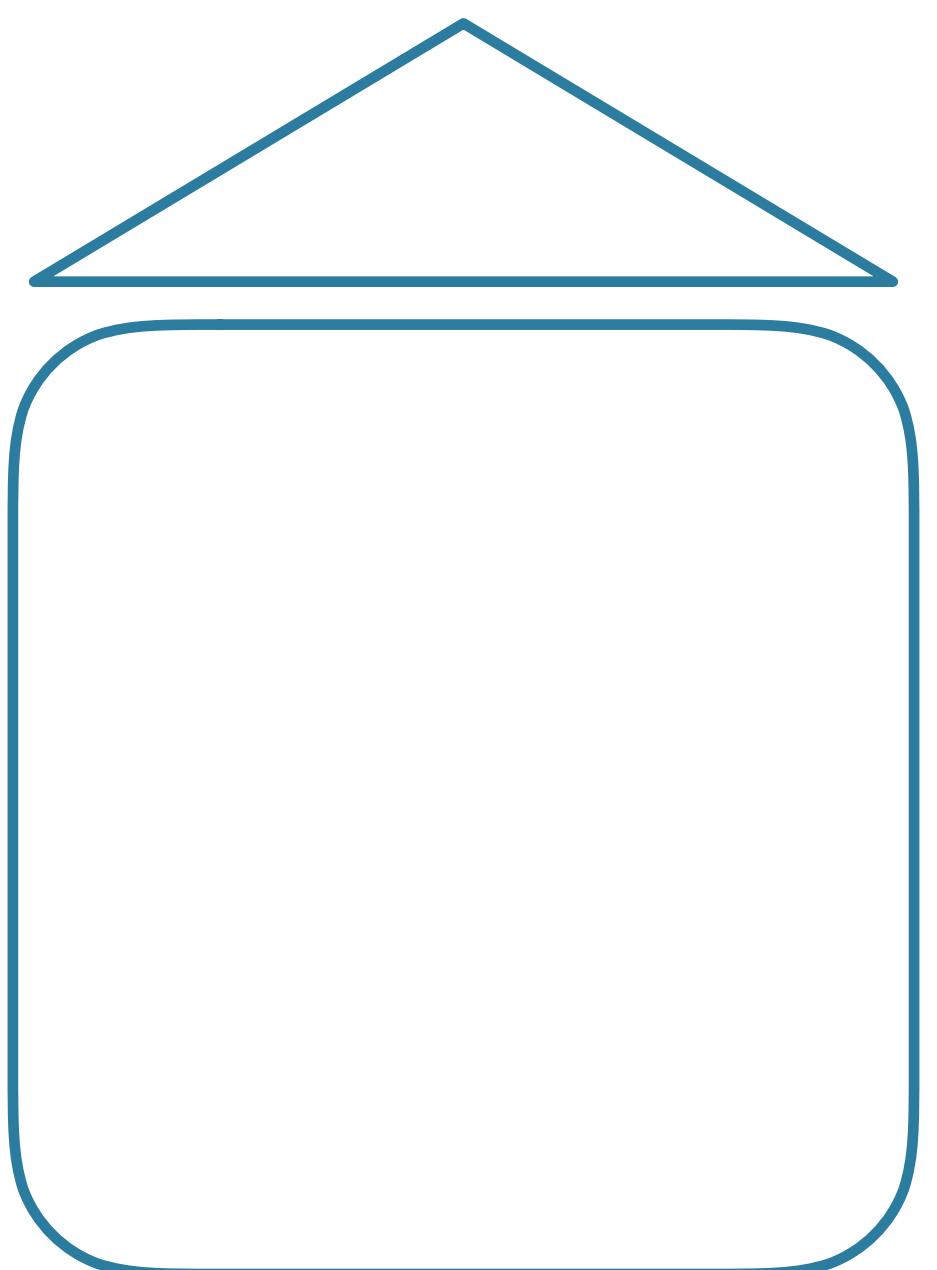
Task 3 inputs

Standard Learning

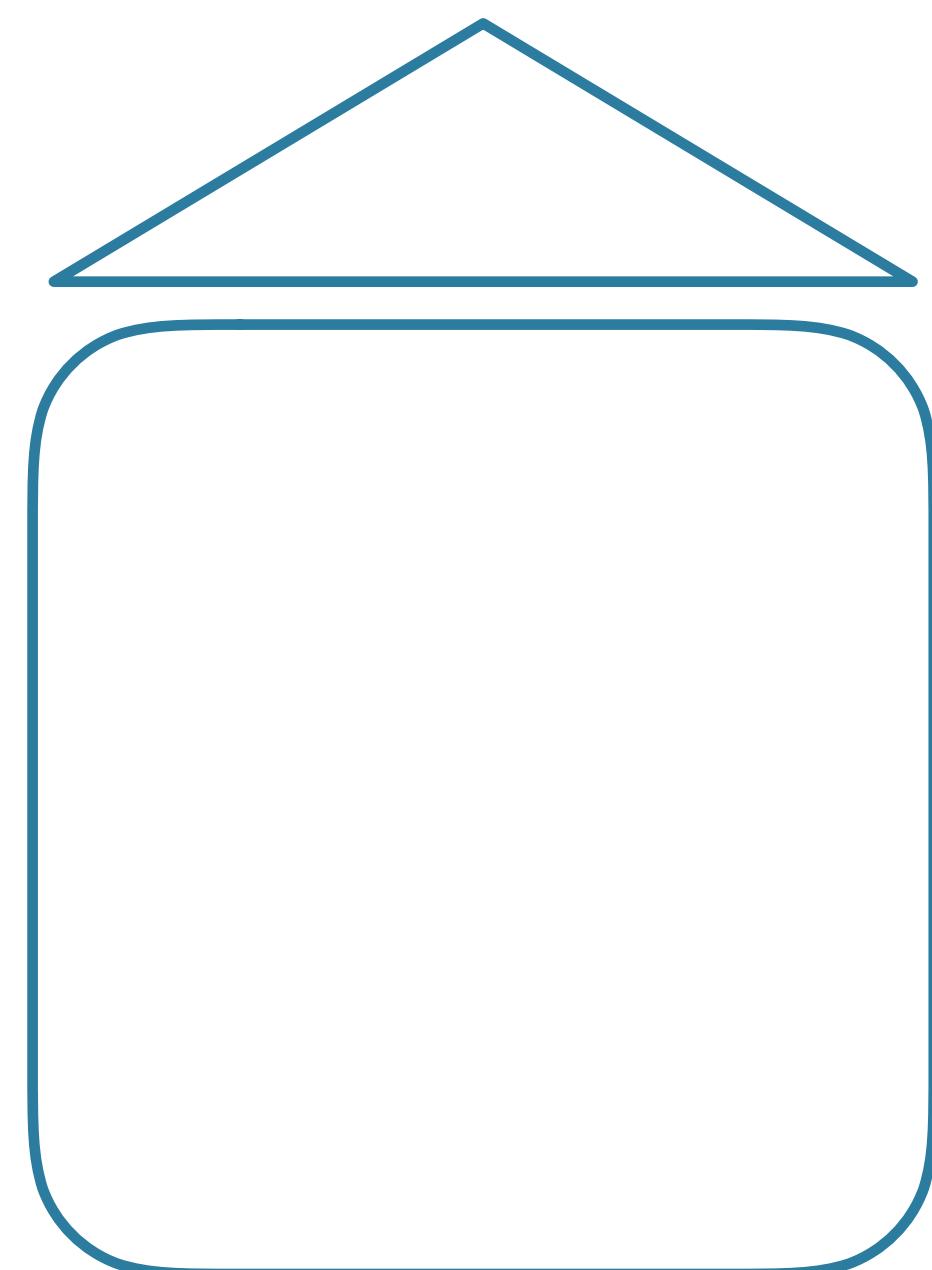
Task 1 outputs



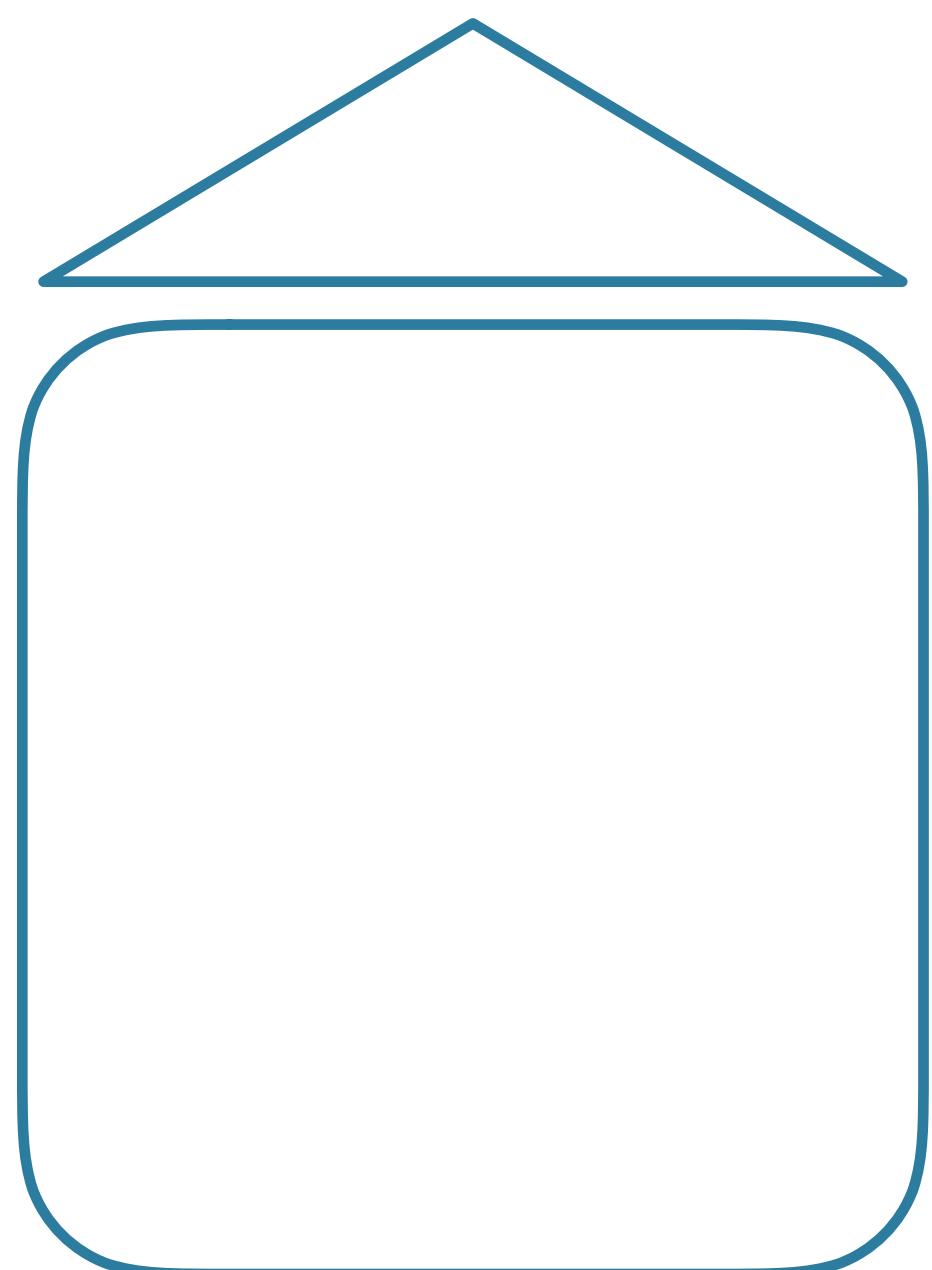
Task 2 outputs



Task 3 outputs



Task 4 outputs



Task 1 inputs

Task 2 inputs

Task 3 inputs

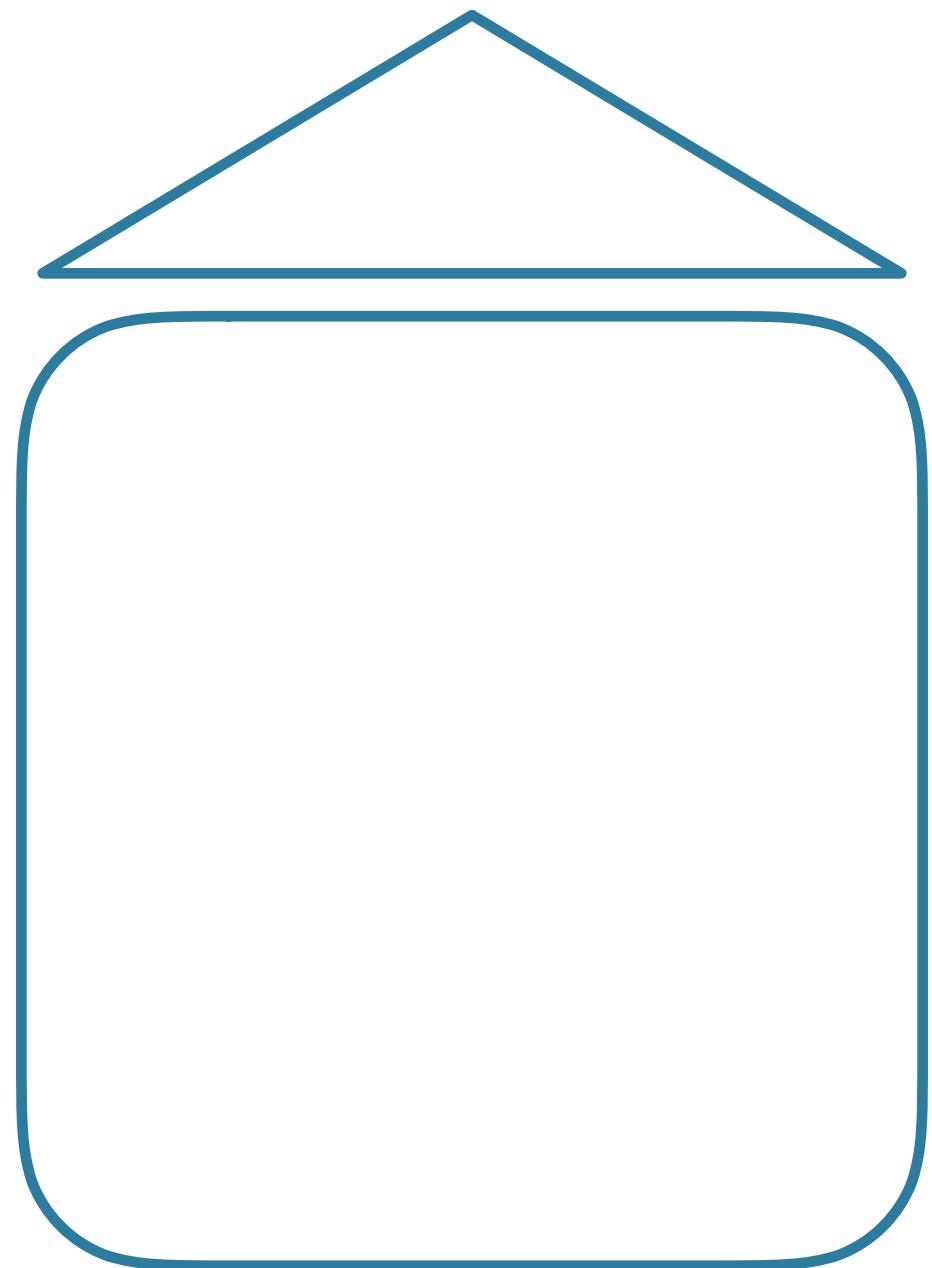
Task 4 inputs

Standard Learning

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

Transfer Learning

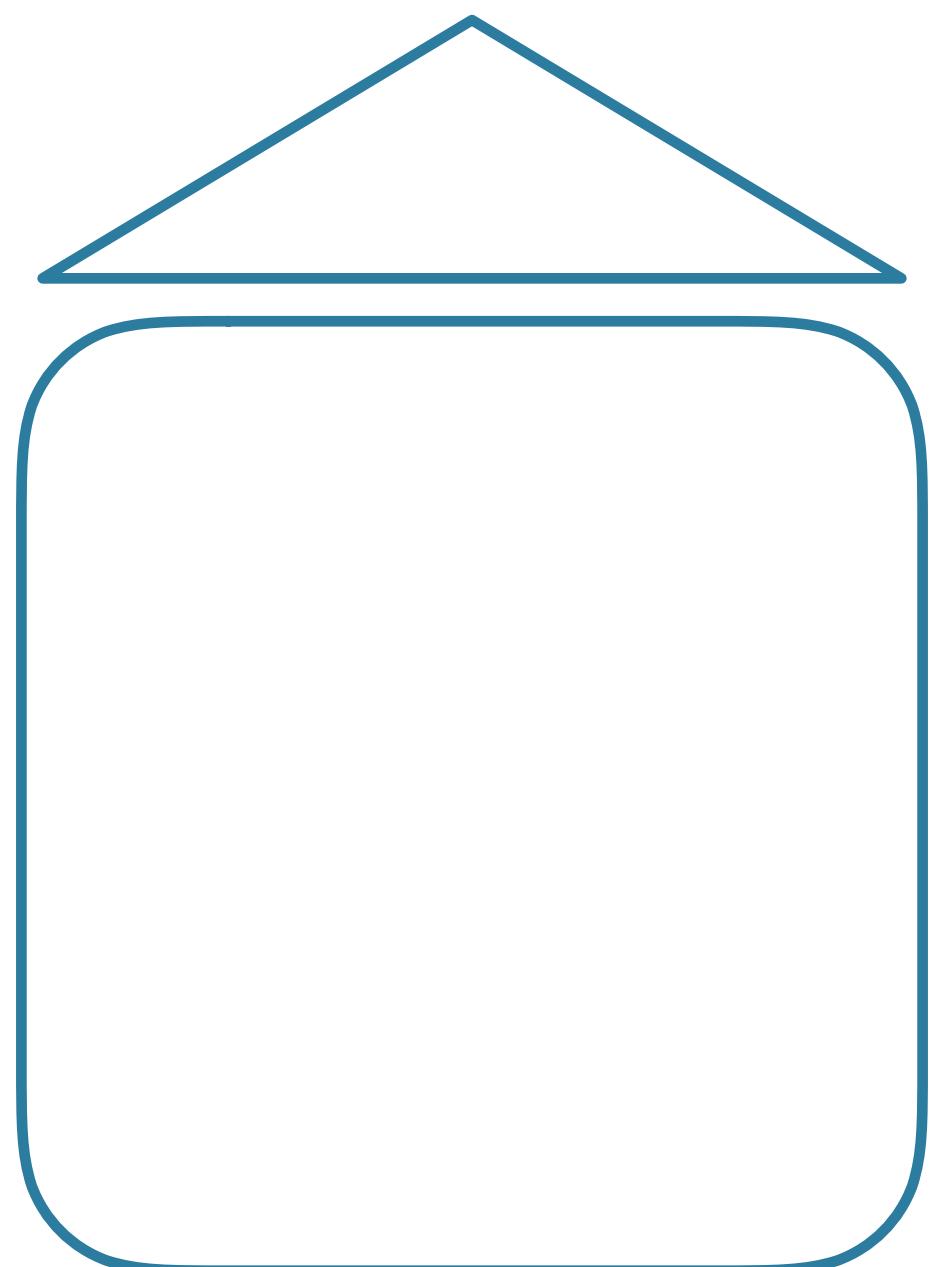
“pre-training” task outputs



“pre-training” task outputs

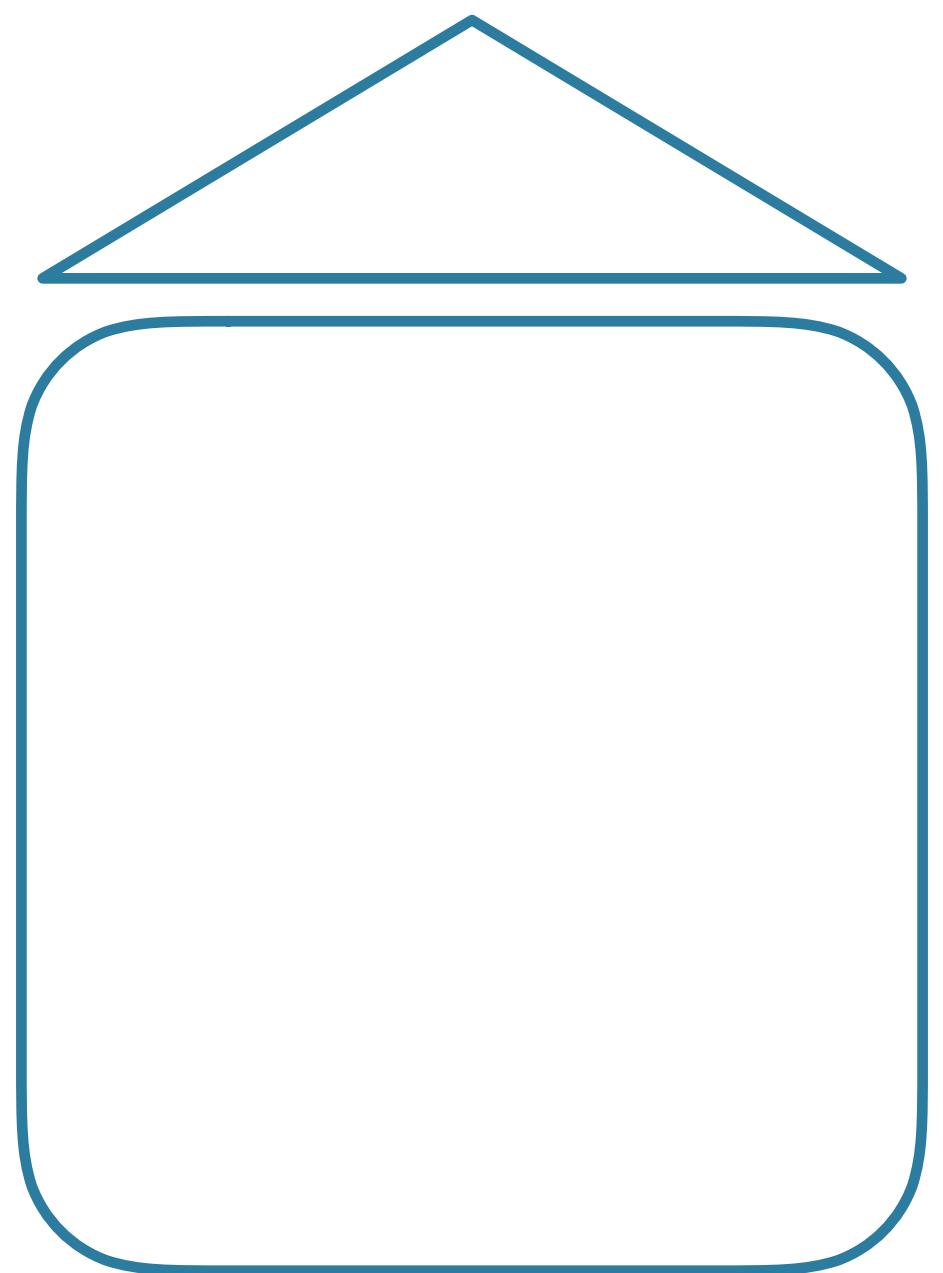
Transfer Learning

“pre-training” task outputs



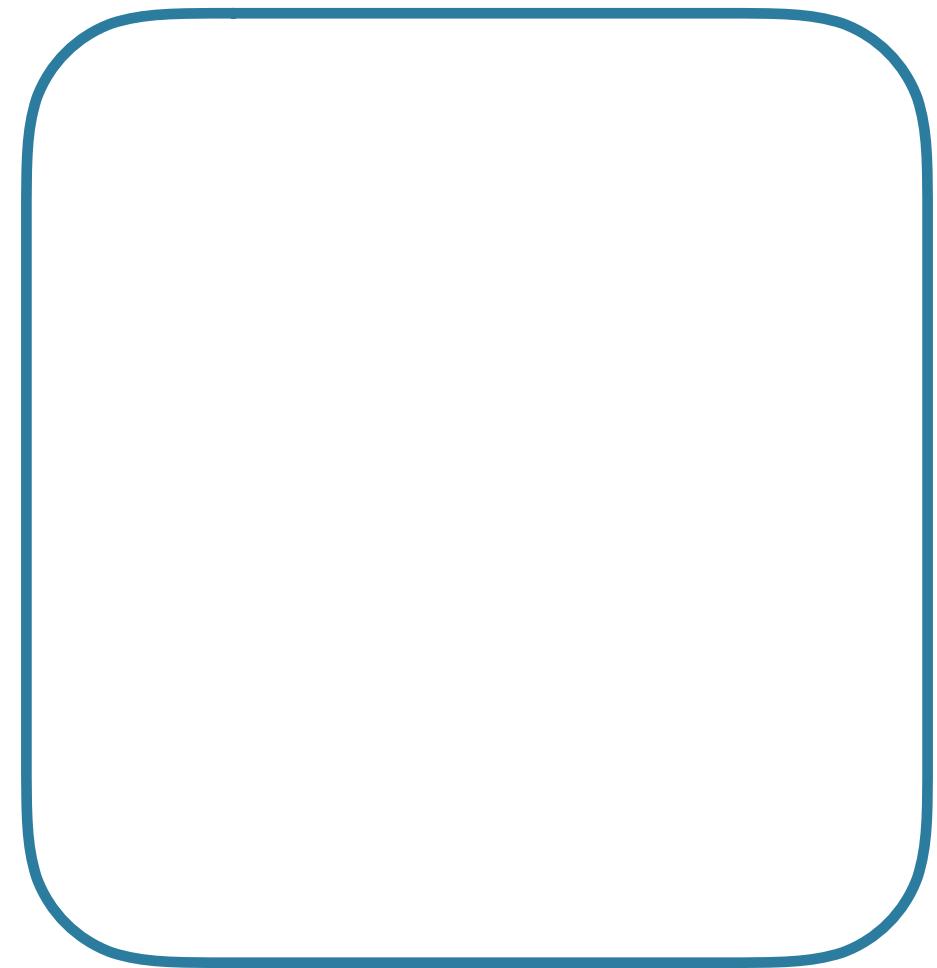
Transfer Learning

“pre-training” task outputs



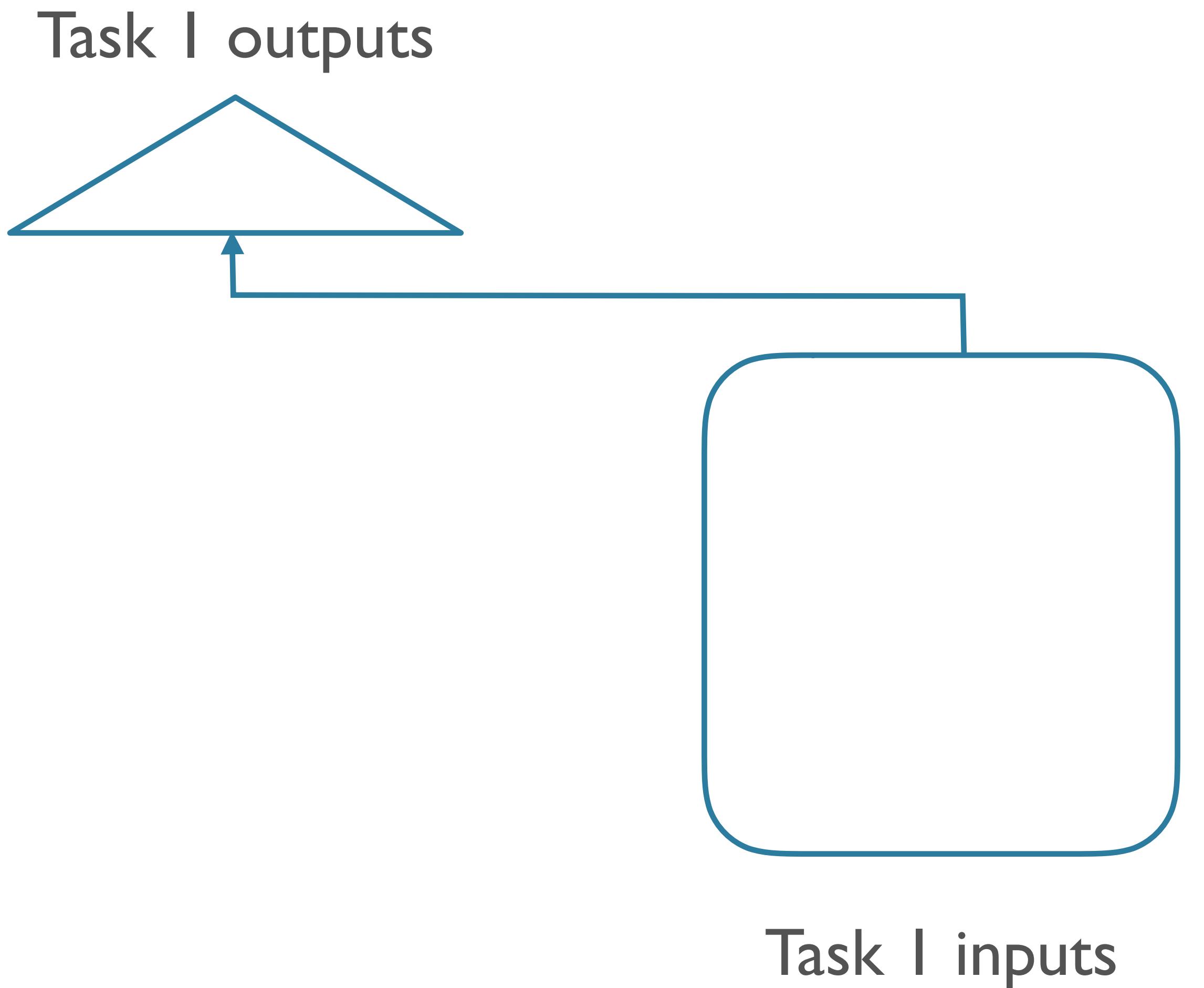
Task I inputs

Transfer Learning

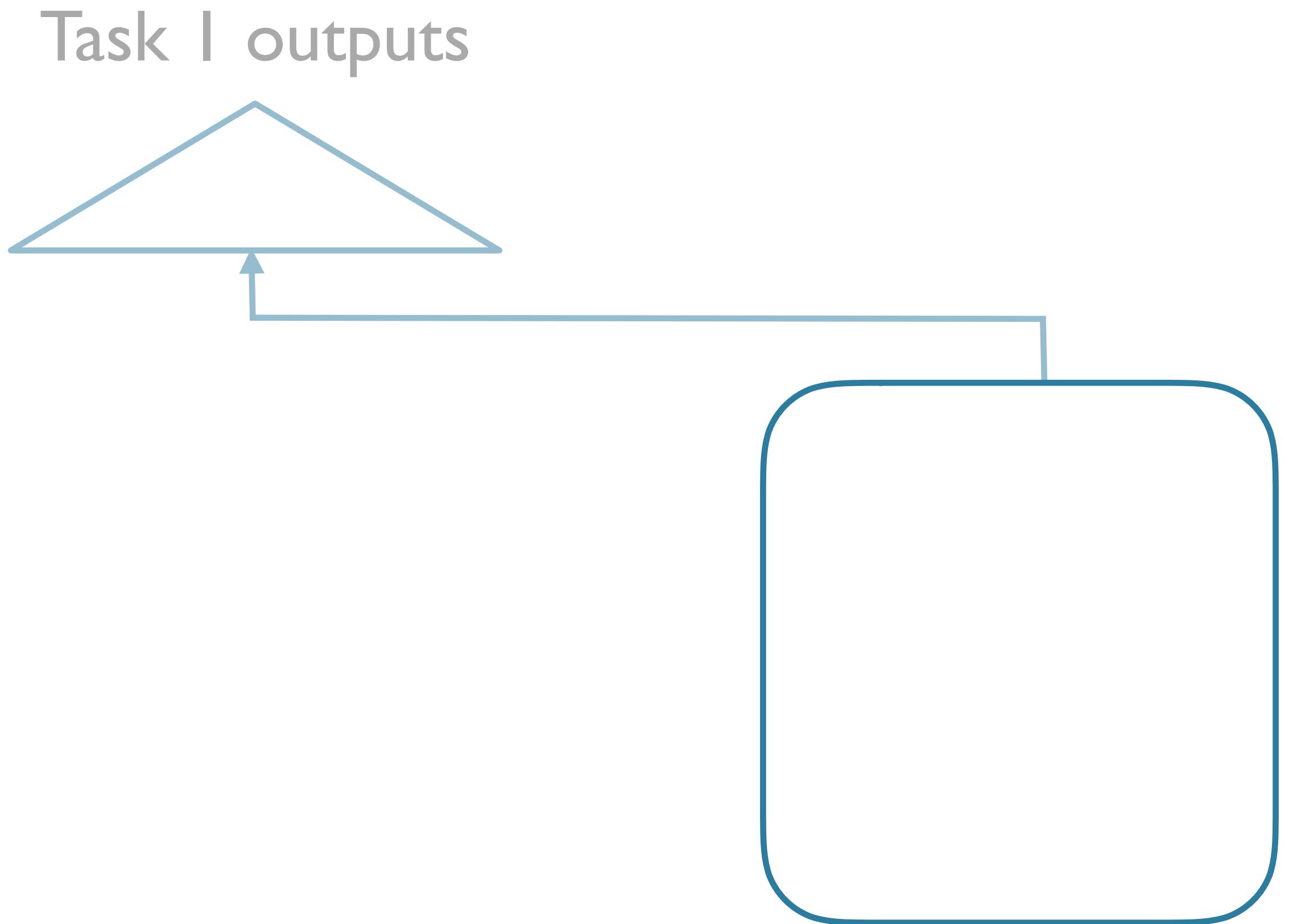


Task I inputs

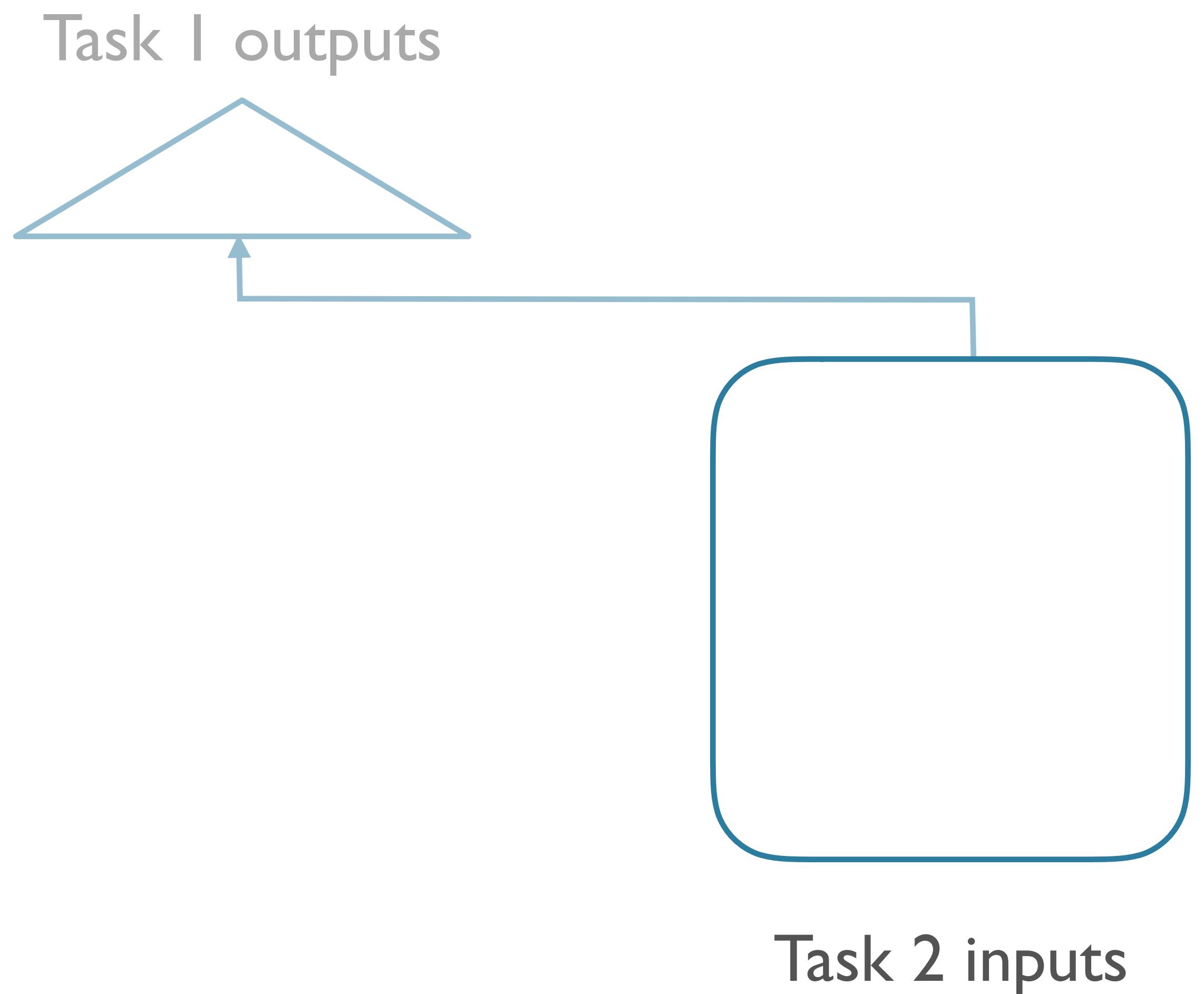
Transfer Learning



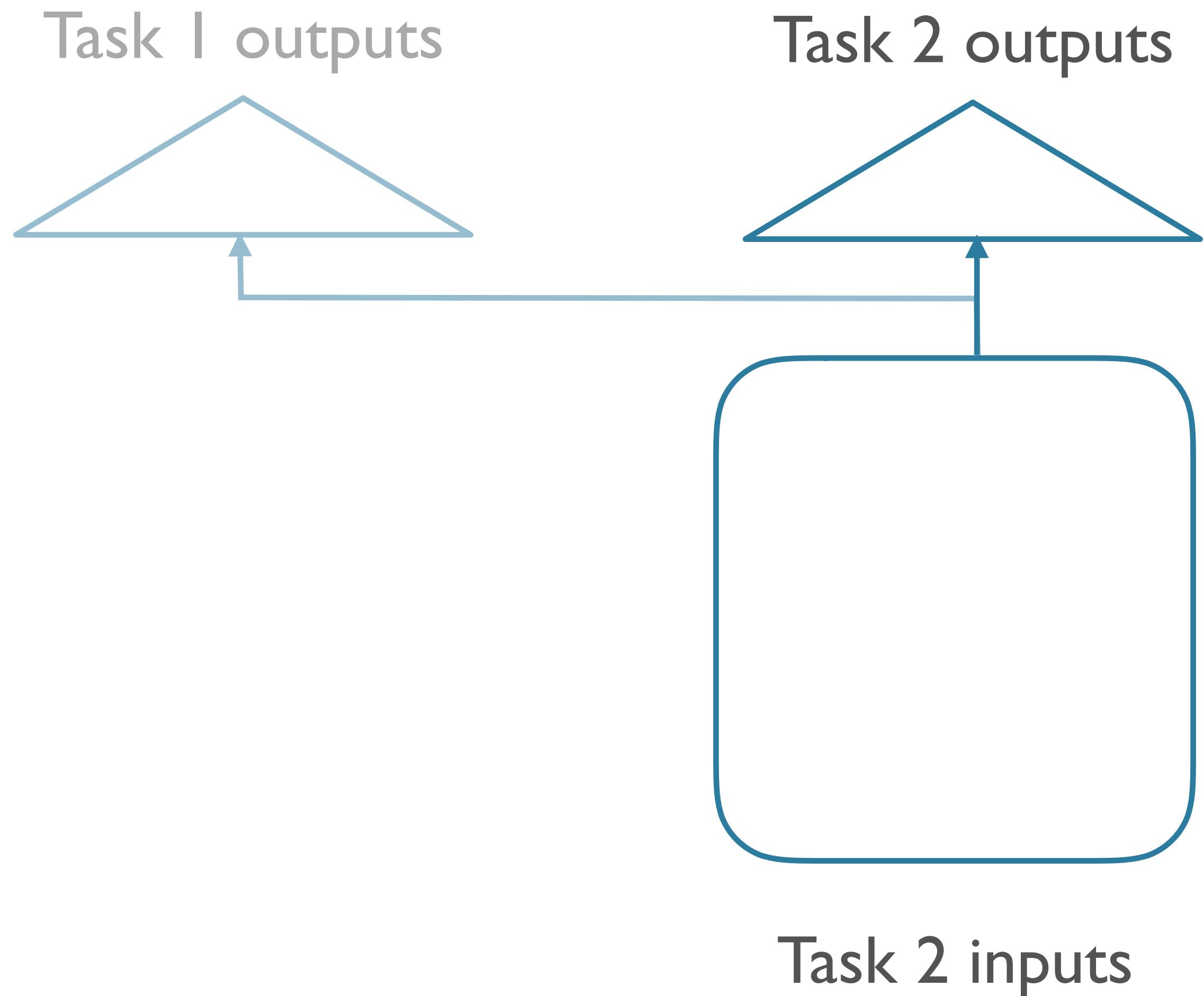
Transfer Learning



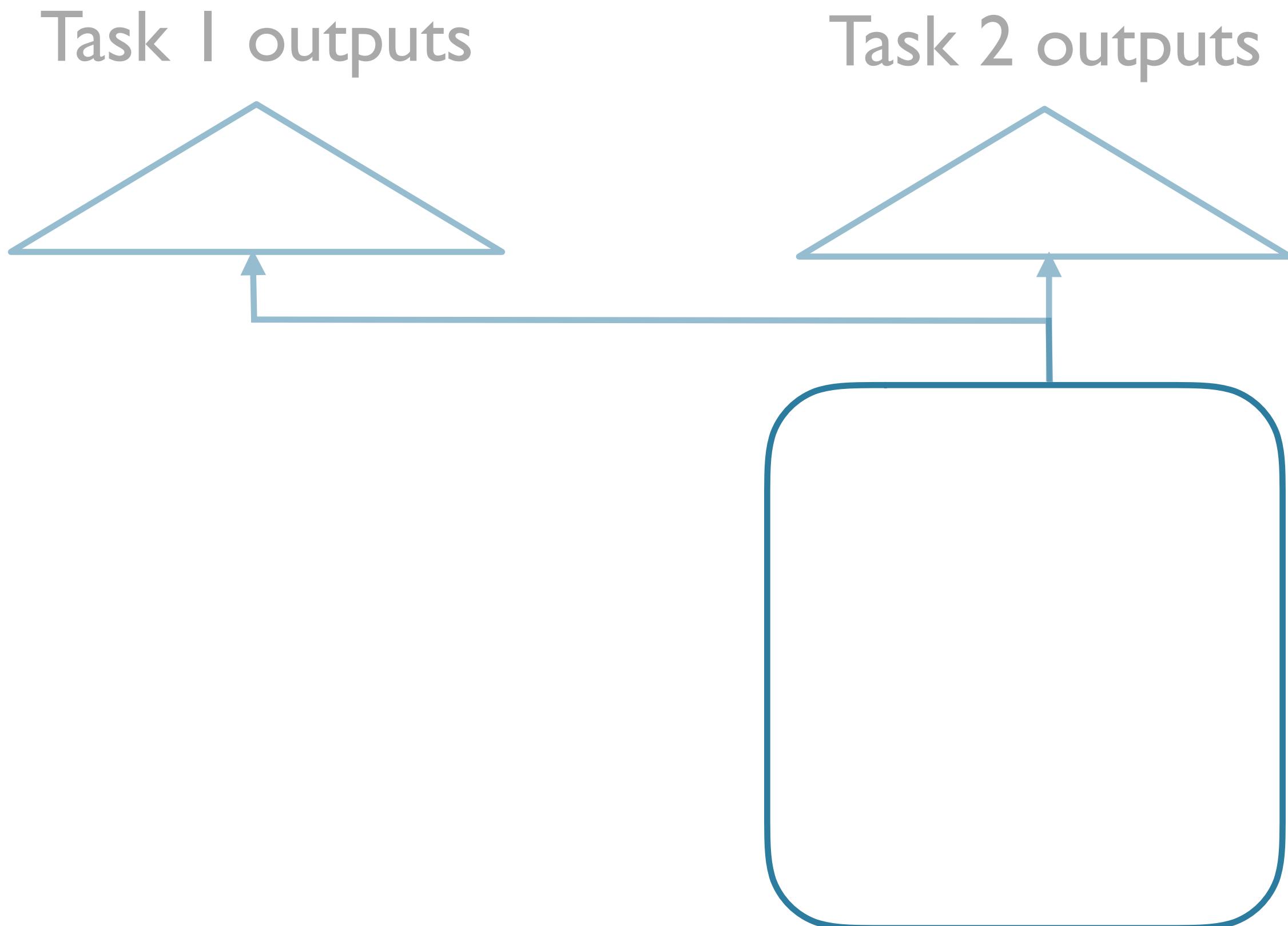
Transfer Learning



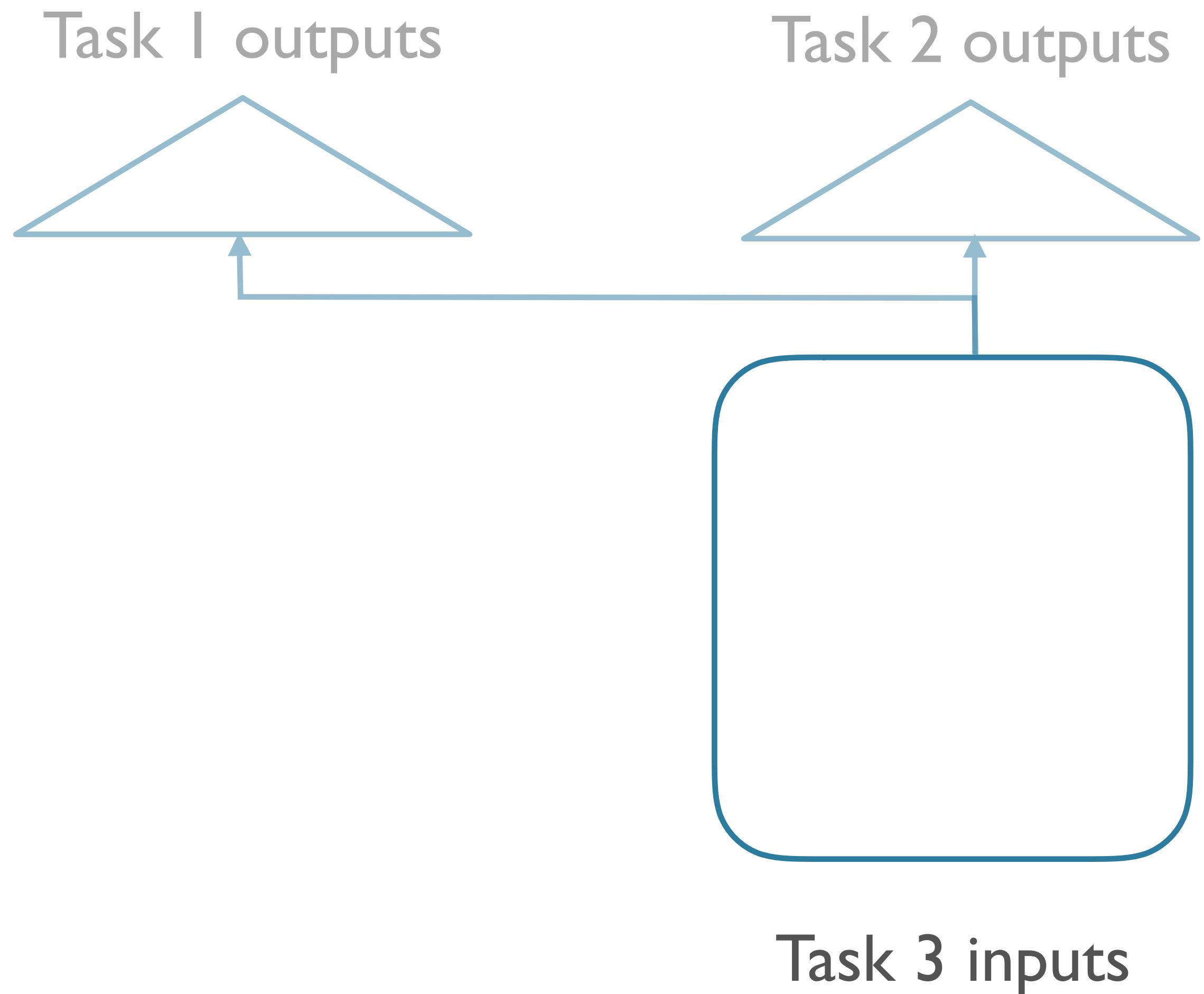
Transfer Learning



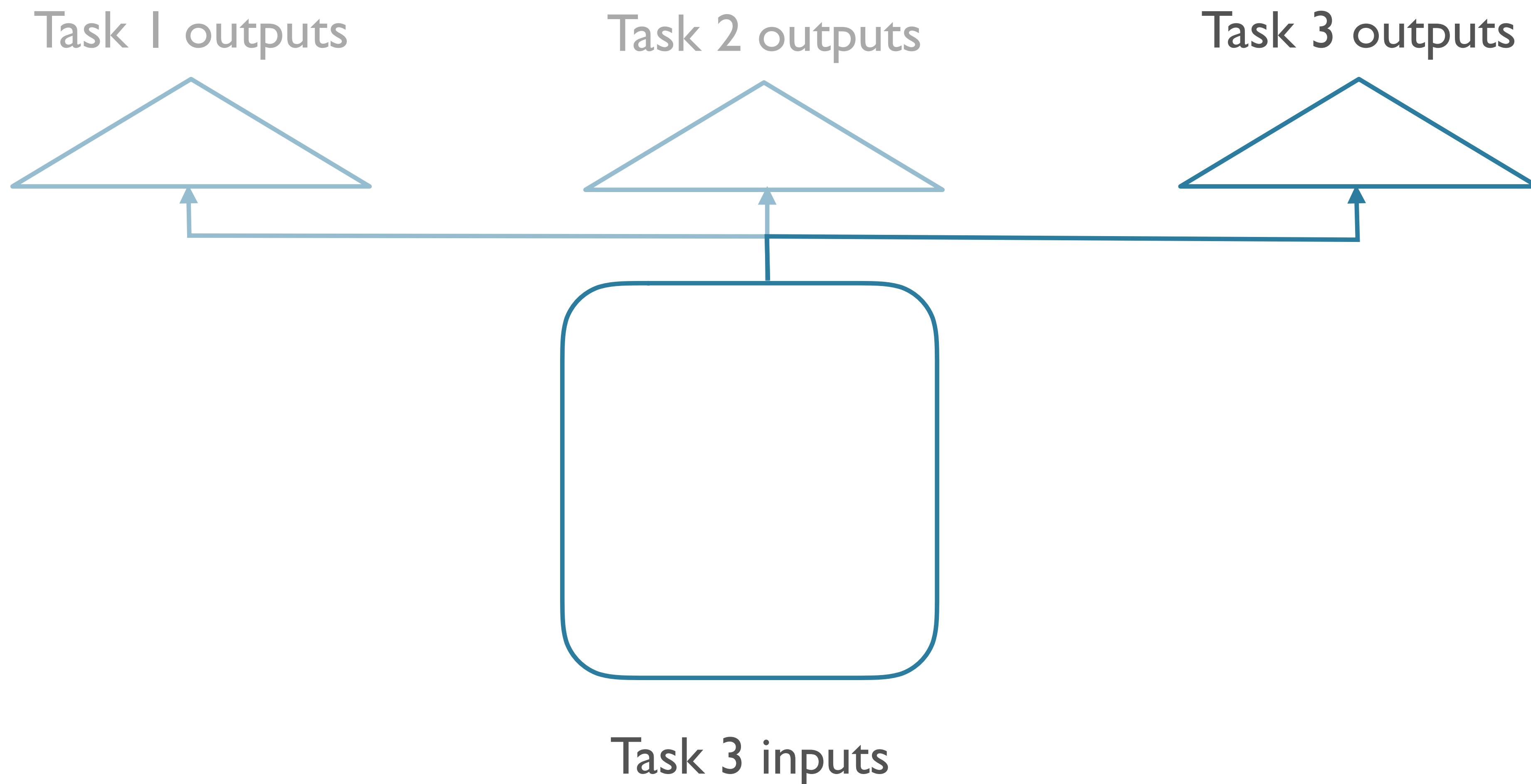
Transfer Learning



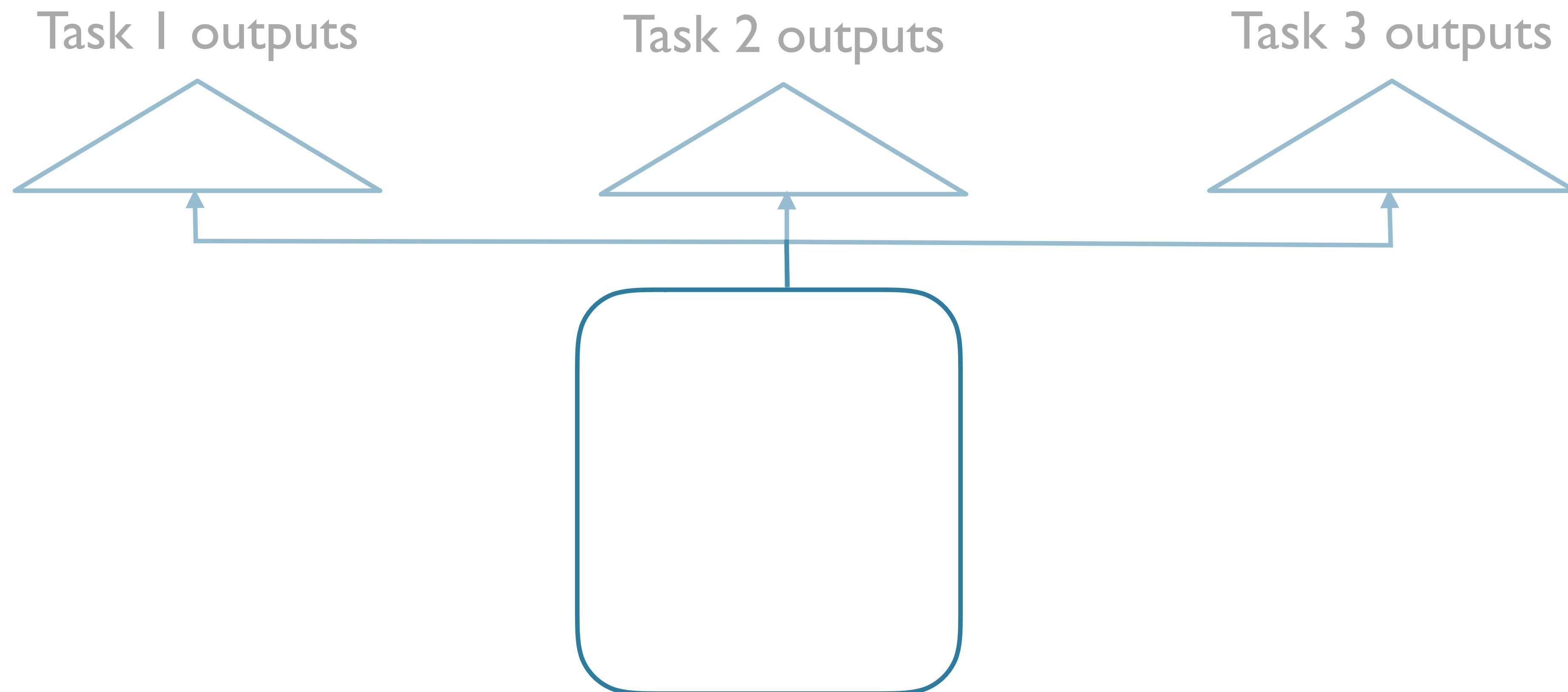
Transfer Learning



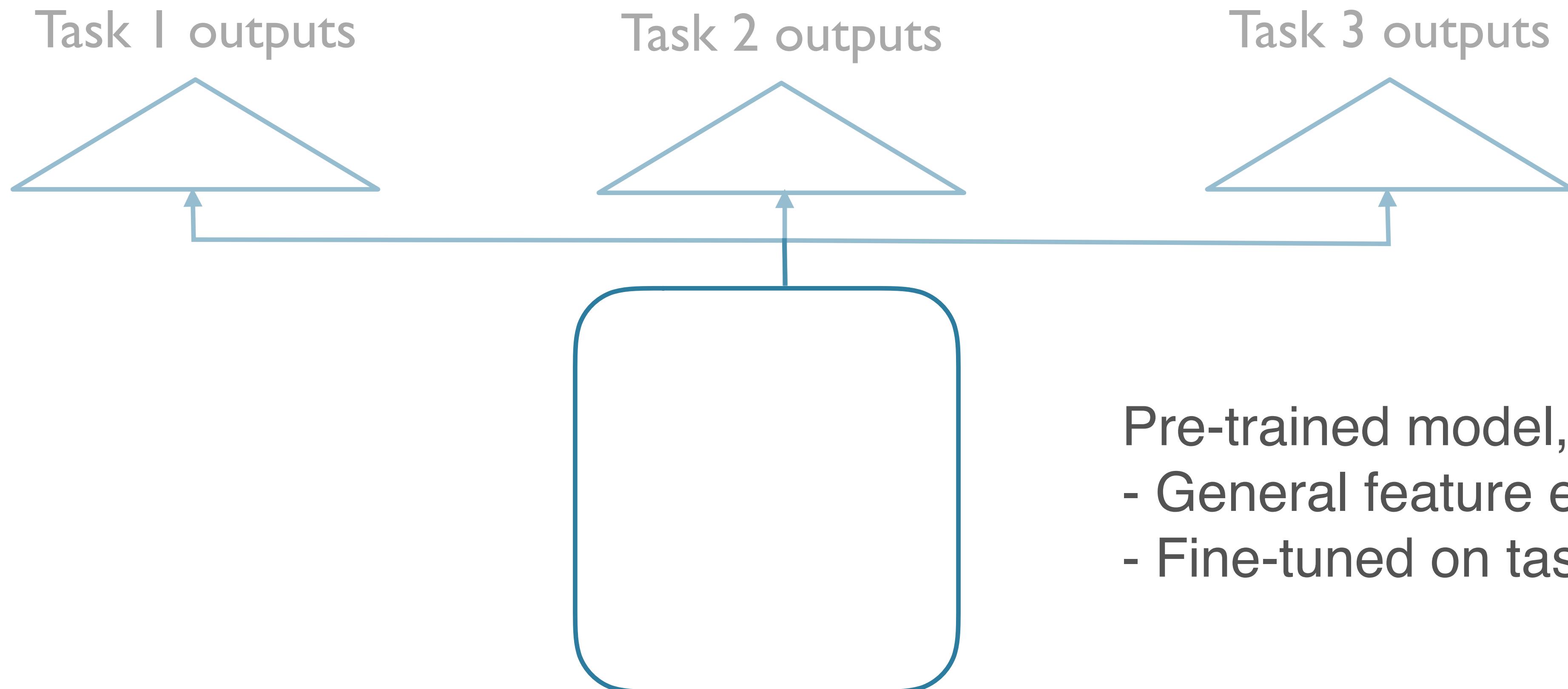
Transfer Learning



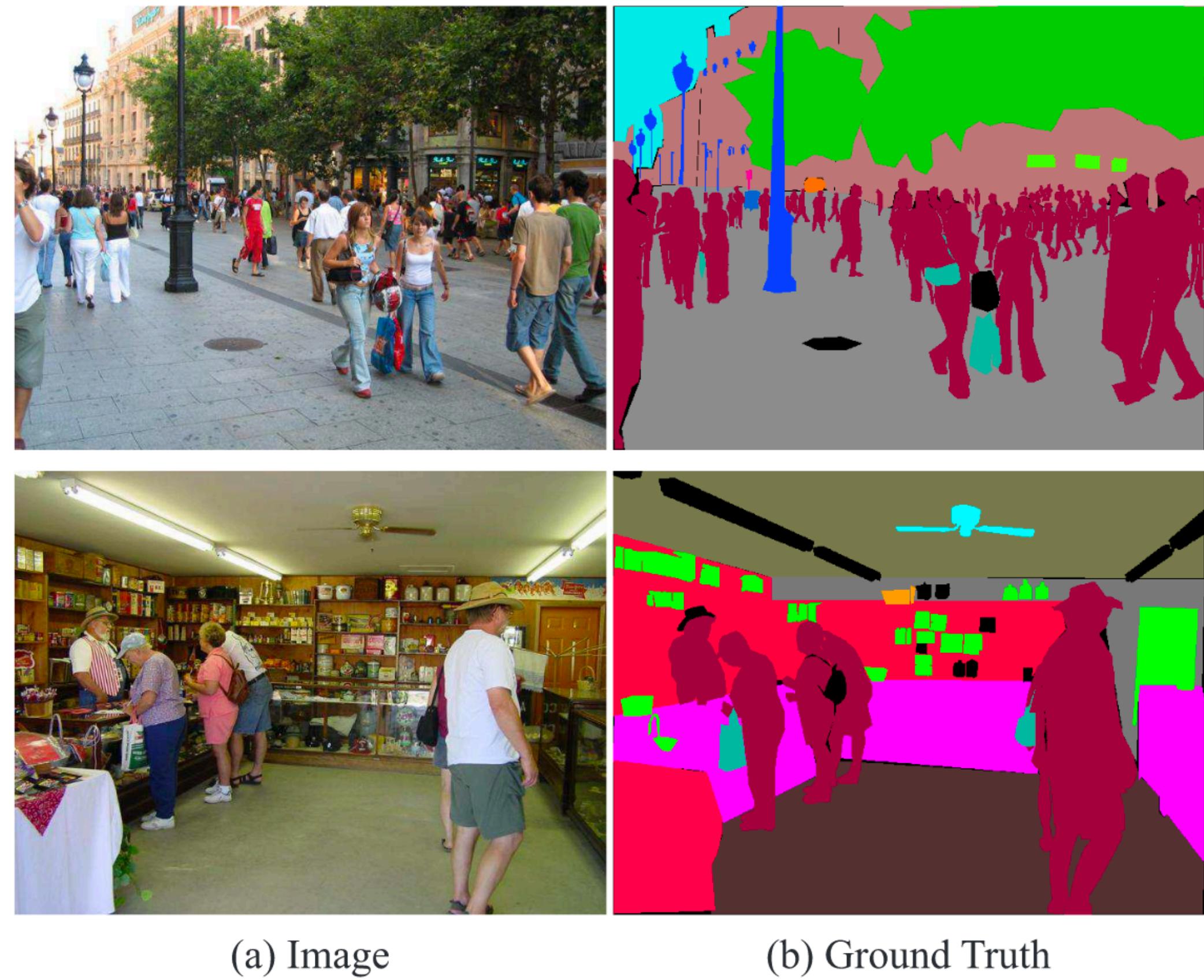
Transfer Learning



Transfer Learning



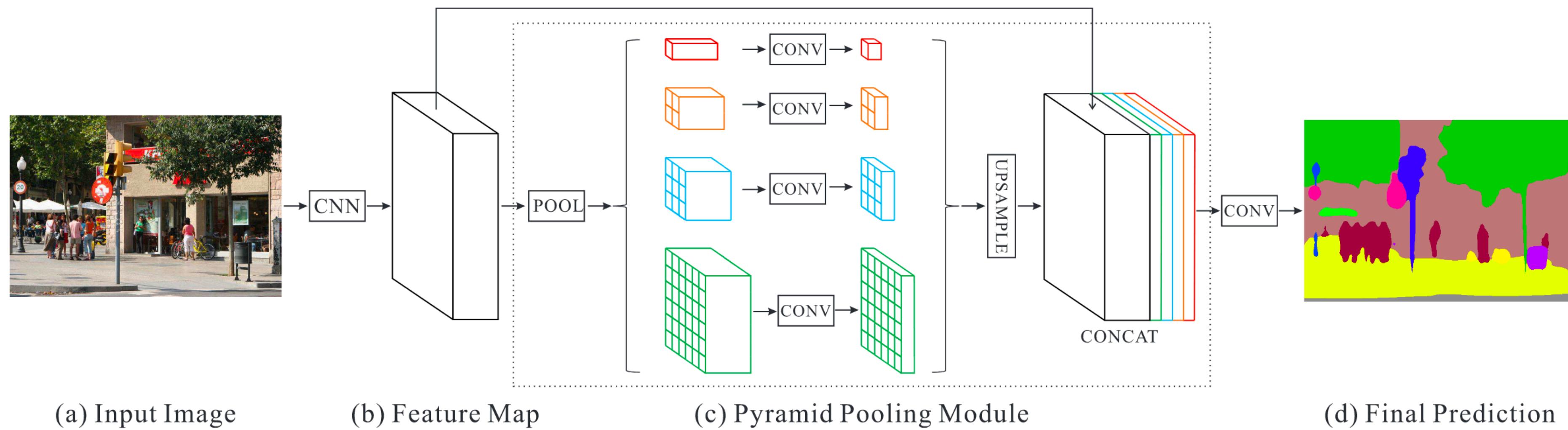
Example: Scene Parsing



(a) Image

(b) Ground Truth

Example: Scene Parsing



(a) Input Image

(b) Feature Map

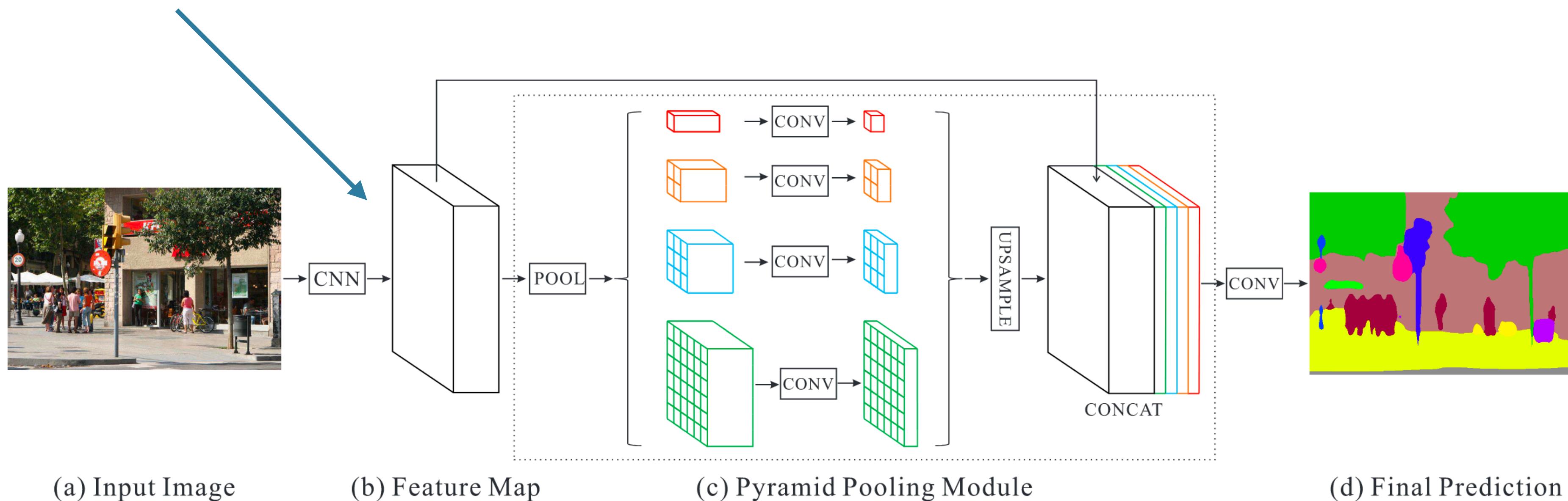
(c) Pyramid Pooling Module

(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)
 - [*: we will talk about variations later in the seminar]

Language Modeling

- Recent innovation: use *language modeling* (a.k.a. next word prediction)
 - [*: we will talk about variations later in the seminar]
- Linguistic knowledge:
 - The students were happy because _____ ...
 - The student was happy because _____ ...

Language Modeling

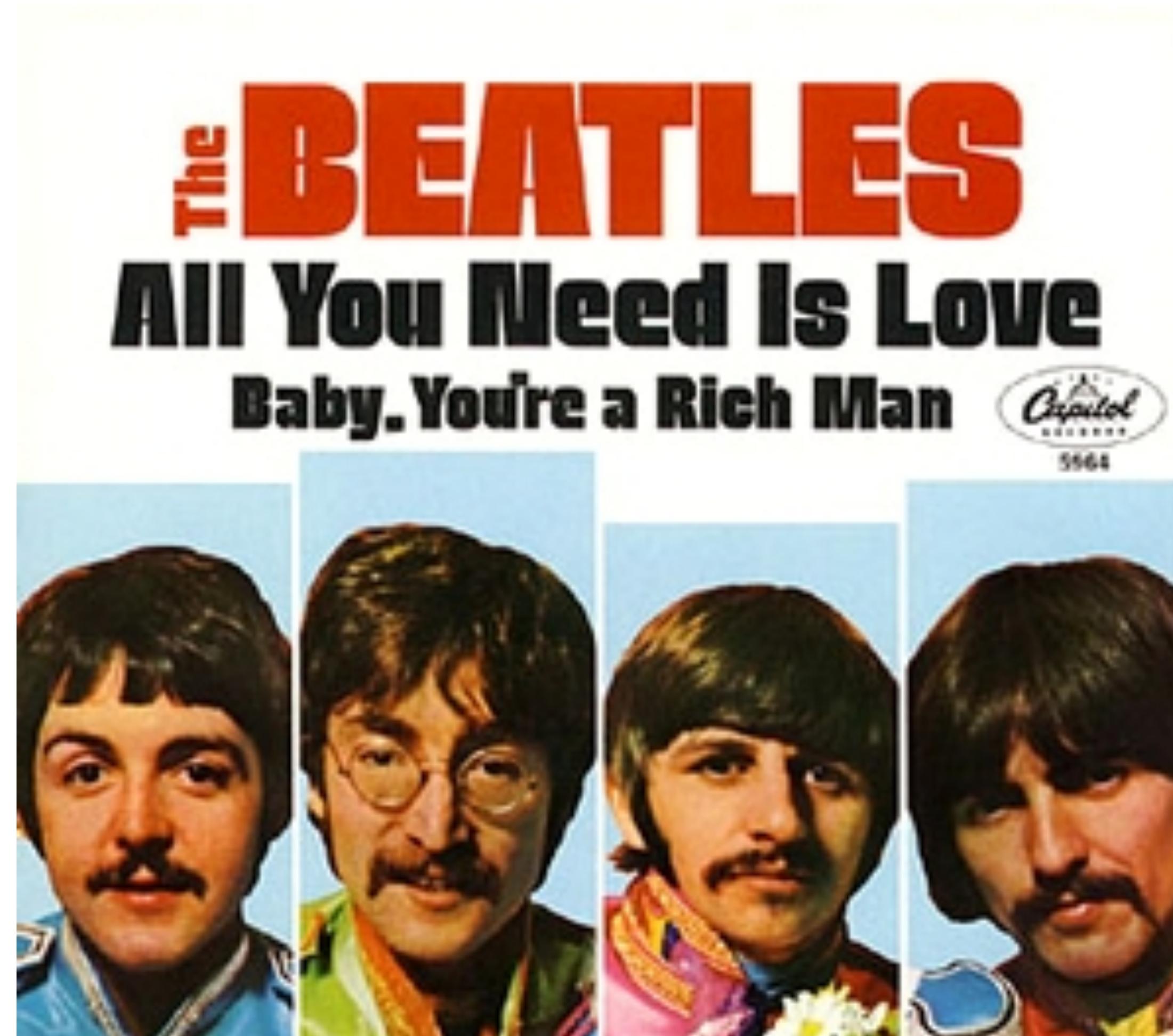
- Recent innovation: use *language modeling* (a.k.a. next word prediction)
 - [*: we will talk about variations later in the seminar]
- Linguistic knowledge:
 - The students were happy because _____ ...
 - The student was happy because _____ ...
- World knowledge:
 - The POTUS gave a speech after missiles 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 is abundant

- News sites (e.g. [Google 1B](#))
- Wikipedia (e.g. [WikiText103](#))
- Reddit
-
- General web crawling:
 - <https://commoncrawl.org/>

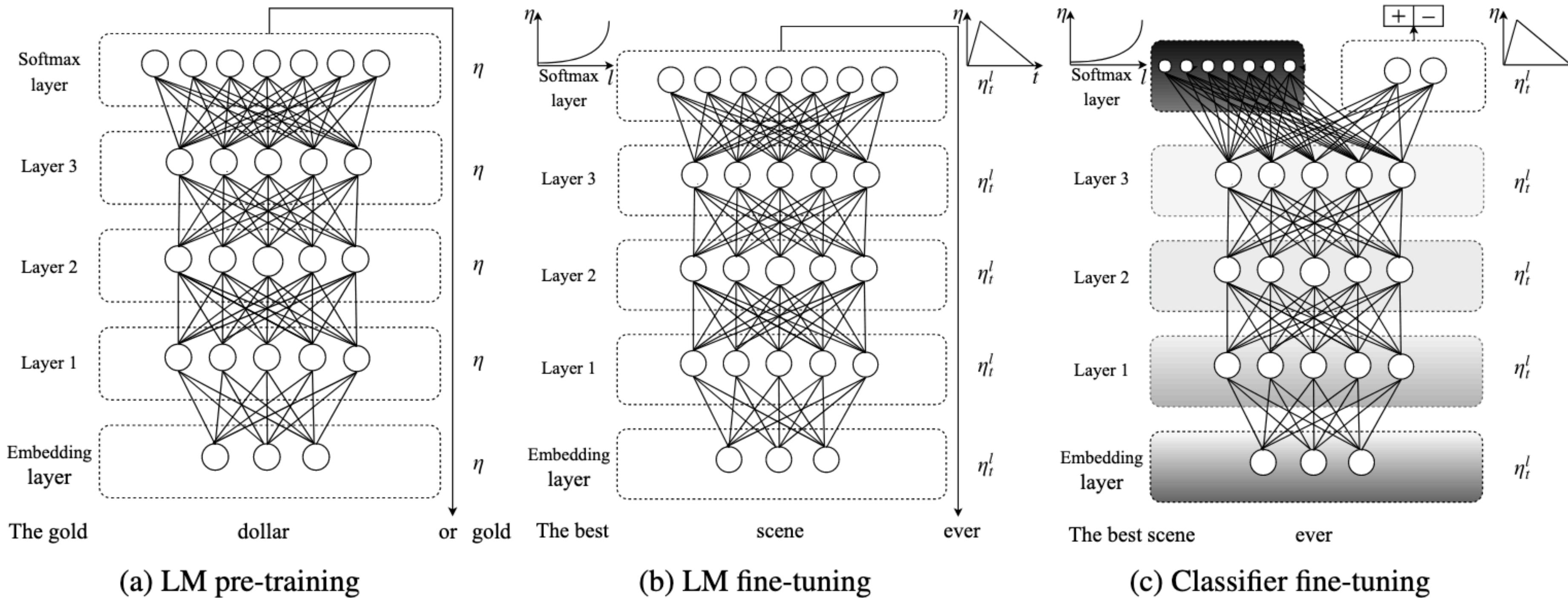
The Revolution will not be [Annotated]

Yann LeCun



<https://twitter.com/rblong/status/916062474545319938?lang=en>

ULMFiT

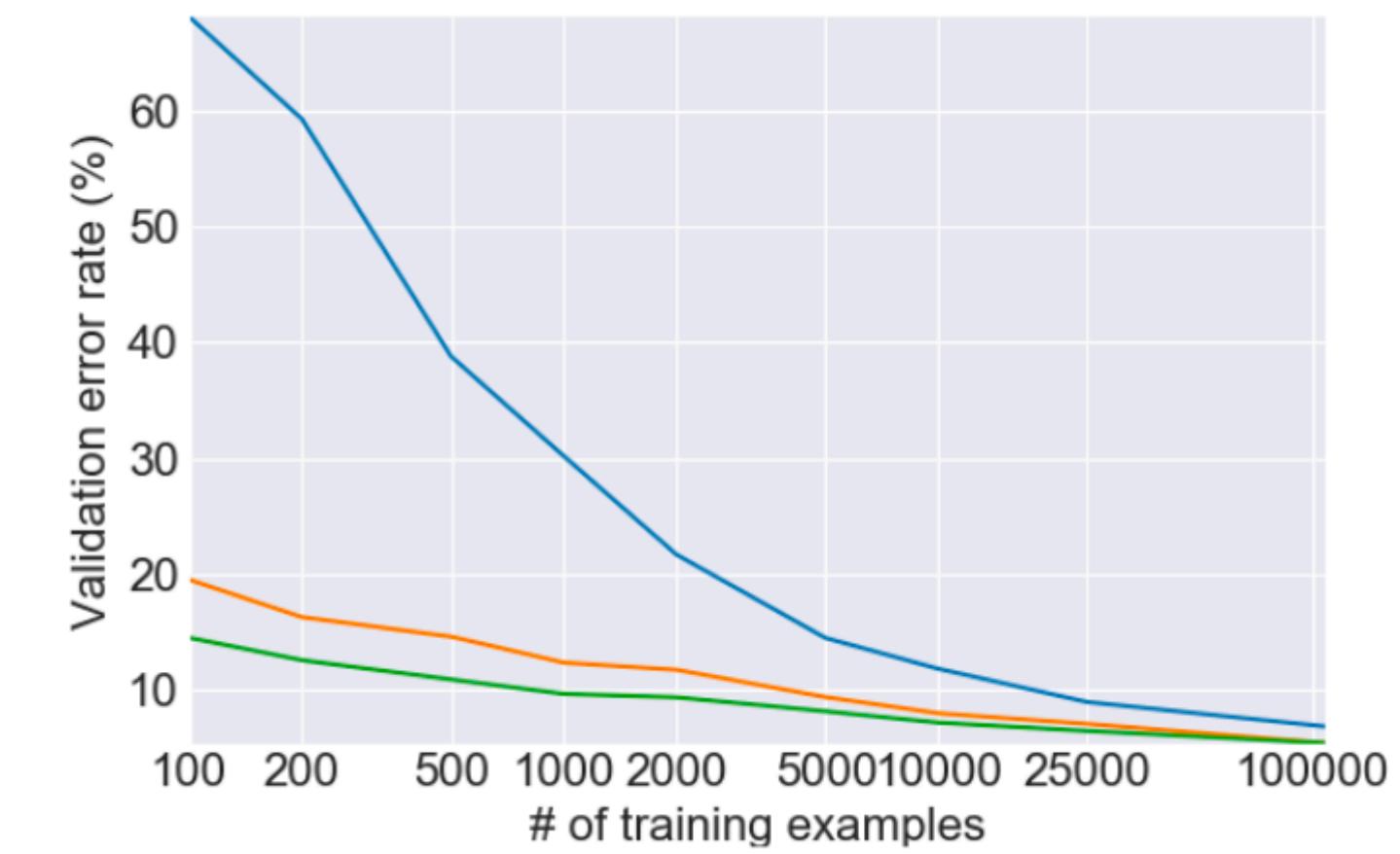
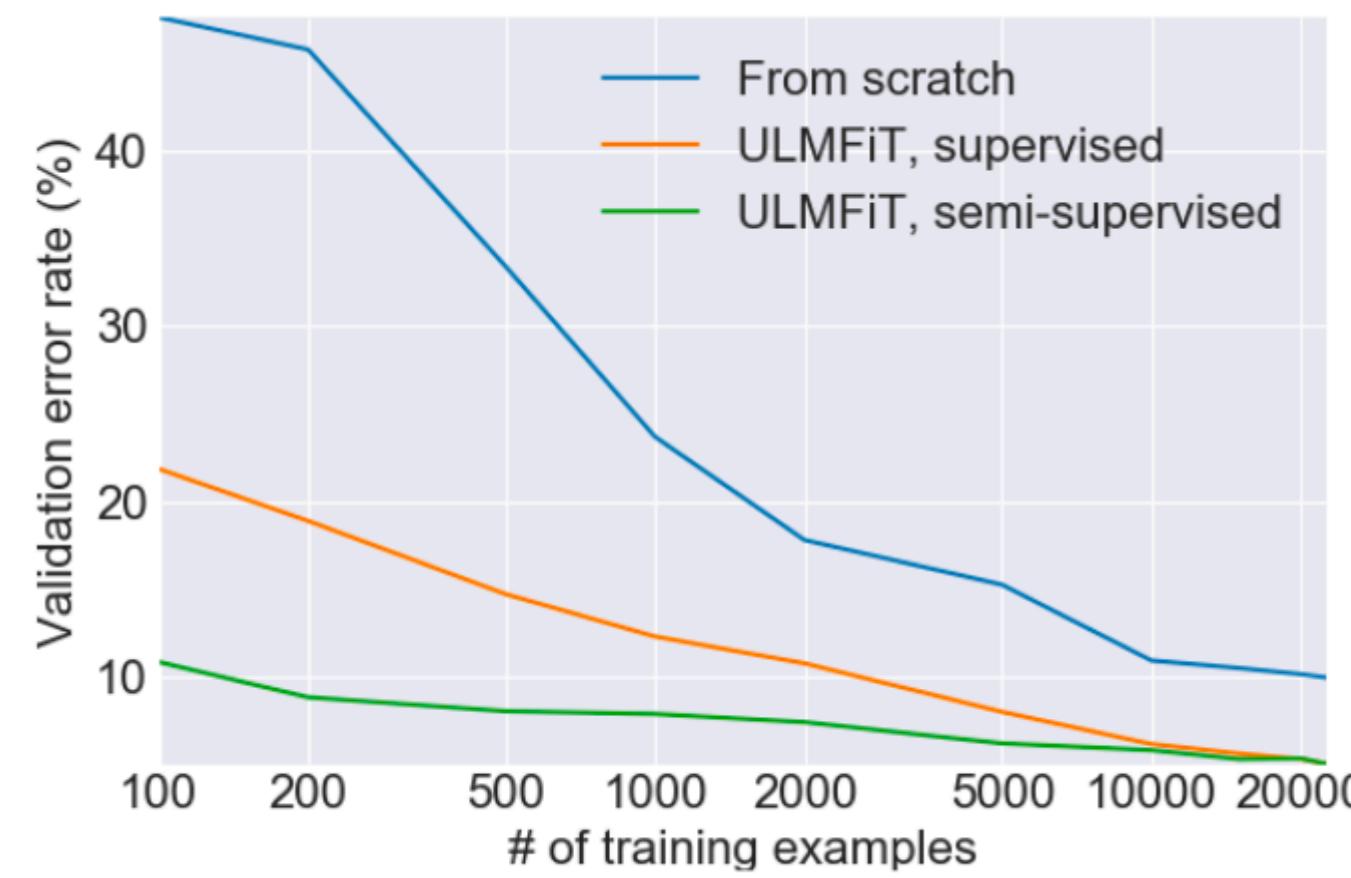


Universal Language Model Fine-tuning for Text Classification (ACL '18)

ULMFiT

| Model | Test | Model | Test | |
|-------|-----------------------------------|------------|------------------------------|------------|
| IMDb | CoVe (McCann et al., 2017) | 8.2 | 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



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
- Embeddings from Language Models (ELMo)
 - [aka the OG NLP Muppet]



Deep Contextualized Word Representations

Peters et. al (2018)

- Comparison to GloVe:

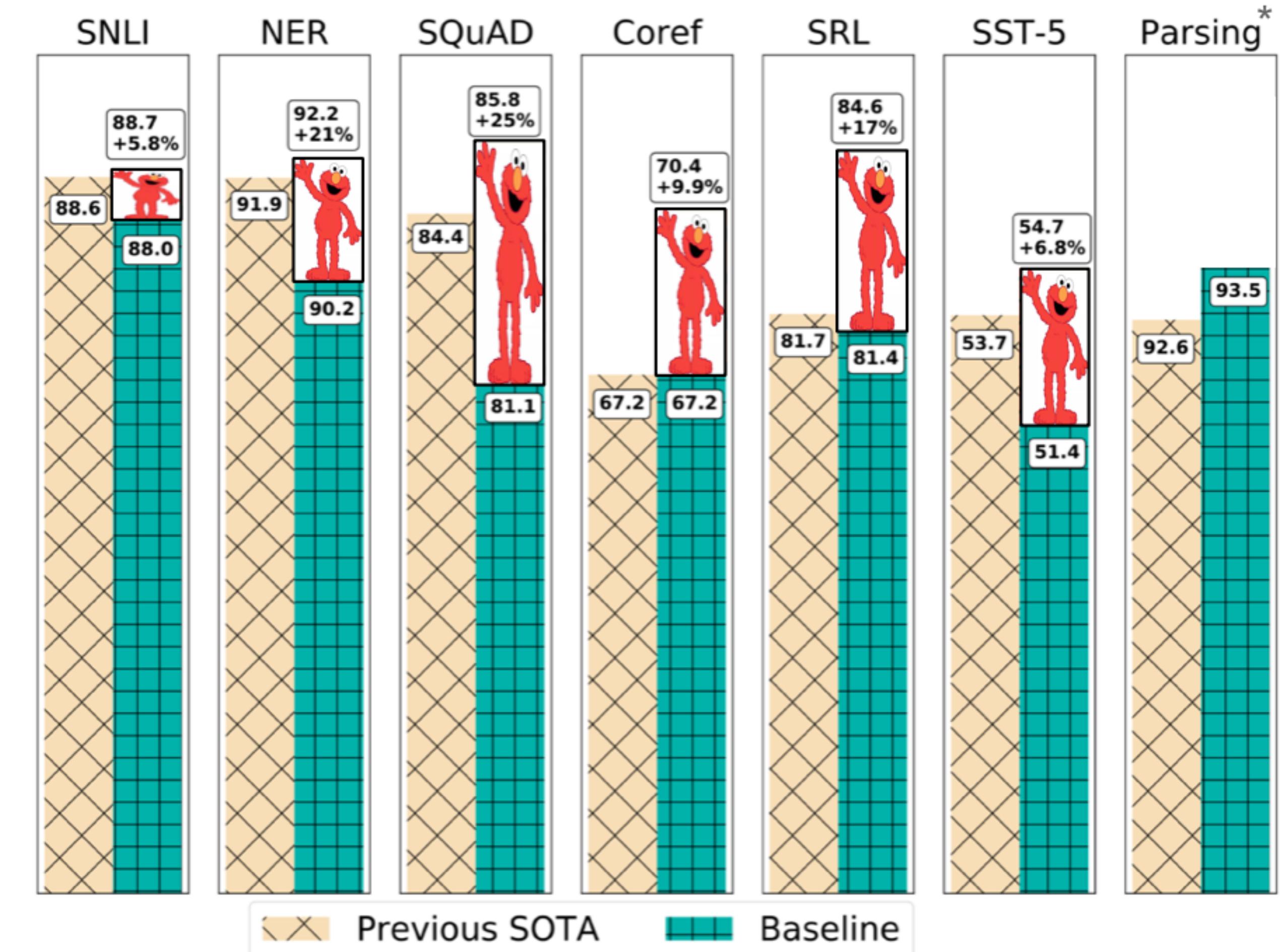
| | Source | Nearest Neighbors |
|-------|--|---|
| GloVe | play | playing, game, games, played, players, plays, player, Play, football, multiplayer |
| biLM | Chico Ruiz made a spectacular play on Alusik's grounder... Olivia De Havilland signed to do a Broadway play for Garson... | Kieffer, the only junior in the group, was commended for his ability to hit in the clutch, as well as his all-round excellent playthey 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:

SQuAD = [Stanford Question Answering Dataset](#)
SNLI = [Stanford Natural Language Inference Corpus](#)
SST-5 = [Stanford Sentiment Treebank](#)



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



BERT

Bidirectional Encoder Representations from Transformers

Devlin et al 2019

Initial Results

| System | MNLI-(m/mm) 392k | QQP 363k | QNLI 108k | SST-2 67k | CoLA 8.5k | STS-B 5.7k | MRPC 3.5k | RTE 2.5k | Average - |
|-----------------------|---------------------|-------------|--------------|--------------|--------------|---------------|--------------|-------------|--------------|
| 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 |
| BERT _{BASE} | 84.6/83.4 | 71.2 | 90.5 | 93.5 | 52.1 | 85.8 | 88.9 | 66.4 | 79.6 |
| BERT _{LARGE} | 86.7/85.9 | 72.1 | 92.7 | 94.9 | 60.5 | 86.5 | 89.3 | 70.1 | 82.1 |

Major Application



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

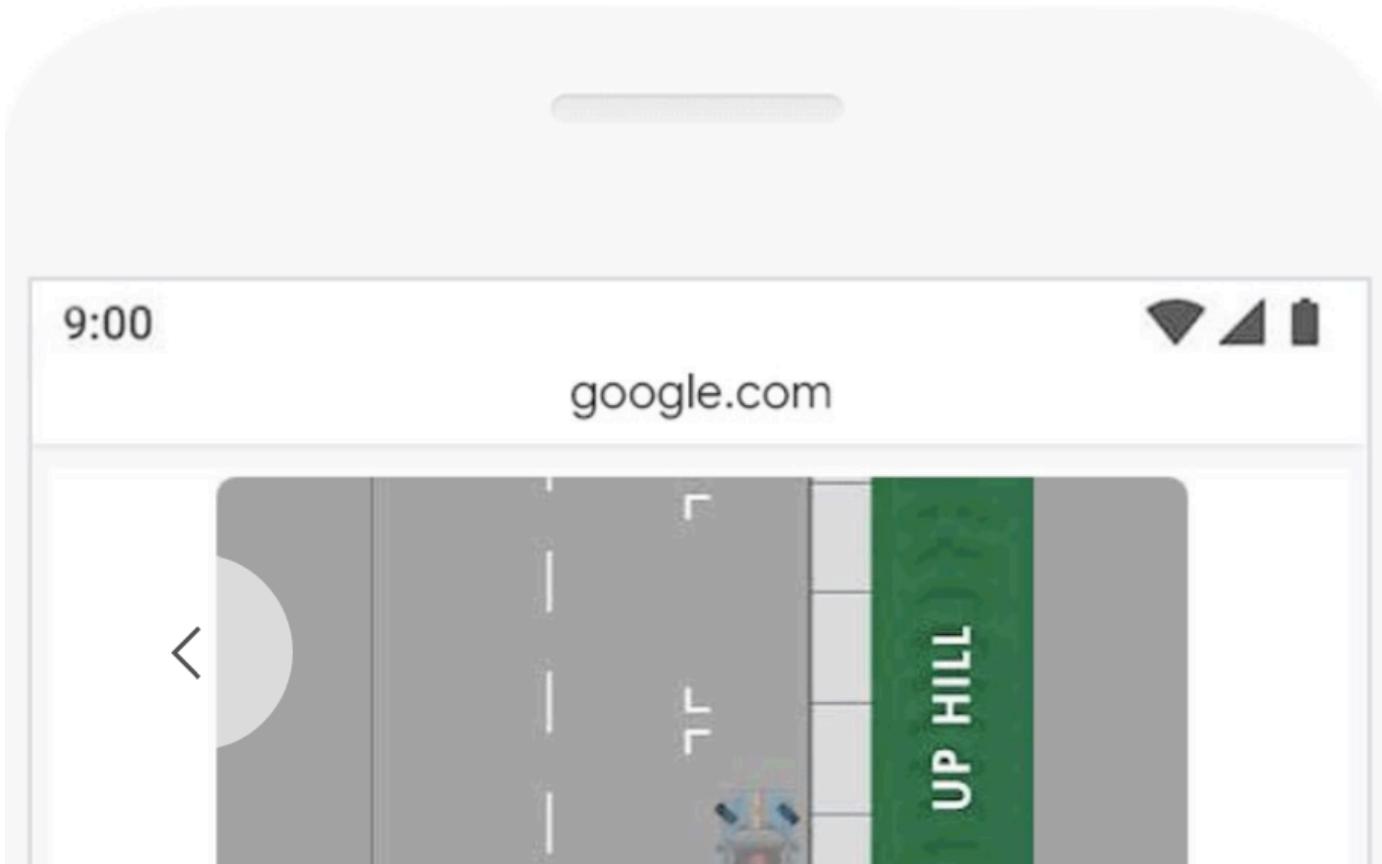
If there's one thing I've learned over the 15 years working on Google Search, it's that people's curiosity is endless. We see billions of searches every day, and 15 percent of those queries are ones we haven't seen before--so we've built ways to return results for queries we can't anticipate.

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

Major Application

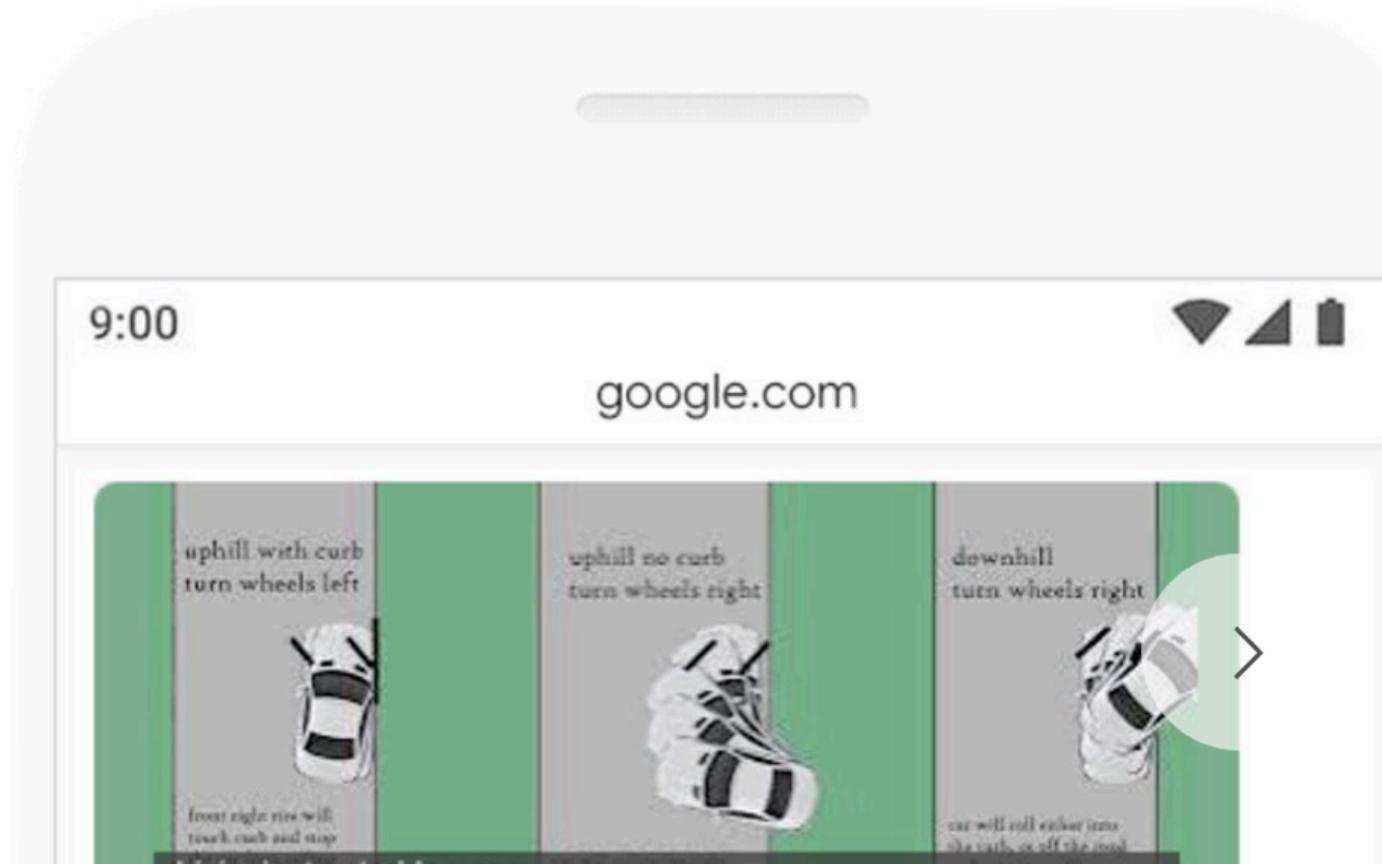
 parking on a hill with no curb

BEFORE



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



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 & GATECHMT-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](#)

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

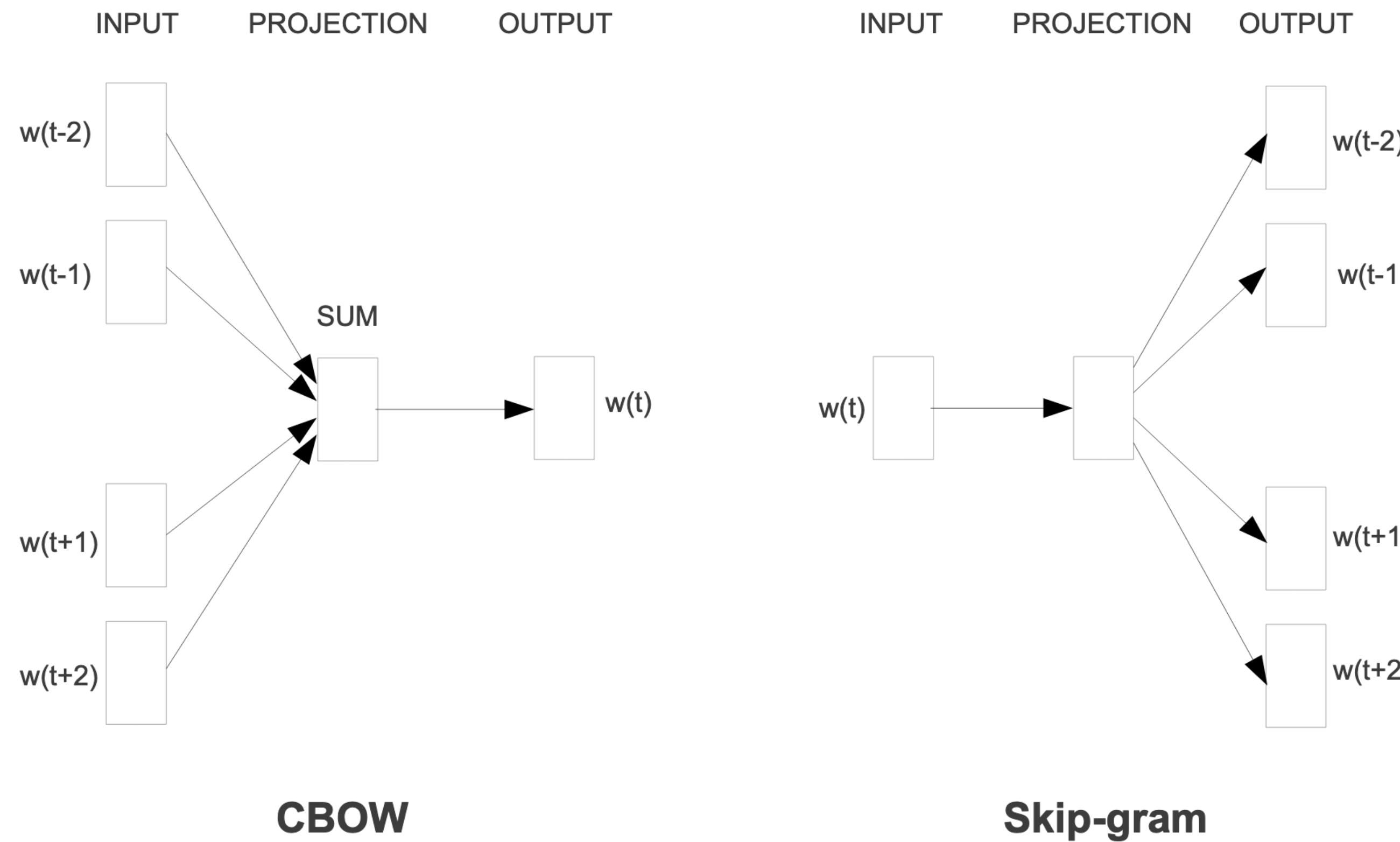
- One distinction:
 - *Global* representations:
 - word2vec, GloVe: *one* vector for each word *type* (e.g. ‘play’)
- *Contextual* representations (from LMs):
 - Representation of word in context, not independently

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

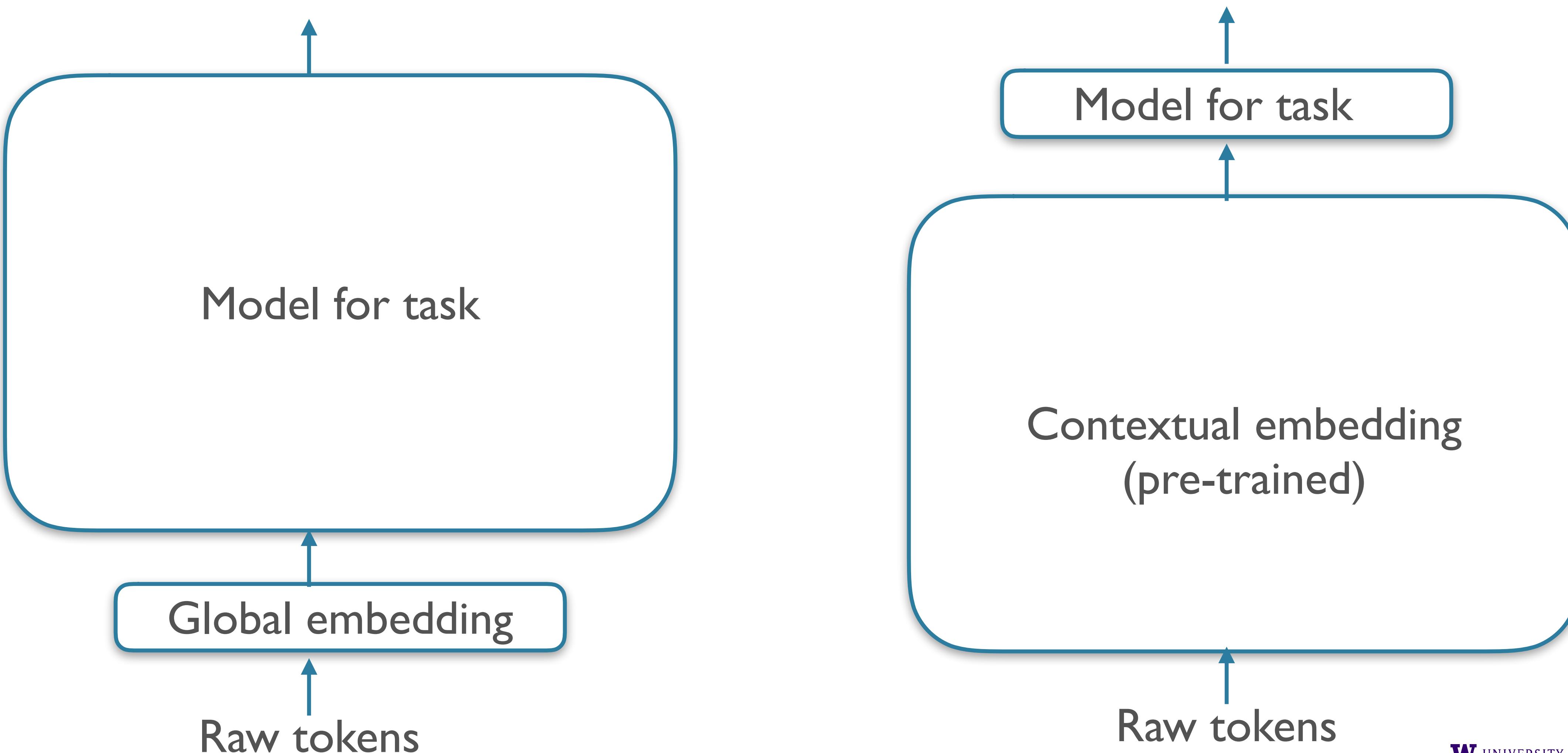
Global Embeddings: Models

Global Embeddings: Models



Mikolov et al 2013a (the OG word2vec paper)

Shallow vs Deep Pre-training



NLP's “Clever Hans Moment”

∇ The Gradient

ME EDITOR'S NOTE OVERVIEWS PERSPECTIVES ABOUT SUBSCRIBE Q

Clever Hans

BERT



NLP's Clever Hans
Moment has Arrived

26.AUG.2019

[link](#)

Clever Hans

- Early 1900s, a horse trained by his owner to do:
 - Addition
 - Division
 - Multiplication
 - Tell time
 - Read German
 - ...
- Wow! Hans is really smart!

Clever Hans Effect

Clever Hans Effect

- Upon closer examination / experimentation...

Clever Hans Effect

- Upon closer examination / experimentation...
- Hans' success:

Clever Hans Effect

- Upon closer examination / experimentation...
- Hans' success:
 - 89% when questioner knows answer

Clever Hans Effect

- Upon closer examination / experimentation...
- Hans' success:
 - 89% when questioner knows answer
 - 6% when questioner doesn't know answer

Clever Hans Effect

- Upon closer examination / experimentation...
- Hans' success:
 - 89% when questioner knows answer
 - 6% when questioner doesn't know answer
- Further experiments: as Hans' taps got closer to correct answer, facial tension in questioner increased

Clever Hans Effect

- Upon closer examination / experimentation...
- Hans' success:
 - 89% when questioner knows answer
 - 6% when questioner doesn't know answer
- Further experiments: as Hans' taps got closer to correct answer, facial tension in questioner increased
- Hans didn't solve the task but exploited *a spuriously correlated cue*

Central question

- Do BERT et al's major successes at solving NLP tasks show that we have achieved robust natural language understanding in machines?
- Or: are we seeing a “Clever BERT” phenomenon?

Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference

R. Thomas McCoy,¹ Ellie Pavlick,² & Tal Linzen¹

¹Department of Cognitive Science, Johns Hopkins University

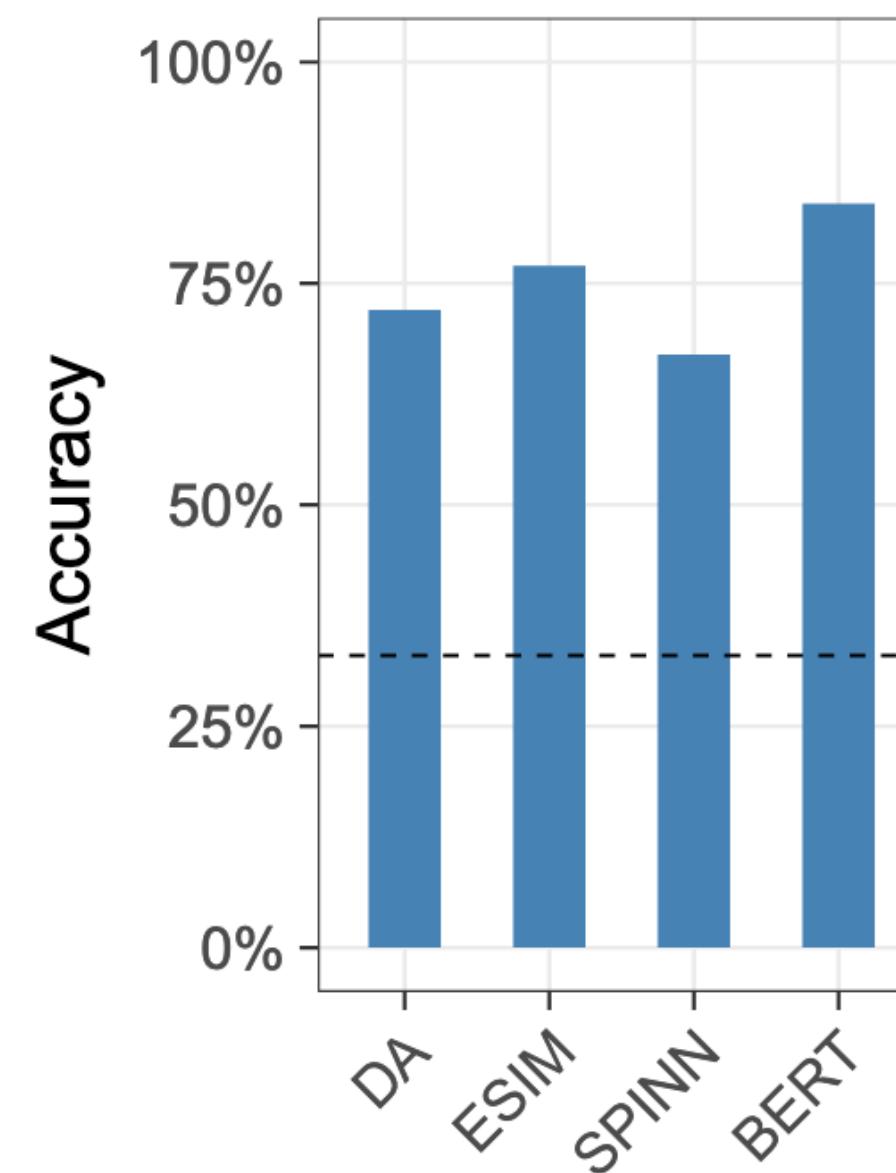
²Department of Computer Science, Brown University

tom.mccoy@jhu.edu, ellie.pavlick@brown.edu, tal.linzen@jhu.edu

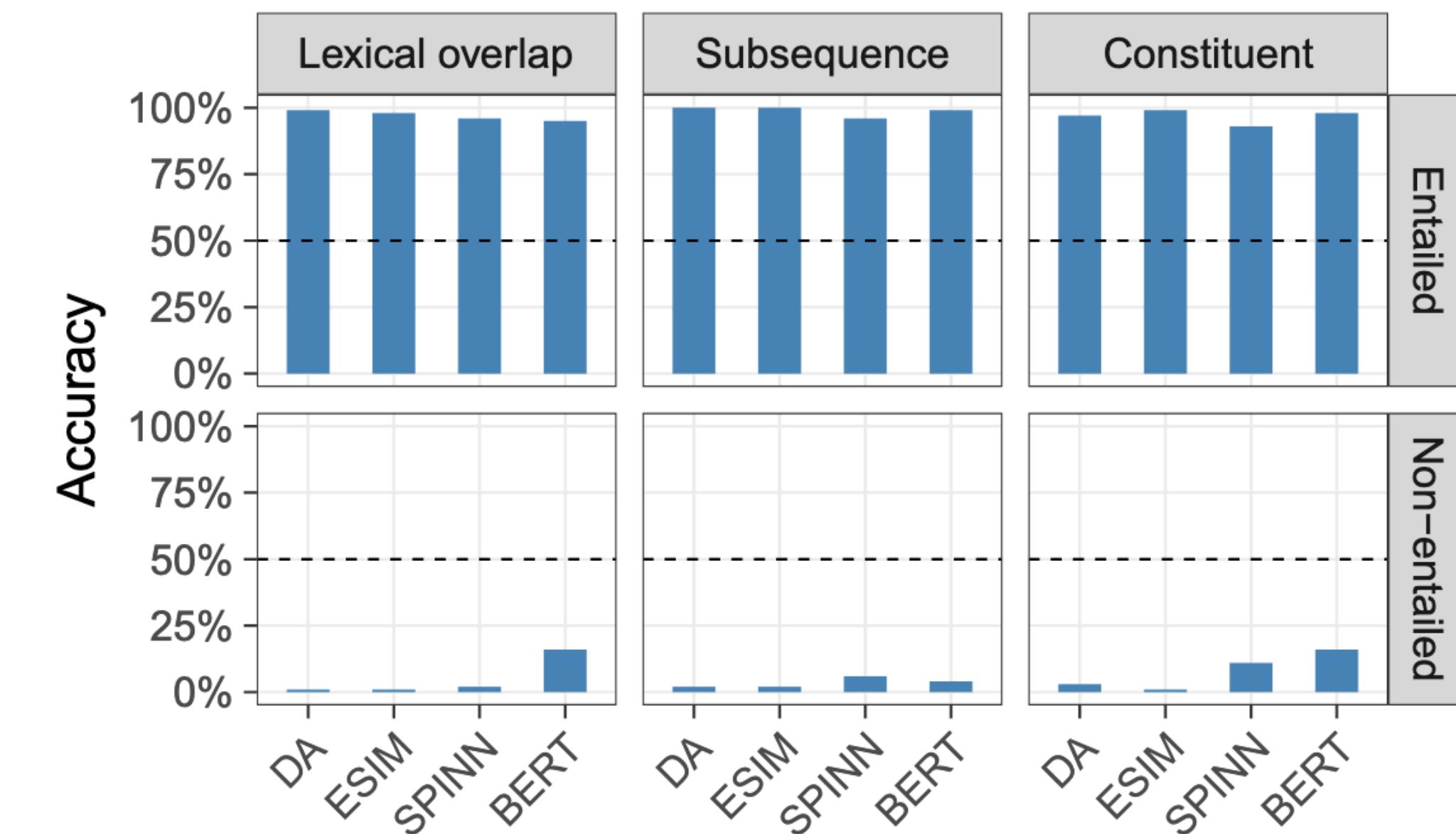
McCoy et al 2019

| Heuristic | Premise | Hypothesis | Label |
|---------------------------------|---|---------------------------------|-------|
| Lexical overlap heuristic | The banker near the judge saw the actor. | The banker saw the actor. | E |
| | The lawyer was advised by the actor. | The actor advised the lawyer. | E |
| | The doctors visited the lawyer. | The lawyer visited the doctors. | N |
| | The judge by the actor stopped the banker. | The banker stopped the actor. | N |
| Subsequence heuristic | The artist and the student called the judge. | The student called the judge. | E |
| | Angry tourists helped the lawyer. | Tourists helped the lawyer. | E |
| | The judges heard the actors resigned. | The judges heard the actors. | N |
| | The senator near the lawyer danced. | The lawyer danced. | N |
| Constituent heuristic | Before the actor slept, the senator ran. | The actor slept. | E |
| | The lawyer knew that the judges shouted. | The judges shouted. | E |
| | If the actor slept, the judge saw the artist. | The actor slept. | N |
| | The lawyers resigned, or the artist slept. | The artist slept. | N |

Results



(a)



(b)

(performance improves if fine-tuned on this challenge set)

Probing Neural Network Comprehension of Natural Language Arguments

Timothy Niven and Hung-Yu Kao

Intelligent Knowledge Management Lab
Department of Computer Science and Information Engineering
National Cheng Kung University
Tainan, Taiwan

tim.niven.public@gmail.com, hykao@mail.ncku.edu.tw

Abstract

We are surprised to find that BERT’s peak performance of 77% on the Argument Reasoning Comprehension Task reaches just three points below the average untrained human baseline. However, we show that this result is entirely accounted for by exploitation of spurious statistical cues in the dataset. We analyze the nature of these cues and demonstrate that a range of models all exploit them. This analysis informs the construction of an adversarial dataset on which all models achieve random accuracy. Our adversarial dataset provides a

| | |
|--------------------|---|
| Claim | Google is not a harmful monopoly |
| Reason | People can choose not to use Google |
| Warrant | Other search engines don’t redirect to Google |
| Alternative | All other search engines redirect to Google |

Reason (and since) Warrant \rightarrow Claim
Reason (but since) Alternative $\rightarrow \neg$ Claim

Figure 1: An example of a data point from the ARCT test set and how it should be read. The inference from R and A to $\neg C$ is by design.

The Argument Reasoning Comprehension Task (ARCT) (Habernal et al., 2018a) defers the problem of disengaging arguments and focuses on in-

[link](#)

Recent Analysis Explosion

- E.g. BlackboxNLP workshop [[2018](#), [2019](#)]
- New “Interpretability and Analysis” track at ACL

Why care?

- Effects of learning what neural language models understand:
 - Engineering: can help build better language technologies via improved models, data, training protocols, ...
 - Trust, critical applications
 - Theoretical: can help us understand biases in different architectures (e.g. LSTMs vs Transformers), similarities to human learning biases
 - Ethical: e.g. do some models reflect problematic social biases more than others?

Stretch Break!

Course Overview / Logistics

Large Scale

- Motivating question: what do neural language models understand about natural language?
 - Focus on *meaning*, where much of the literature has focused on *syntax*
- A *research seminar*: in groups, you will carry out and execute a novel analysis project.
 - Think of it as a proto-conference-paper, or the seed of a conference paper.

Course structure

- First half: learning about the tools and techniques required
 - Wk 2: language models
[architectures, tasks, data, ...]
 - Wk 3: analysis methods
[visualization, probing classifiers, artificial data, ...]
 - Wk 4: resources / datasets
[guest lecture by Rachel Rudinger]
 - Wk 5: technical resources / writing tips
- Be active! Reading, participating, planning ahead

Course structure

- Second half: *presentations*
 - Each group will give one “special topic” presentation and lead a discussion, e.g.:
 - reading a paper or two on a topic related to your final project
 - explaining a method you are using in project, issues, etc.
 - Final week: project presentation festival!
 - “Mini conference”, incl. reception

Evaluation

- Proposal: 10%
- Special topic presentation: 20%
- Final project presentation: 10%
- Final paper: 50%
- Participation: 10%

Reading List

- Semi- but not fully comprehensive list of recent papers on website
- Browse, get ideas/inspiration
- Deep dive on a few later
- NB: I'll add key-words tomorrow for sorting

Group Formation (HW1)

Three Tasks

- Form groups (more next)
- Set up repository
 - GitHub, GitLab, patas Git server ...
 - Make it private for now!
 - Don't put private or sensitive data in the repo! (incl LDC corpora)
- Add ACL paper template to repository
 - <https://acl2020.org/calls/papers/#paper-submission-and-templates>
 - Format for final paper

Groups

- There will be *eight* groups
 - Sized 4-6 people
- Unified grade
- Group decides how to divide work, but reports who did what at the end.
- Aim to diversify talents / interests in the group.
 - Experimental design
 - Data work
 - Implementation
 - Experiment running / analyzing
 - Writing
 - Speaking (presentations)

Communication

- CLMS Student Slack
 - Useful, since a majority of students in this seminar are on it already
 - Self-organize (575 channel?), based on interests, background competences, etc
 - For students not on it yet:
 - Canvas thread for requesting access
 - CLMS students: please add ASAP
- For general / non-group discussions, still use Canvas discussions.
- NB: I am not on that Slack (nor are other faculty)

Registering Groups

- List your groups here:
 - https://docs.google.com/spreadsheets/d/1ziyww5J49iQ7iE8ElgzMX6rl2cR_cbcKxK89pvZDPF4/edit?usp=sharing
- On Canvas, upload “readme.pdf” with:
 - Group #, screenshot of repository

Thanks! Looking forward to a great quarter!