Audio Processing, Internet-Scale Training Data, and Scaling Laws

Danna Gurari
University of Colorado Boulder
Spring 2025

https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2025/

Review

- Last lecture
  - Multimodal problems
  - Image captioning: pioneering dataset and model
  - Visual question answering: pioneering dataset and model
  - LXMERT: multimodal representations
- Assignments (Canvas)
  - Lab assignment 2 grades are out
    - Email all regrade requests to our TA, Nick Cooper (a comment in Canvas is not sufficient)
  - Problem set 4 due (final one) earlier today
  - Lab assignment 3 (last one!) due in 1.5 weeks
- Questions?

Today’s Topics

- Other data modalities
- Internet-scale trained models
- Scaling laws
- Lab assignment 3: multimodal task
- Programming tutorial

Today’s Topics

- Other data modalities
- Internet-scale trained models
- Scaling laws
- Lab assignment 3: multimodal task
- Programming tutorial

Recall: Data Types We Have Focused On

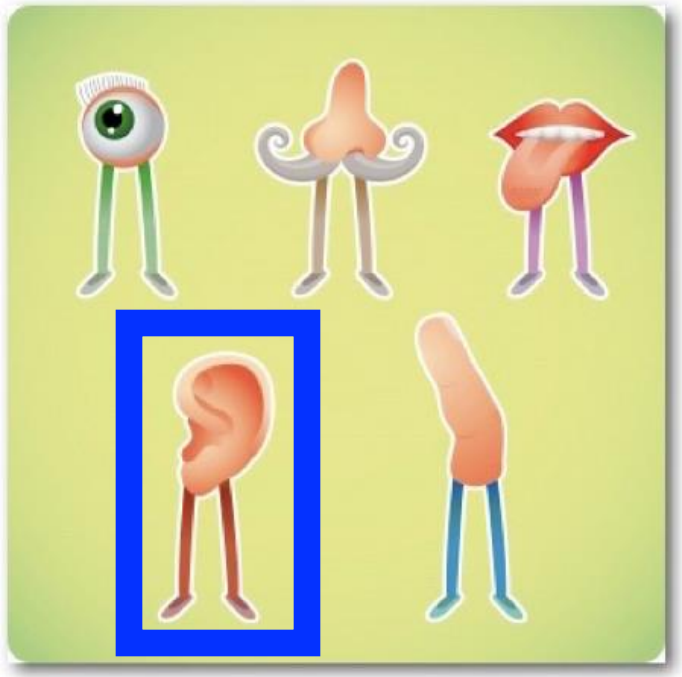
Images

157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

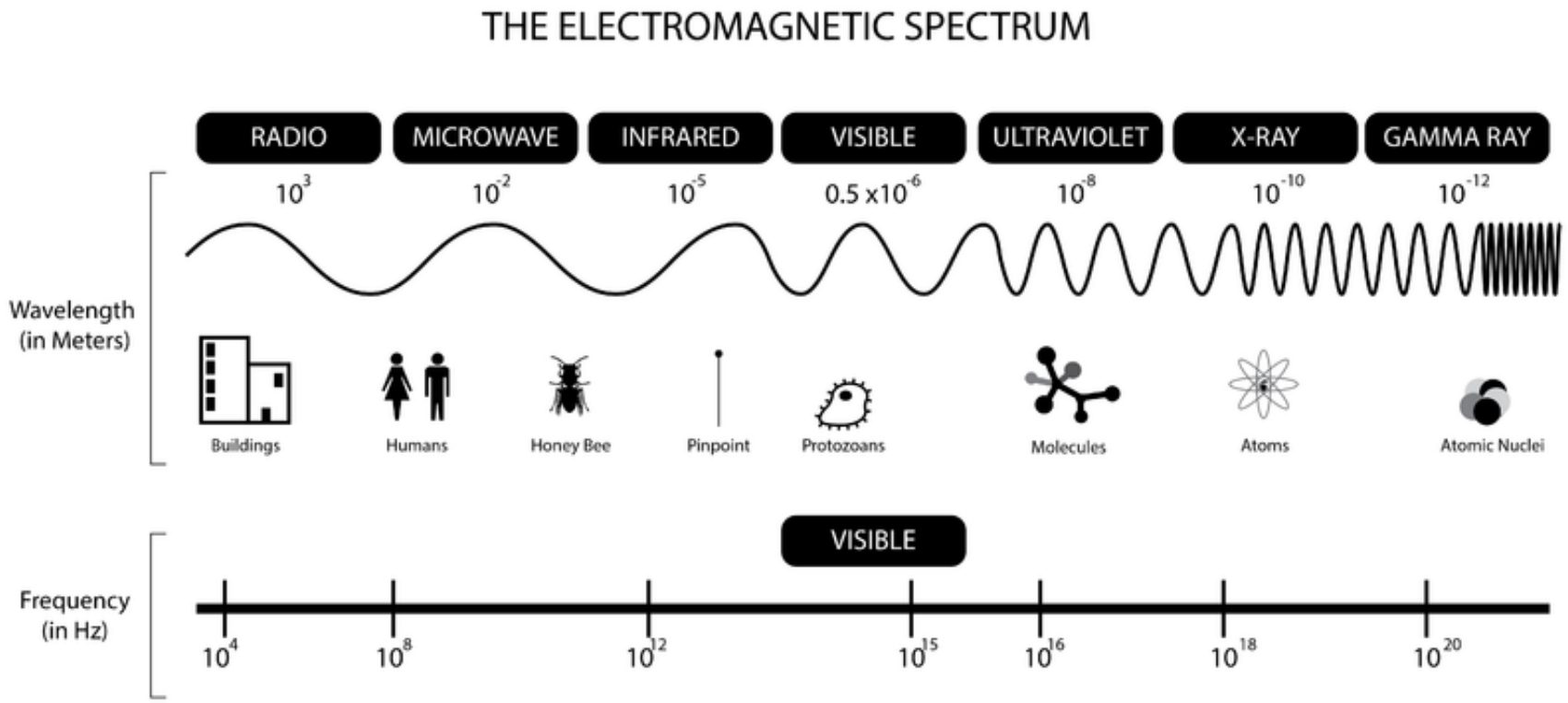
Text

Many Data Modalities Can Be Explored!

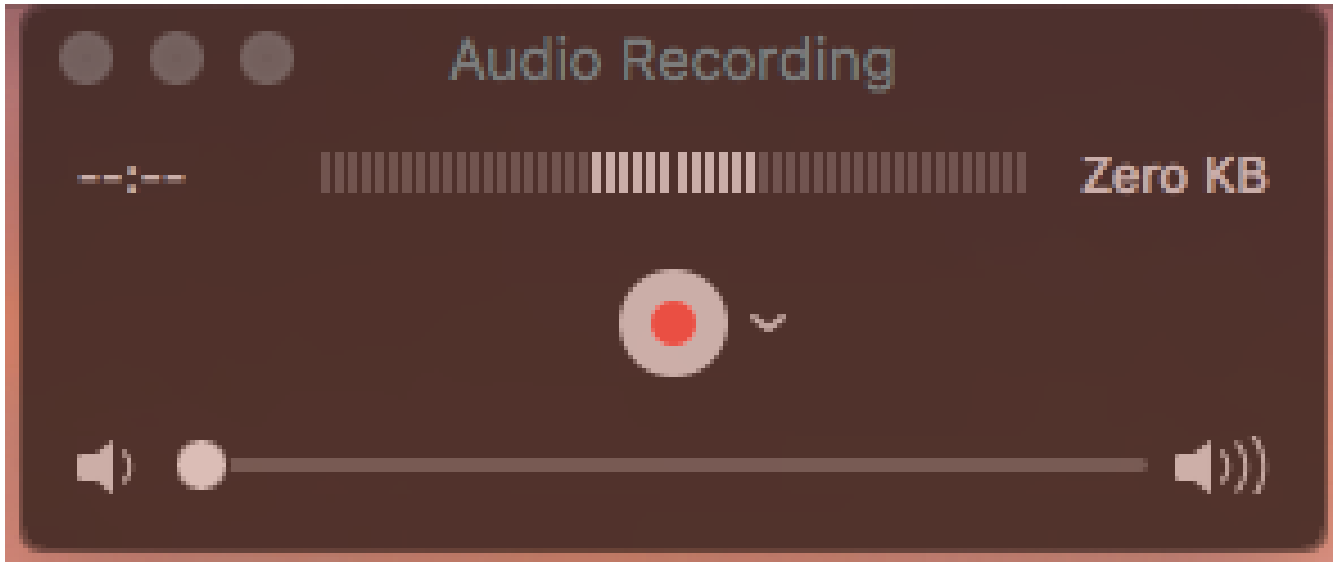
e.g., All Human Senses!



e.g., Beyond Human Senses!



Another Popular Modality: Audio

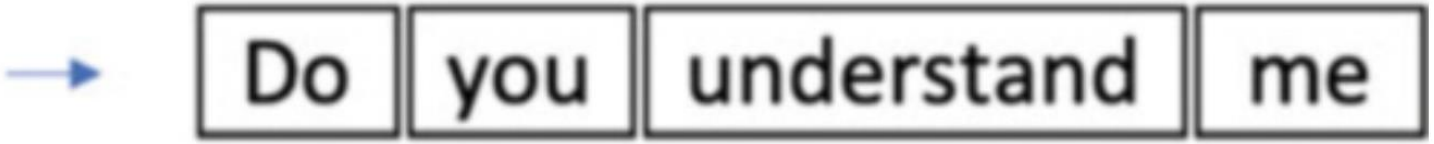


Speech Processing: Problem Definition

Input: spoken language

Output: machine readable text

Raw Speech Signal



Transcription

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

What Is Speech?

Raw Speech Signal

Compression waves created by pushing air from one's lungs and modulating it using one's tongue, teeth, and lips

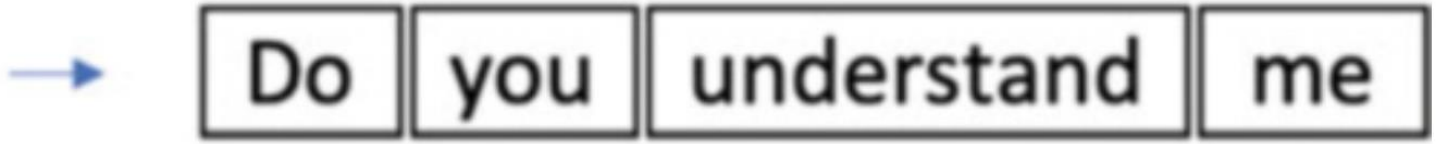
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Why Is Speech Processing Challenging?

Input can be diverse
including different accents,
volumes, pace, and cadence

Temporal data needs to be
segmented into distinct words

Raw Speech Signal



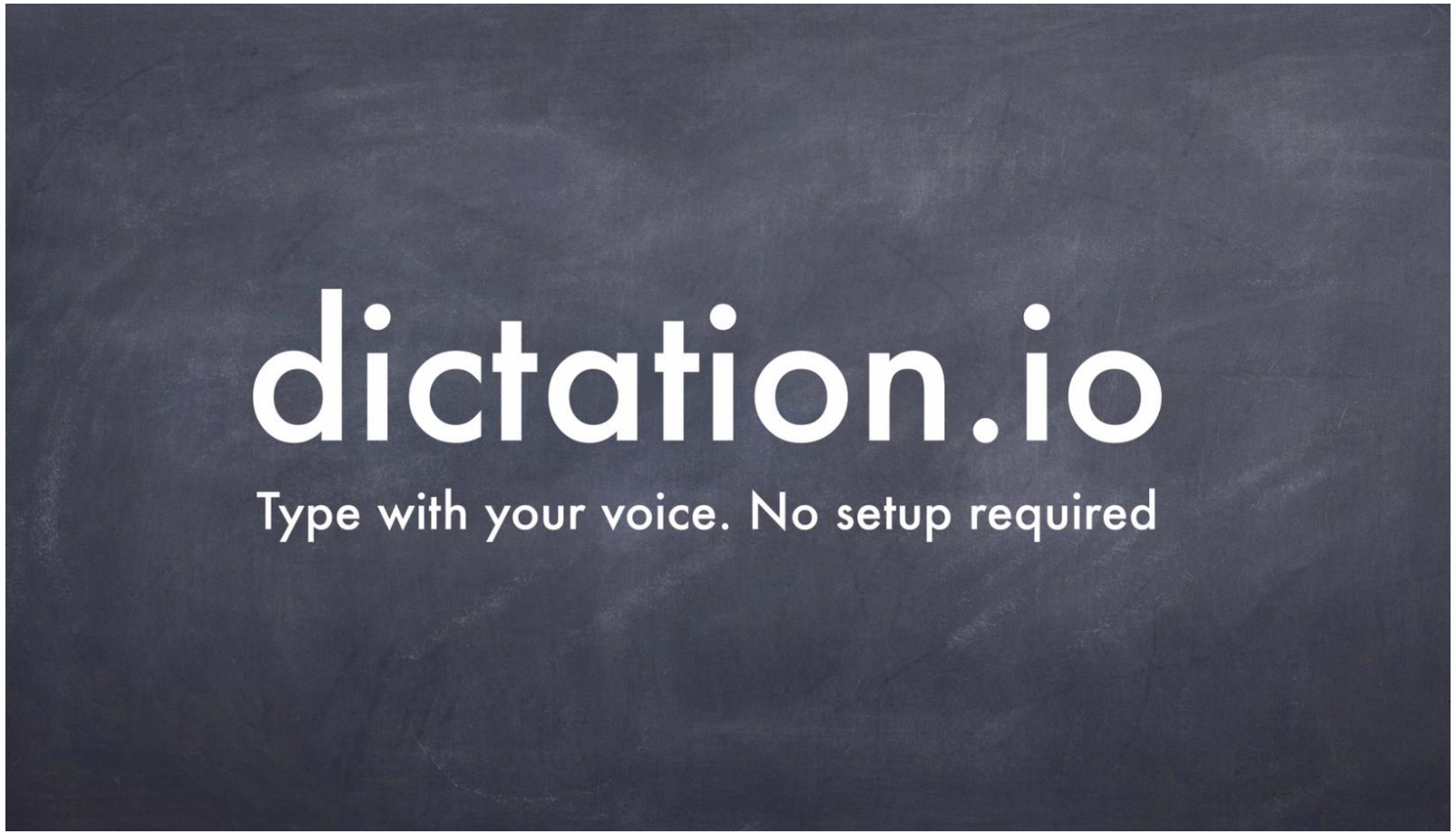
Transcription

Technology can introduce artifacts
including varying quality, echos,
and background noise

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

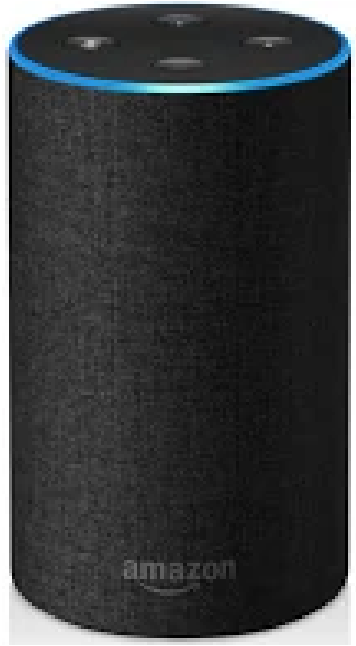
Application: Voice Typing on Mobile Devices

Application: Voice Typing for Productivity Apps



Demo starting at 2:00: https://www.youtube.com/watch?v=5UK4vLzU9co&t=76s

Application: Virtual Assistant



e.g., Amazon's Echo with Alexa

e.g., Google Home

e.g., Baidu DuerOS

https://dueros.baidu.com/en/html/dueros/index.html

Application: Audio Transcription (e.g., Analysis & Situational/Permanent Hearing Impairments)

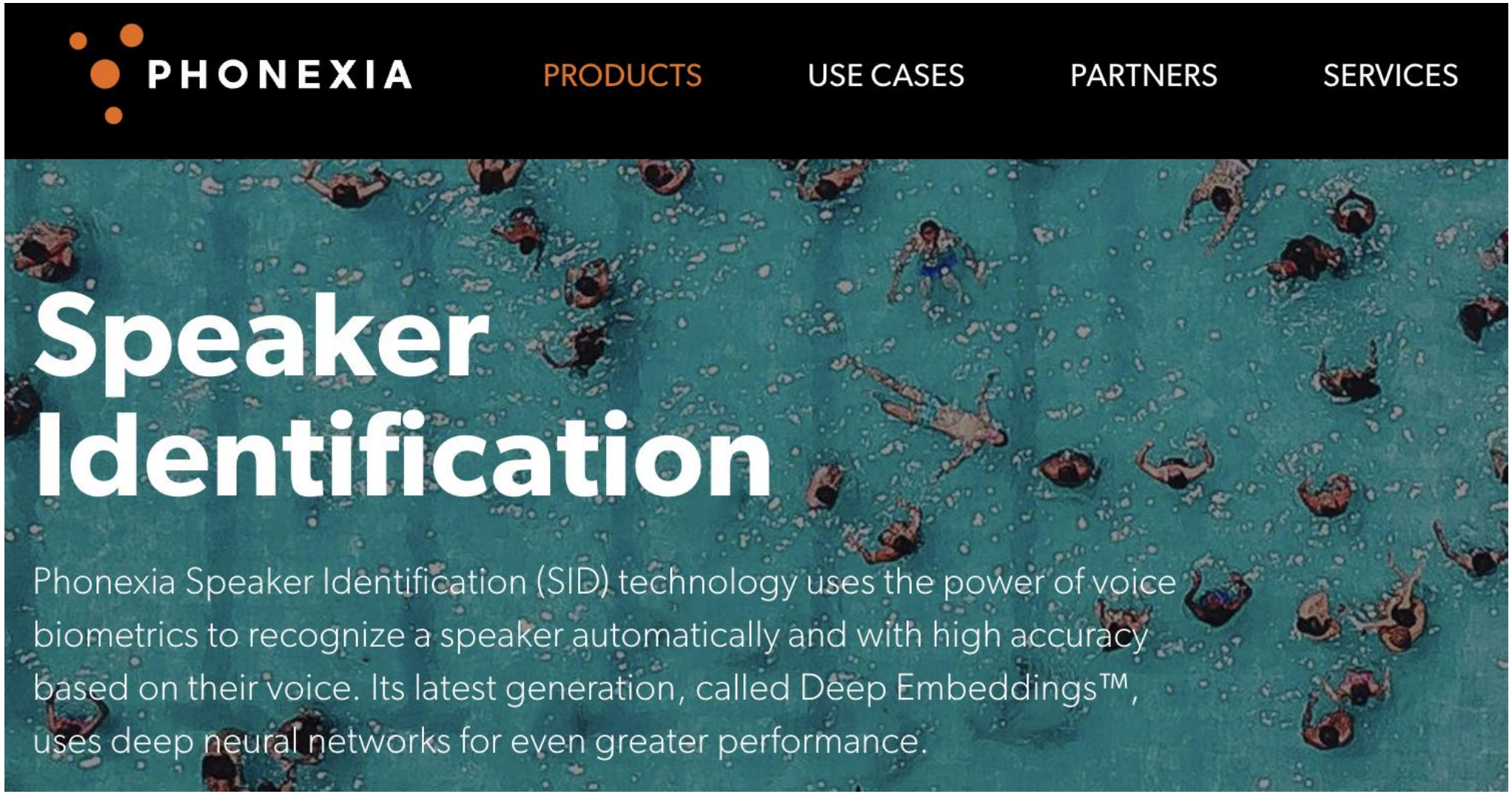
https://www.gmrtranscription.com/blog/podcast-transcription

Application: Video Captioning (e.g., Analysis & Situational/Permanent Hearing Impairments)

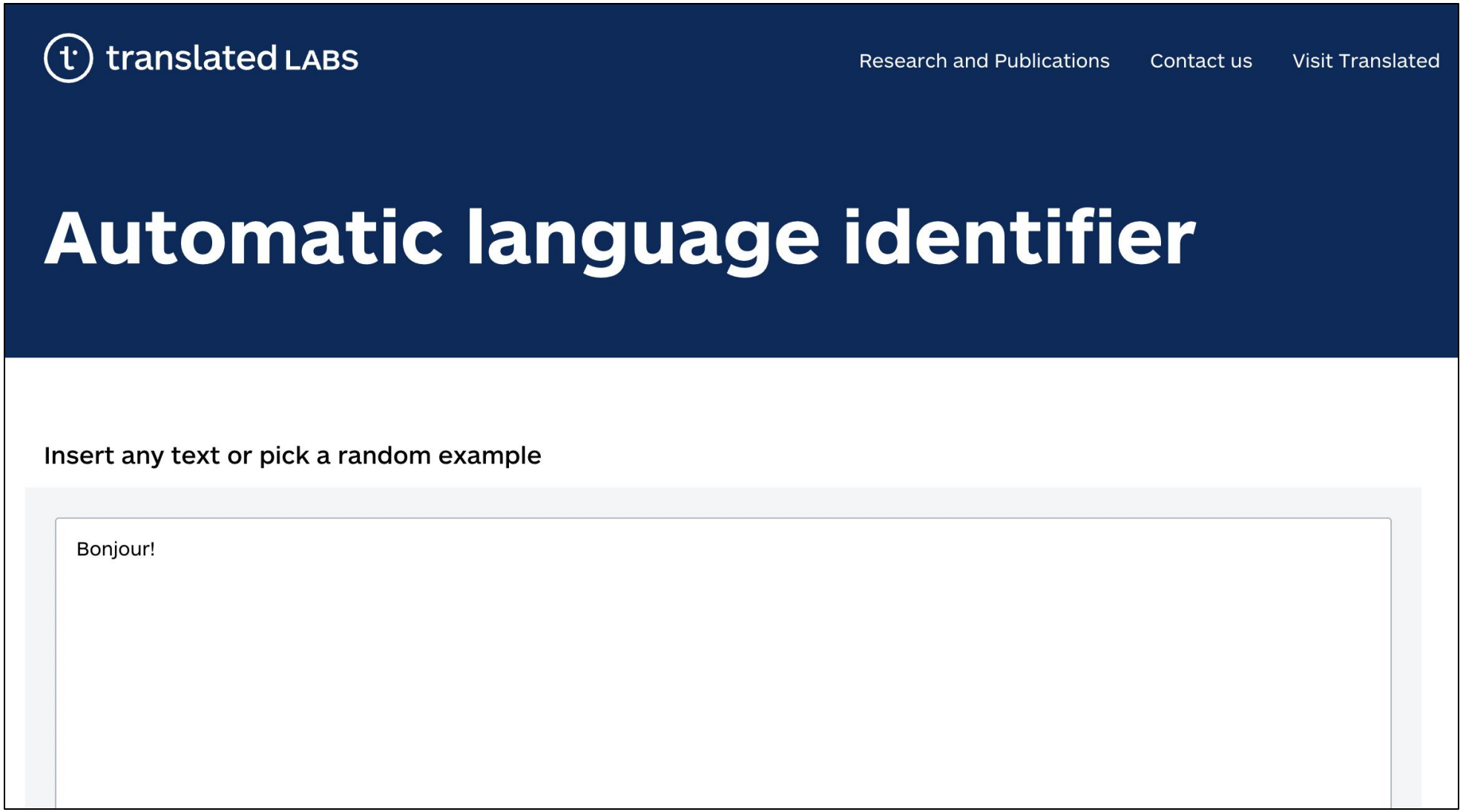
https://www.techsmith.com/blog/add-captions-subtitles-video/

Application: Speech Emotion Recognition (e.g., for Help Desks and Negotiators)

Application: Speaker Identification (e.g., Security)



Application: Language Identification



Application: Speech Enhancement

Evaluation: Spectrum of Tasks

Commands — Discourse

(brief with small vocabulary) — (lengthy with large vocabulary)

Evaluation: Word Error Rate

- Indicates edit distance between the prediction and the target as follows:

word replaced (e.g., "goose" transcribed as "choose")

word added (e.g., "caboose" transcribed as "the goose")

word omitted (e.g., "it is a fun" transcribed as "it is fun")

$$\frac{\text{Substitutions} + \text{Insertions} + \text{Deletions}}{\text{Number of Words Spoken}}$$

- What indicates better performance: larger or smaller values?

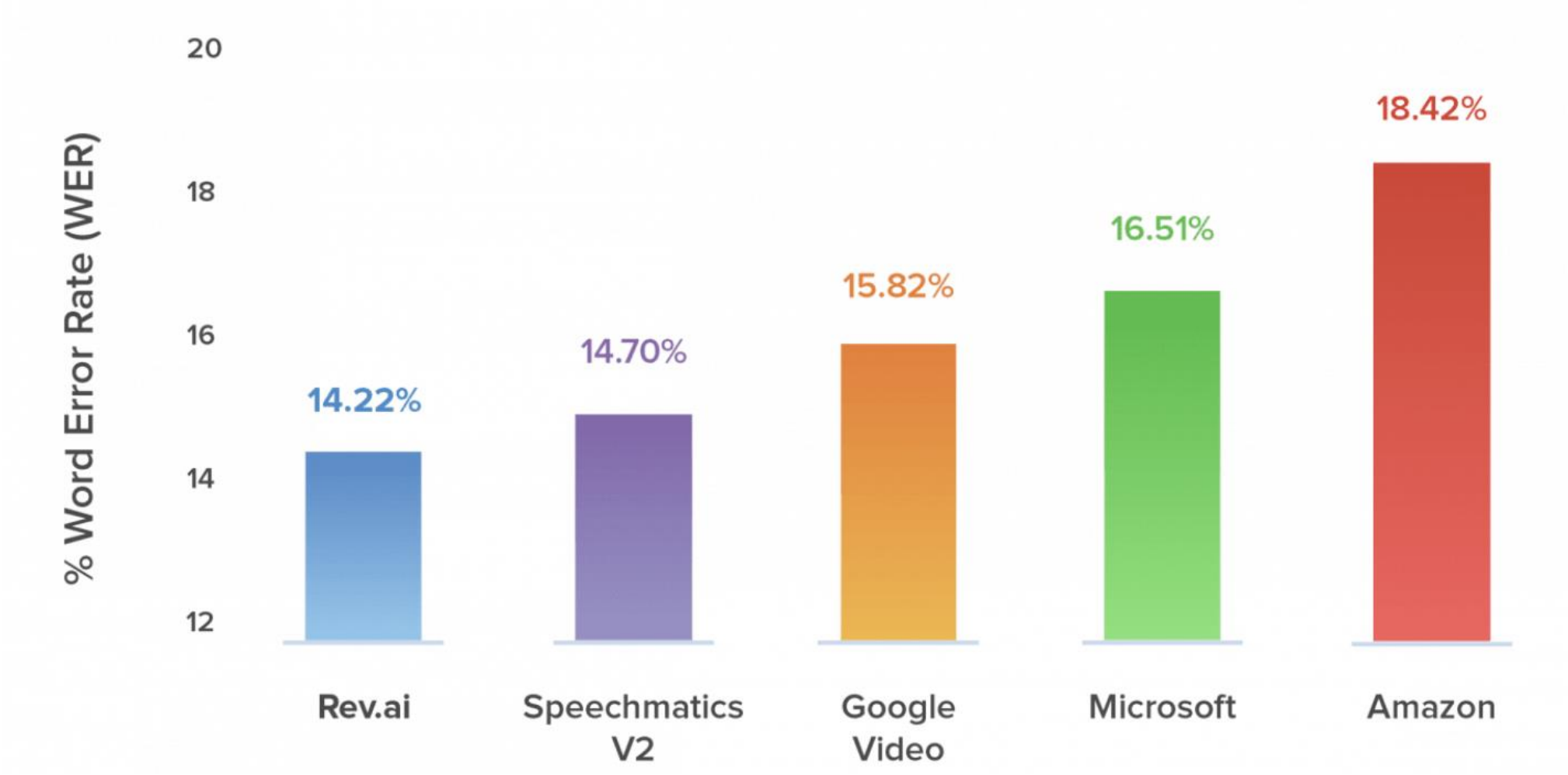
Evaluation: Word Error Rate Example

- Correct: The sun makes it look like uh a warm, day to go outside to adventure.
- Predicted: The son makes it to bike with a swarm to go outside to Denver today.
- Number of words spoken?
  - 15
- WER?

$$\frac{\text{Substitutions + Insertions + Deletions}}{\text{Number of Words Spoken}}$$

$$\frac{6 + 1 + 1}{15} = 0.53$$

Evaluation: Word Error Rate Comparison



https://www.rev.com/blog/resources/what-is-wer-what-does-word-error-rate-mean

Evaluation: What Are Limitations of Word Error Rate as an Evaluation Metric?

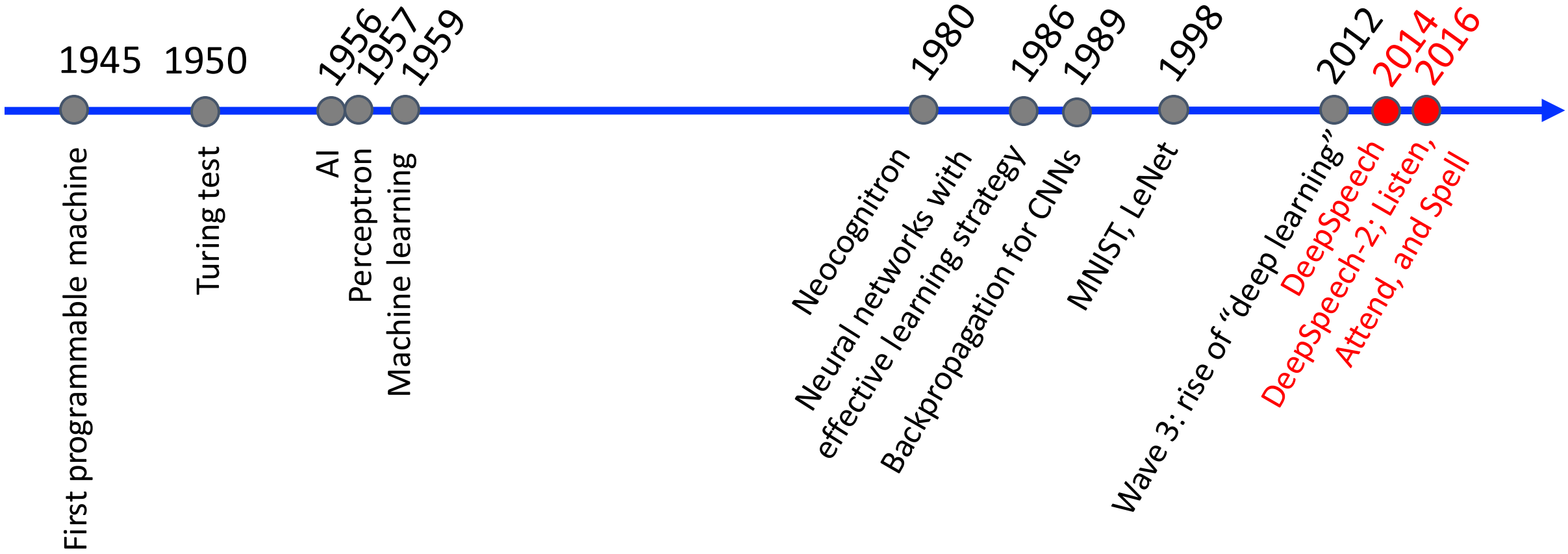
- Does not indicate why errors occur
  - Background noise (e.g., music, other talking)
  - Specialized language (i.e., words reflecting domain expertise)
  - Speaker pronunciations/accent

- Does not reflect whether transcription correctly captures:
  - Capitalization
  - Punctuation
  - Numbers
  - Paragraphs

- May indicate poor quality when humans could understand the content

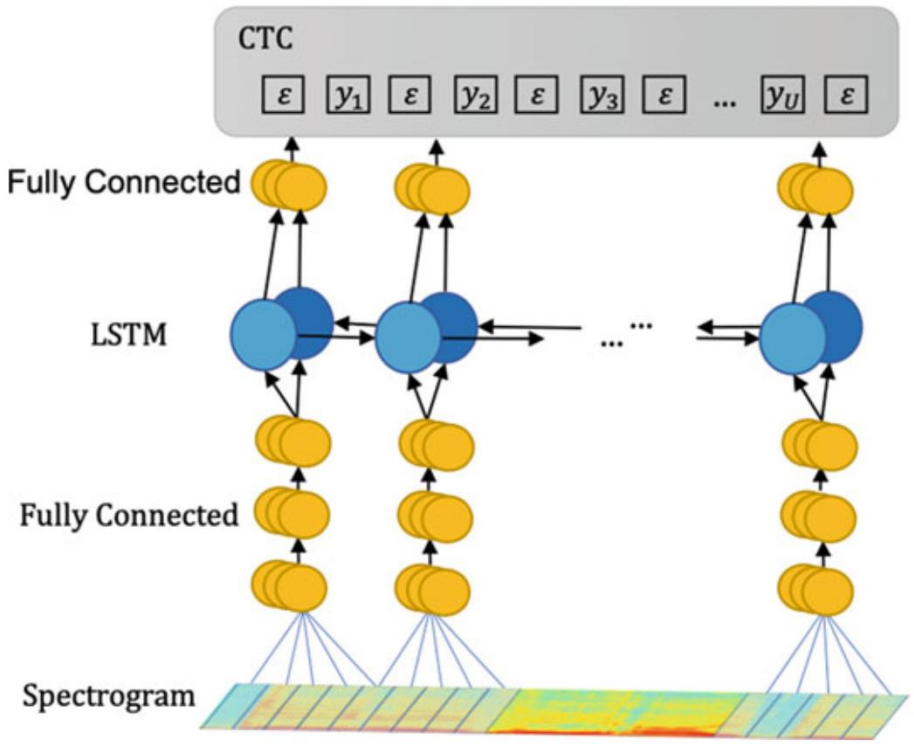
- Weights all word errors equally

Popular Methods: Historical Context

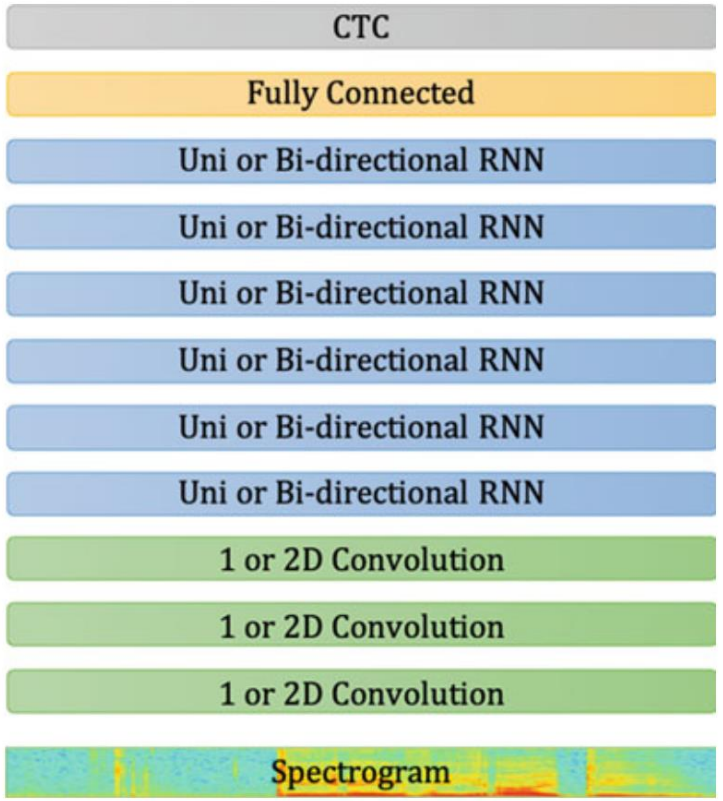


Popular Methods: Resembles What We Learned

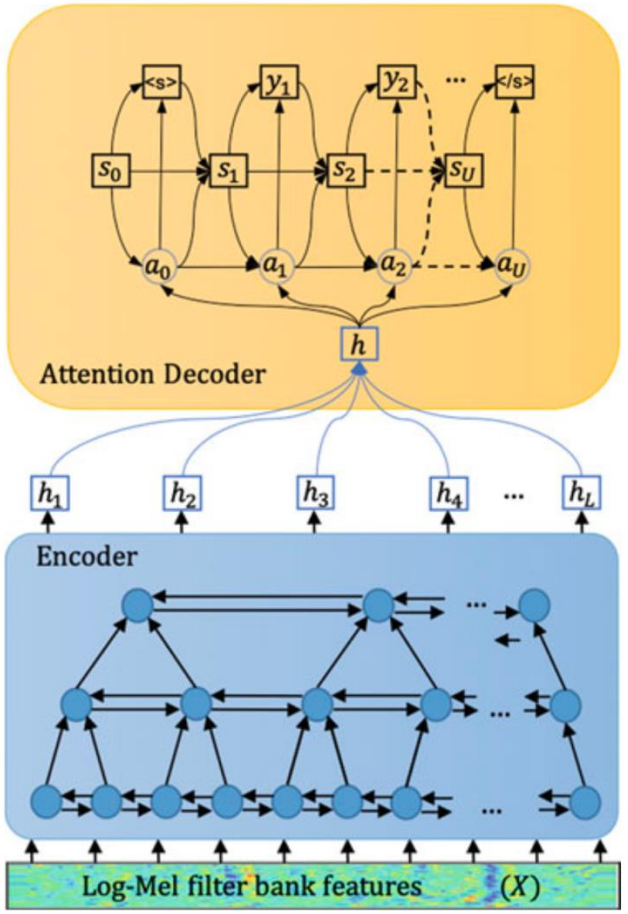
DeepSpeech



DeepSpeech2



Listen, Attend, and Spell

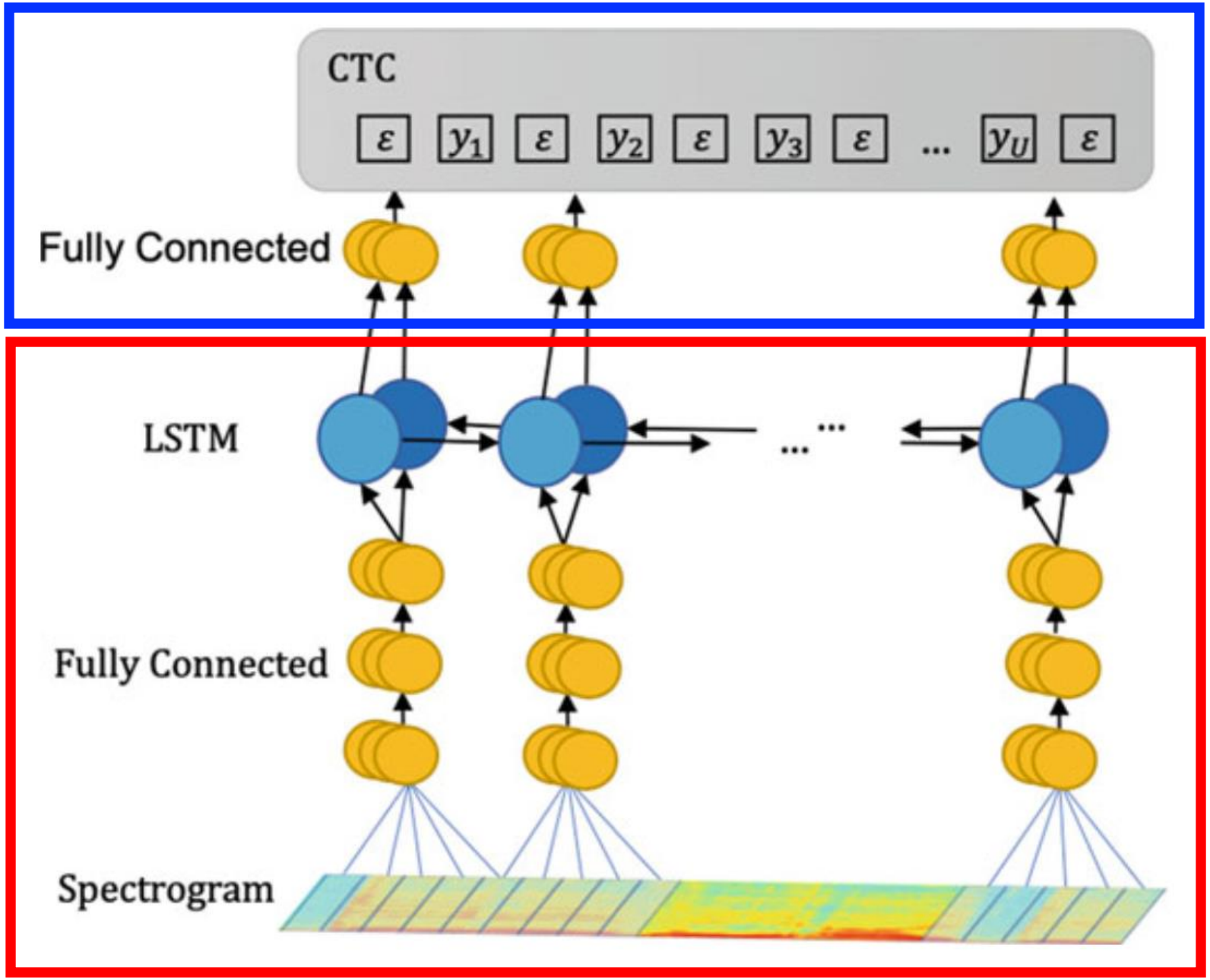


Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech

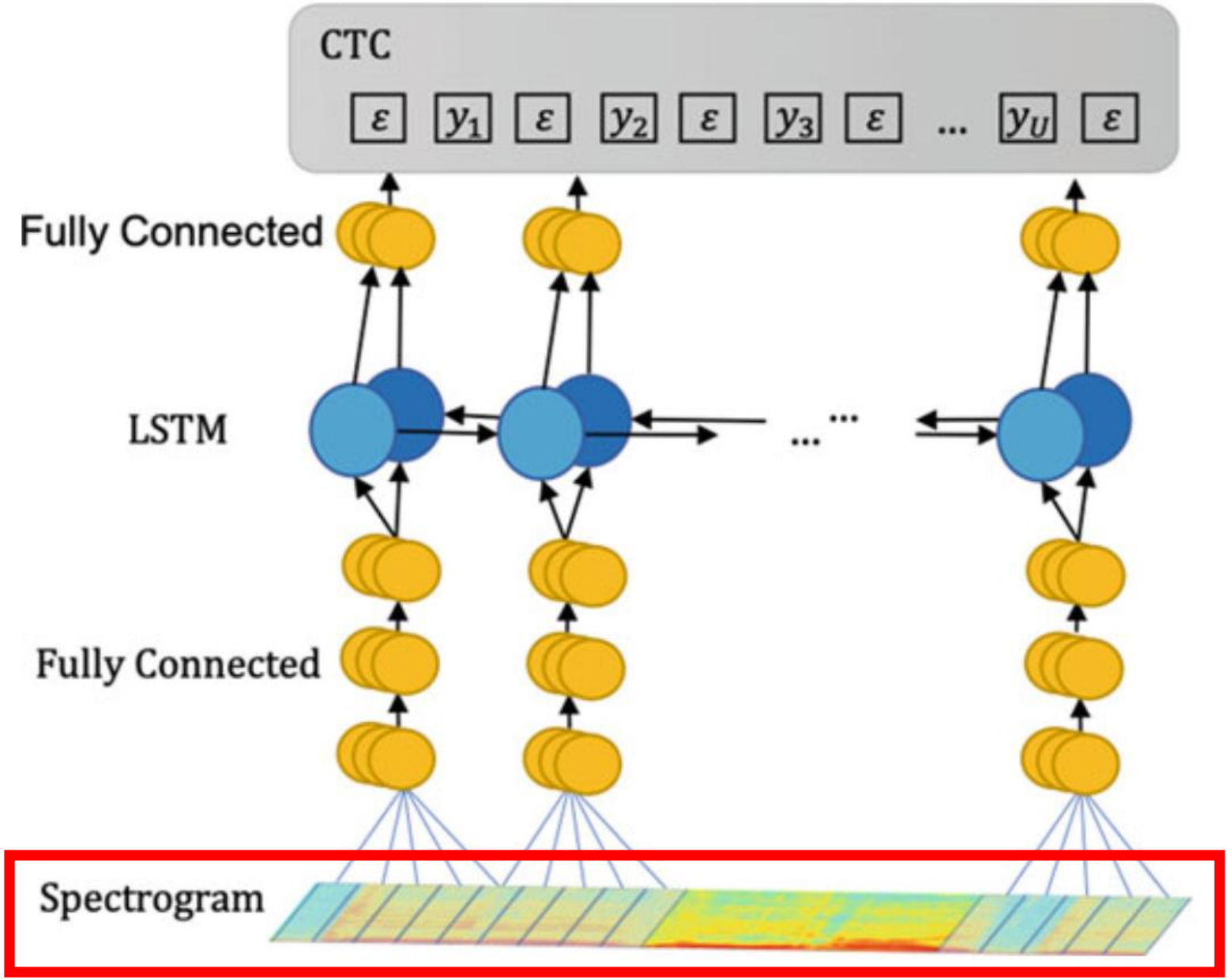
Decoder:

Encoder:



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech

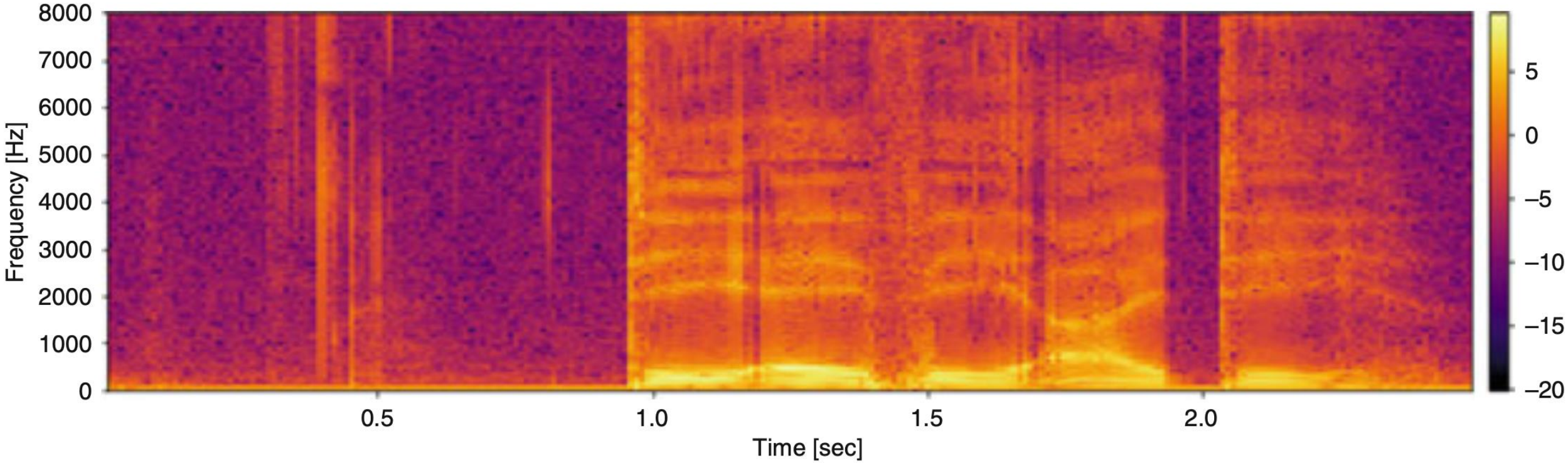


Input:

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Spectrogram: Visual Representation of Audio

Color: amplitude of frequency at a given time point

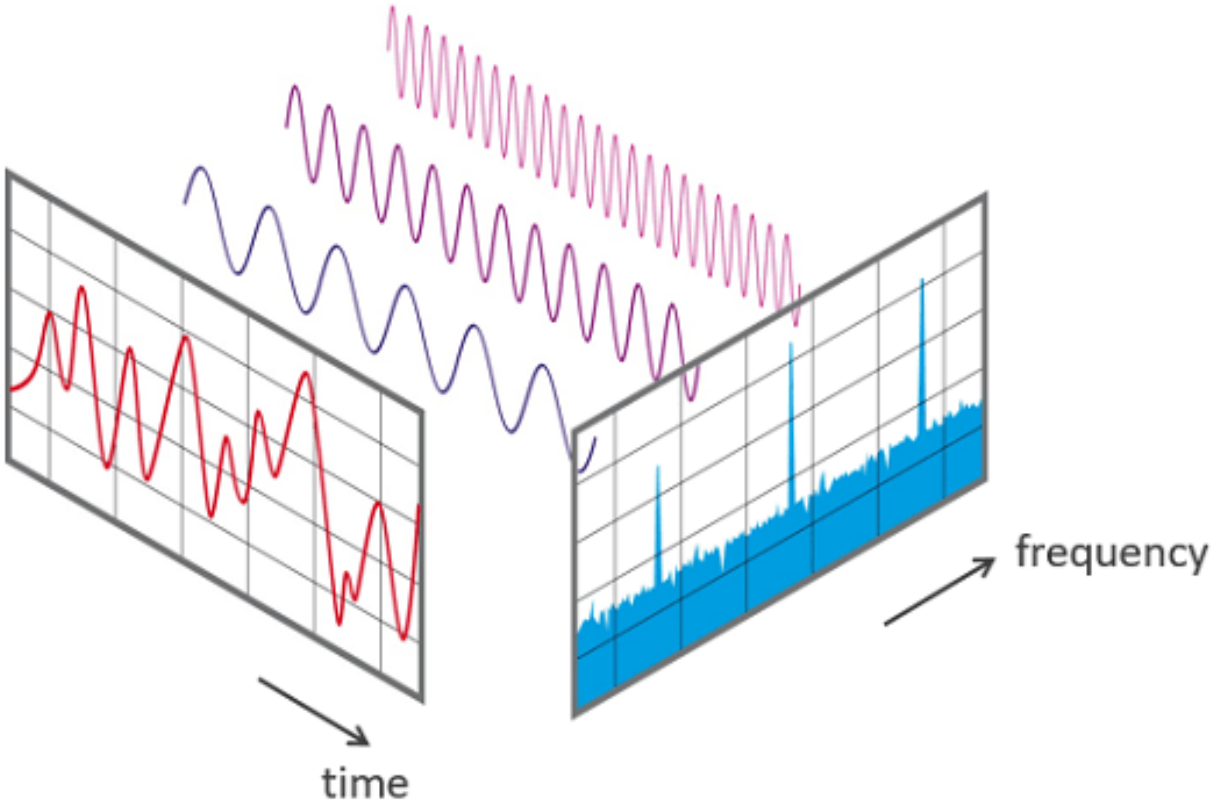


Created by sliding a short window across the audio signal and applying a Fourier transform to each window

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Background: Frequency Analysis of Audio Clip

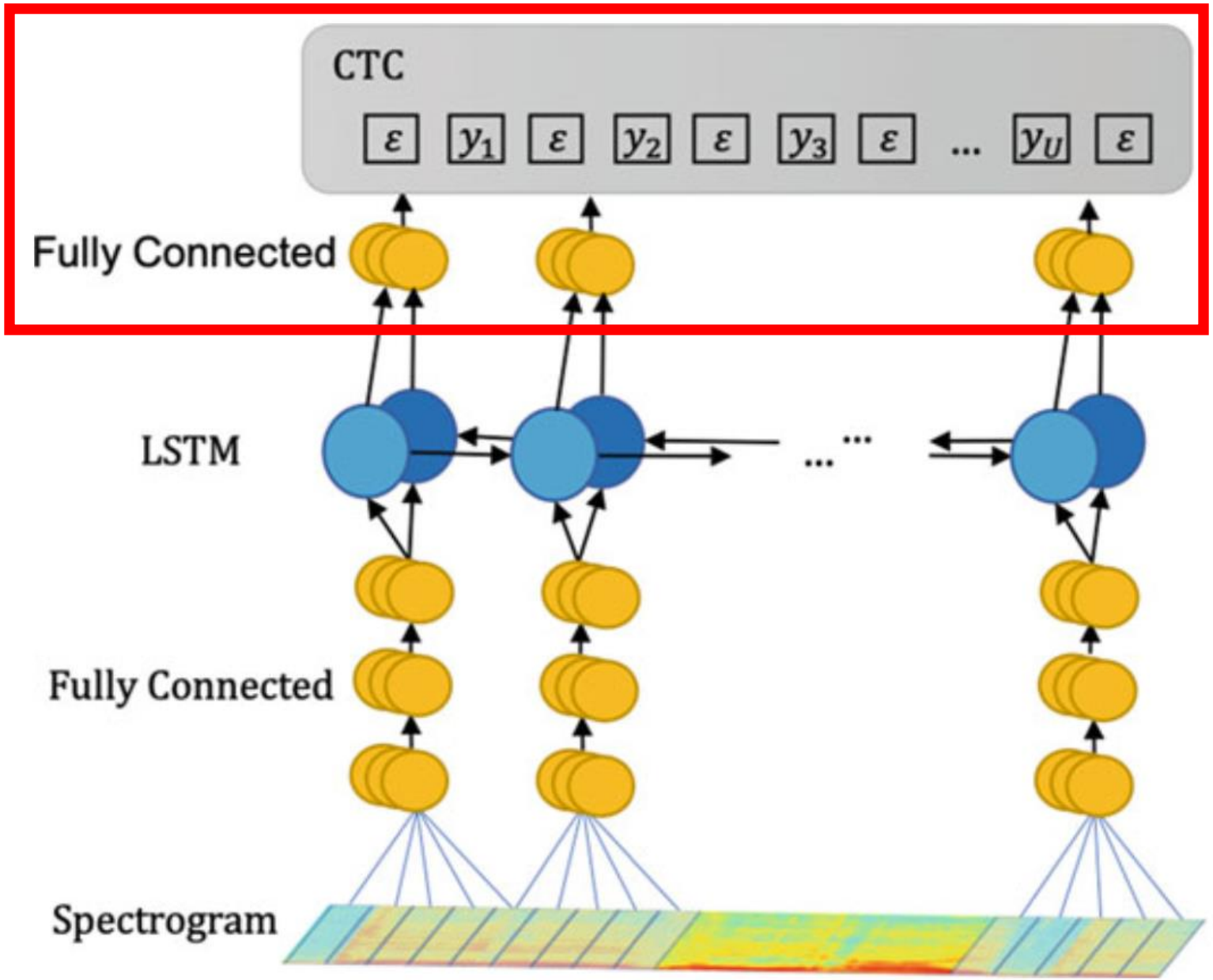
Fourier transform: represents a signal as a sum of sines and cosines (*frequency-domain*):



https://dev.to/trekhleb/playing-with-discrete-fourier-transform-algorithm-in-javascript-53n5

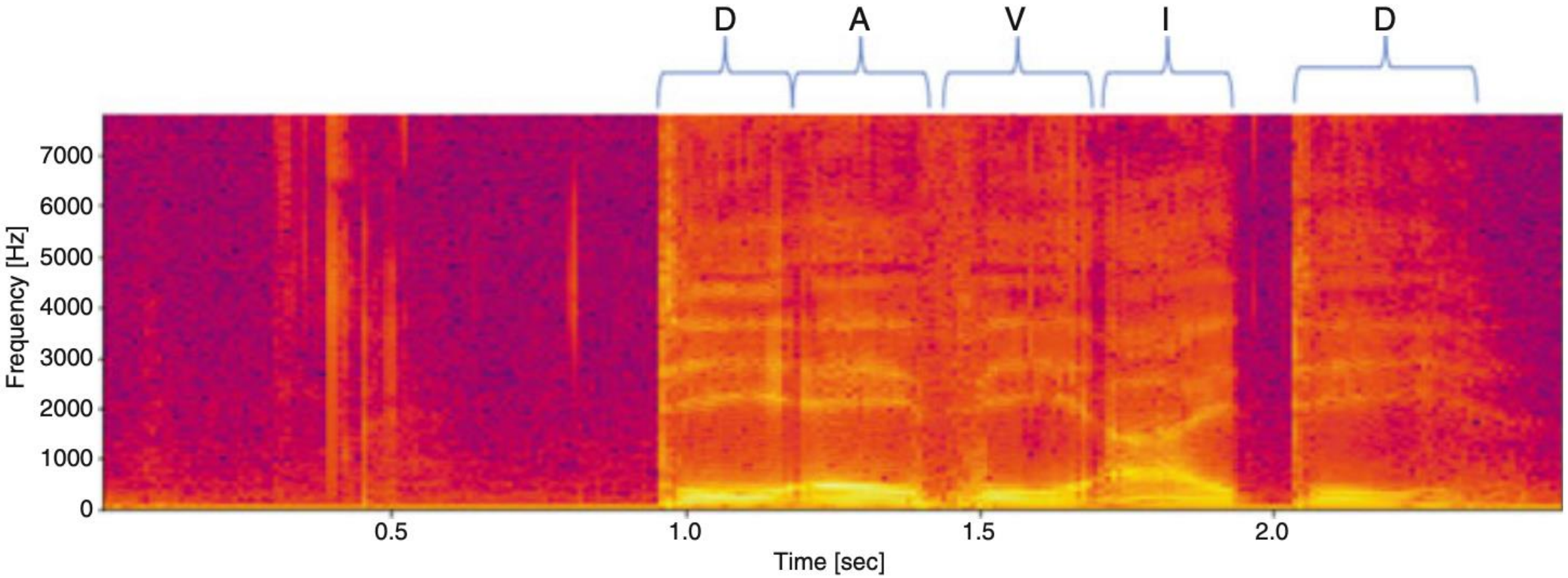
DeepSpeech

Output: character sequence predicted by a softmax layer



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

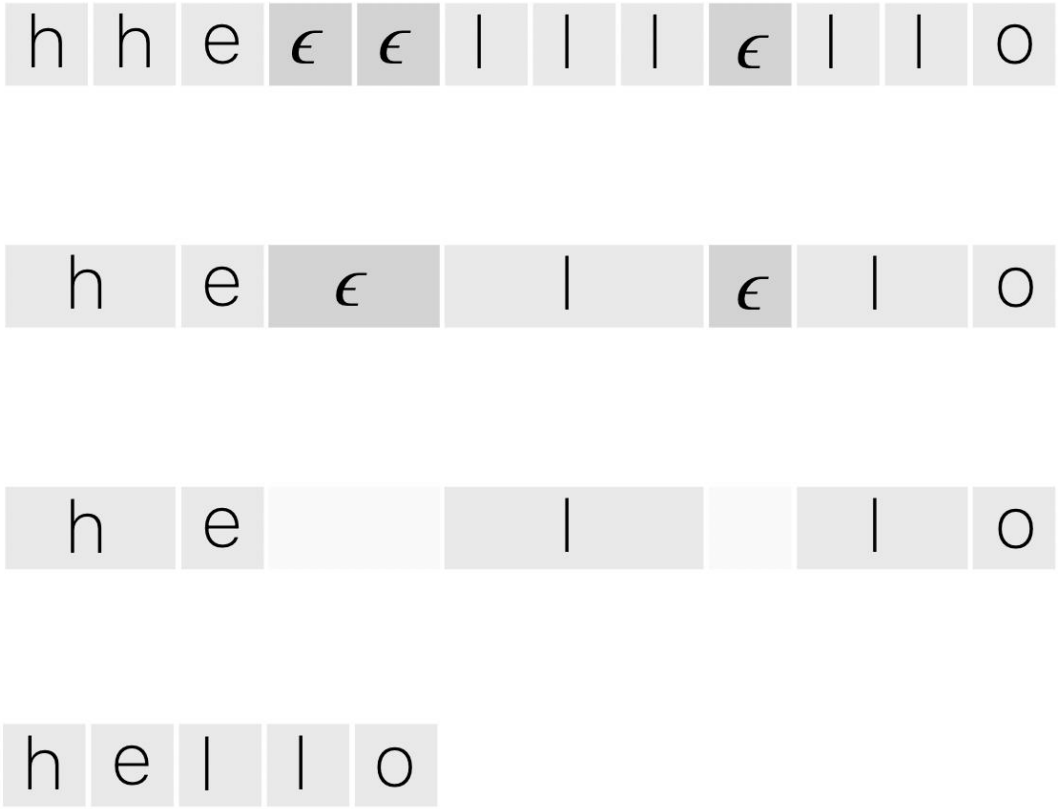
CTC: Input-Output Representation



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

CTC: Input-Output Representation

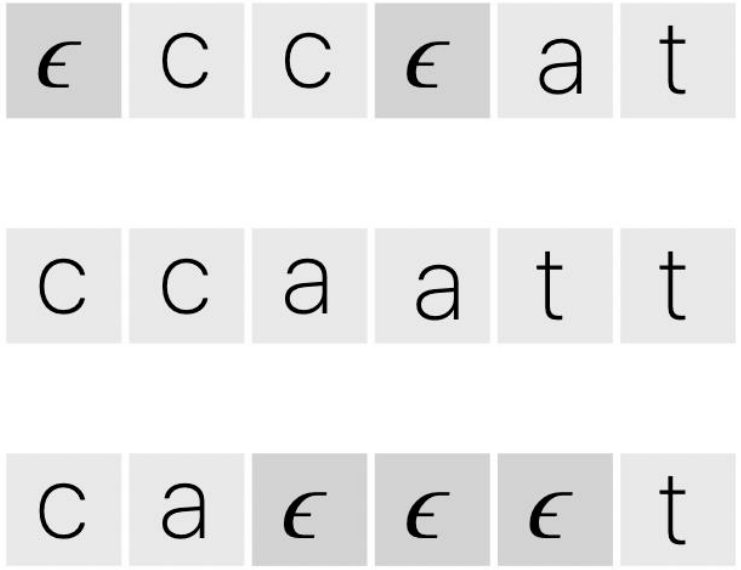
Key idea: blank token supports silent stretches and letter repeats (e.g., “hello” vs “helo”)



https://distill.pub/2017/ctc/

CTC: Input-Output Representation

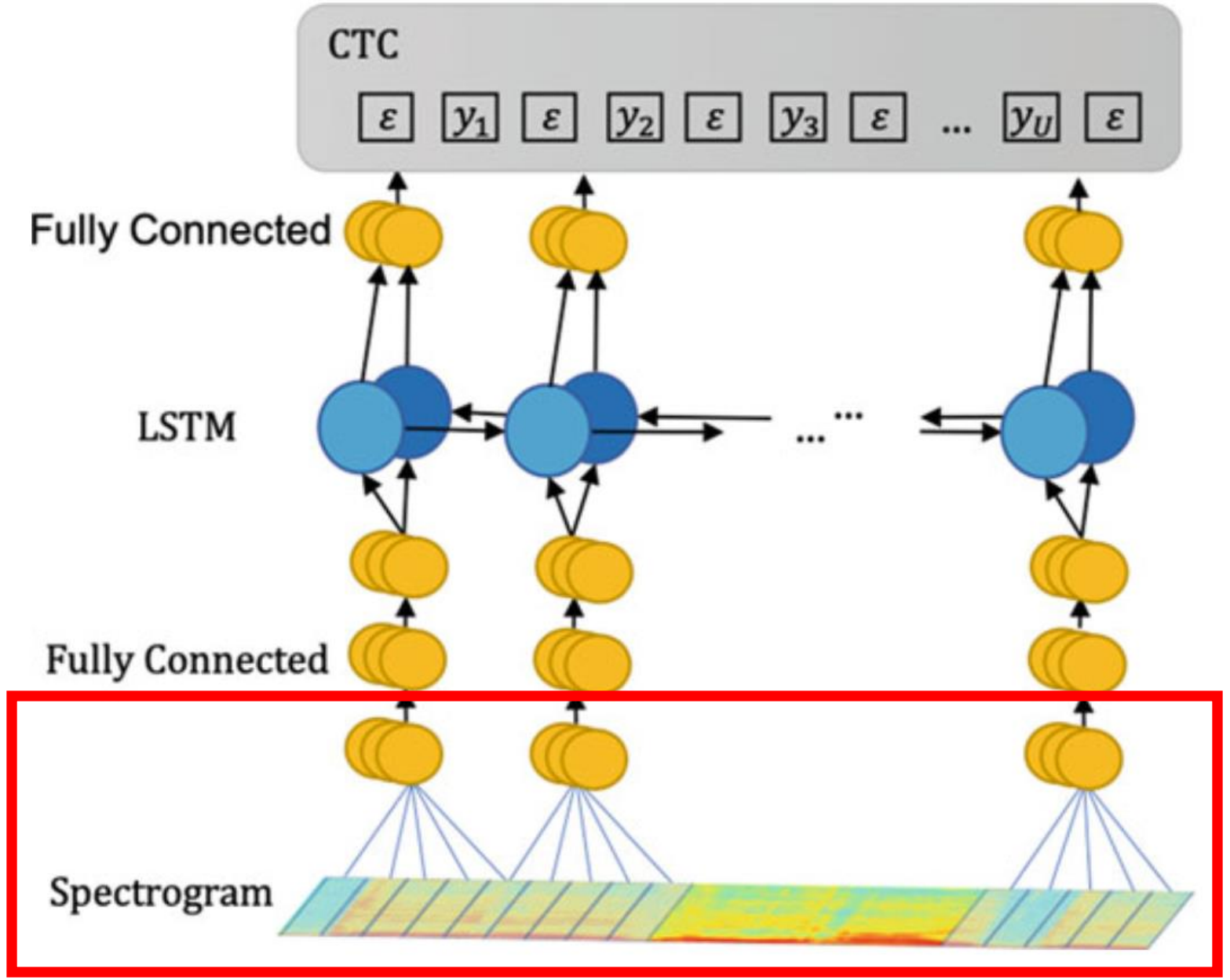
Key idea: blank token supports silent stretches and letter repeats (e.g., “hello” vs “helo”)



Supports recognizing the same word when spoken differently!

https://distill.pub/2017/ctc/

DeepSpeech



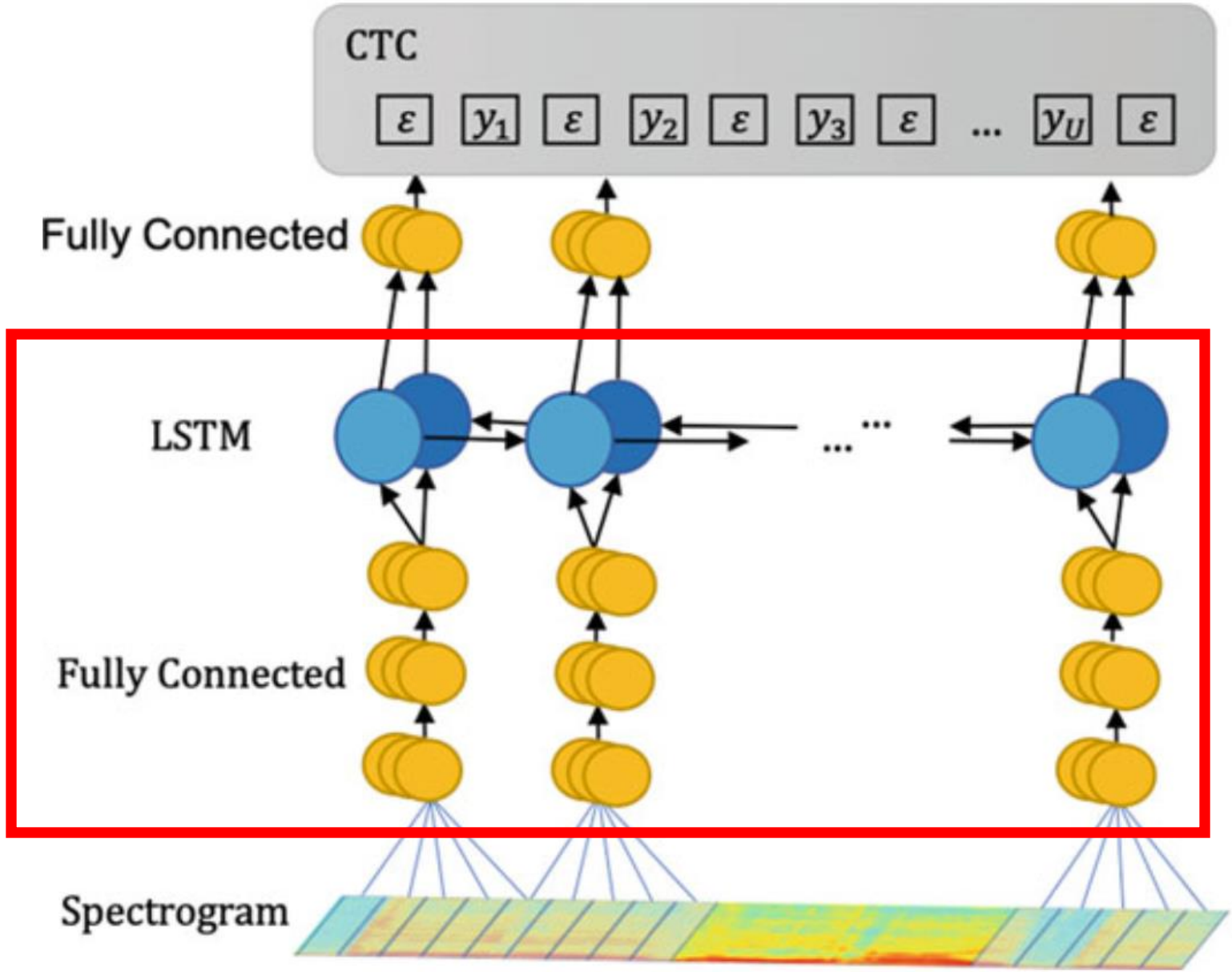
First hidden layer looks at context around input:

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech

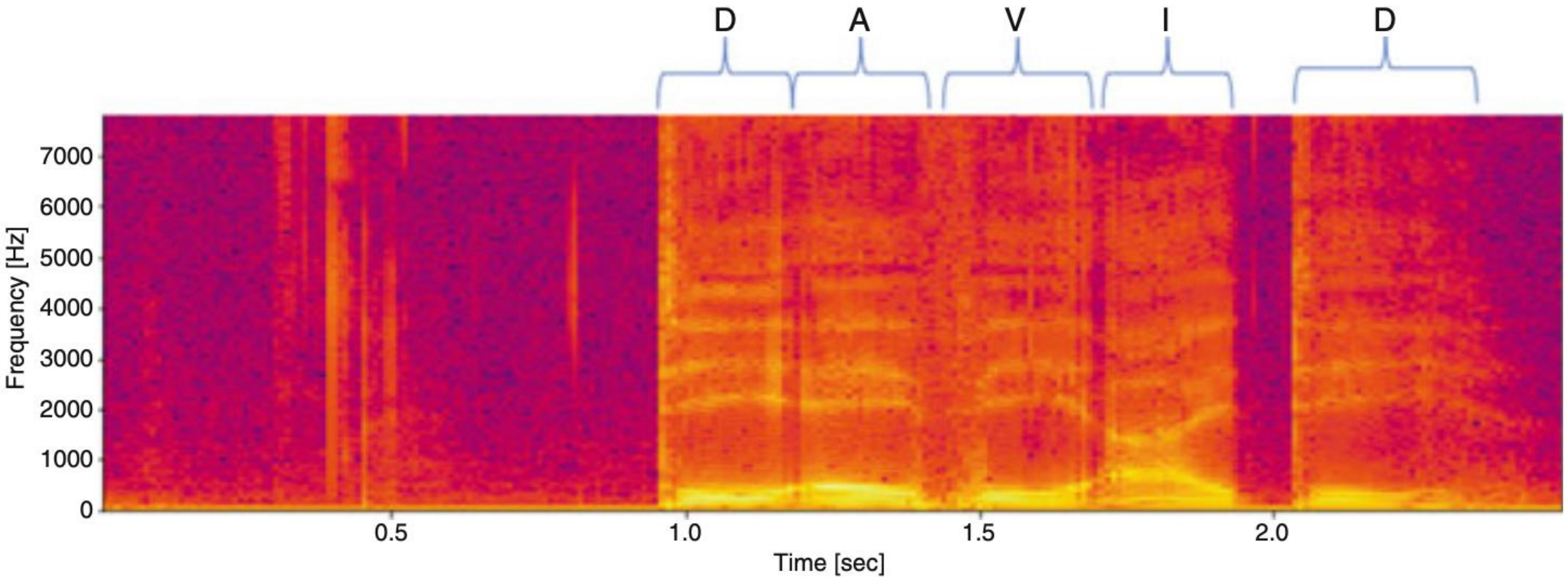
3 fully-connected layers
followed by bidirectional LSTM:

Can this run in real-time?
- no, it must hear everything



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech: Optimization Function (CTC)



The CTC loss function enables learning output alignment without a per input label

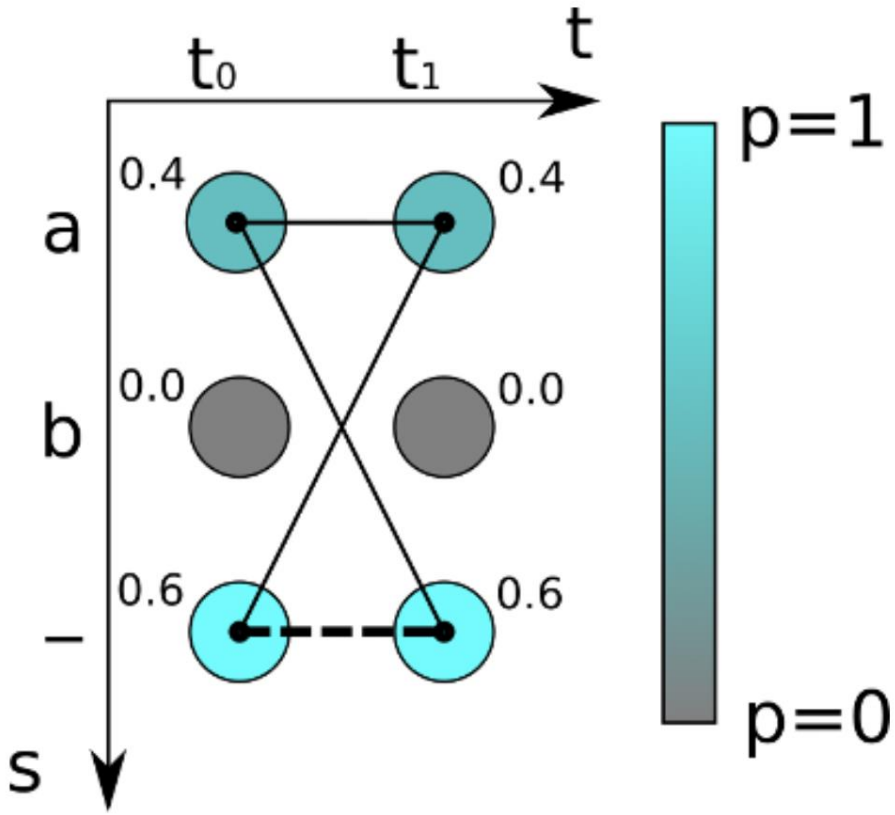
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech: Optimization Function (CTC)

- How many timesteps (t) are in this example?
  - 2
- How many token options (s) in this example?
  - 3: 2 characters (a, b) and blank (“-”)

Predicts most plausible from all possible alignments:

- Probability of “a” is sum of all “a” representations
  - Probability of “aa”?
    - 0.4 x 0.4 = 0.16
  - Probability of “a-”?
    - 0.4 x 0.6 = 0.24
  - Probability of “-a”?
    - 0.6 x 0.4 = 0.24
  - Sum: 0.16 + 0.24 + 0.24 = 0.64



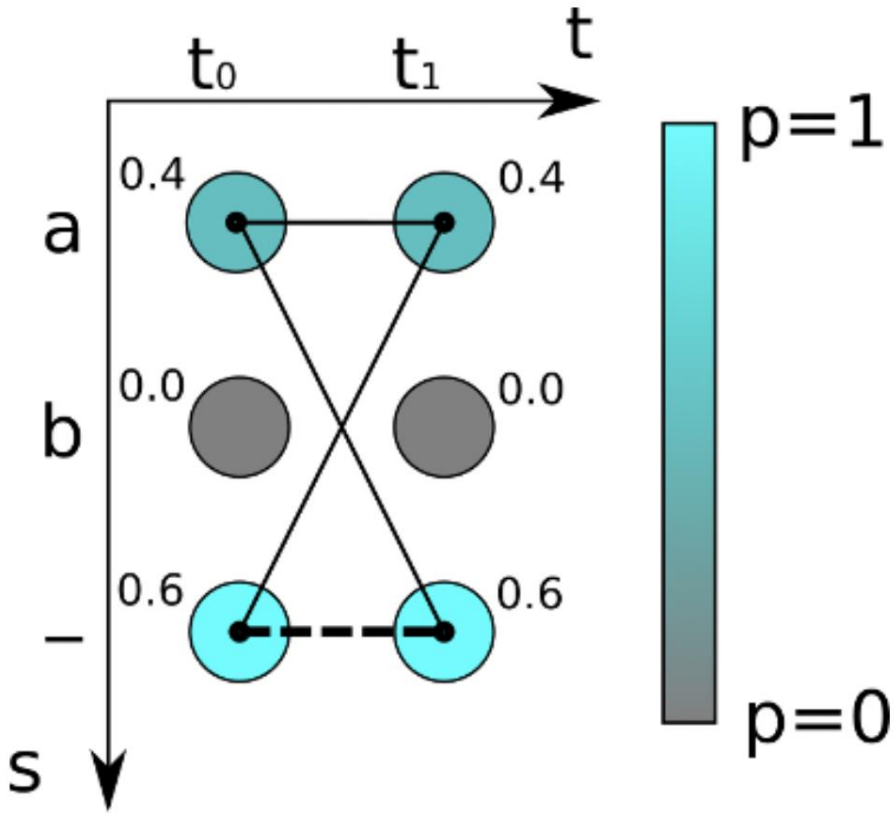
https://towardsdatascience.com/intuitively-understanding-connectionist-temporal-classification-3797e43a86c

DeepSpeech: Optimization Function (CTC)

- How many timesteps (t) are in this example?
  - 2
- How many token options (s) in this example?
  - 3: 2 characters (a, b) and blank (“-”)

Predicts most plausible from all possible alignments:

- Probability of “a”: 0.64
- Probability of “-” is sum of all “-” representations
  - Probability of “--”?
    - 0.6 x 0.6 = 0.36



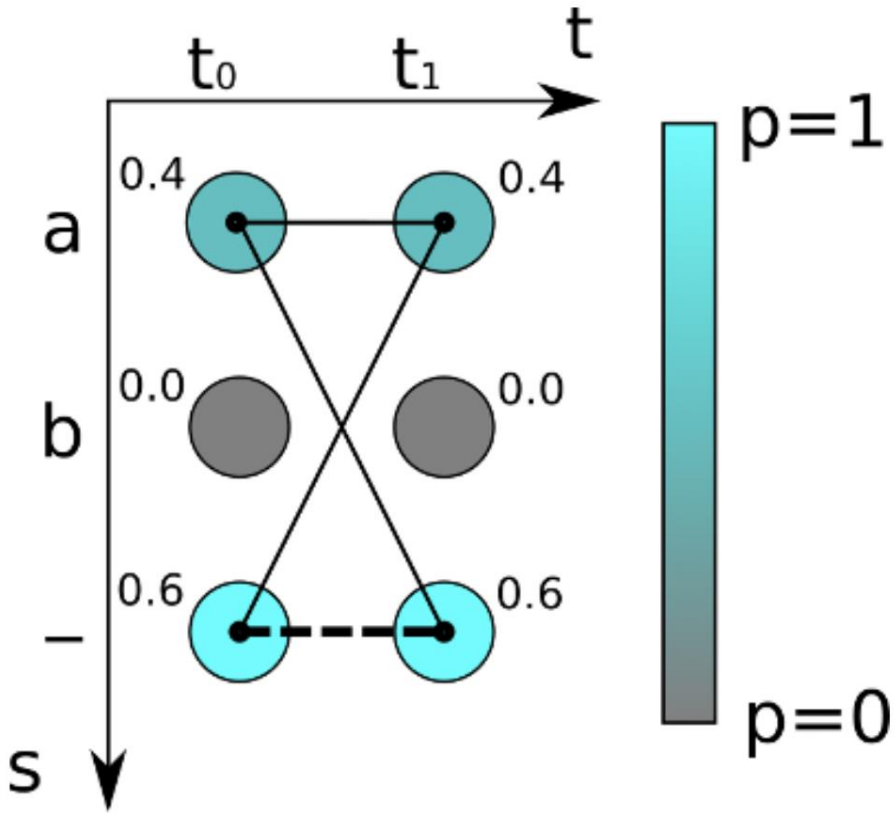
https://towardsdatascience.com/intuitively-understanding-connectionist-temporal-classification-3797e43a86c

DeepSpeech: Optimization Function (CTC)

- How many timesteps (t) are in this example?
  - 2
- How many token options (s) in this example?
  - 3: 2 characters (a, b) and blank (“-”)

Predicts most plausible from all possible alignments:

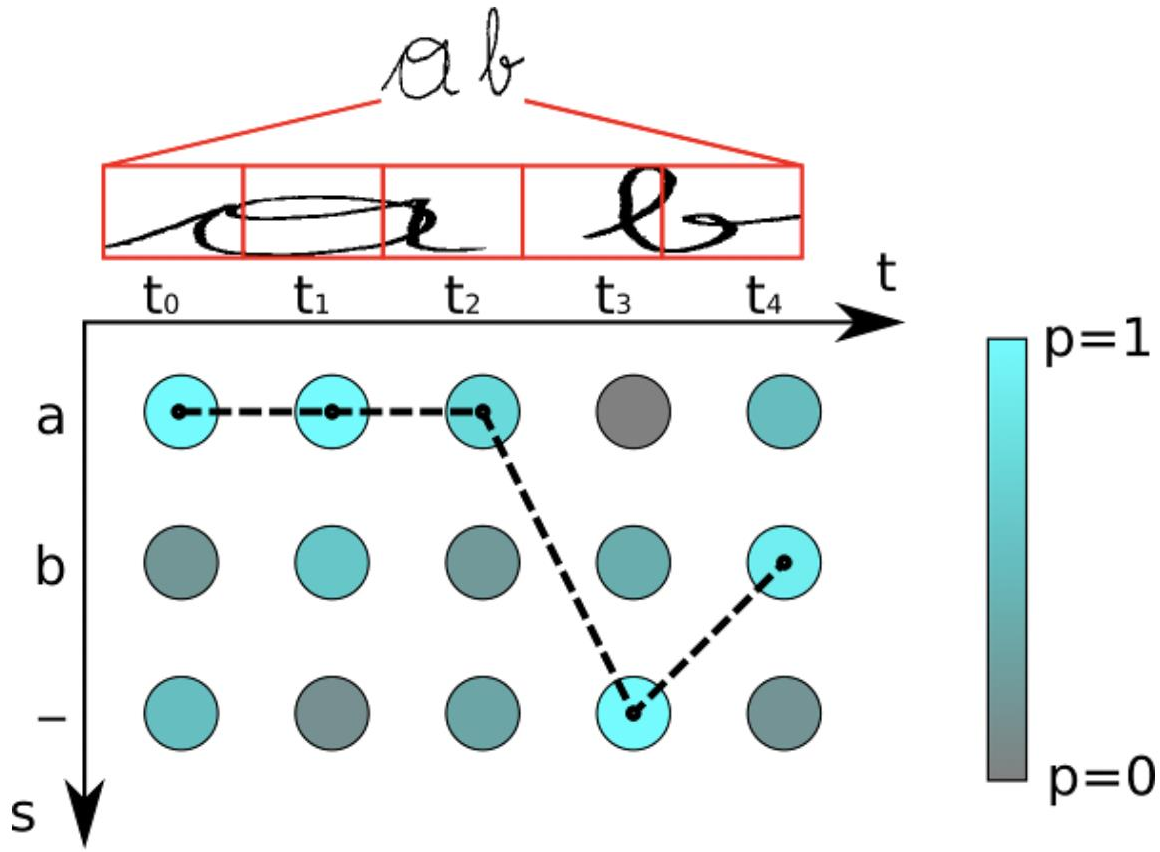
- Probability of “a”: 0.64
- Probability of ““: 0.36
- And so on for all possible alignments…



https://towardsdatascience.com/intuitively-understanding-connectionist-temporal-classification-3797e43a86c

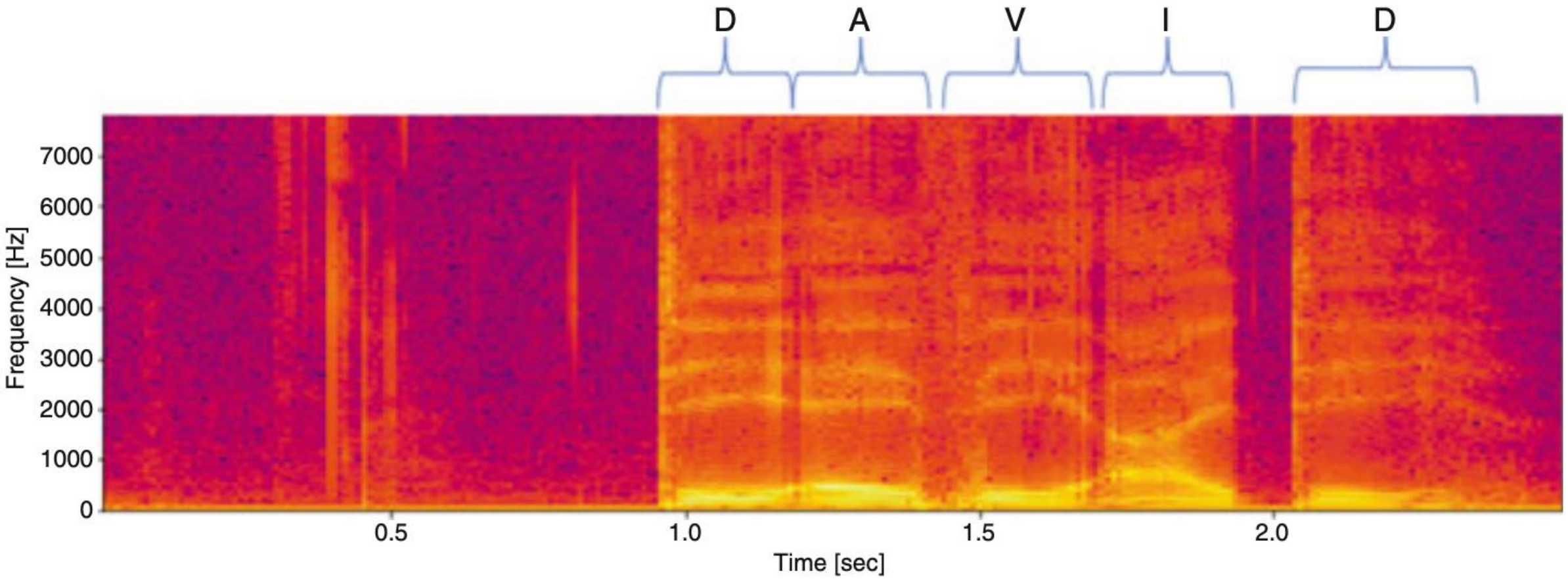
DeepSpeech: Optimization Function (CTC)

Most plausible from all possible alignments learned with best path decoding



https://towardsdatascience.com/intuitively-understanding-connectionist-temporal-classification-3797e43a86c

DeepSpeech: Optimization Function (CTC)



CTC uses dynamic programming to accelerate computation and is differentiable

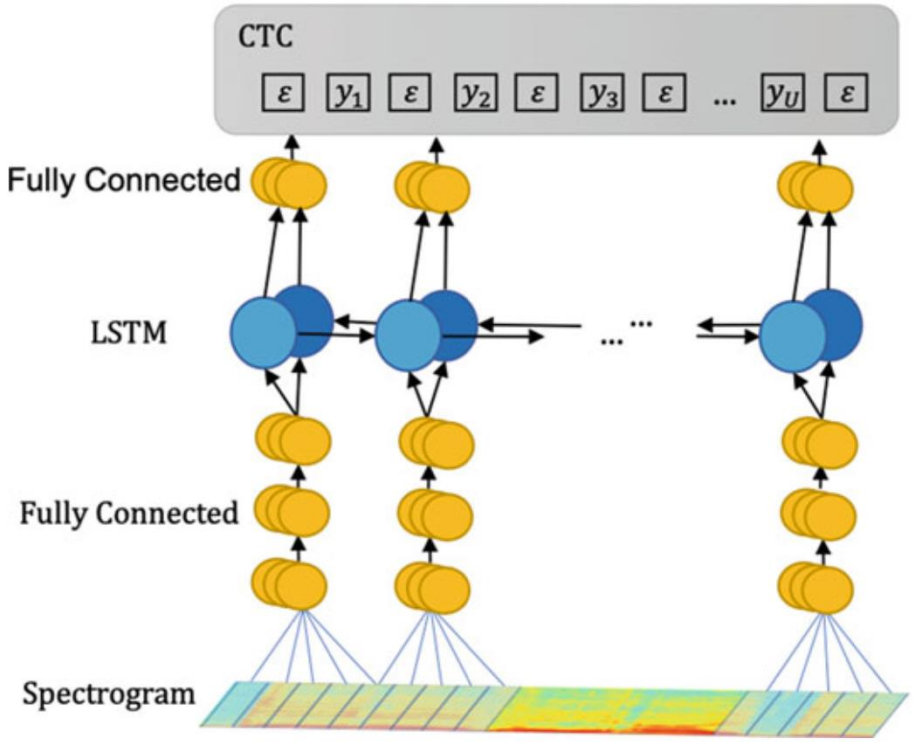
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech: Training (Key Ideas)

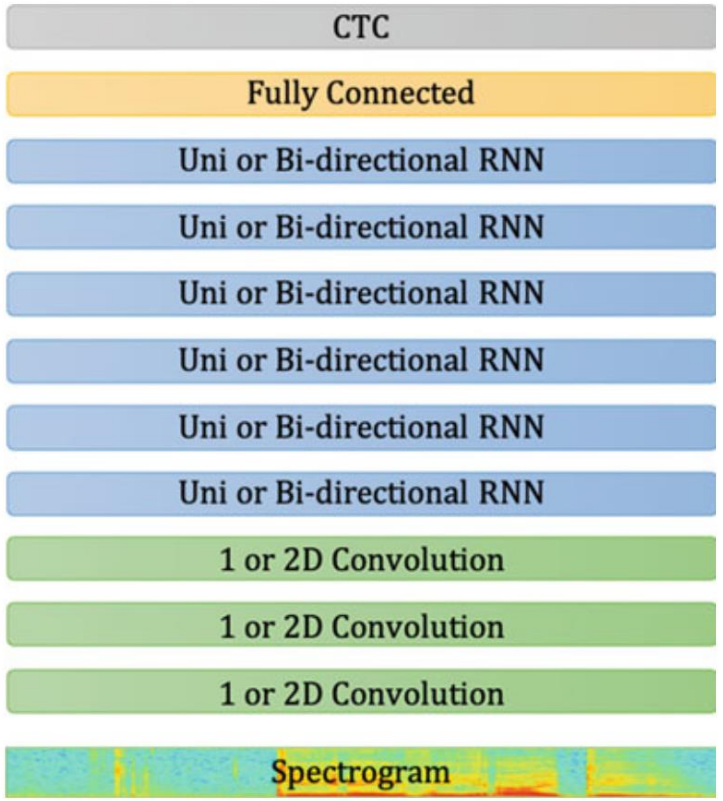
- 5,000 hours from 9,600 speakers
- Regularization
  - Dropout
  - Data augmentation: audio file translated 5 ms forward and backward
- Results boosted by incorporating a language model

Popular Methods: Resembles What We Learned

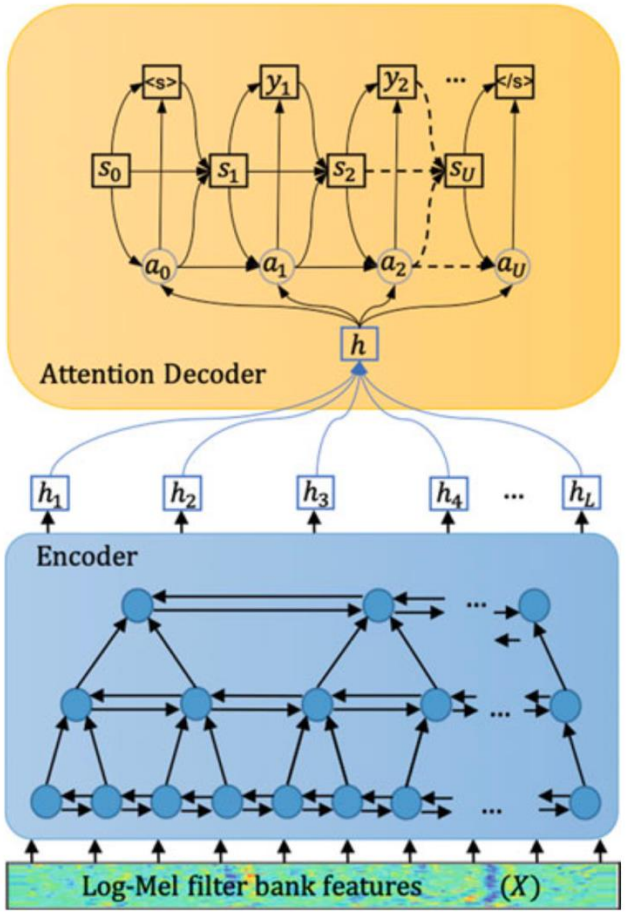
DeepSpeech



DeepSpeech2



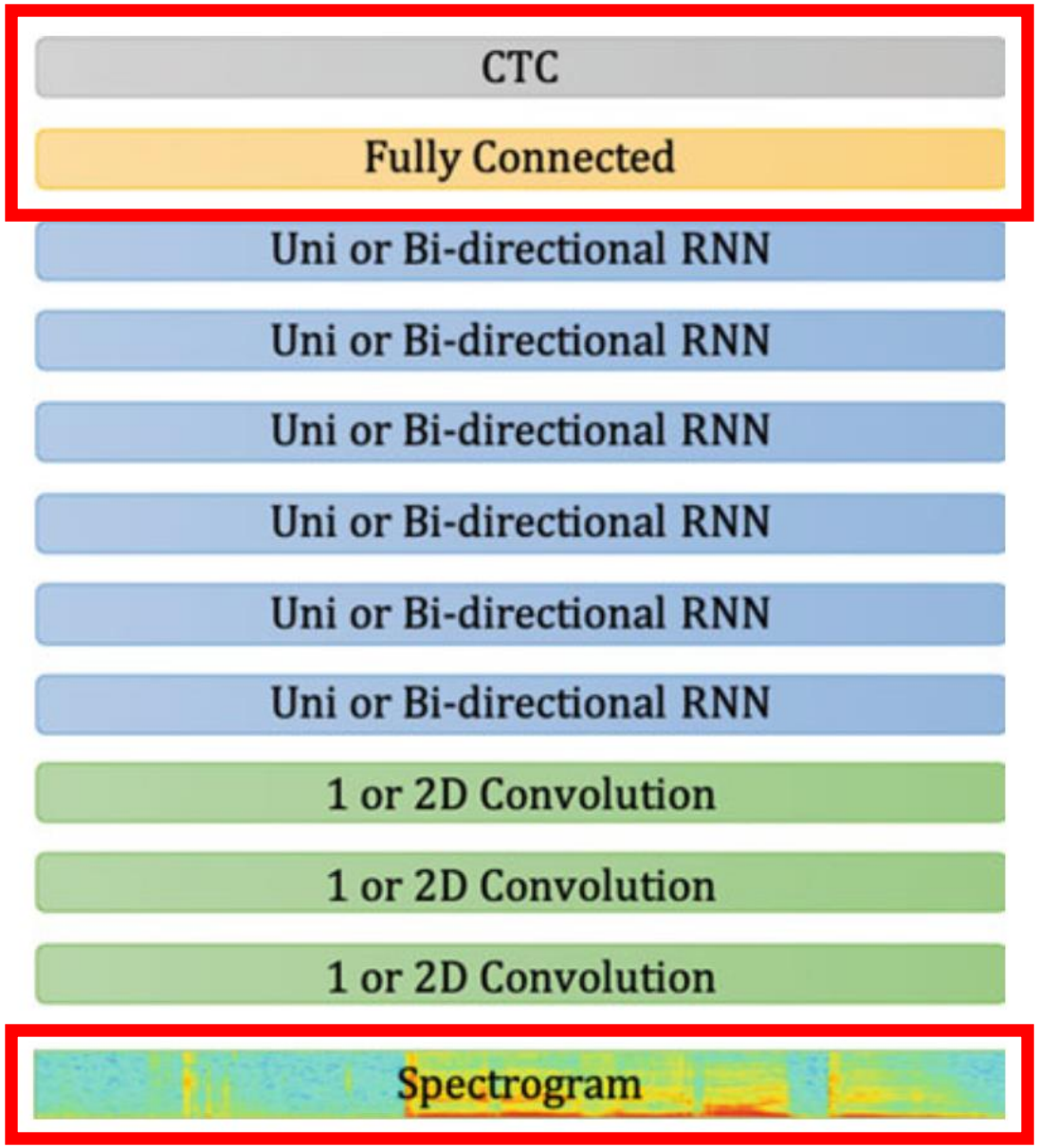
Listen, Attend, and Spell



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech2

Similar output:
(two architectures for
English and Mandarin)



Extension of DeepSpeech that
achieves a 7x speed-up and
43.4% relative WER improvement
with a deeper architecture

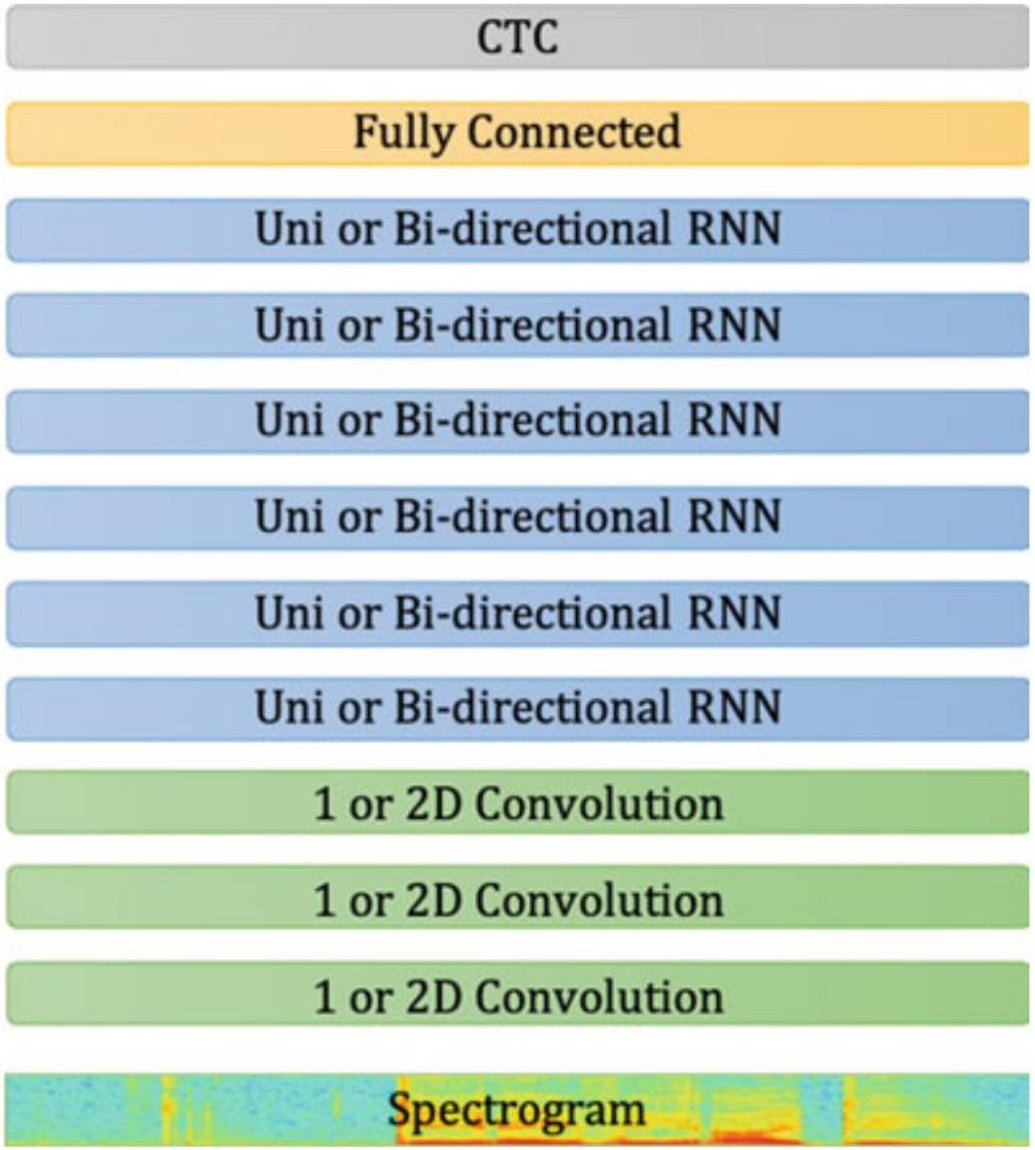
Same input:

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech2

Training protocol difference from DeepSpeech:

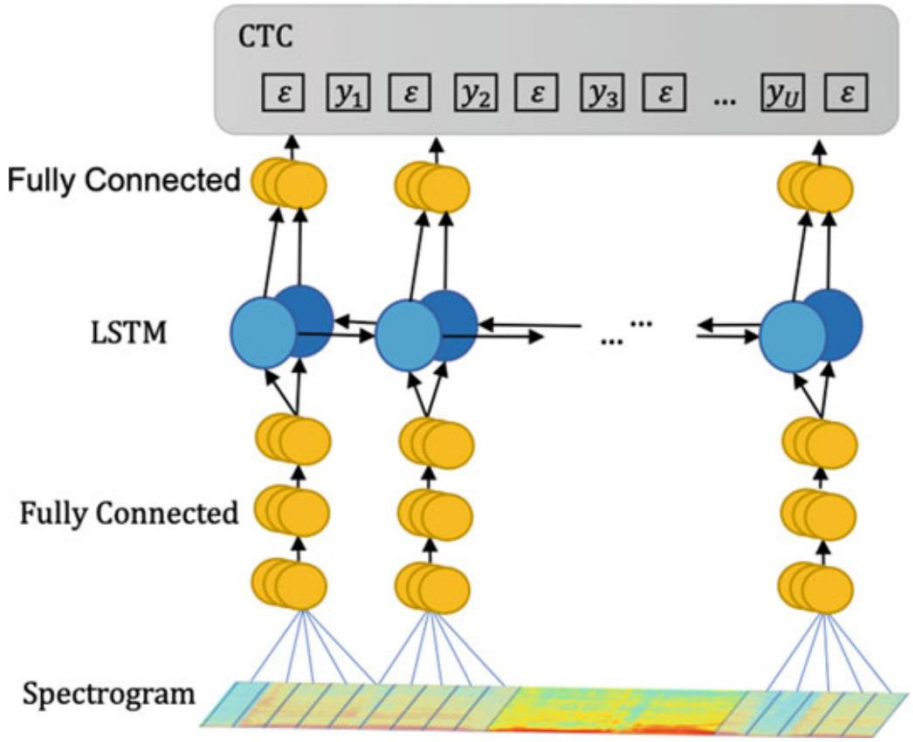
- More training data (11,940 hours for English and 9,400 hours for Mandarin)
- Curriculum learning: trains based on length of utterances for first epoch with shorter ones first (improves WER by over 1 point)



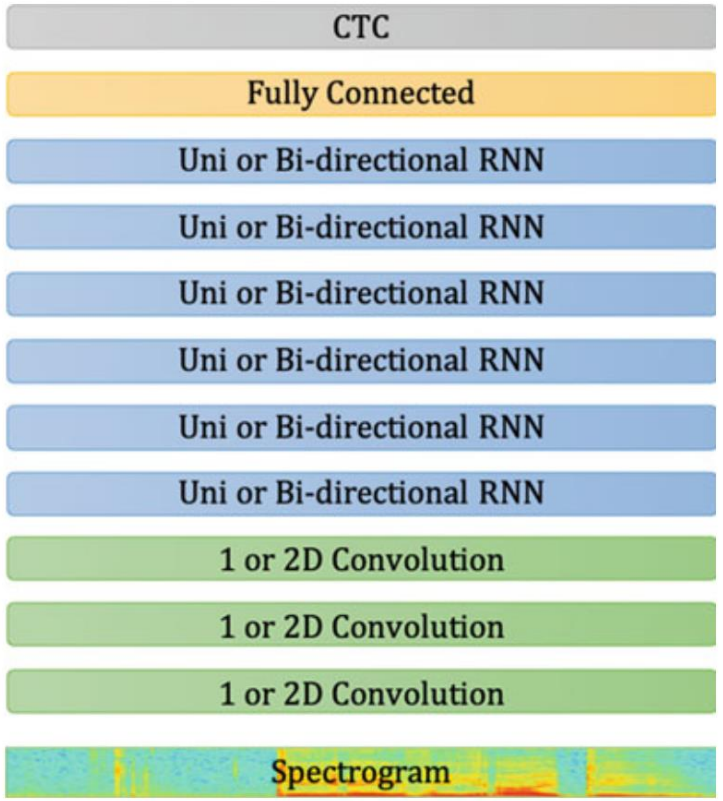
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Popular Methods: Resembles What We Learned

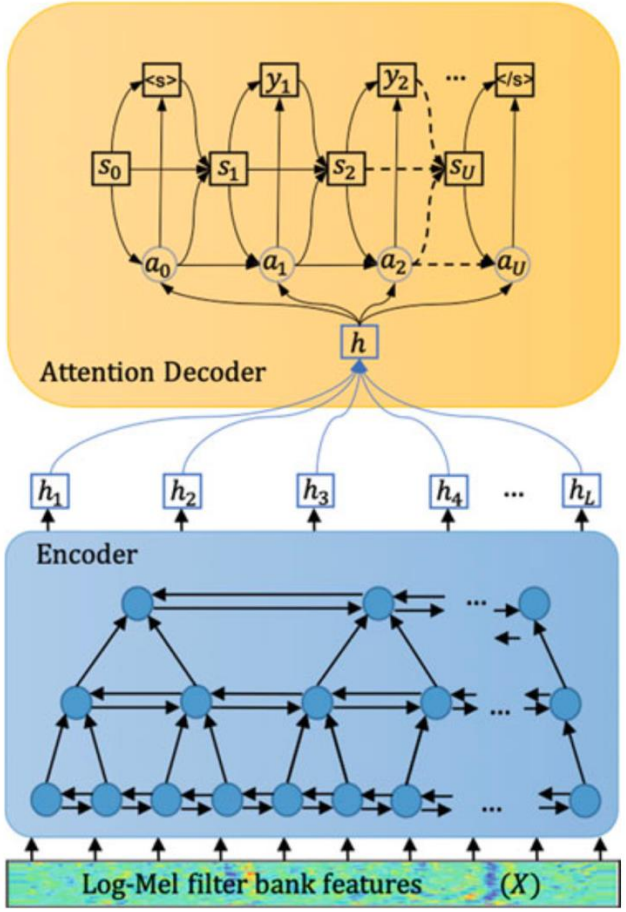
DeepSpeech



DeepSpeech2



Listen, Attend, and Spell

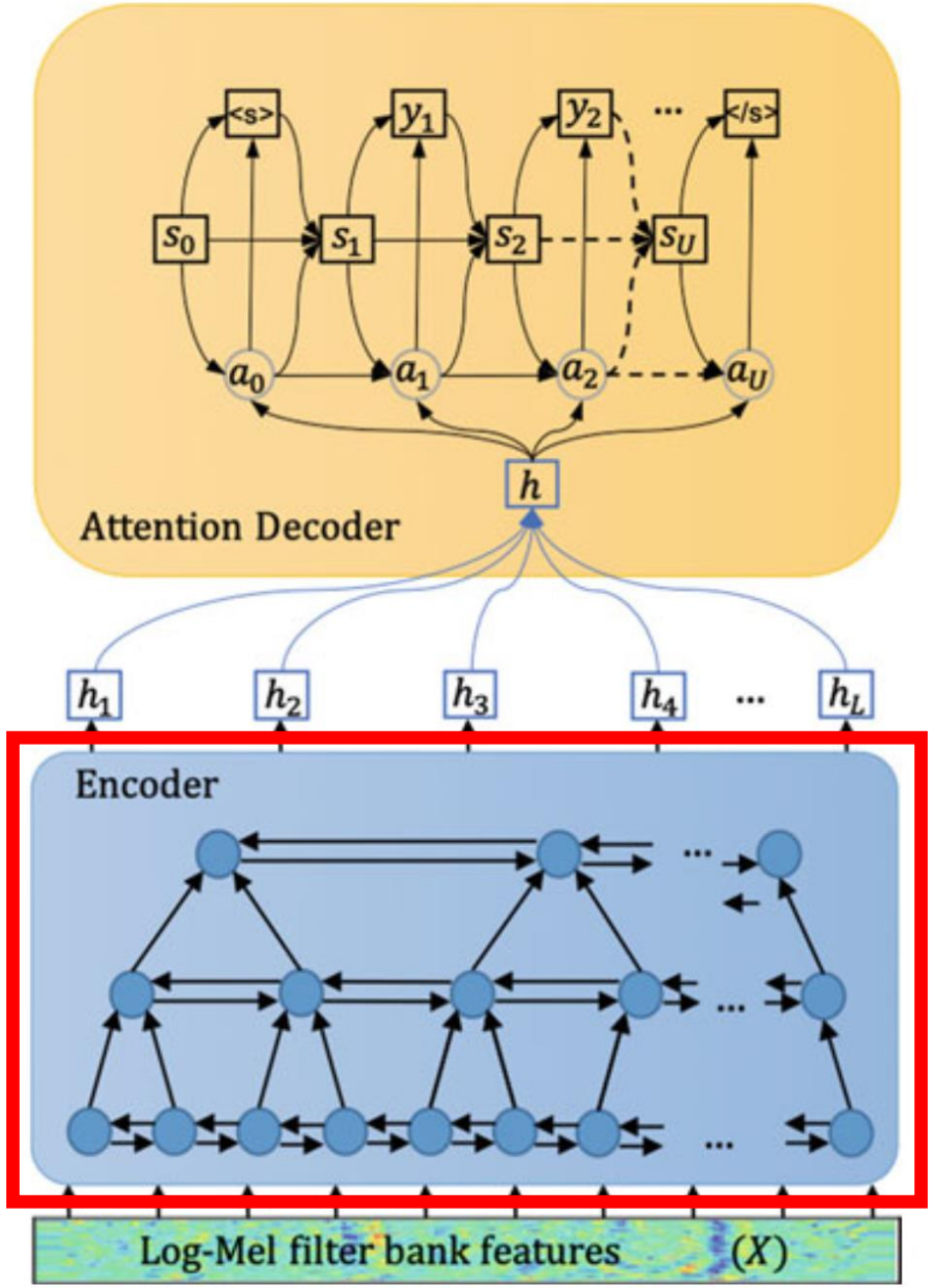


Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Listen, Attend, and Spell

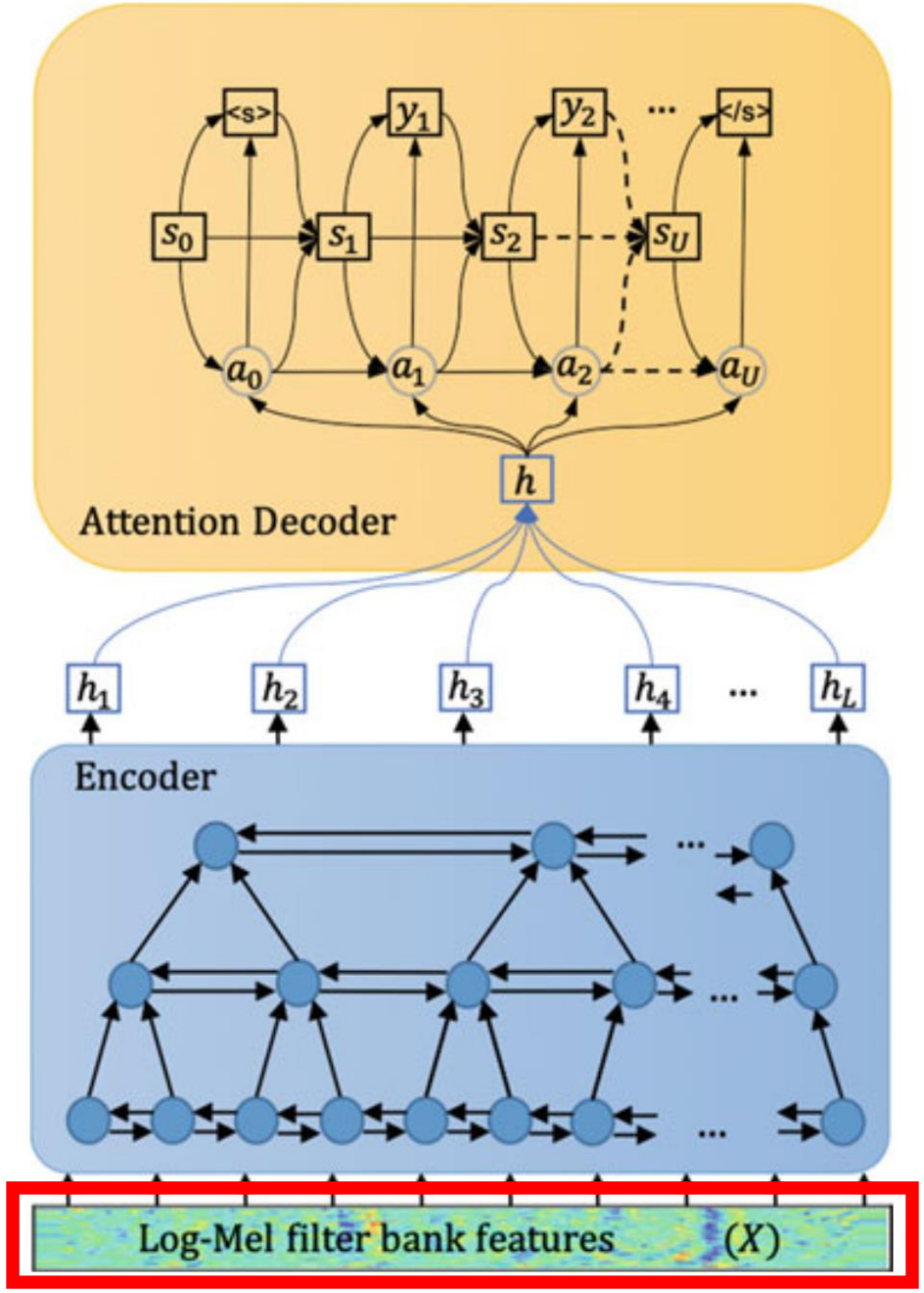
Mimics original paper on sequence to sequence learning with attention where the decoder learns what to attend to in the encoded representation

Pyramid structure reduces number of input time steps



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Listen, Attend, and Spell

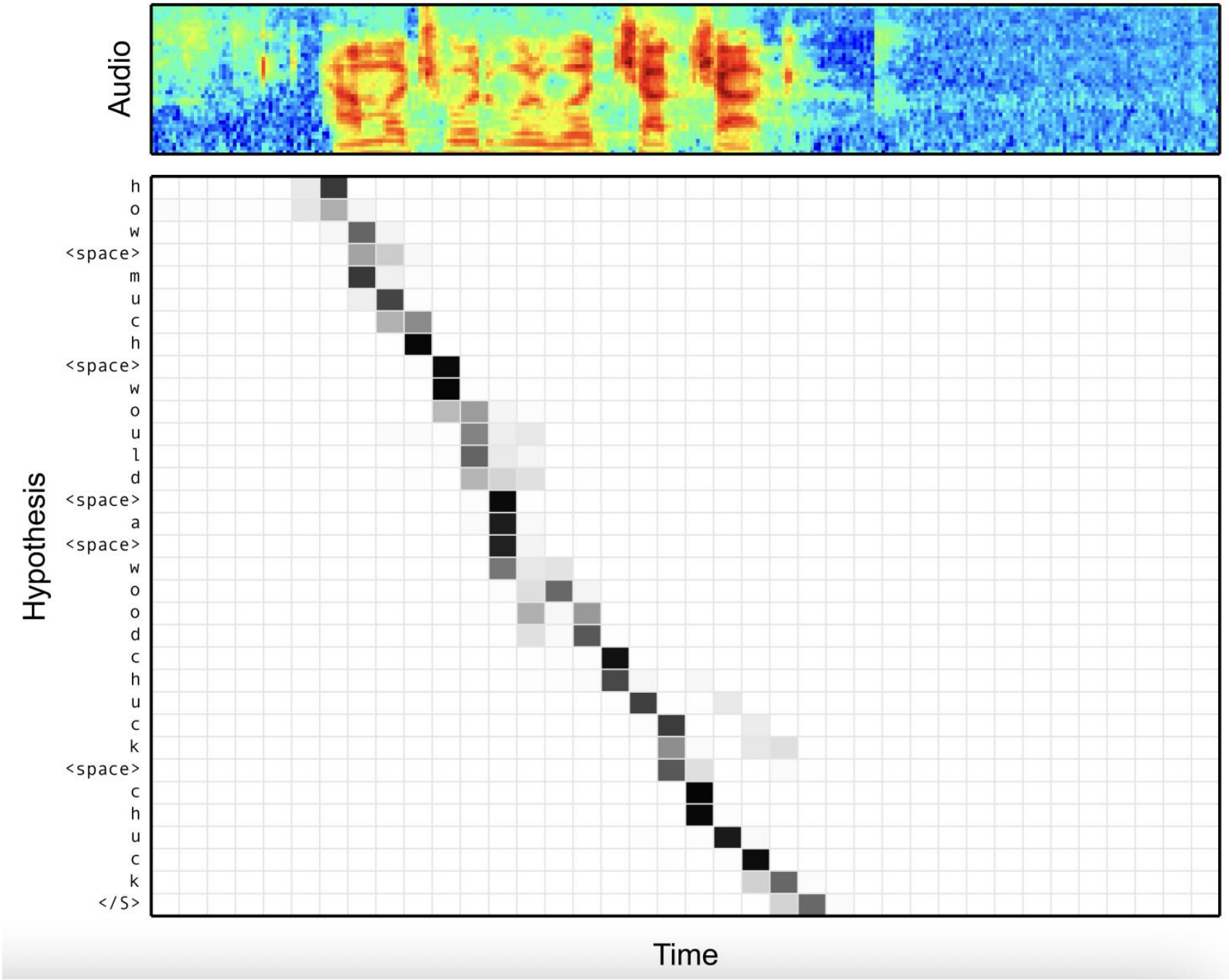


Input: more sophisticated hand-crafted representation than spectrogram

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Result

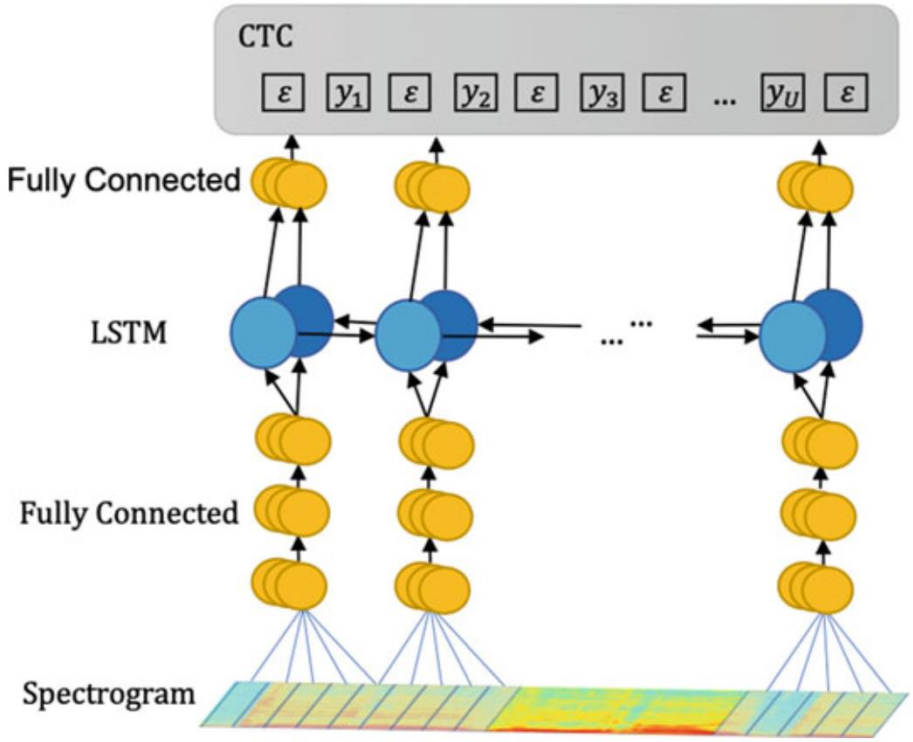
Attention enables visualizing alignment between audio signal and characters



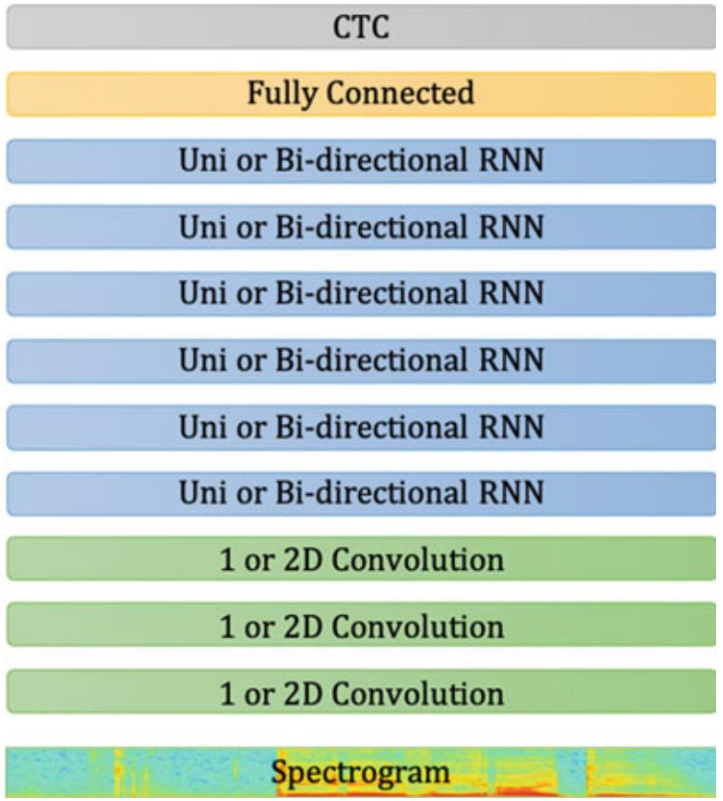
Chan et al. Listen, Attend and Spell. ICASSP 2016.

Popular Methods: Resembles What We Learned

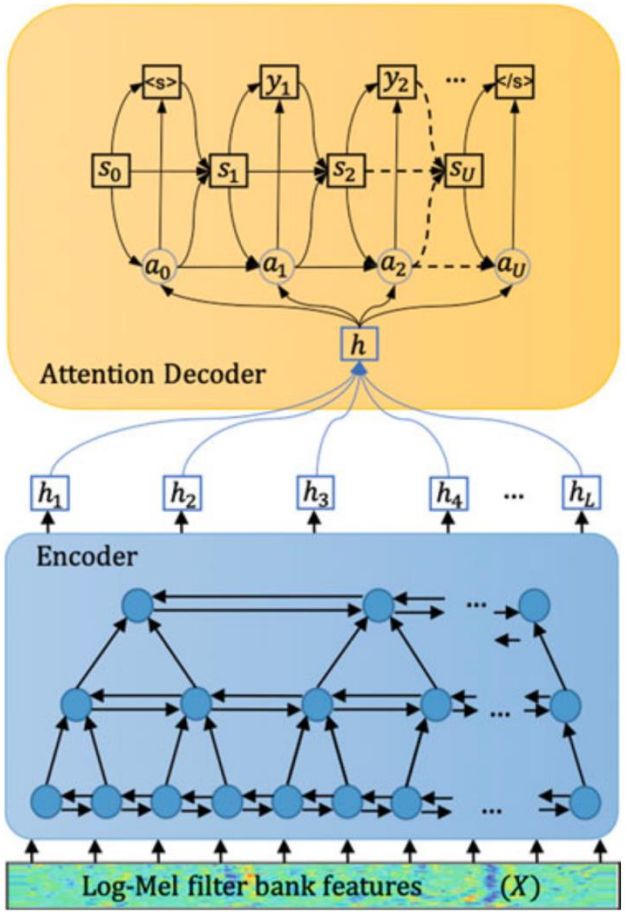
DeepSpeech



DeepSpeech2



Listen, Attend, and Spell



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Trend: Multimodal Models with Audio Support

- Generate audio for videos
- Fake video detection
- Detangling different audio signals in a video
- Visual question answering (question asked audibly)

Today’s Topics

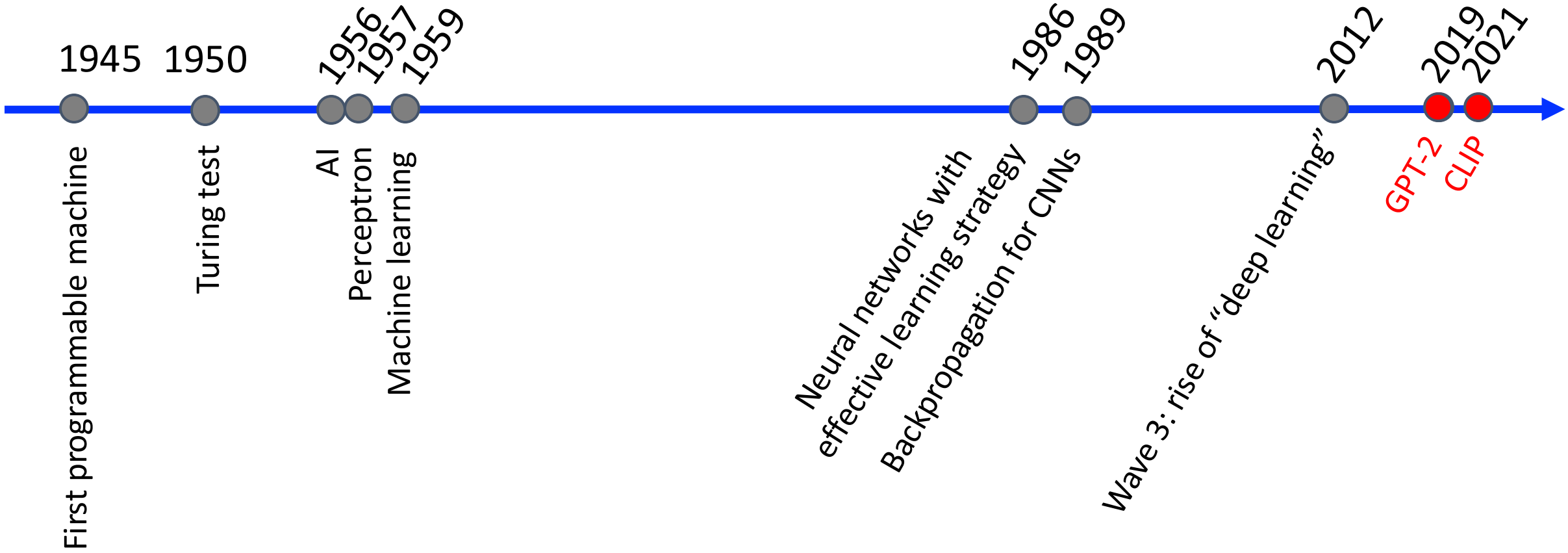
- Other data modalities
- Internet-scale trained models
- Scaling laws
- Lab assignment 3: multimodal task
- Programming tutorial

Shift Around 2019: Internet-Scale Data

Key ingredients for deep learning:

1. Lots of training data
2. Sufficient hardware with modern GPUs
3. Transformer model architecture

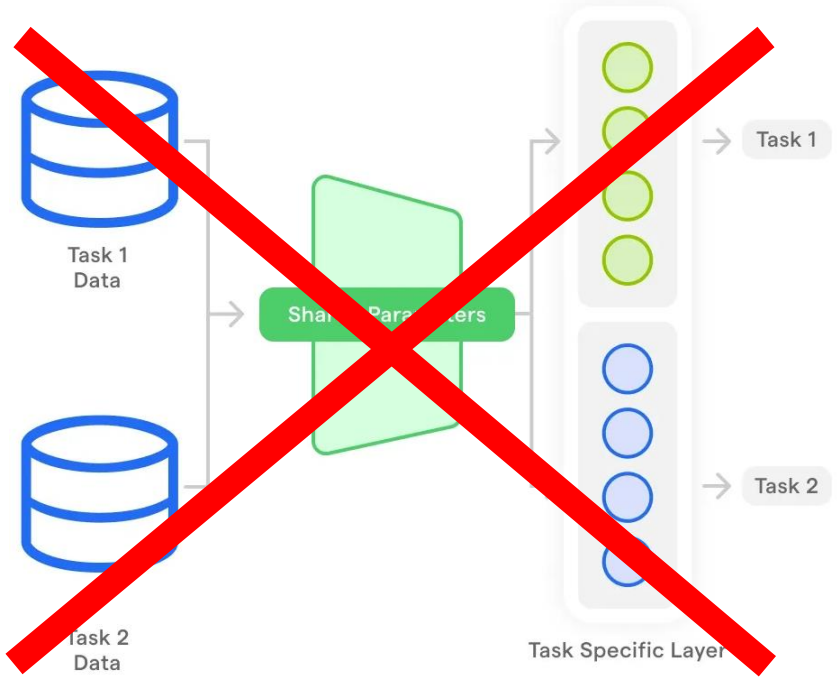
Historical Context



GPT-2: Motivating Argument for NLP

Goal: multi-task learning so a model supports diverse domains and tasks

(e.g., recall fine-tuning with special symbols for BERT and GPT and LXMERT)



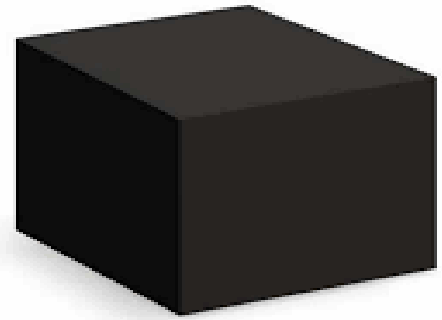
https://botpenguin.com/glossary/multi-task-learning

Idea: use language for diverse domains and tasks

(i.e., no architectural change or special fine-tuning symbols; this is called zero-shot learning)

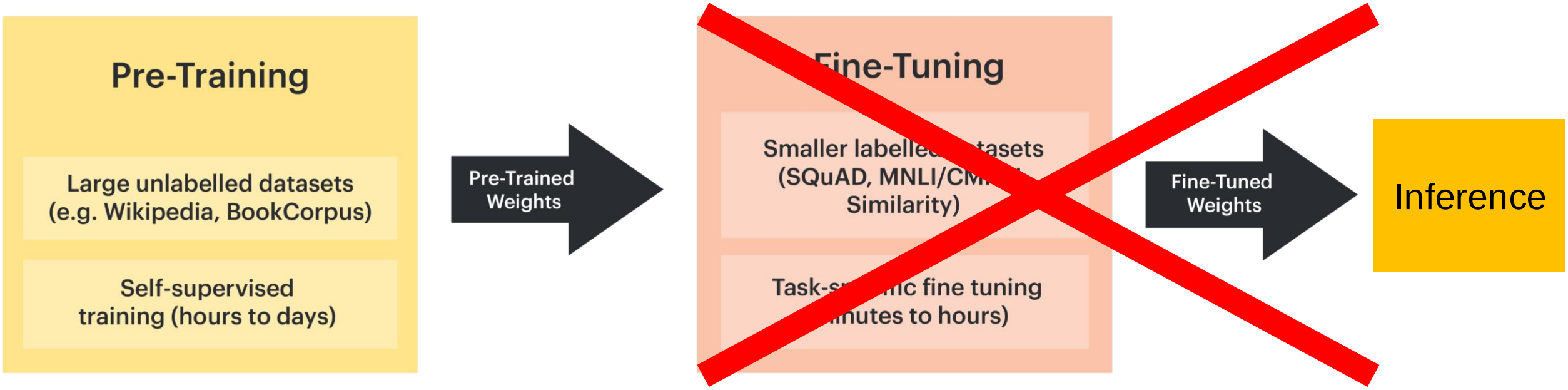
e.g., Input string "Translate to French: [english text]" →

e.g., Input string "Summarize: [text]; TL;DR:" →



Radford et al. Language Models are Unsupervised Multitask Learners. OpenAI Blog 2019.

GPT-2 Approach: Contain Supervised Fine-Tuning Objectives Within Pre-Training Objective



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

GPT-2: Unsupervised Multi-Task Learning

- Learn from natural language on the Internet, since it already represents diverse domains and tasks
- Which source?
  -  Common Crawl: free monthly scrapes of Internet since 2008

    from California non-profit organization (petabytes in size = ~10^{15} bytes)
    - Full dataset poor for training (e.g., spam, broken links, biased content)
  -  reddit: they used a subset of human-curated content in outbound links validated as high quality through “3 karma” ratings
    - 45 million links through Dec 2017, de-duplicated and pre-processed
    - Result: 8M documents = 40 GB

Radford et al. Language Models are Unsupervised Multitask Learners. OpenAI Blog 2019.

GPT-2: Architecture

Modified GPT:

- Layer norm layers re-positioned and added
- Parameter initialization modified
- Vocabulary expanded to 50,257 tokens
- Context size increased from 512 to 1,024 tokens
- More layers (i.e., 48 vs 12)
- More parameters (1,542M vs 124M*)

Architecture code is open-source (no training code)

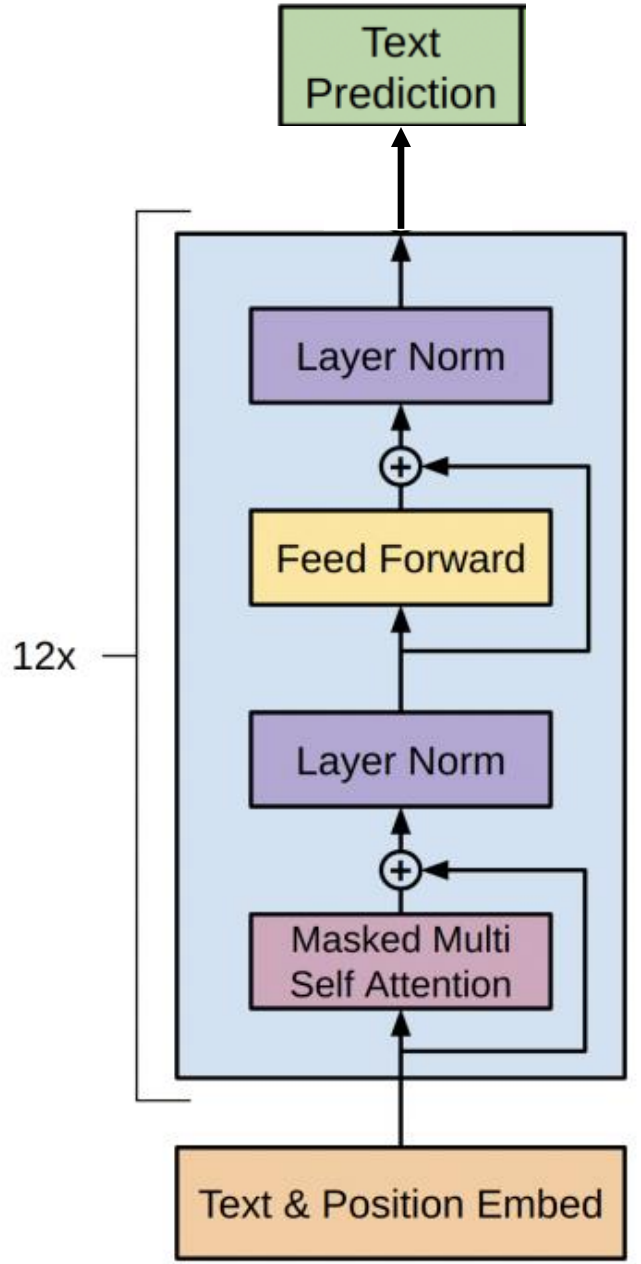


Image source: Radford et al. Technical Report 2018.
Radford et al. Language Models are Unsupervised Multitask Learners. OpenAI Blog 2019.

* Incorrect # reported in paper

GPT-2: Training

Same objective as GPT (predict next word)

Learning rate tuned manually based on results from training on 5% of training data

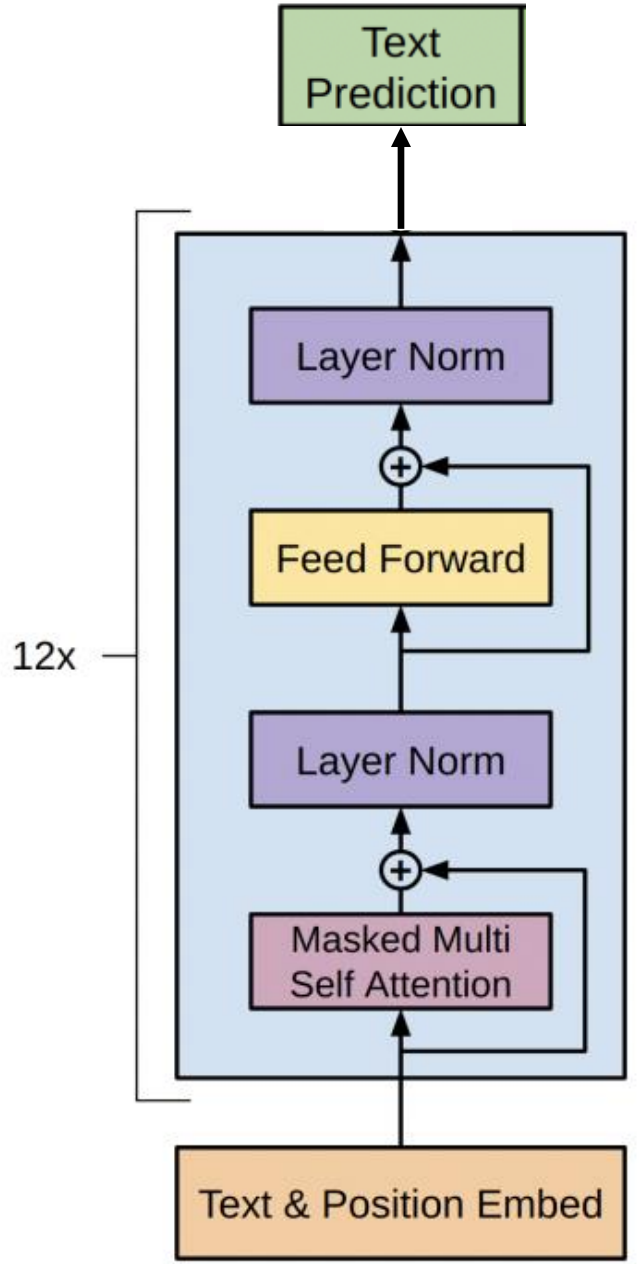
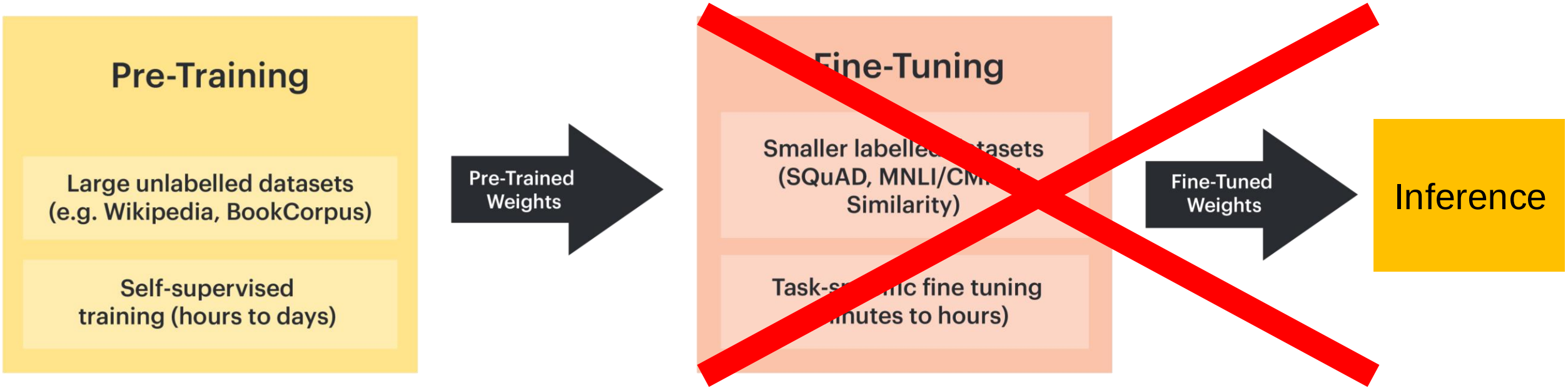


Image source: Radford et al. Technical Report 2018.
Radford et al. Language Models are Unsupervised Multitask Learners. OpenAI Blog 2019.

GPT-2 Experiments: Can Model Generalize to Novel Tasks and Domains Without Fine-Tuning?



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

GPT-2 Experimental Findings

Achieved state-of-the-art performance on 7 of 8 tested NLP dataset challenges in zero-shot setting

Acknowledgments: “Thanks to everyone who wrote the text, shared the links, and upvoted the content in WebText. Many millions of people were involved in creating the data that GPT-2 was trained on.”

Radford et al. Language Models are Unsupervised Multitask Learners. OpenAI Blog 2019.

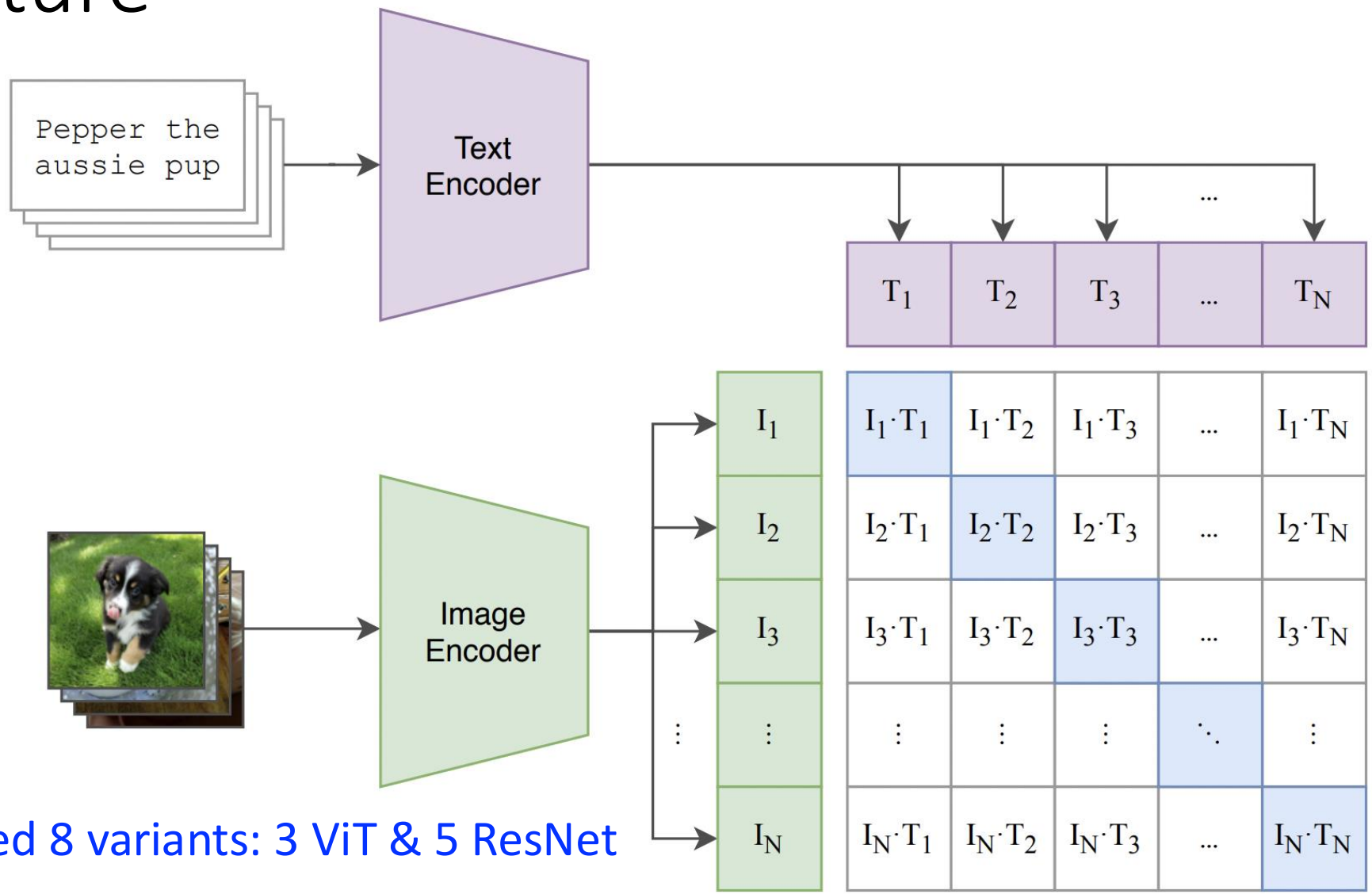
CLIP: Done Again by Radford and OpenAI Team

Named after the proposed technique: Contrastive Language Image Pre-training

> Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

Novelty: image analysis models trained with natural language supervision using the vast amounts of publicly available data on the Internet

CLIP Architecture



Tried 8 variants: 3 ViT & 5 ResNet

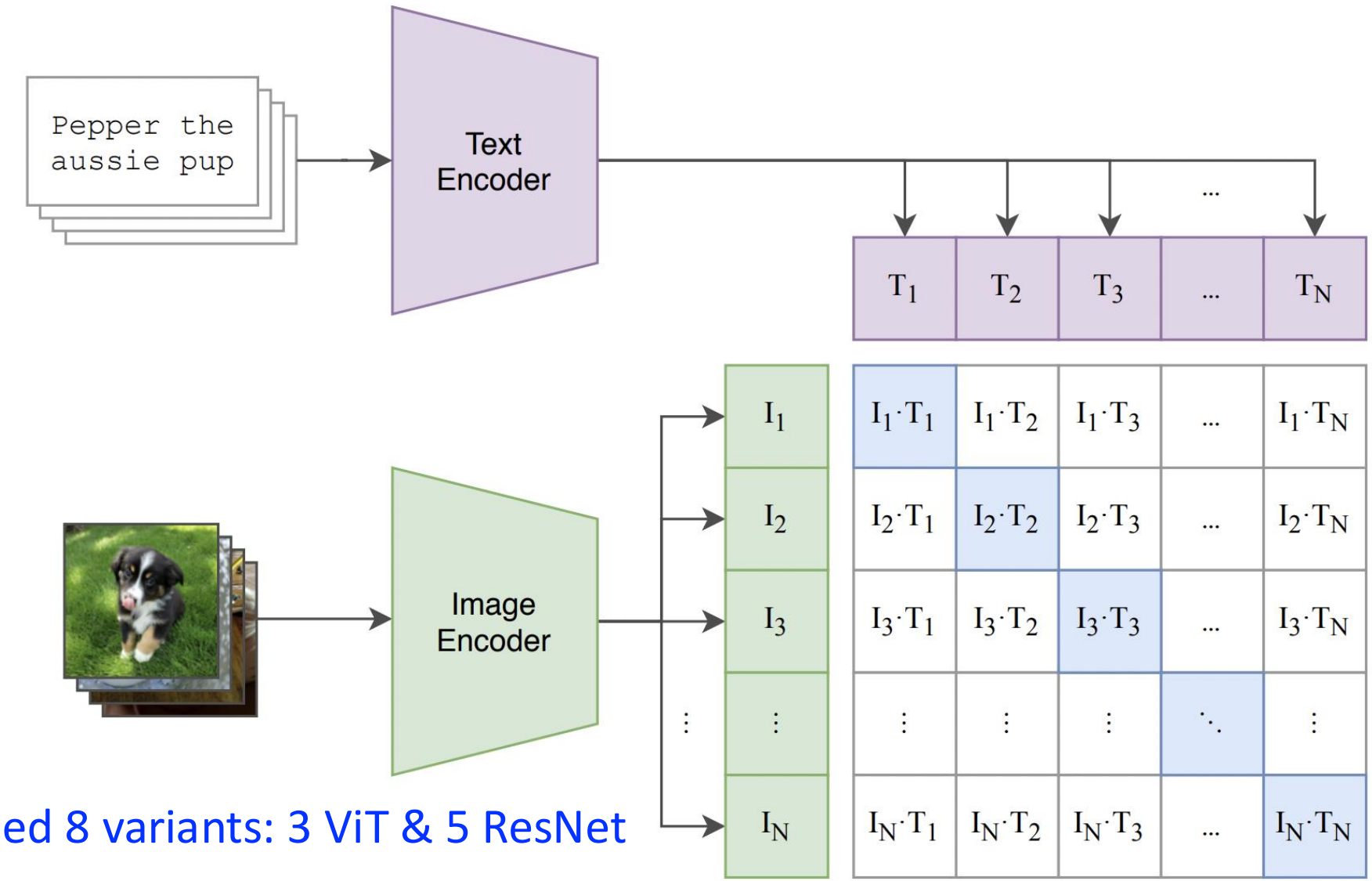
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021

CLIP Training

Task: predict which image-text pairs match

Data: 400 million image-text pairs from Internet from 500,000 queries (e.g., words occurring 100+ times in English version of Wikipedia and all WordNet synonyms)

Training: with largest ResNet, took 18 days on 592 V100 GPUs; with largest ViT, took 12 days on 256 V100 GPUs

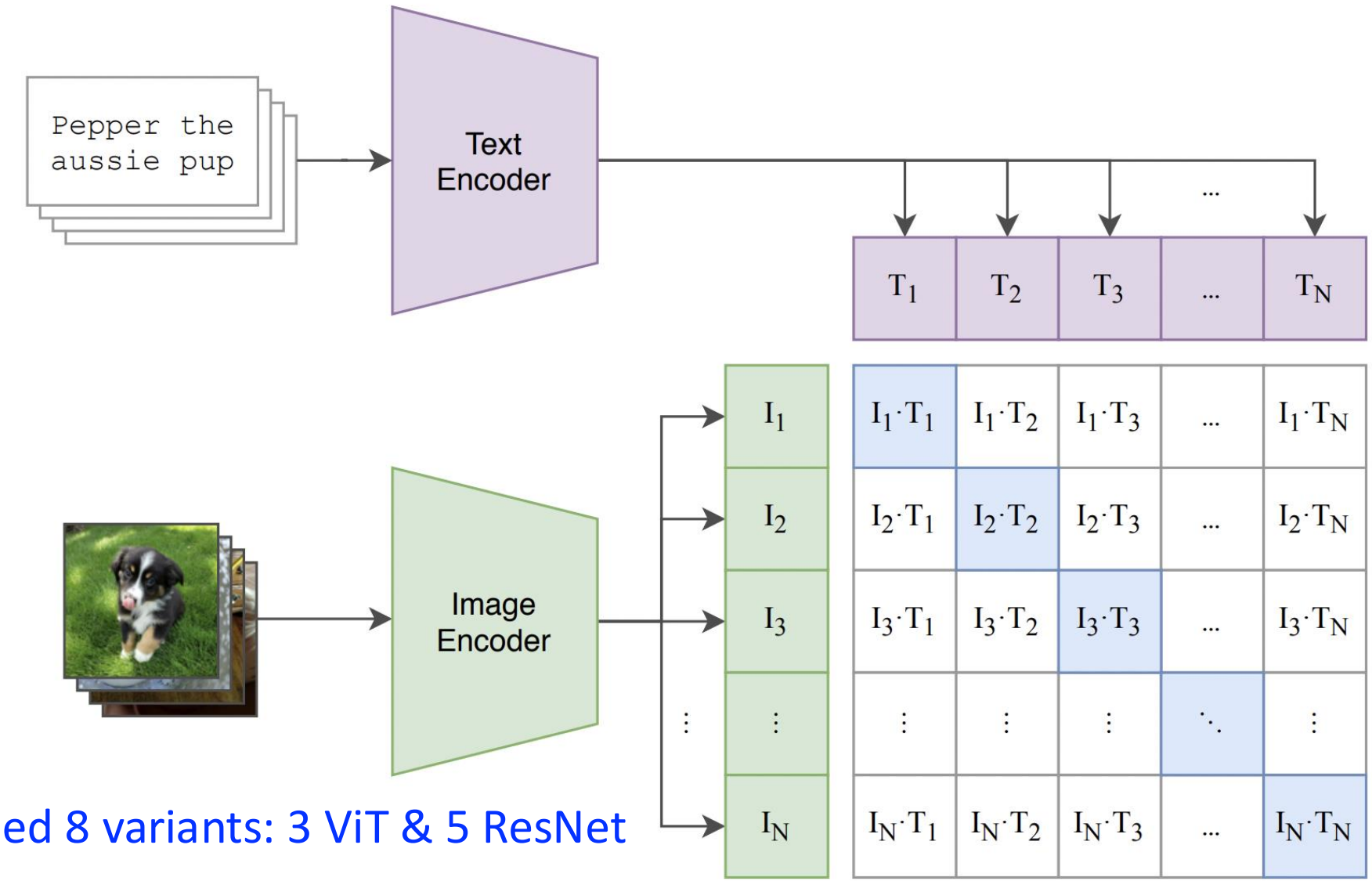


Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021

CLIP Training

- Learns feature embeddings for image and text encoders that push correct image-text pairs together and incorrect image-text pairs apart.

- Learns nouns, verbs, adjectives, and more!



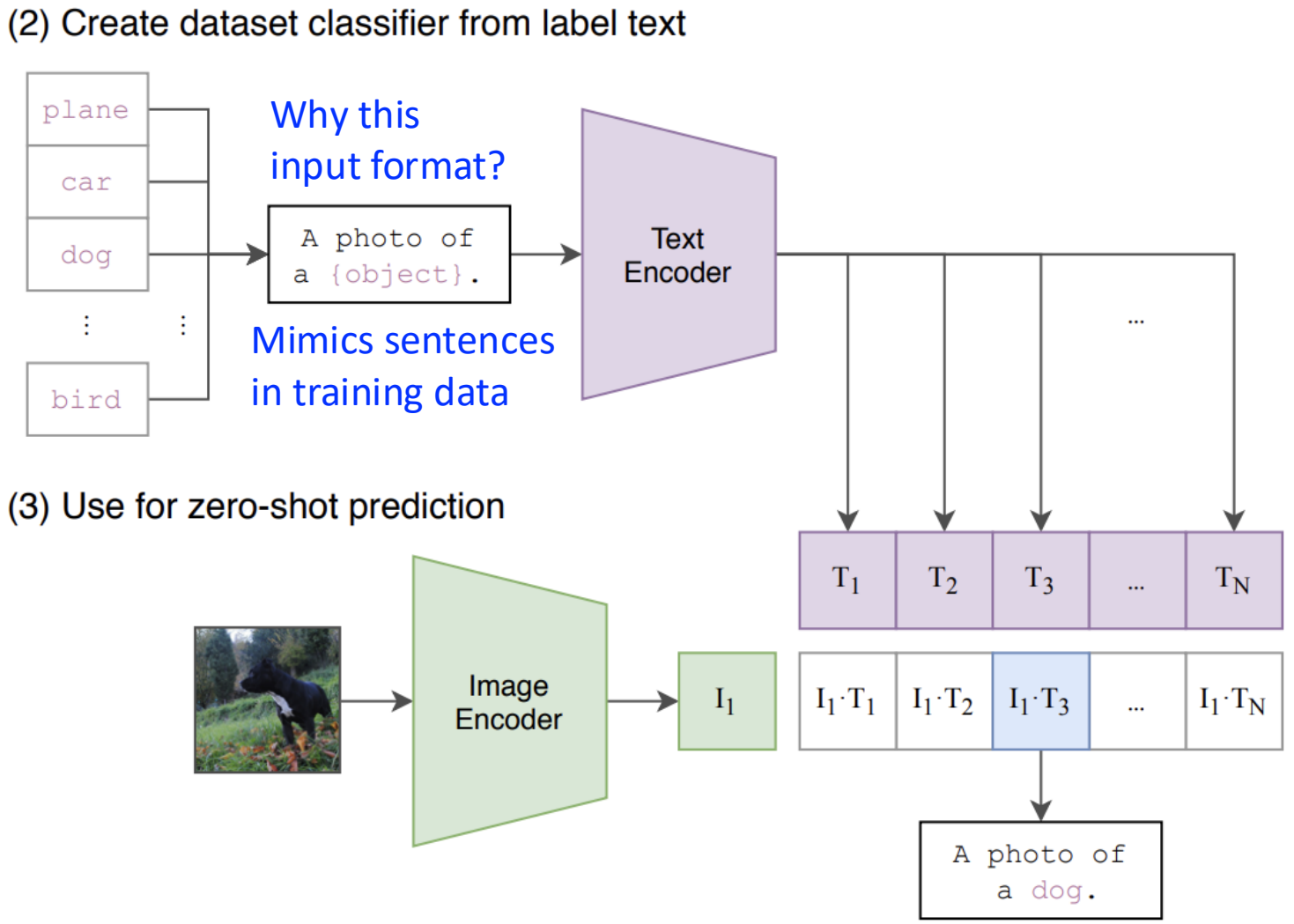
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021

Zero-Shot Performance Evaluated on Over 30 Datasets

CLIP Inference

e.g., zero-shot classification:

1. Compute feature embedding for names of all classes in the dataset by its encoder
2. Compute feature embedding of novel image
3. Compute cosine similarity of each image-text pair embedding
4. Apply softmax to identify most probable match (i.e., highest score)



https://towardsdatascience.com/understanding-zero-shot-learning-making-ml-more-human-4653ac35ccab

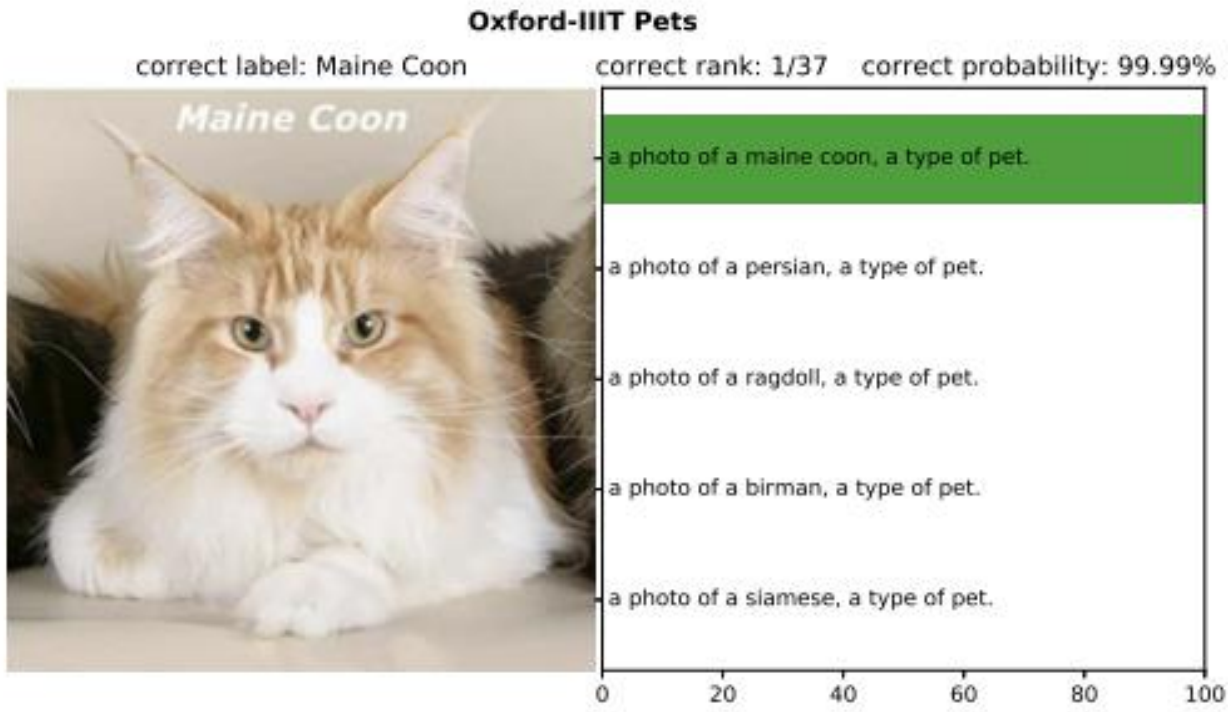
CLIP Benchmark Datasets

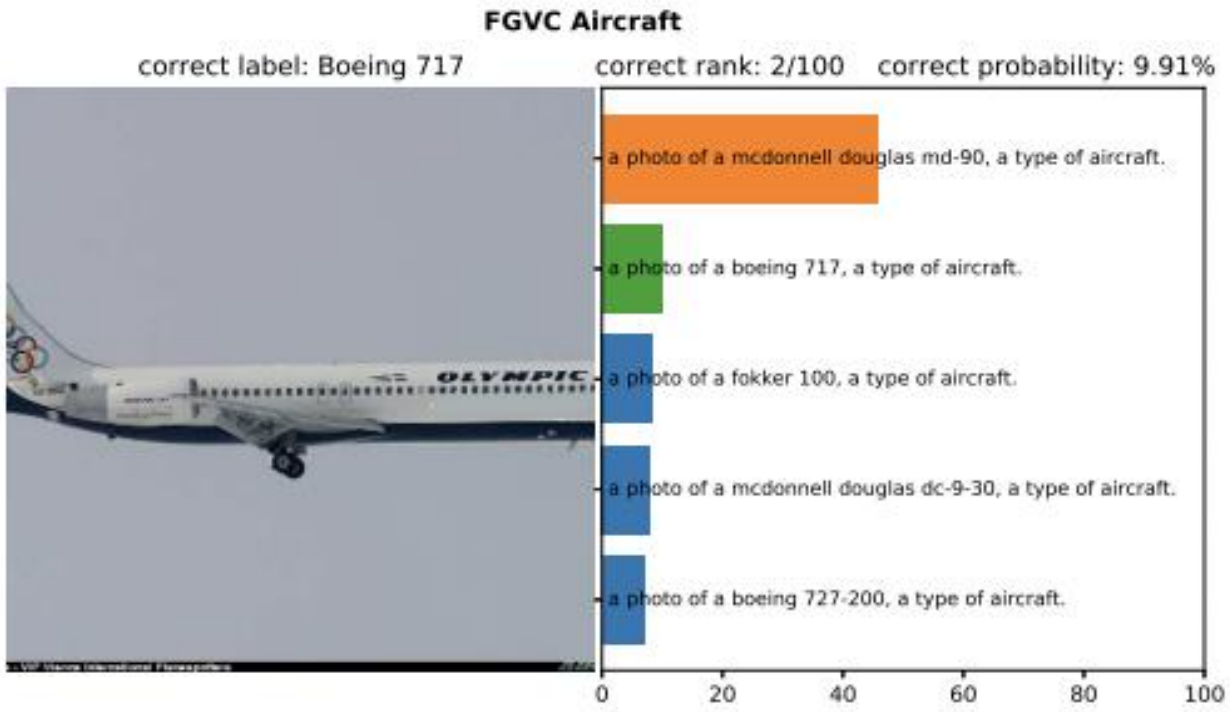
Subset of datasets shown here:

Classification evaluation spanned fine-grained classification (e.g., food, bird, aircraft, and car categories), distribution shifts for ImageNet categories (e.g., corrupted images), and more

Dataset	Classes	Train size	Test size	Evaluation metric
Food-101	102	75,750	25,250	accuracy
CIFAR-10	10	50,000	10,000	accuracy
CIFAR-100	100	50,000	10,000	accuracy
Birdsnap	500	42,283	2,149	accuracy
SUN397	397	19,850	19,850	accuracy
Stanford Cars	196	8,144	8,041	accuracy
FGVC Aircraft	100	6,667	3,333	mean per class
Pascal VOC 2007 Classification	20	5,011	4,952	11-point mAP
Describable Textures	47	3,760	1,880	accuracy
Oxford-IIIT Pets	37	3,680	3,669	mean per class
Caltech-101	102	3,060	6,085	mean-per-class
Oxford Flowers 102	102	2,040	6,149	mean per class
MNIST	10	60,000	10,000	accuracy
Facial Emotion Recognition 2013	8	32,140	3,574	accuracy
STL-10	10	1000	8000	accuracy
EuroSAT	10	10,000	5,000	accuracy
RESISC45	45	3,150	25,200	accuracy
GTSRB	43	26,640	12,630	accuracy
KITTI	4	6,770	711	accuracy
Country211	211	43,200	21,100	accuracy
PatchCamelyon	2	294,912	32,768	accuracy
UCF101	101	9,537	1,794	accuracy
Kinetics700	700	494,801	31,669	mean(top1, top5)
CLEVR Counts	8	2,000	500	accuracy
Hateful Memes	2	8,500	500	ROC AUC
Rendered SST2	2	7,792	1,821	accuracy
ImageNet	1000	1,281,167	50,000	accuracy

CLIP: Fine-Grained Classification Predictions



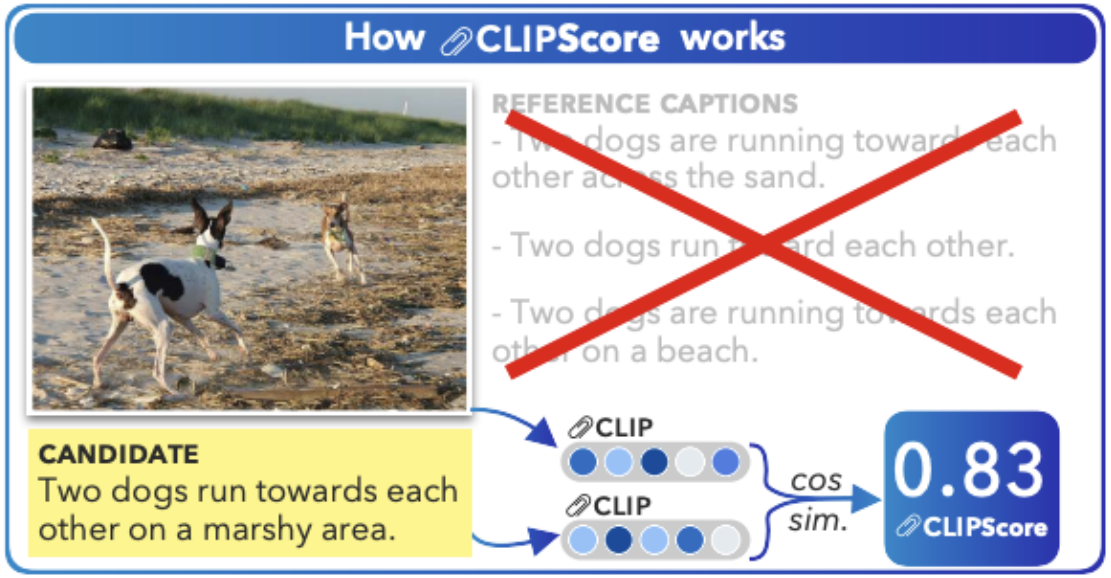


Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021

CLIP Is WIDELY Used; e.g.,

CLIPScore:
A Reference-free Evaluation Metric for Image Captioning

Jack Hessel[†] **Ari Holtzman**[‡] **Maxwell Forbes**[‡] **Ronan Le Bras**[†] **Yejin Choi**[†‡]
[†]Allen Institute for AI
[‡]Paul G. Allen School of Computer Science & Engineering, University of Washington
{jackh,ronanlb}@allenai.org {ahai,mbforbes,yejin}@cs.washington.edu

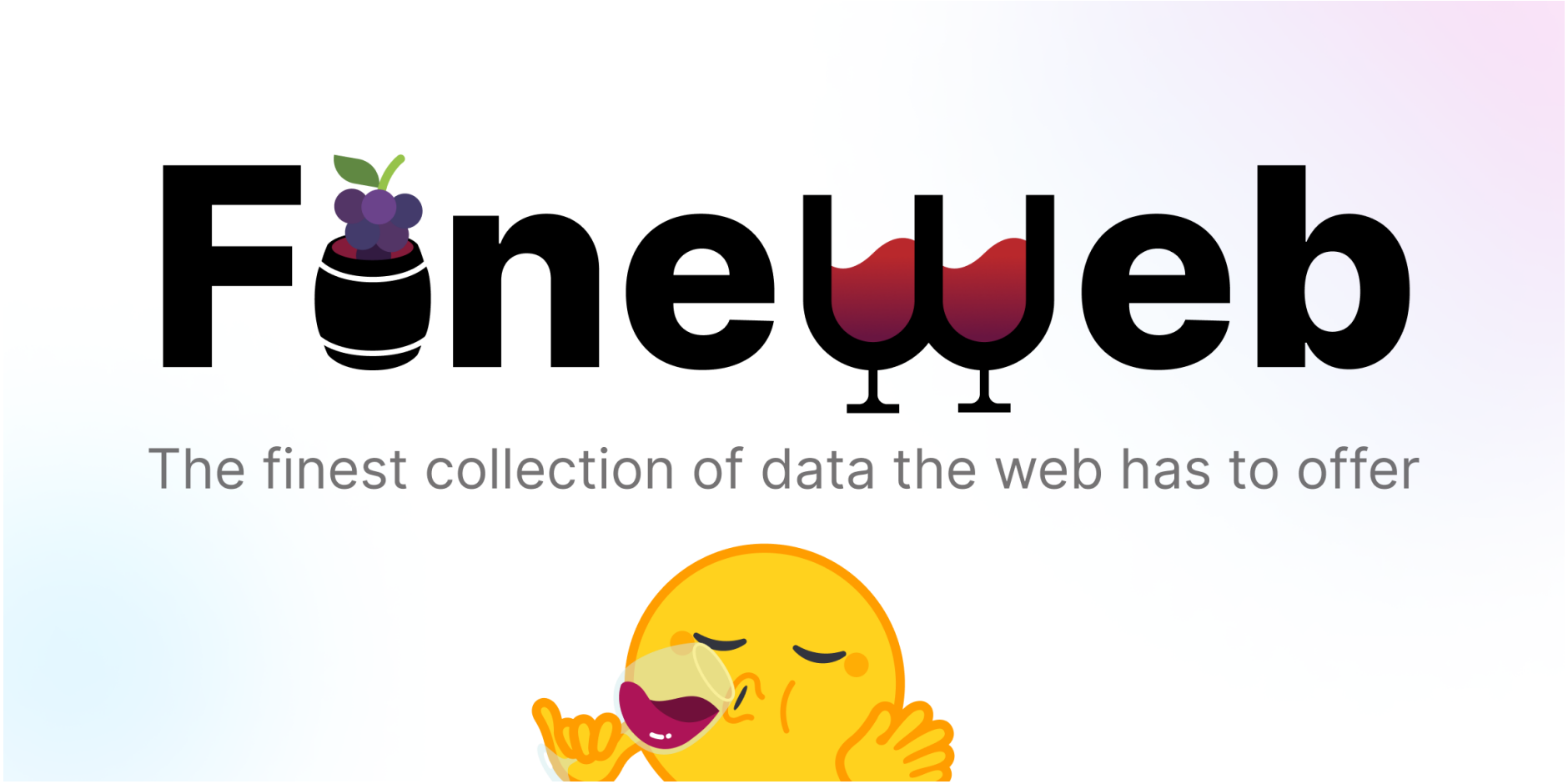


Many Other Internet-Scale Datasets Exist; e.g.,

	Tokens	Open source	Curated data sources	Deduplication level
SlimPajama	627B	Yes	Yes	Extensive
RedPajama	1.21T	Yes	Yes	Partial
RefinedWeb-600B	600B	Yes	No	Extensive
RefinedWeb-5T	5T	No	No	Extensive
LLaMA	1.4T	No	Yes	Partial
MPT	1T	No	Yes	Partial
MassiveText	1.4T	No	Yes	Extensive

https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama

Many Other Internet-Scale Datasets Exist; e.g.,



https://huggingface.co/datasets/HuggingFaceFW/fineweb

Today's Topics

- Other data modalities
- Internet-scale trained models
- Scaling laws
- Lab assignment 3: multimodal task
- Programming tutorial

(arXiv 2020)

Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com

Tom Henighan
OpenAI
henighan@openai.com

Tom B. Brown
OpenAI
tom@openai.com

Benjamin Chess
OpenAI
bchess@openai.com

Rewon Child
OpenAI
rewon@openai.com

Scott Gray
OpenAI
scott@openai.com

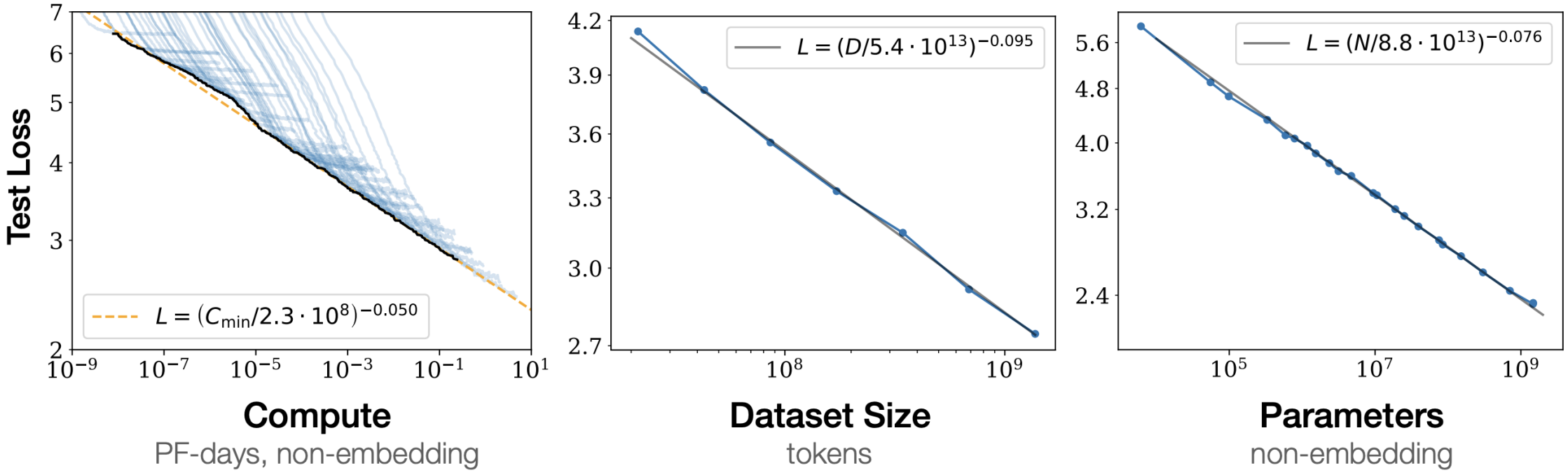
Alec Radford
OpenAI
alec@openai.com

Jeffrey Wu
OpenAI
jeffwu@openai.com

Dario Amodei
OpenAI
damodei@openai.com

Scaling Laws: Empirical Observations

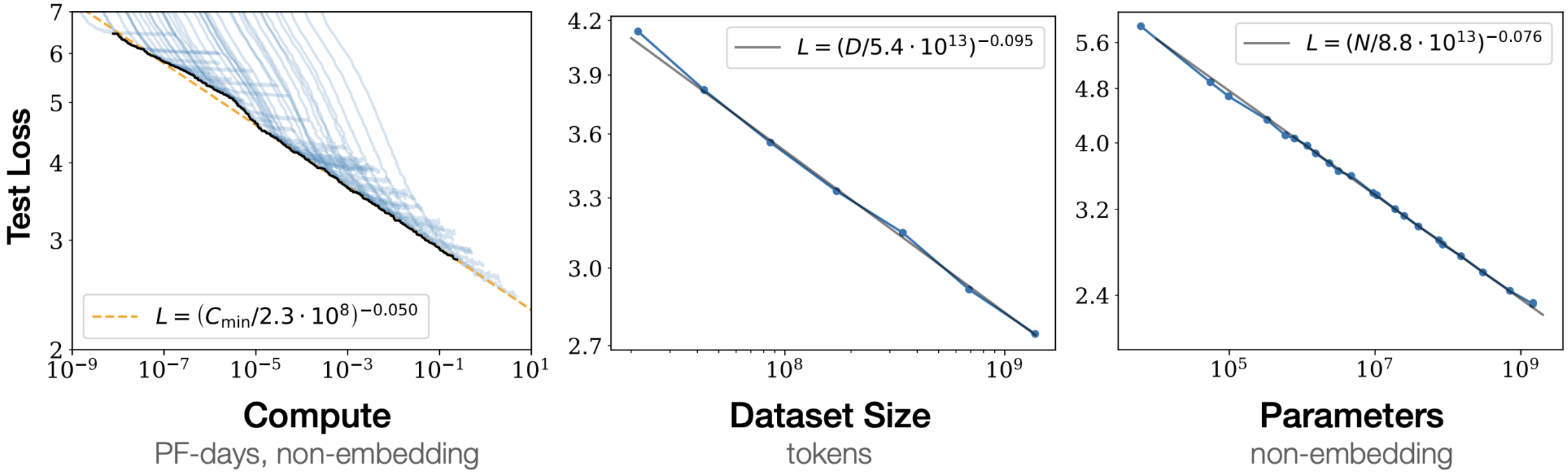
- Paper helped inspire examining how model performance is influenced by (1) Available compute, (2) Training data size, and (3) Model size
- What "law" do you see?



Kaplan et al. Scaling Laws for Neural Language Models. arXiv 2020

Scaling Laws: Empirical Observations

Key observations: (1) Power-law relationship with each factor: changing one causes the other to change proportionally as a power (exponent) of it and (2) Increases to each improves performance smoothly (as long as all three are scaled together)



Kaplan et al. Scaling Laws for Neural Language Models. arXiv 2020

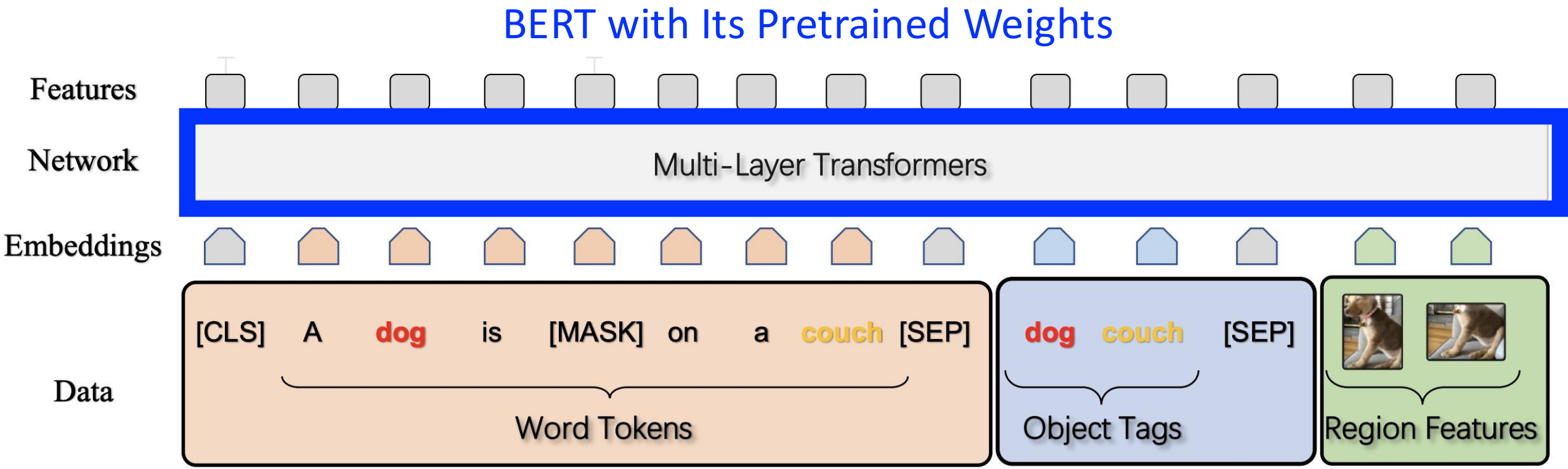
(CVPR 2022)

Scaling Up Vision-Language Pre-training for Image Captioning

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang,
Zicheng Liu, Yumao Lu, Lijuan Wang
Microsoft

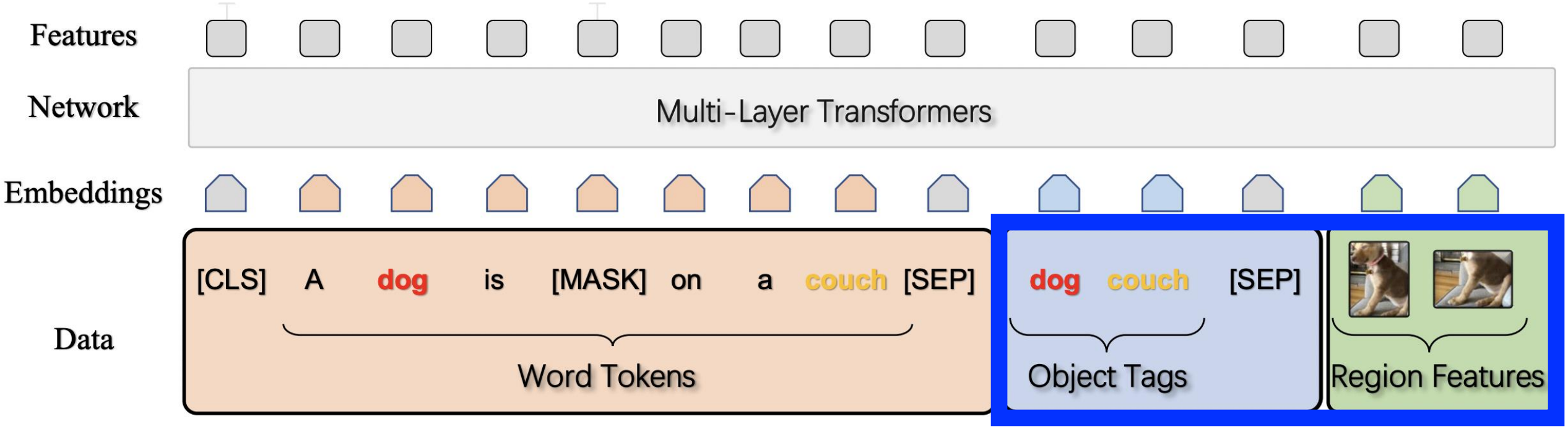
Novelty: first empirical analysis of how a model's image captioning performance is influenced by scaling up vision-language pretraining

Architecture



Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

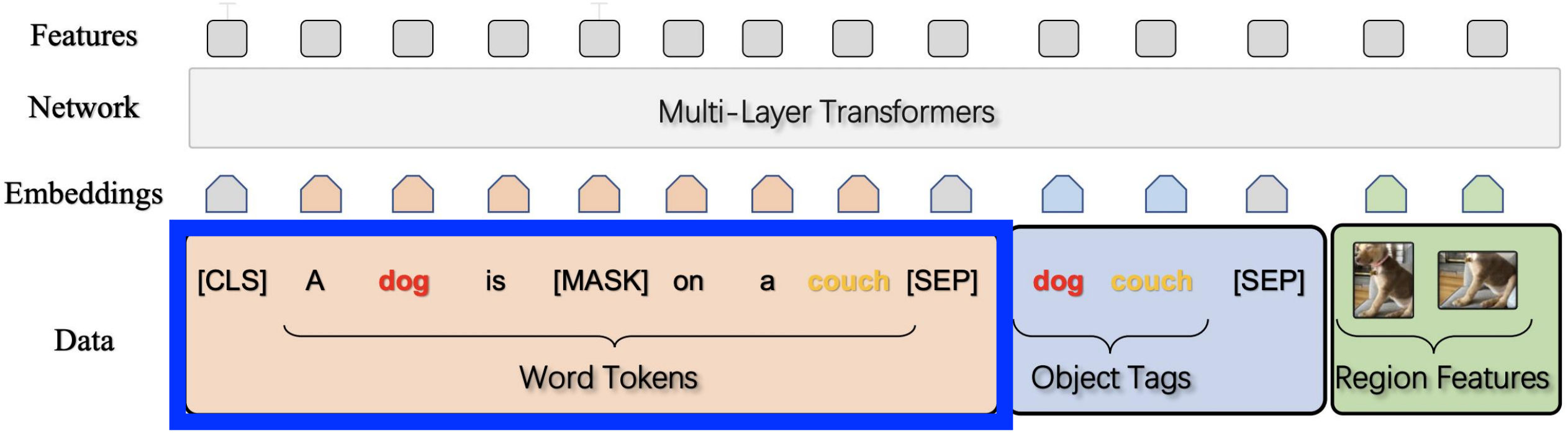
Architecture: Input (Caption + Objects)



Objects and their extracted features from Faster R-CNN

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

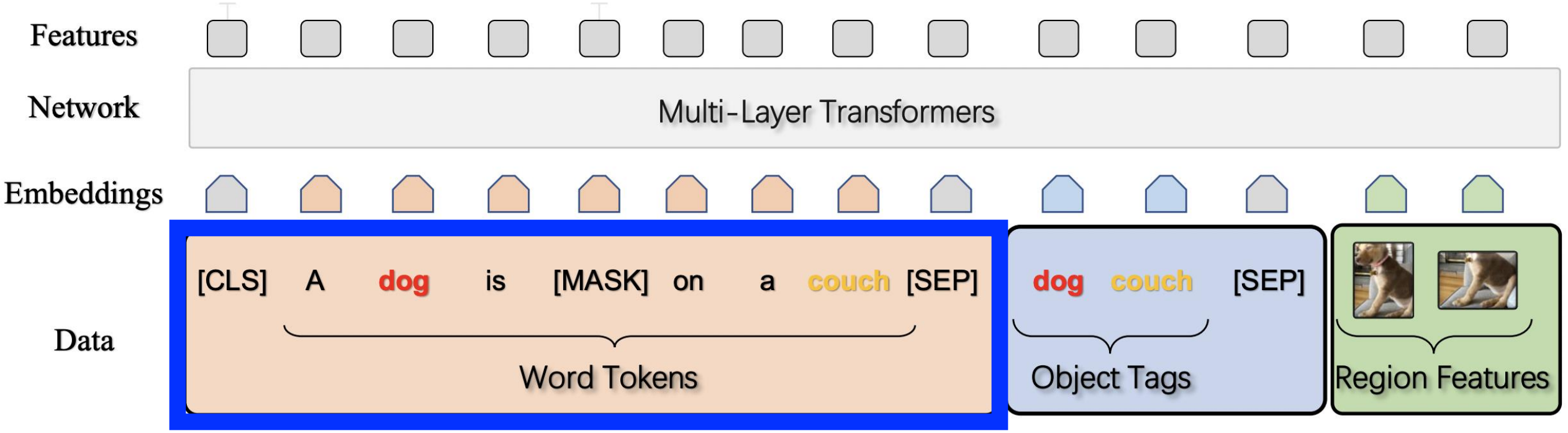
Pretraining Objective: Masked Token Loss



Like BERT, predicts masked tokens for 15% randomly masked
(uses context of surrounding words, image tags, and image features)

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

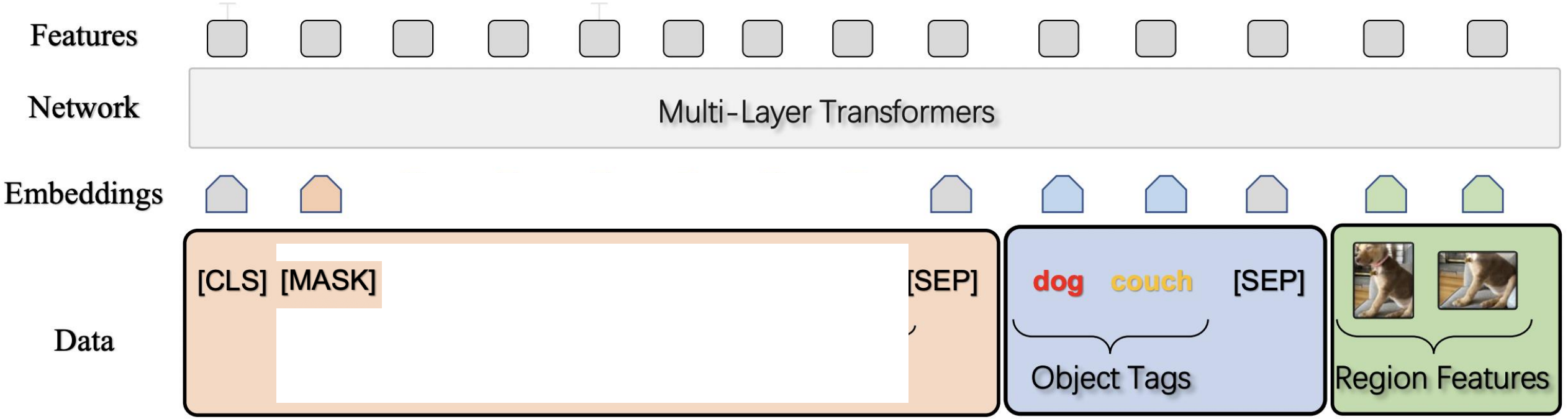
Fine-Tuning Task: Masked Token Loss



Same as pre-training: predict randomly masked tokens

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

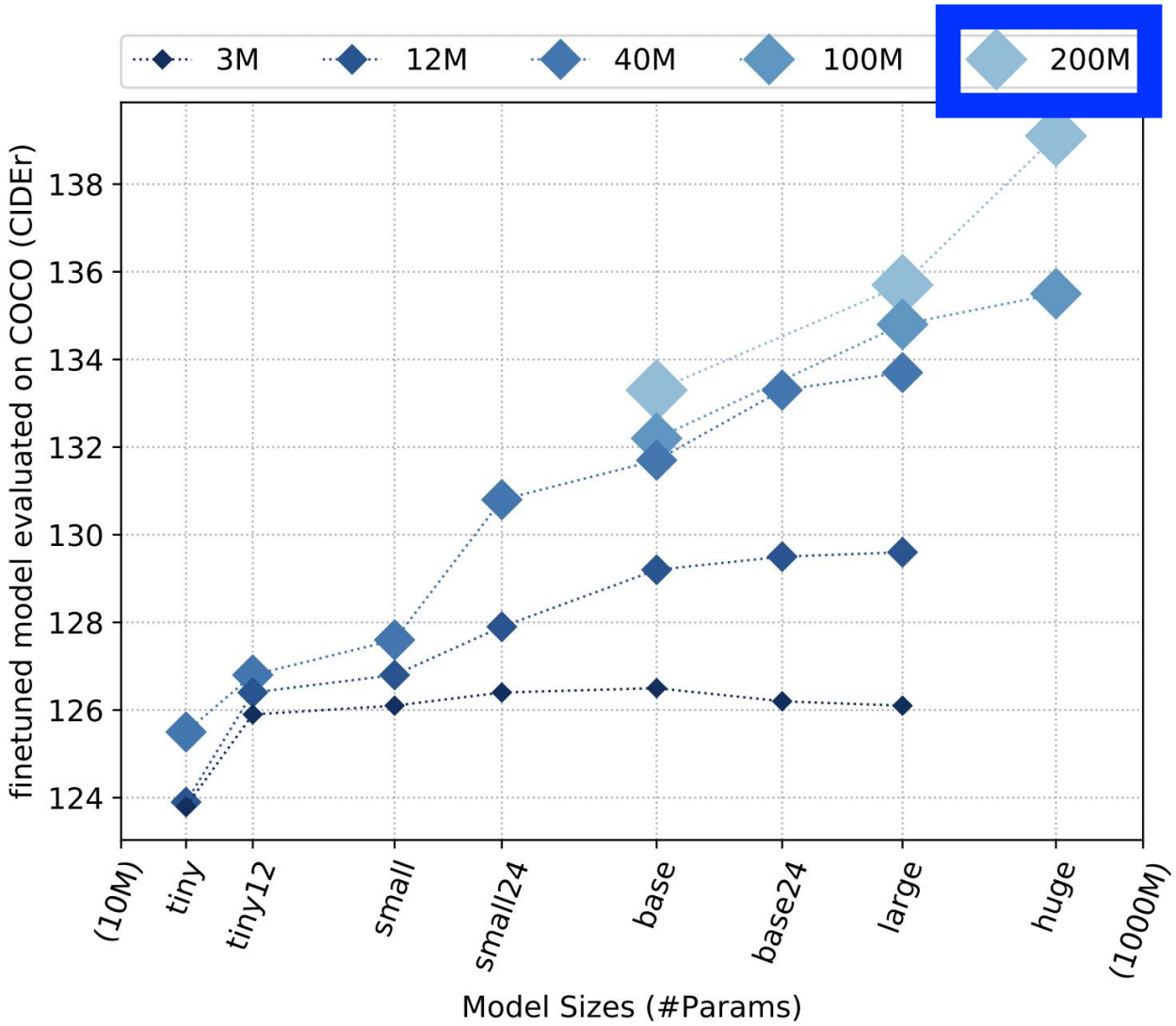
Inference Time



Repeatedly predict a new [MASK] token, incorporating the predicted word into the sequence, until [STOP] is predicted.

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

Influence of Model and Dataset Sizes



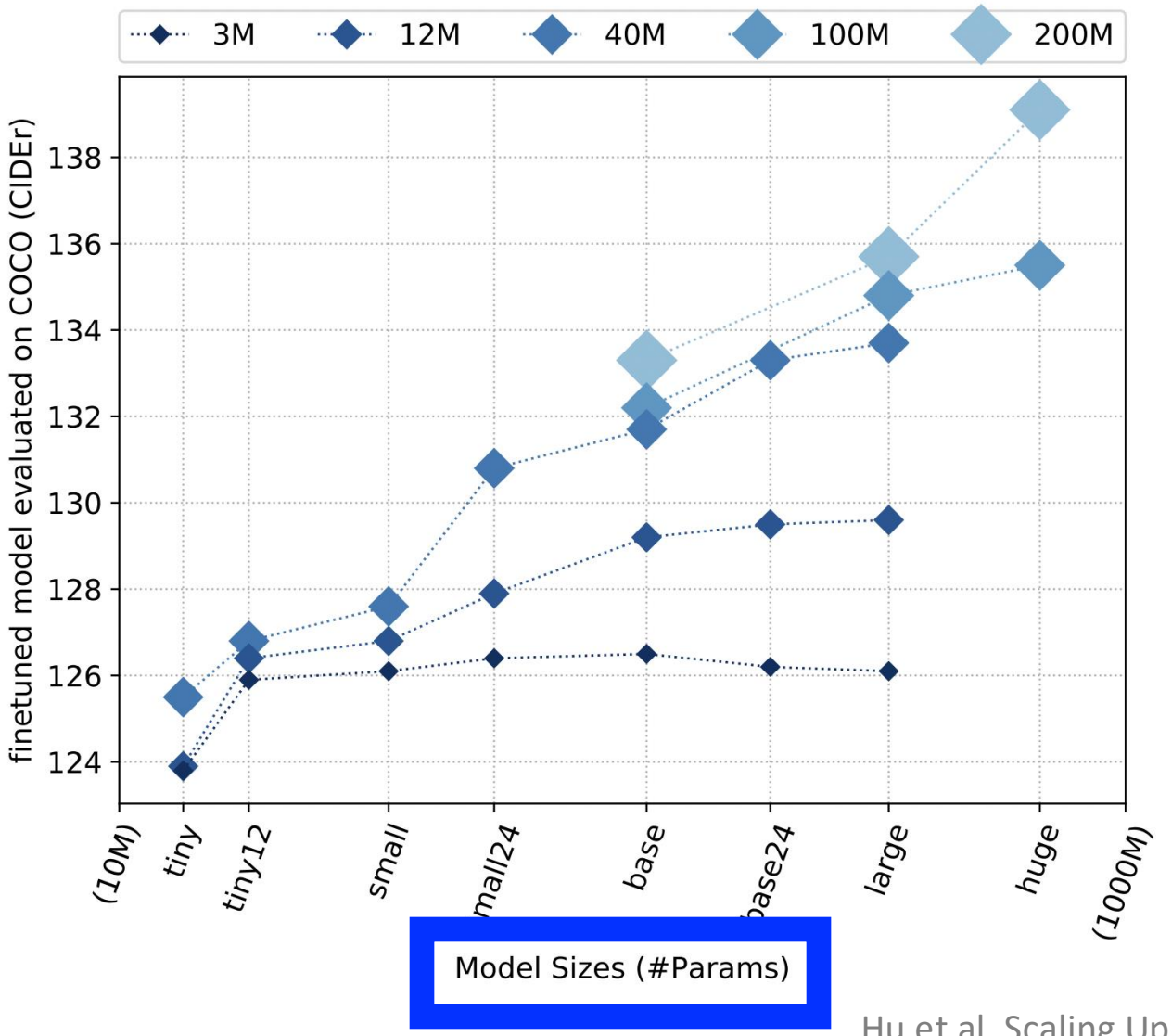
5 pre-training dataset sizes tested, using images from Internet (each with 1 alt text description)

(Most prior work had pre-trained on up to 4M images)

Fine-tuned to target dataset; e.g., COCO-Captions

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022

Influence of Model and Dataset Sizes

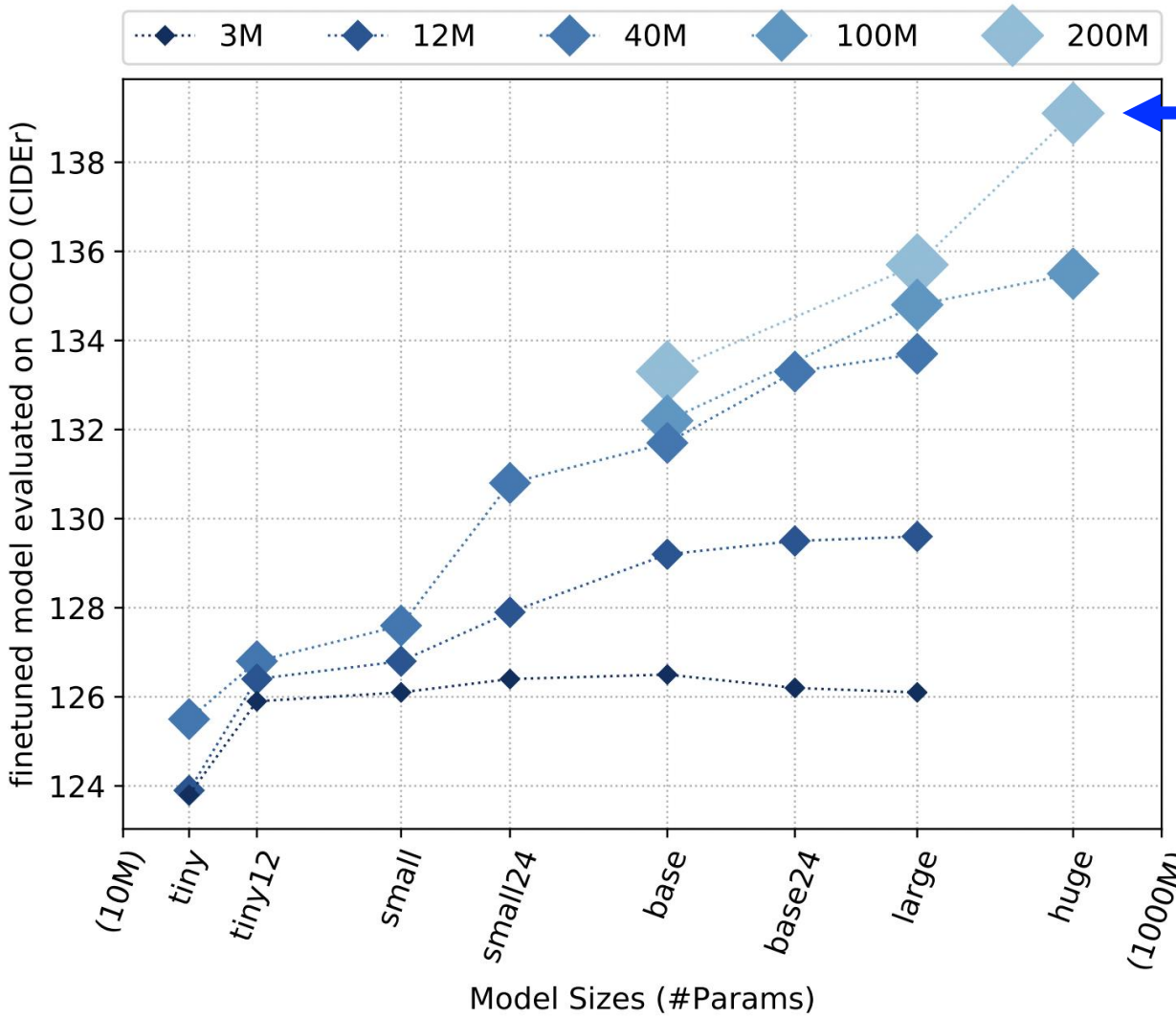


8 model sizes tested on COCO dataset

Model	Layers	Width	MLP	Heads	Param (M)
tiny	6	256	1024	4	13.4
tiny12	12	256	1024	4	18.1
small	12	384	1536	6	34.3
small24	24	384	1536	6	55.6
base	12	768	3072	12	111.7
base24	24	768	3072	12	196.7
large	24	1024	4096	16	338.3
huge	32	1280	5120	16	675.4

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022

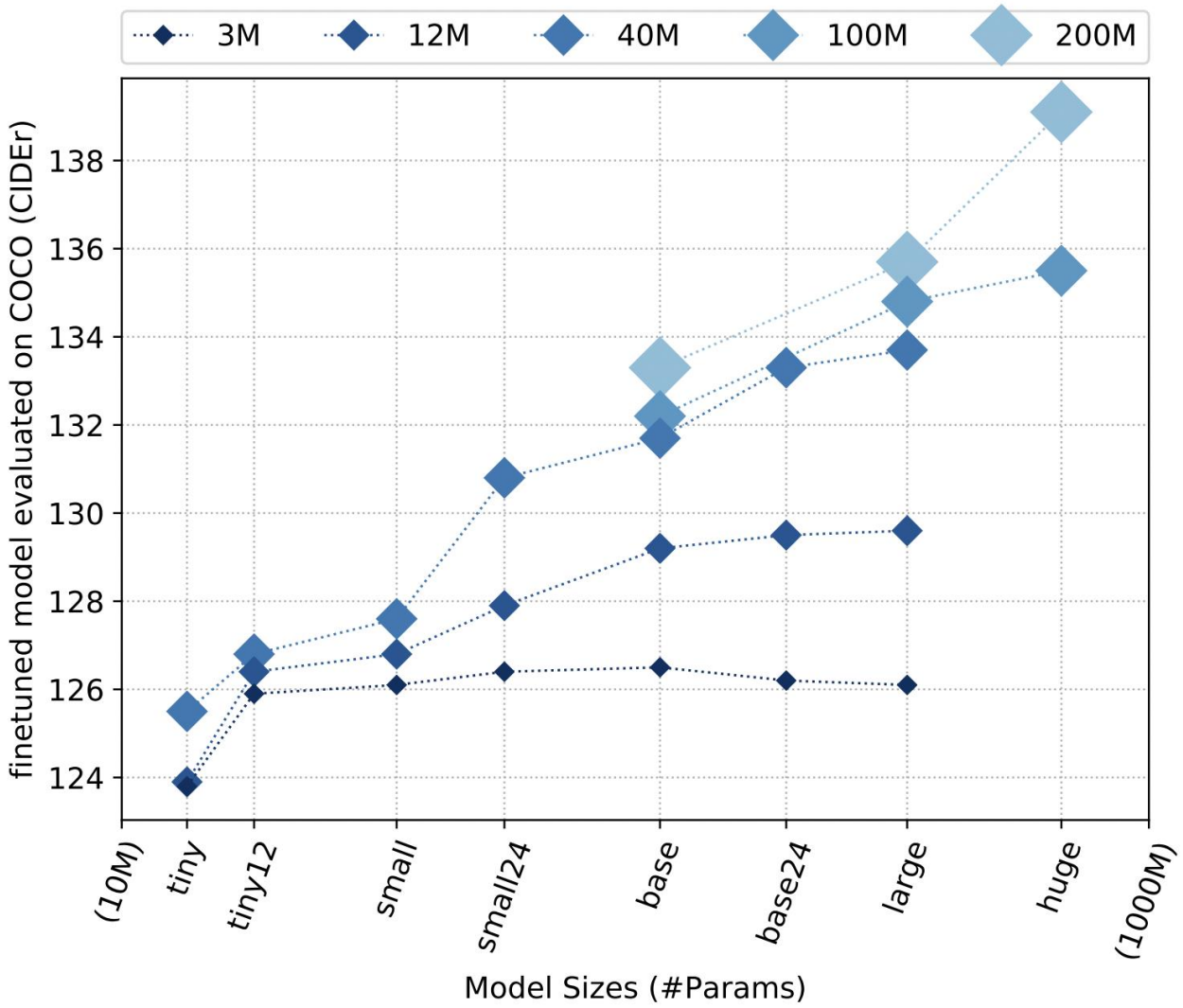
Influence of Model and Dataset Sizes



State-of-the-art performance
on 3 captioning datasets!

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022

Influence of Model and Dataset Sizes



What trend(s) do you observe?

Larger models WITH large training datasets yields better performance!

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022

Today's Topics

- Other data modalities
- Internet-scale trained models
- Scaling laws
- Lab assignment 3: multimodal task
- Programming tutorial

VizWiz-VQA Grand Challenge (7th year in 2025)

VizWiz

Home Browse Dataset Tasks & Datasets Workshops Acknowledgments

2025 VizWiz Grand Challenge Workshop

Overview

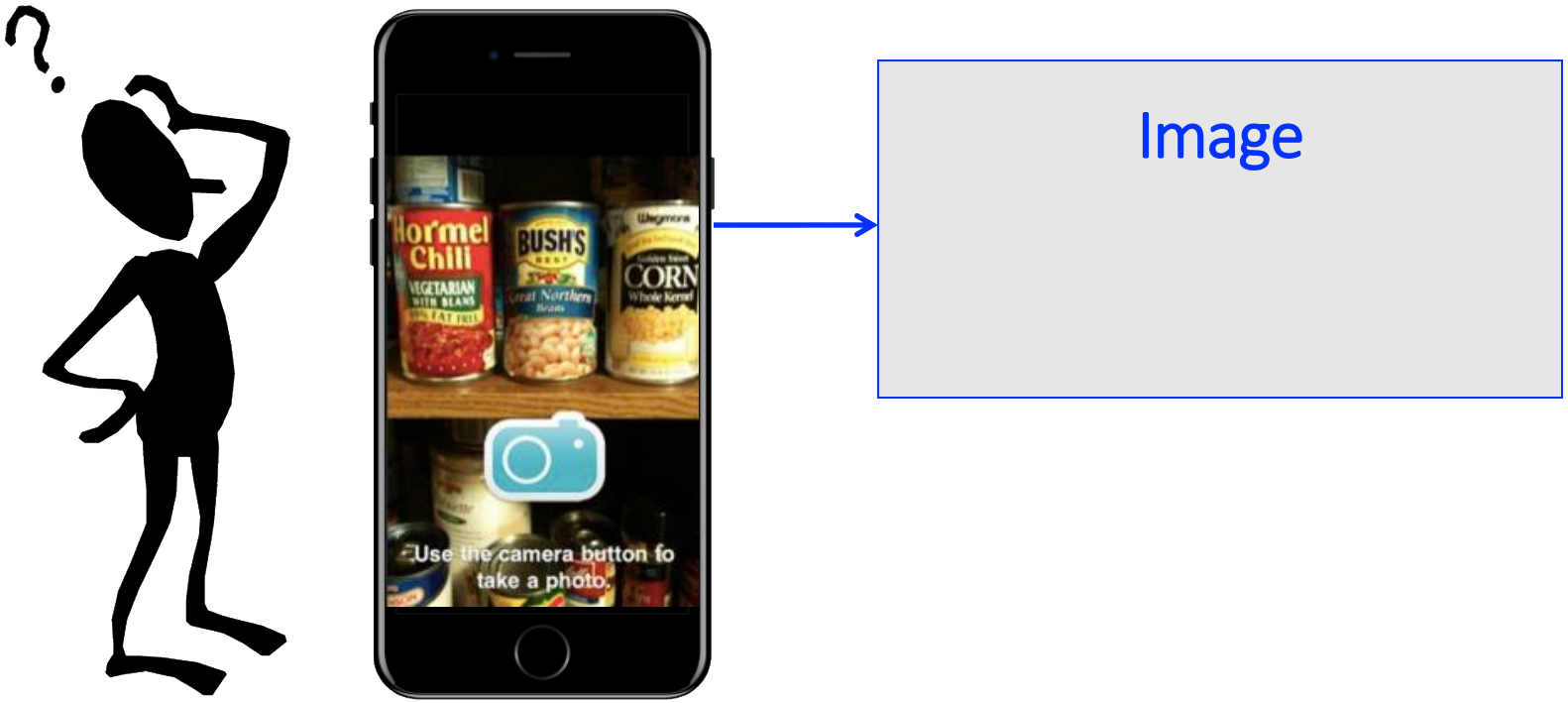
Our goal for this workshop is to educate researchers about the technological needs of people with vision impairments while empowering researchers to improve algorithms to meet these needs. A key component of this event will be to **track progress on five dataset challenges**, where the tasks are to recognize objects in few-shot learning scenarios, answer visual questions, ground answers, recognize visual questions with multiple answer groundings, locate objects in few-shot learning scenarios, classify images in a zero-shot setting. The second key component of this event will be a discussion about current research and application issues, including invited speakers from both academia and industry who will share their experiences in building today's state-of-the-art assistive technologies as well as designing next-generation tools.

https://vizwiz.org

VizWiz: Authentic Use Case

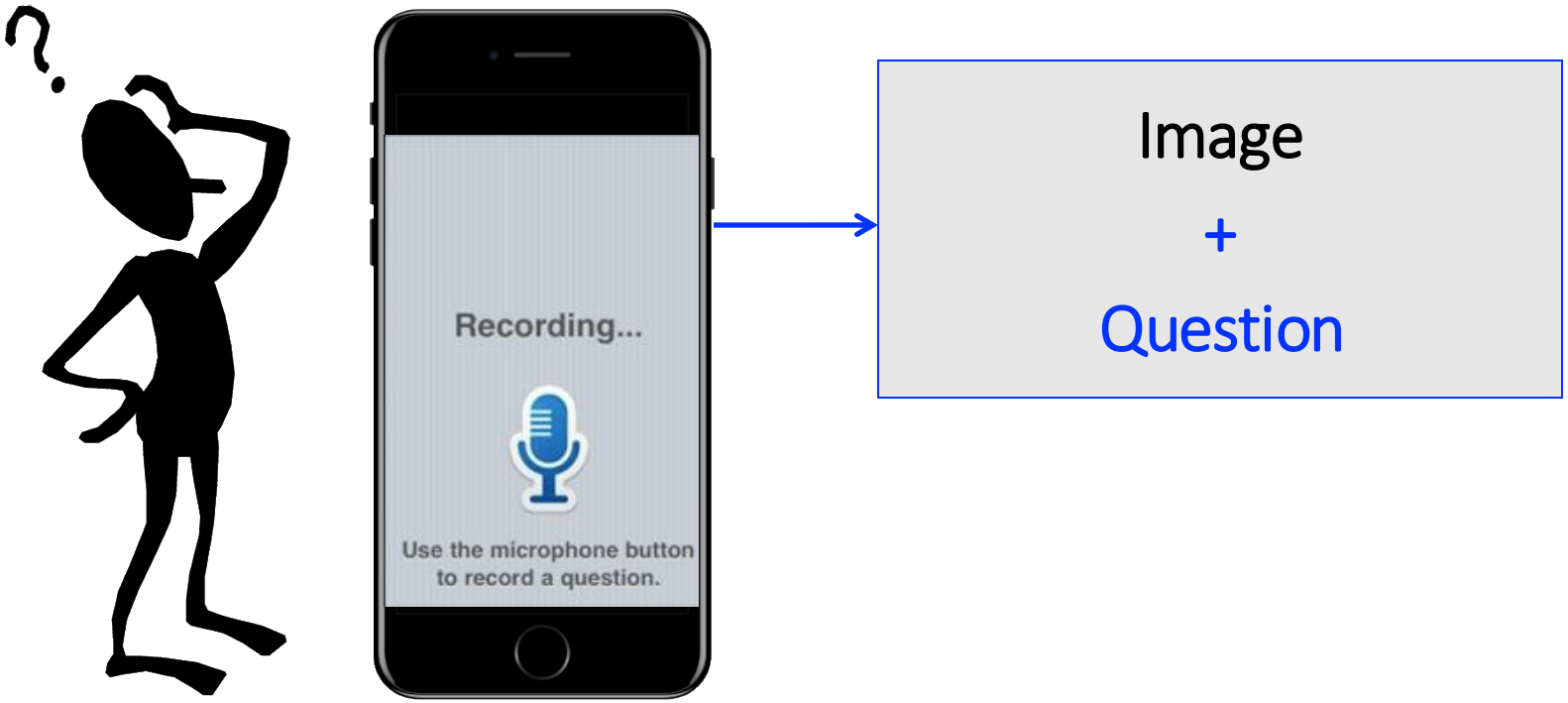
Jeffrey Bigham et al. VizWiz: Nearly Real-time Answers to Visual Questions. UIST 2010

VizWiz: Authentic Use Case



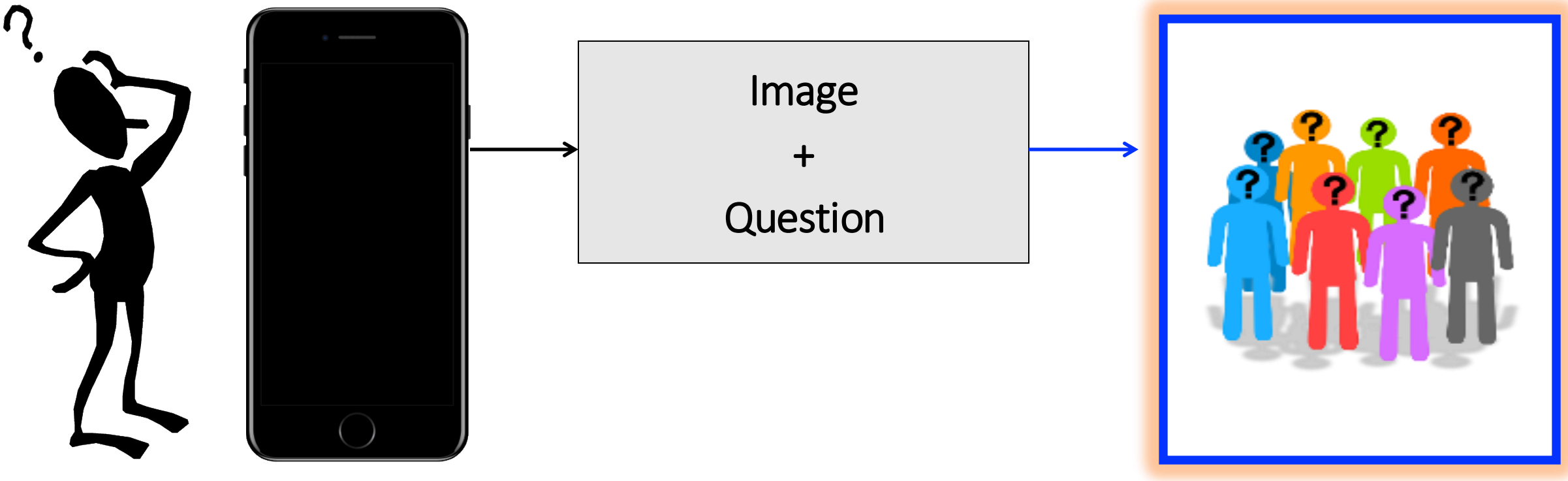
Jeffrey Bigham et al. VizWiz: Nearly Real-time Answers to Visual Questions. UIST 2010

VizWiz: Authentic Use Case



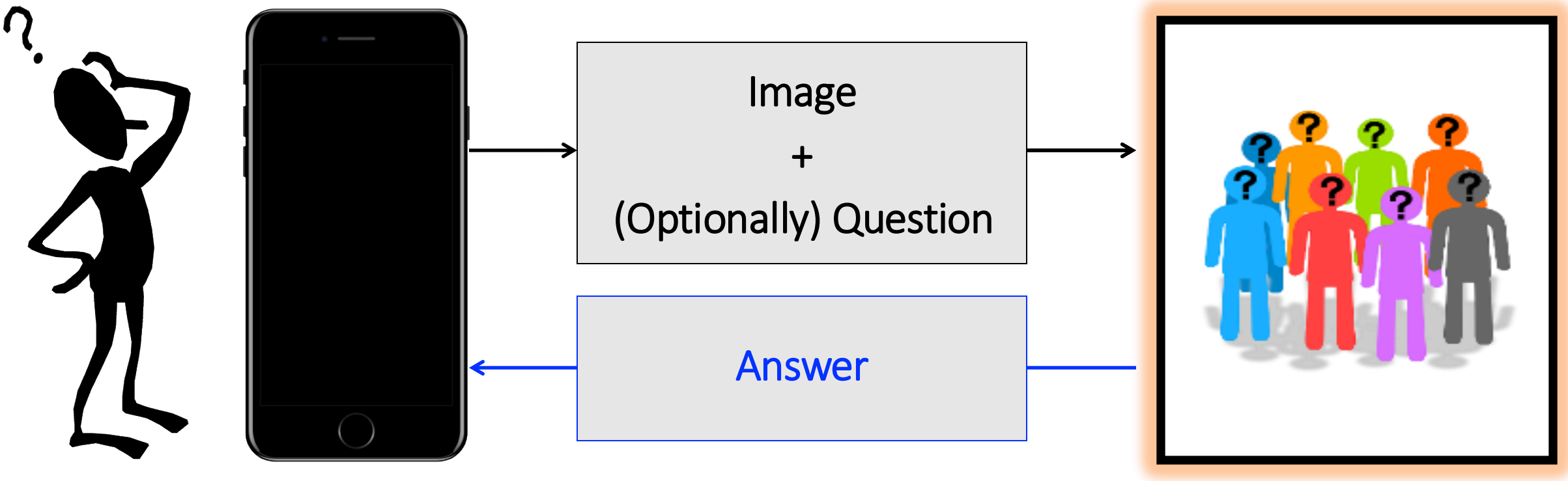
Jeffrey Bigham et al. VizWiz: Nearly Real-time Answers to Visual Questions. UIST 2010

VizWiz: Authentic Use Case



Jeffrey Bigham et al. VizWiz: Nearly Real-time Answers to Visual Questions. UIST 2010

VizWiz: Authentic Use Case



Jeffrey Bigham et al. VizWiz: Nearly Real-time Answers to Visual Questions. UIST 2010

VizWiz: Authentic Use Case

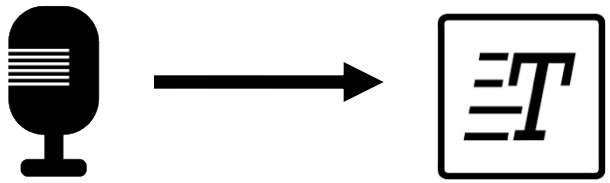
Users agreed to share 44,799 (62%) of requests for dataset creation

Jeffrey Bigham et al. VizWiz: Nearly Real-time Answers to Visual Questions. UIST 2010

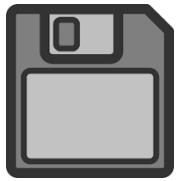
VizWiz: Authentic Use Case

Anonymization

1. Transcribe questions (removes voice)



2. Re-save images (removes metadata)

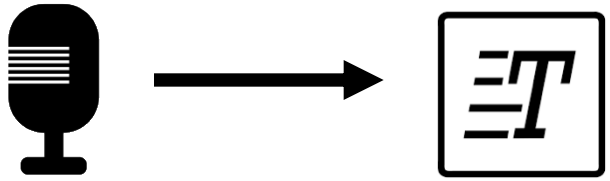


Gurari et al. CVPR 2018

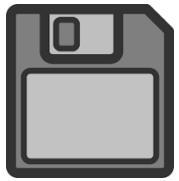
VizWiz: Authentic Use Case

Anonymization

1. Transcribe questions

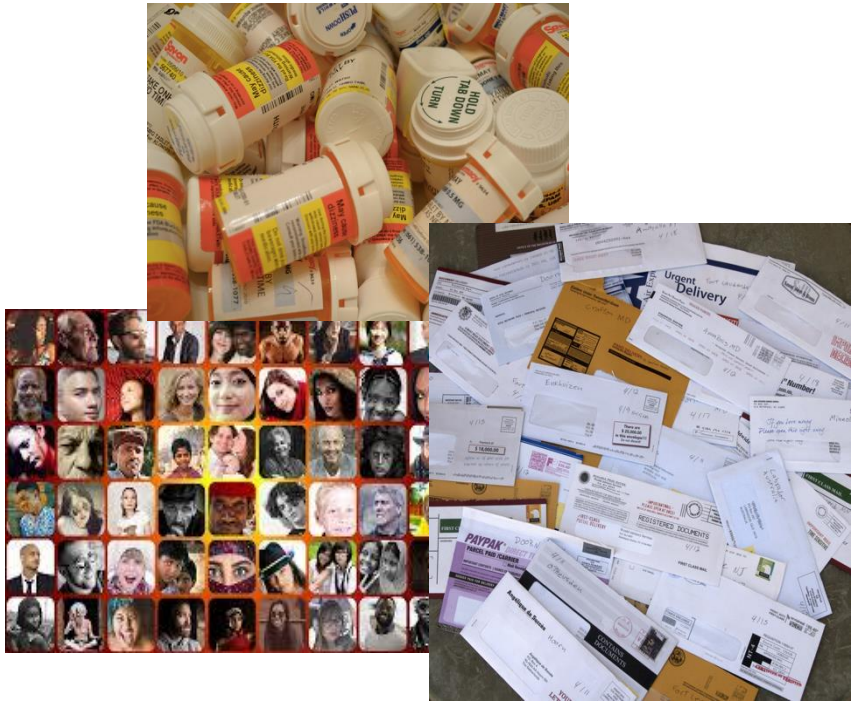


2. Re-save images



In-House Filtering

(personally identifying information)

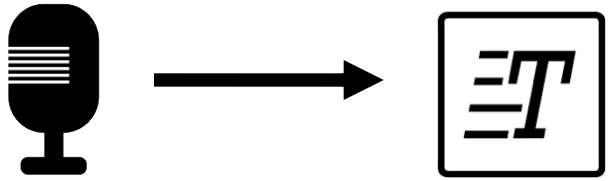


Gurari et al. CVPR 2018

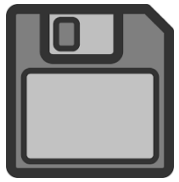
VizWiz: Authentic Use Case

Anonymization

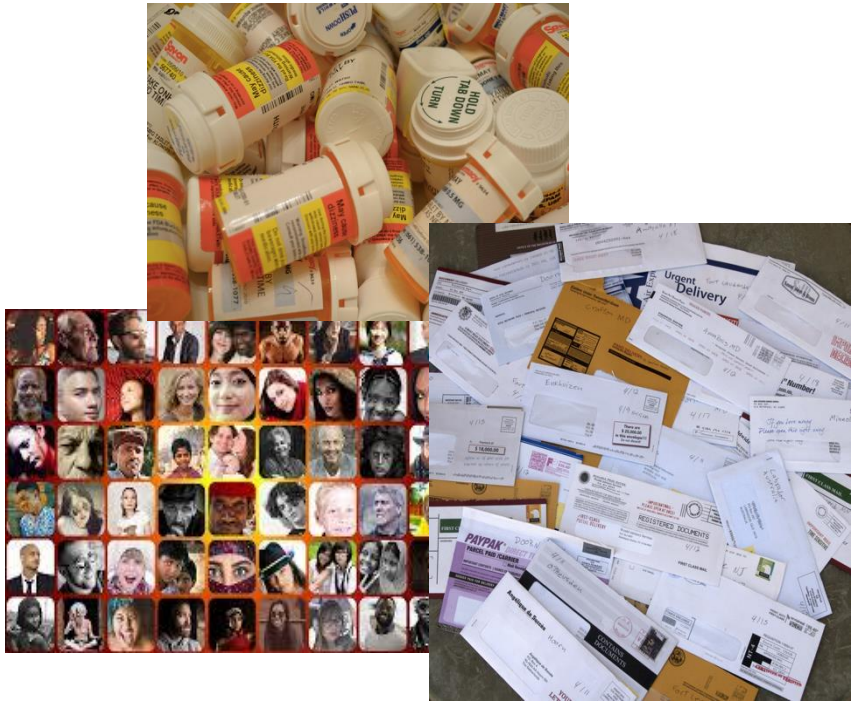
1. Transcribe questions



2. Re-save images



In-House Filtering



Data Labeling

(high quality answers)

amazon mechanical turk

Gurari et al. CVPR 2018

VizWiz: Authentic Use Case

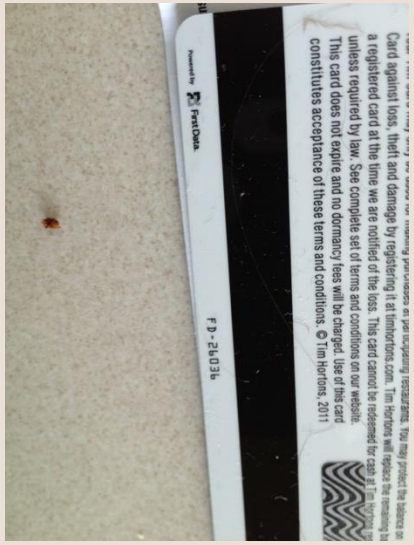
VQA: 32,842 image/question pairs → 328,420 answers

Gurari et al. CVPR 2018

Key Difference of Real-World Use Case from Status Quo: VQs Can Be Unanswerable!

Q: What is the expiration date?

A: unanswerable



Q: What is this a gift card for?

A: unanswerable

Q: What temperature is the dial set to?

A: unanswerable

Gurari et al. CVPR 2018

Today's Topics

- Other data modalities
- Internet-scale trained models
- Scaling laws
- Lab assignment 3: multimodal task
- Programming tutorial

Today’s Topics

- Other data modalities
- Internet-scale trained models
- Scaling laws
- Lab assignment 3: multimodal task
- Programming tutorial


The End