Multi-Task and Transfer Learning in Low-Resource Speech Recognition

Joshua Meyer

PhD Candidate

Department of Linguistics

University of Arizona

Roadmap

- Overview of Transfer Learning
  - Multi-Task Learning
  - Copy-Paste Transfer

- Multi-Task Learning Studies
  - Linguistic Tasks
  - Engineered Tasks
  - Discovered Tasks

- Copy-Paste Transfer Studies
  - Multilingual Transfer
  - Model Interpretability

- Conclusion

Introduction

Motivation

Current training methods
for automatic speech recognition
require massive collections of data.

However, most use-cases have
little — if any — available data.

Motivation

Current training methods
for automatic speech recognition
require massive collections of data.

However, most use-cases have
little — if any — available data.

Motivation

Current training methods
for automatic speech recognition
require massive collections of data.

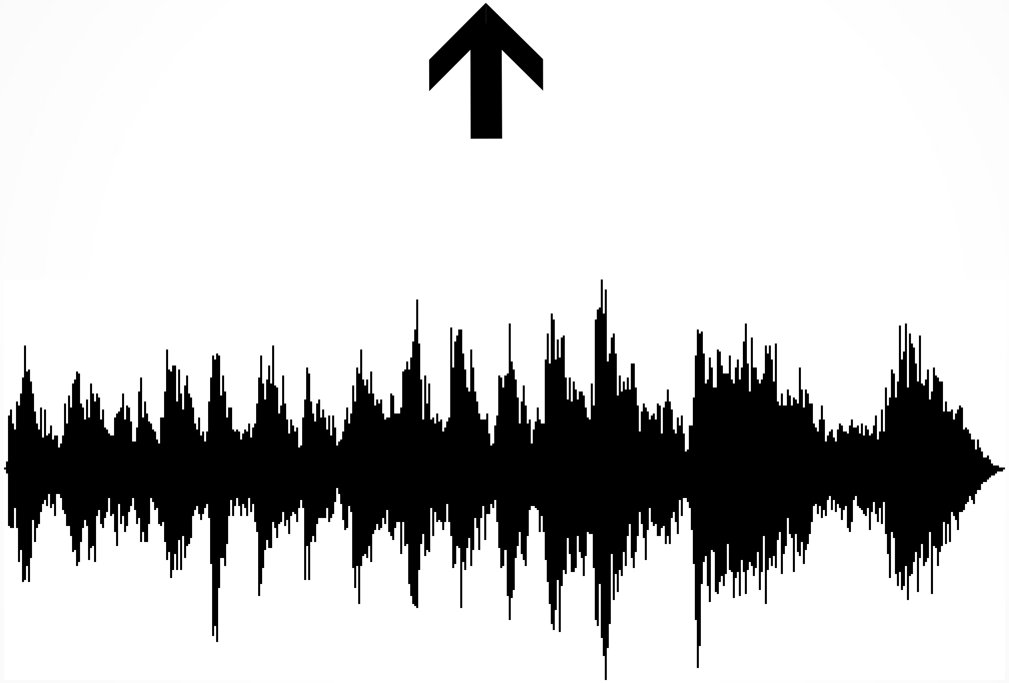
However, most use-cases have
little — if any — available data.

But we can exploit similar use-cases!

Automatic Speech Recognition (ASR)

Automatic Speech Recognition

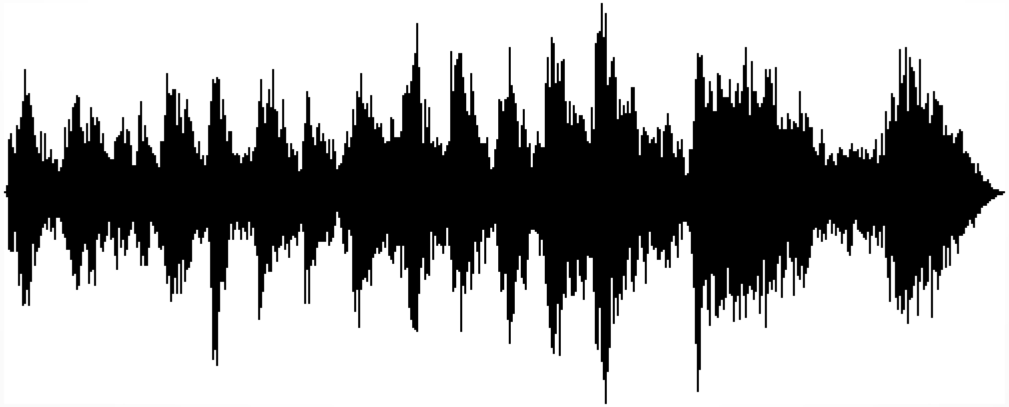
"THE DOG"



Automatic Speech Recognition

"THE DOG"

↑



TOO HARD!

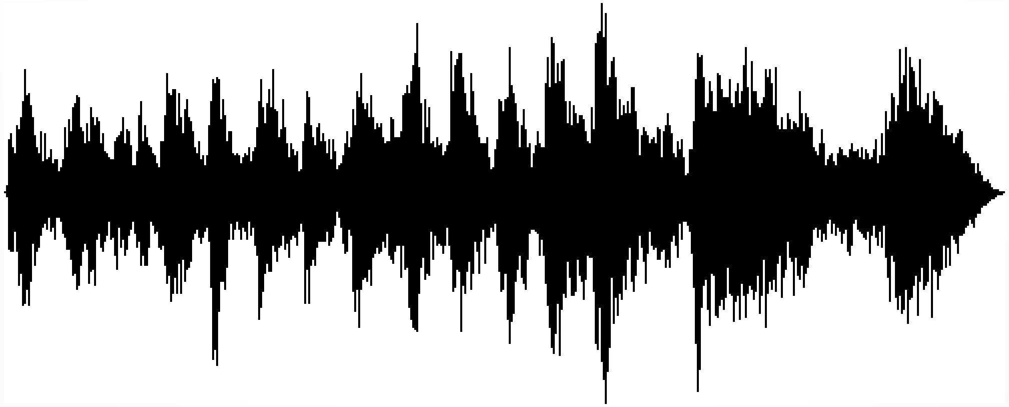
Automatic Speech Recognition

"THE DOG"

↑

T H E D O G

↑



EASIER!

Automatic Speech Recognition

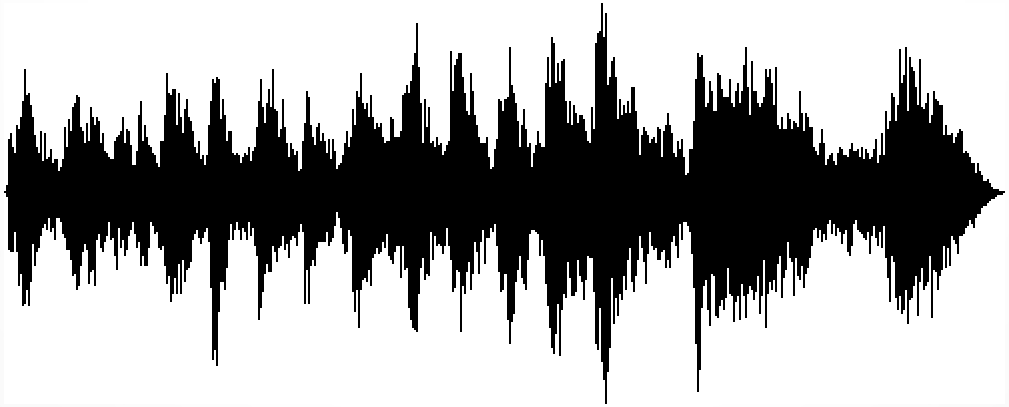
"THE DOG"

↑

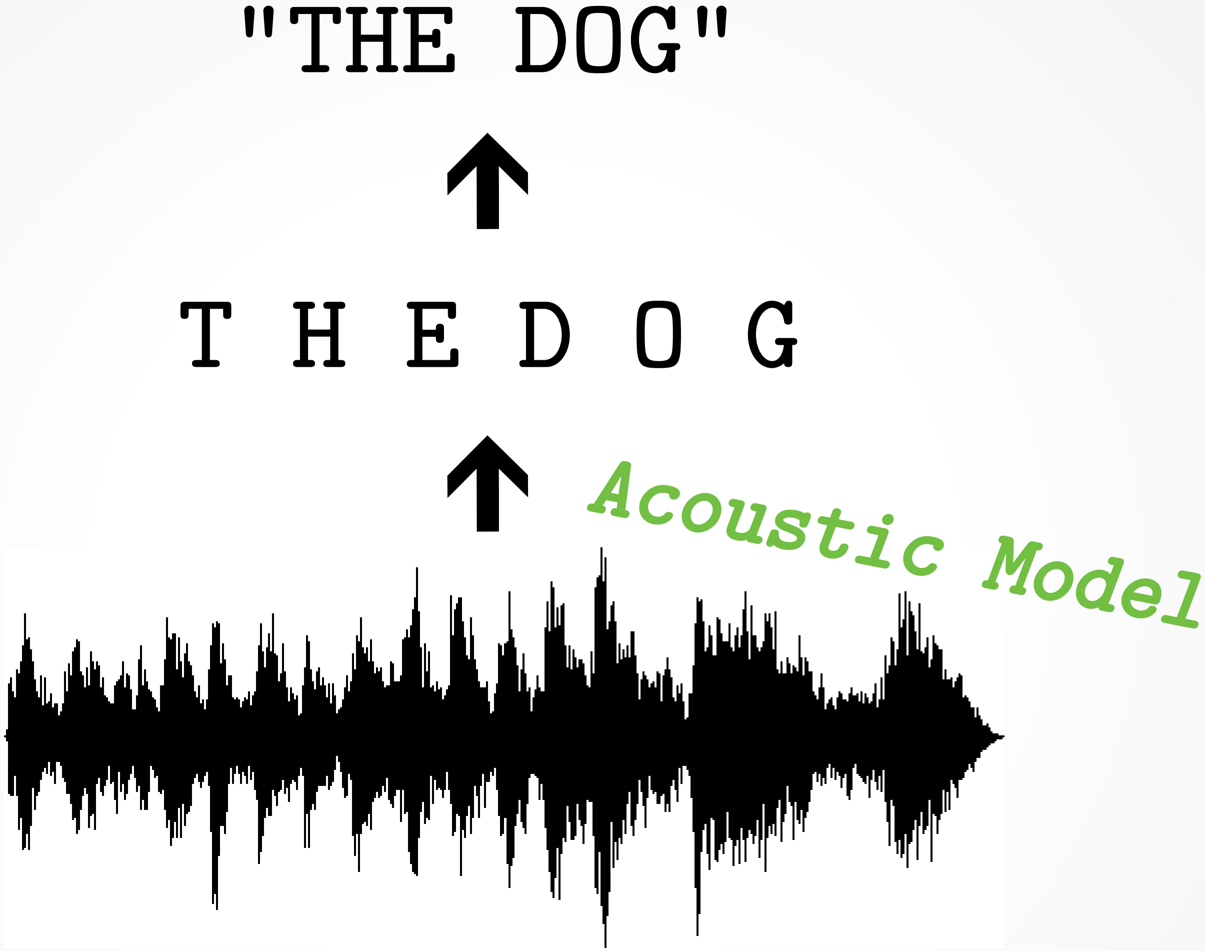
T H E D O G

"Phoneme-like units"

↑



Automatic Speech Recognition



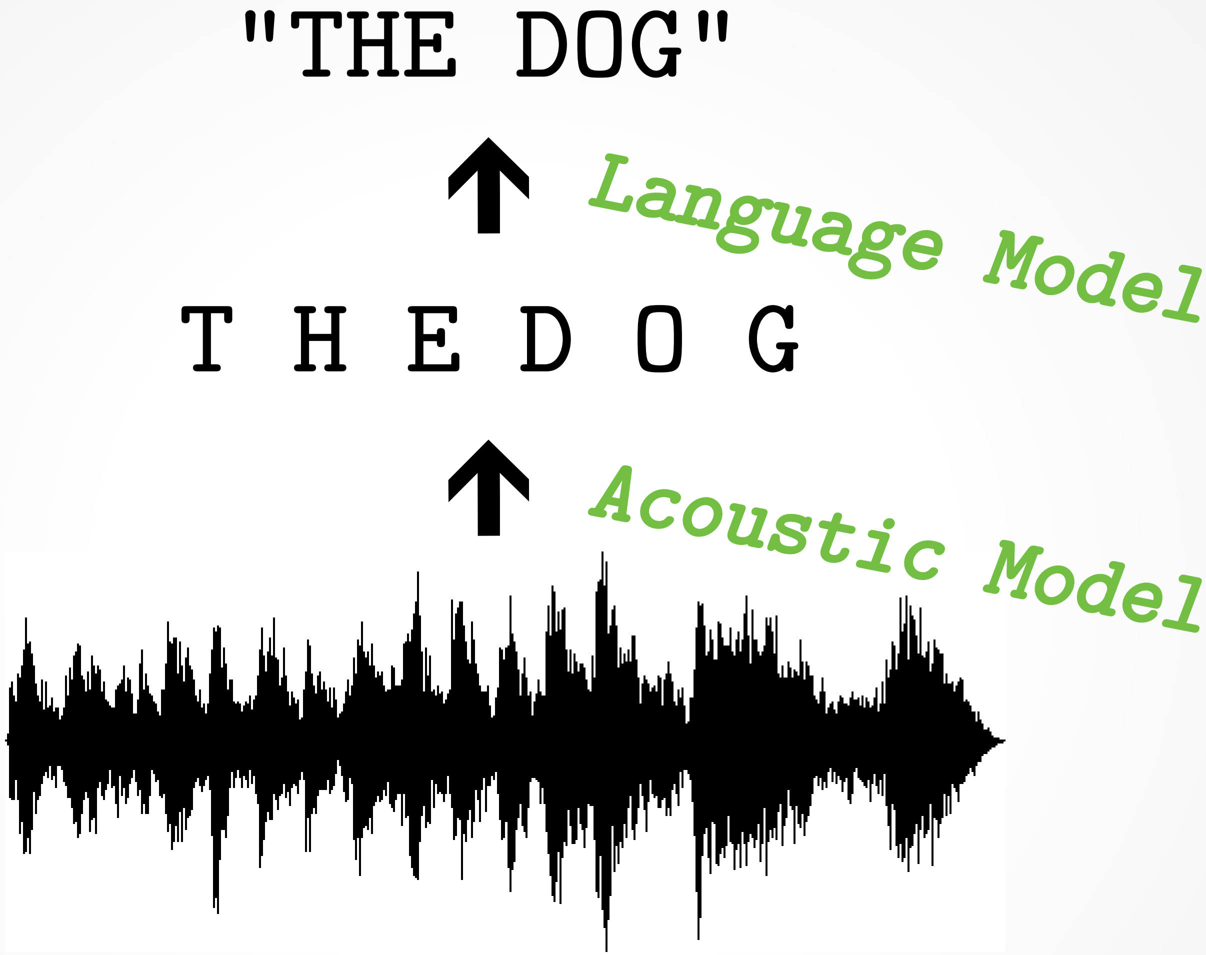
"THE DOG"

↑

T H E D O G

↑ Acoustic Model

Automatic Speech Recognition



"THE DOG"

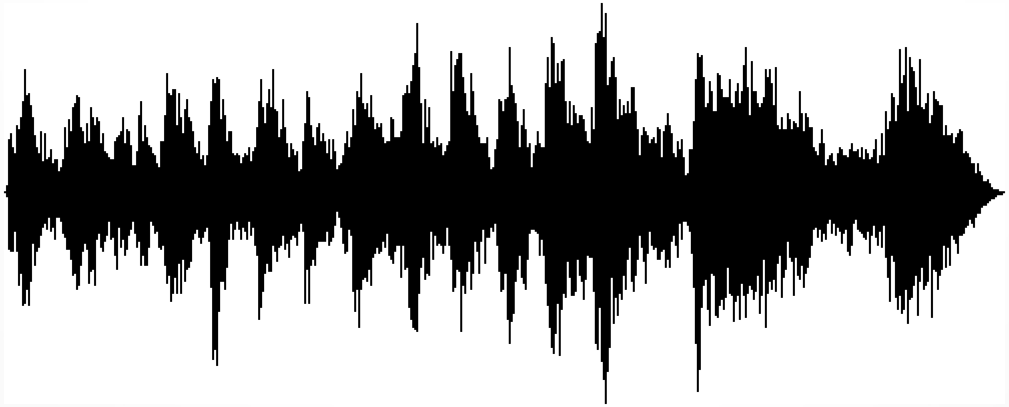
Language Model

T H E D O G

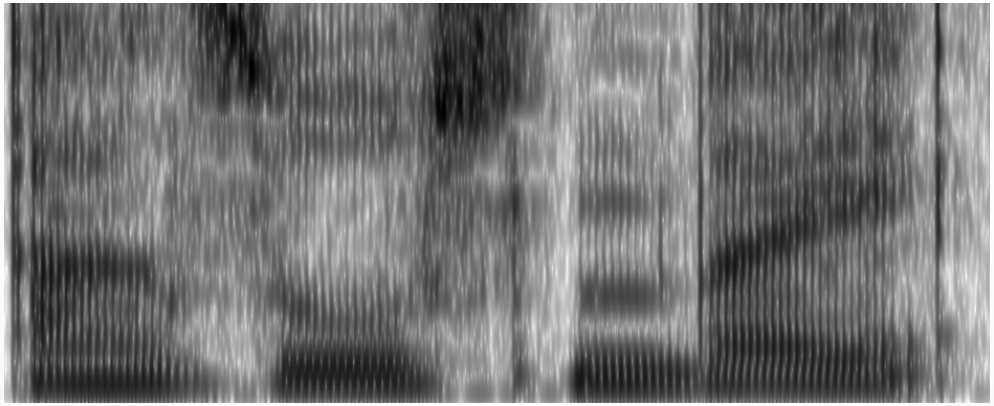
Acoustic Model

ASR Acoustic Modeling

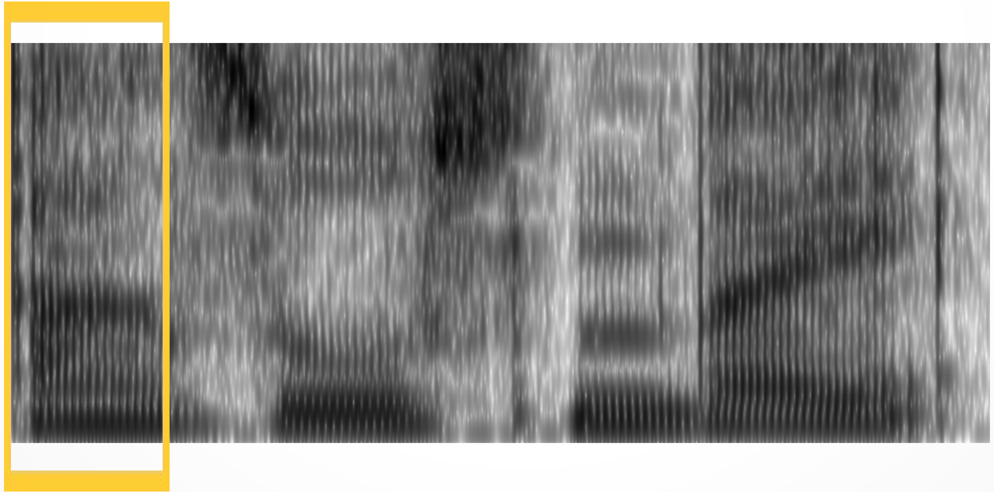
Acoustic Model



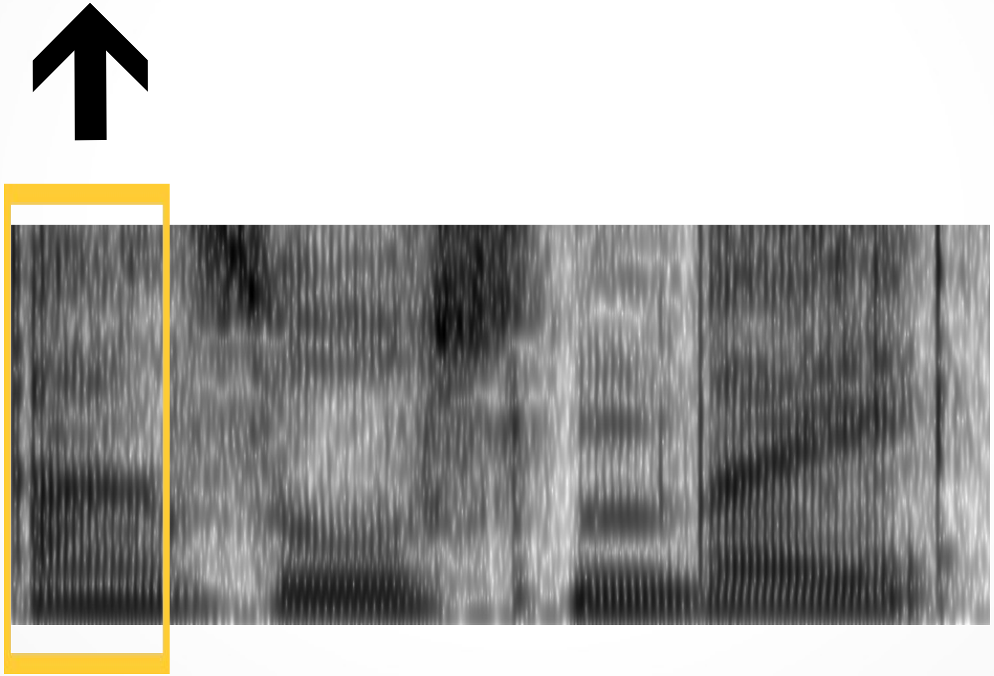
Acoustic Model



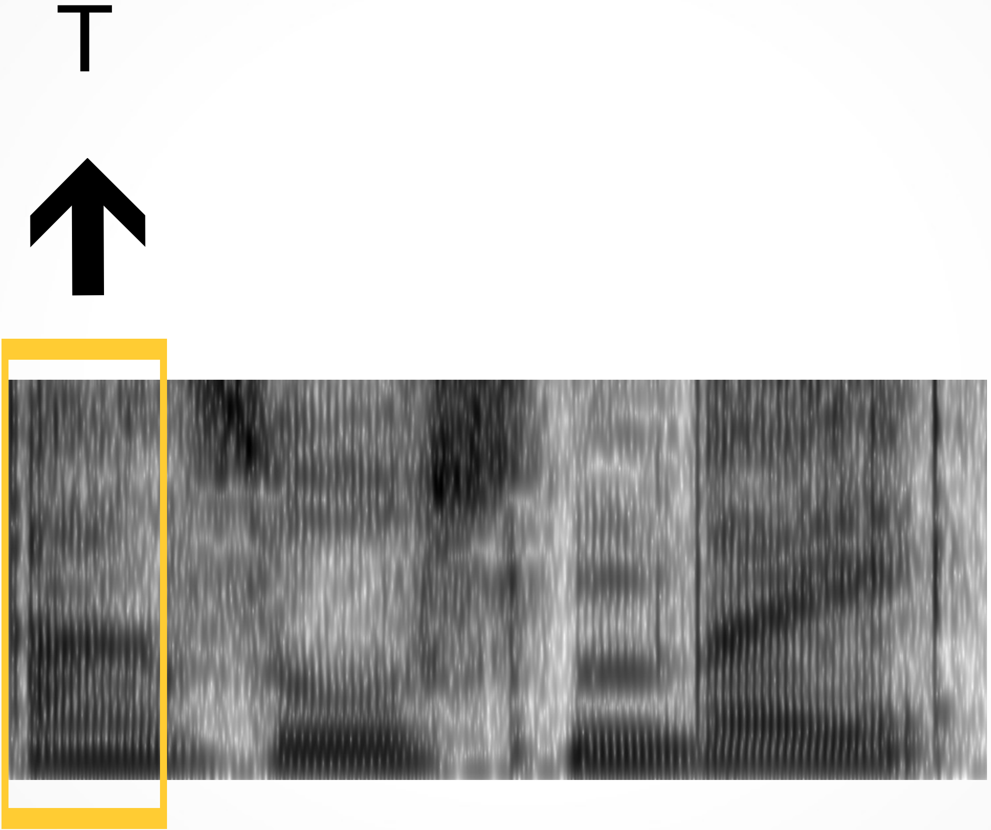
Acoustic Model



Acoustic Model

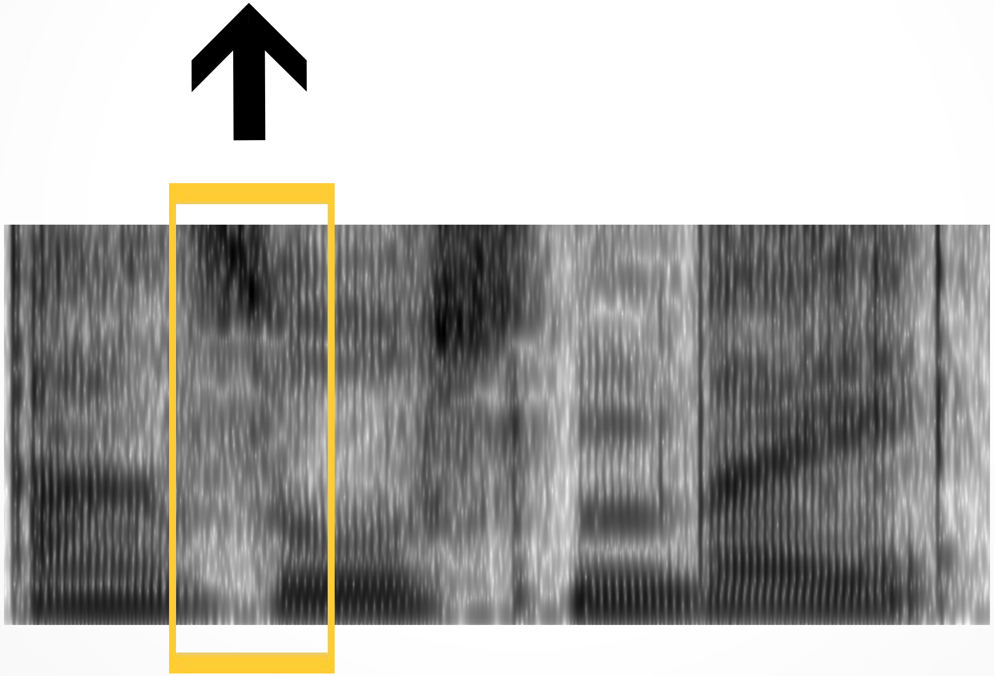


Acoustic Model

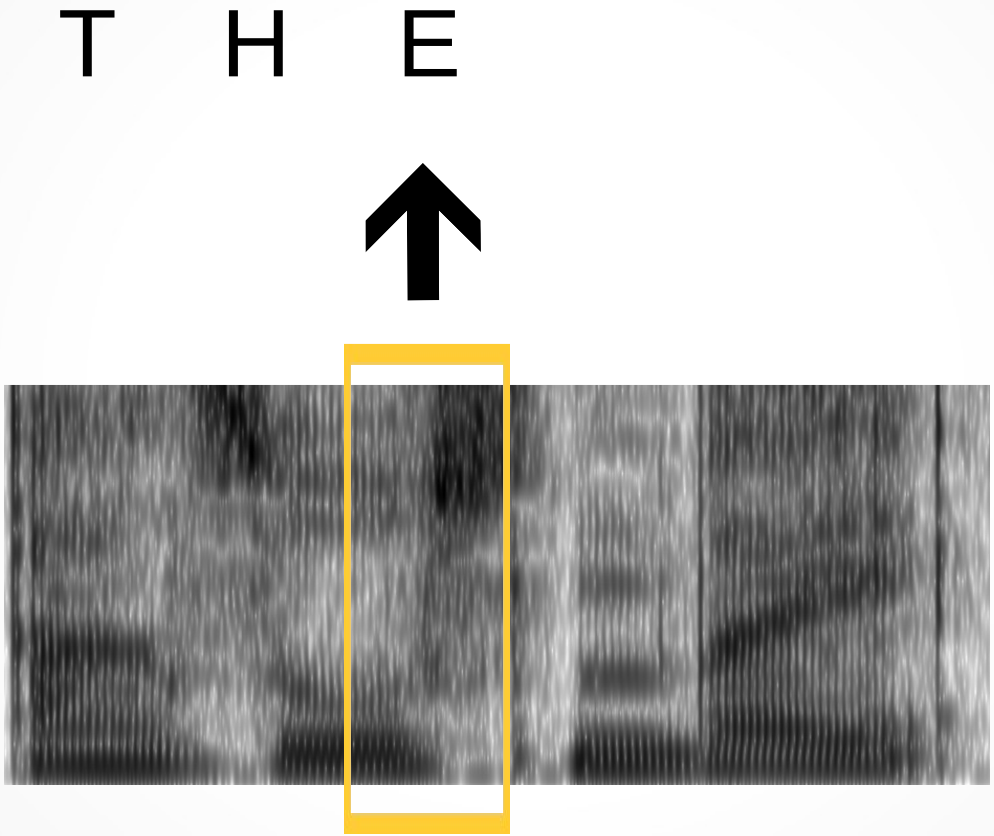


Acoustic Model

T H

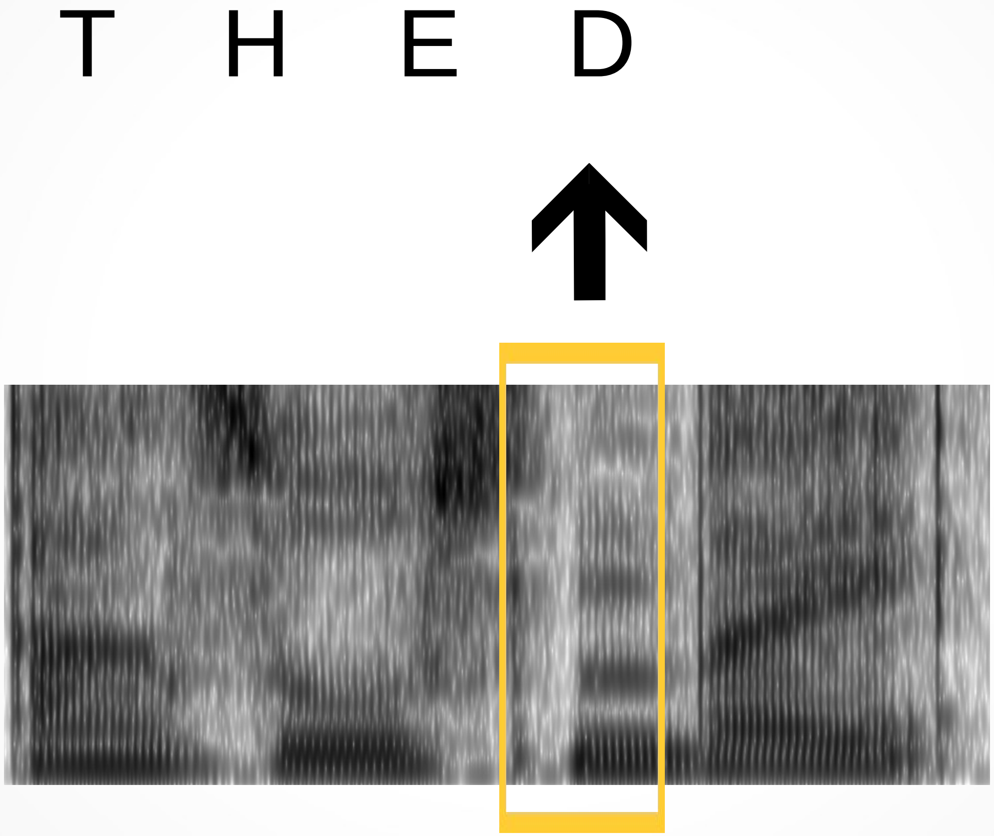


Acoustic Model



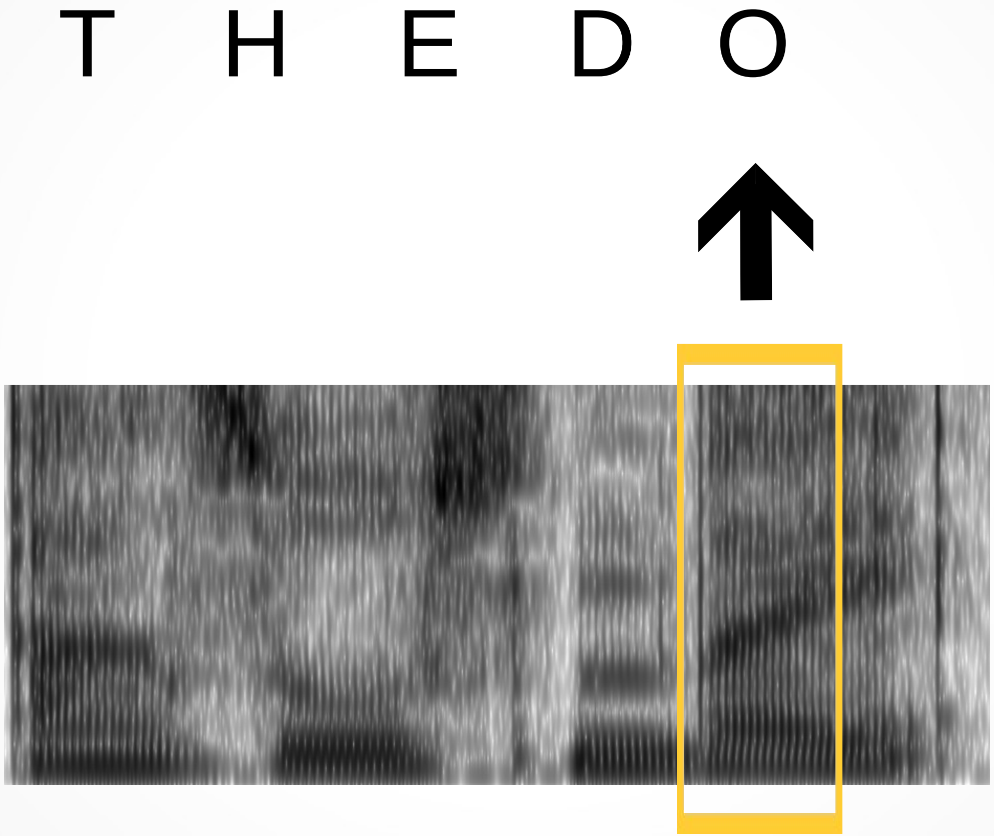
T H E

Acoustic Model



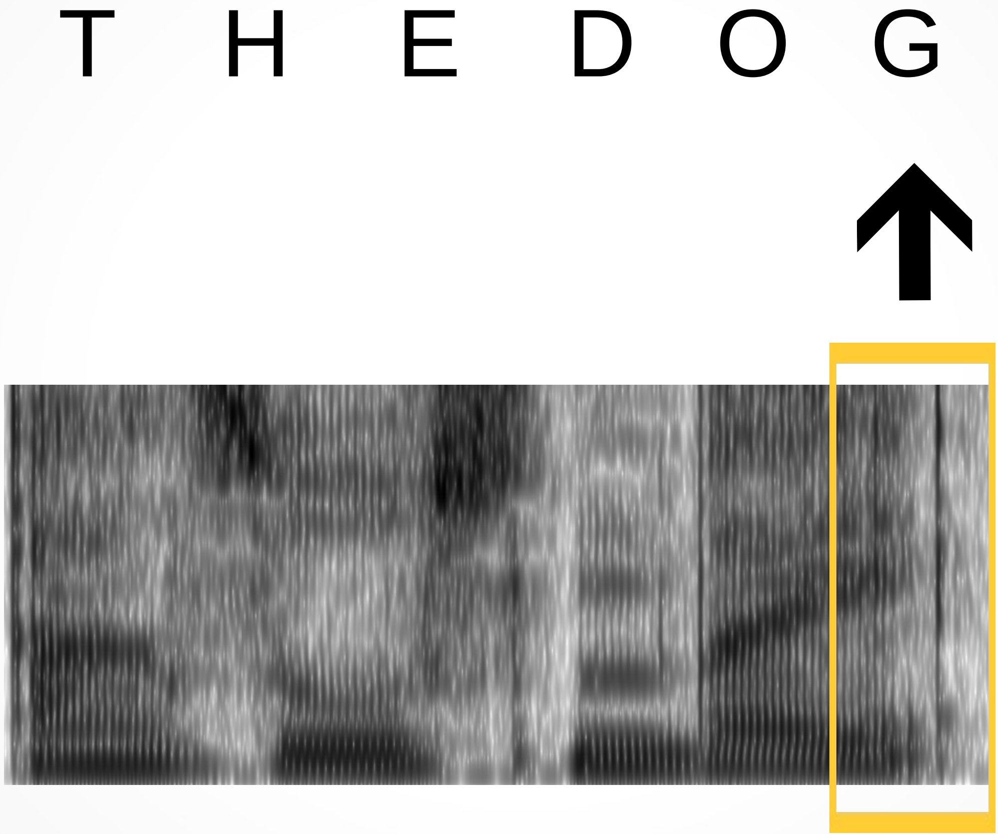
T H E D

Acoustic Model

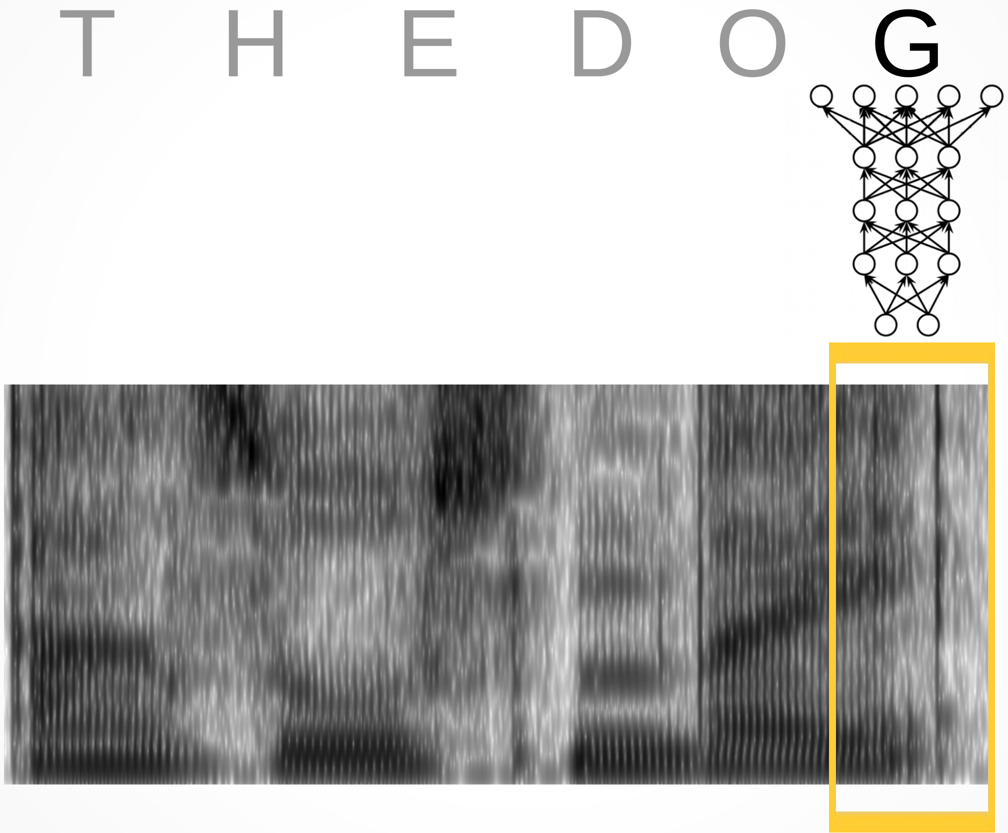


T H E D O

Acoustic Model

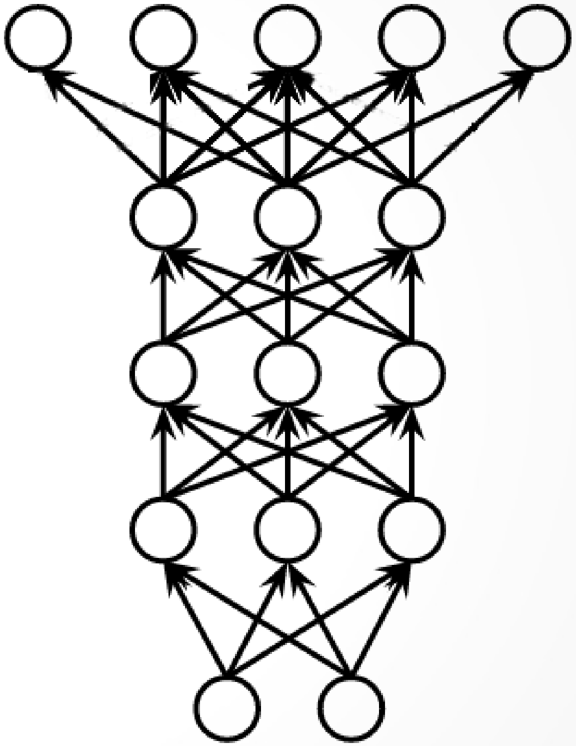


Acoustic Model



T H E D O G

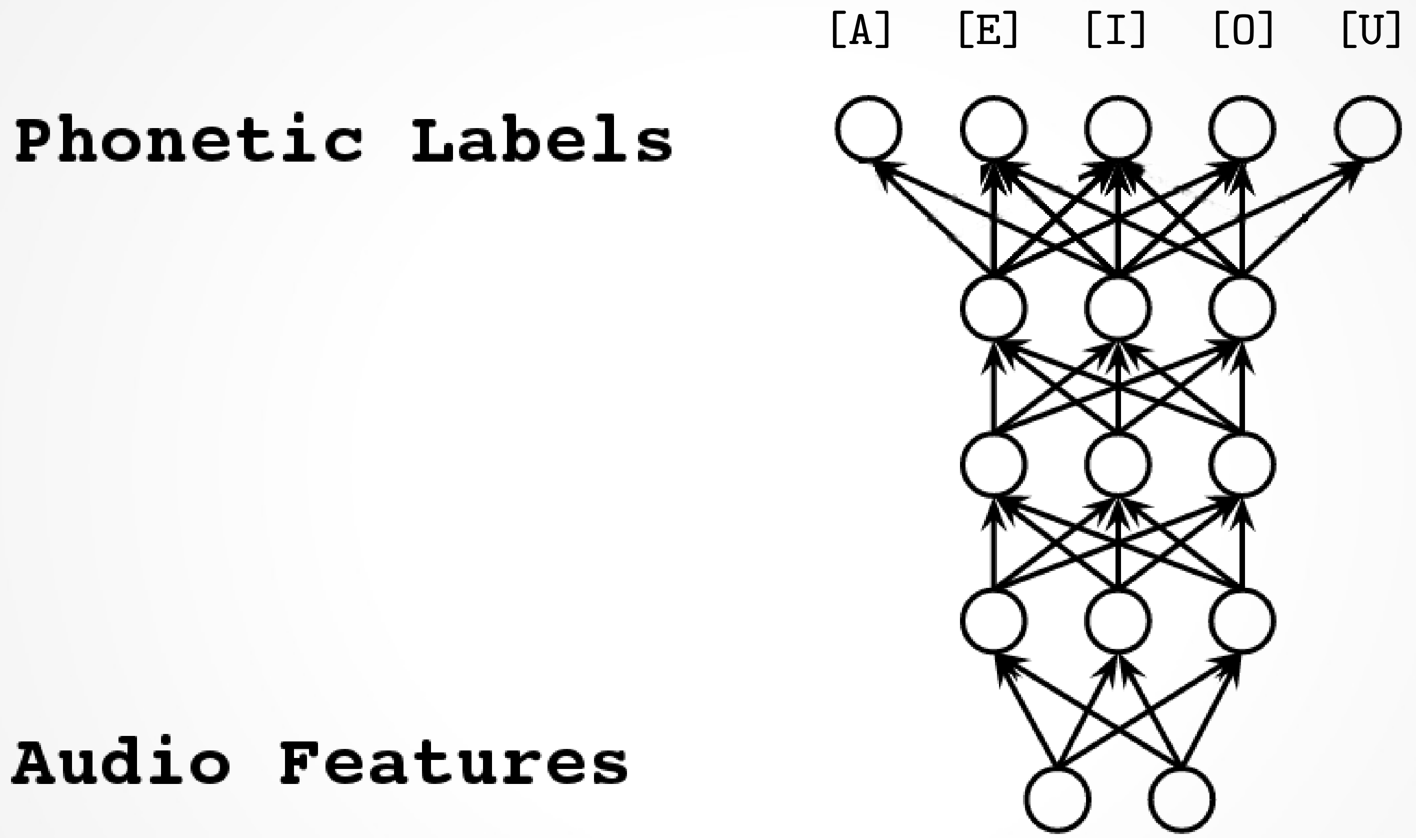
Acoustic Model



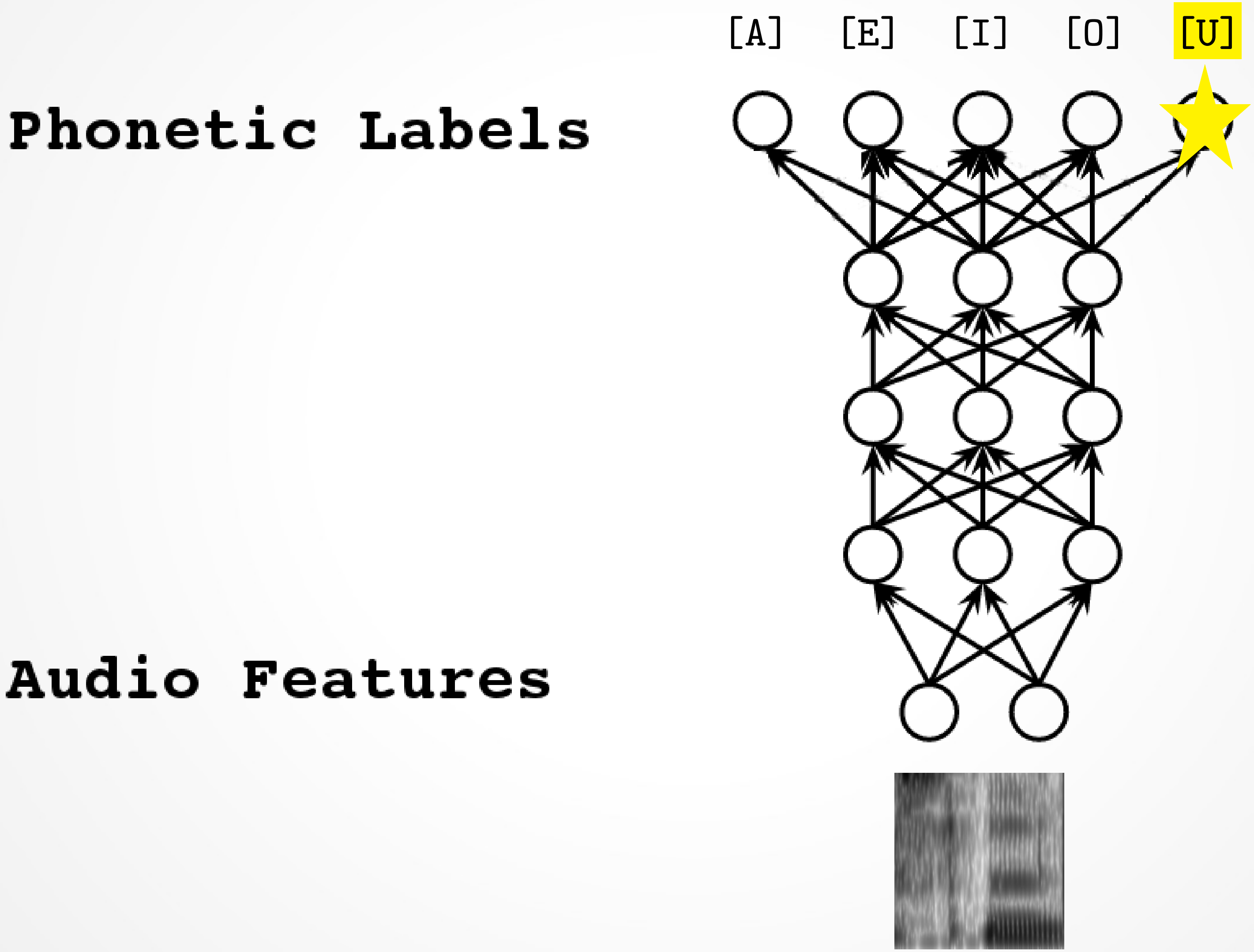
Phonetic Labels

Audio Features

Acoustic Model



Acoustic Model

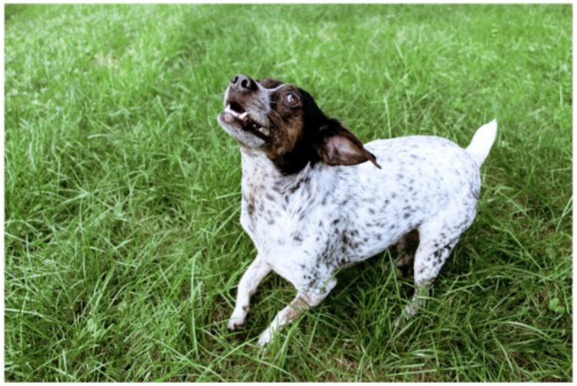


Multi-Task Learning Studies

Overview of MTL

{rottweiler}

{collie}

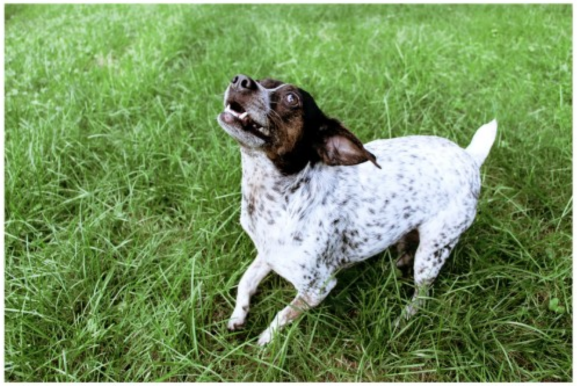


{terrier}

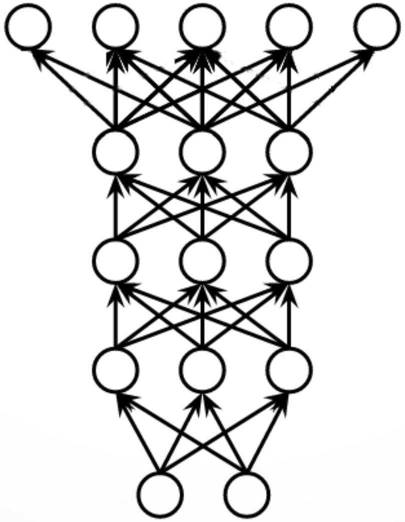
Overview of MTL

{rottweiler}

{collie}



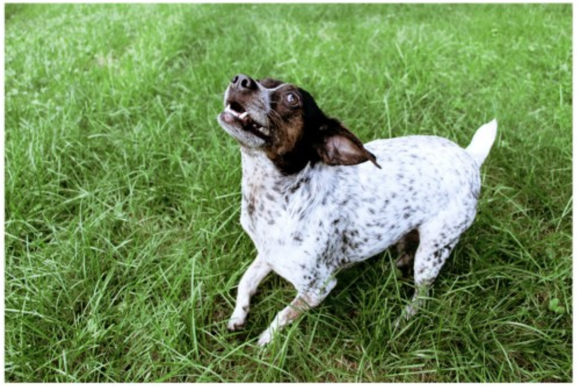
{terrier}



Overview of MTL

{rottweiler, large}

{collie, large}

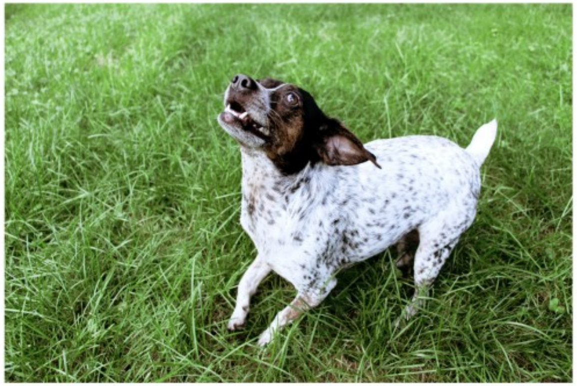


{terrier, small}

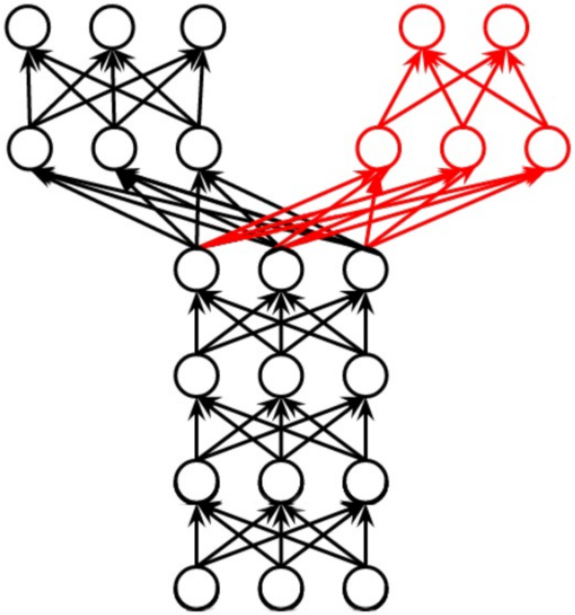
Overview of MTL

{rottweiler, large}

{collie, large}



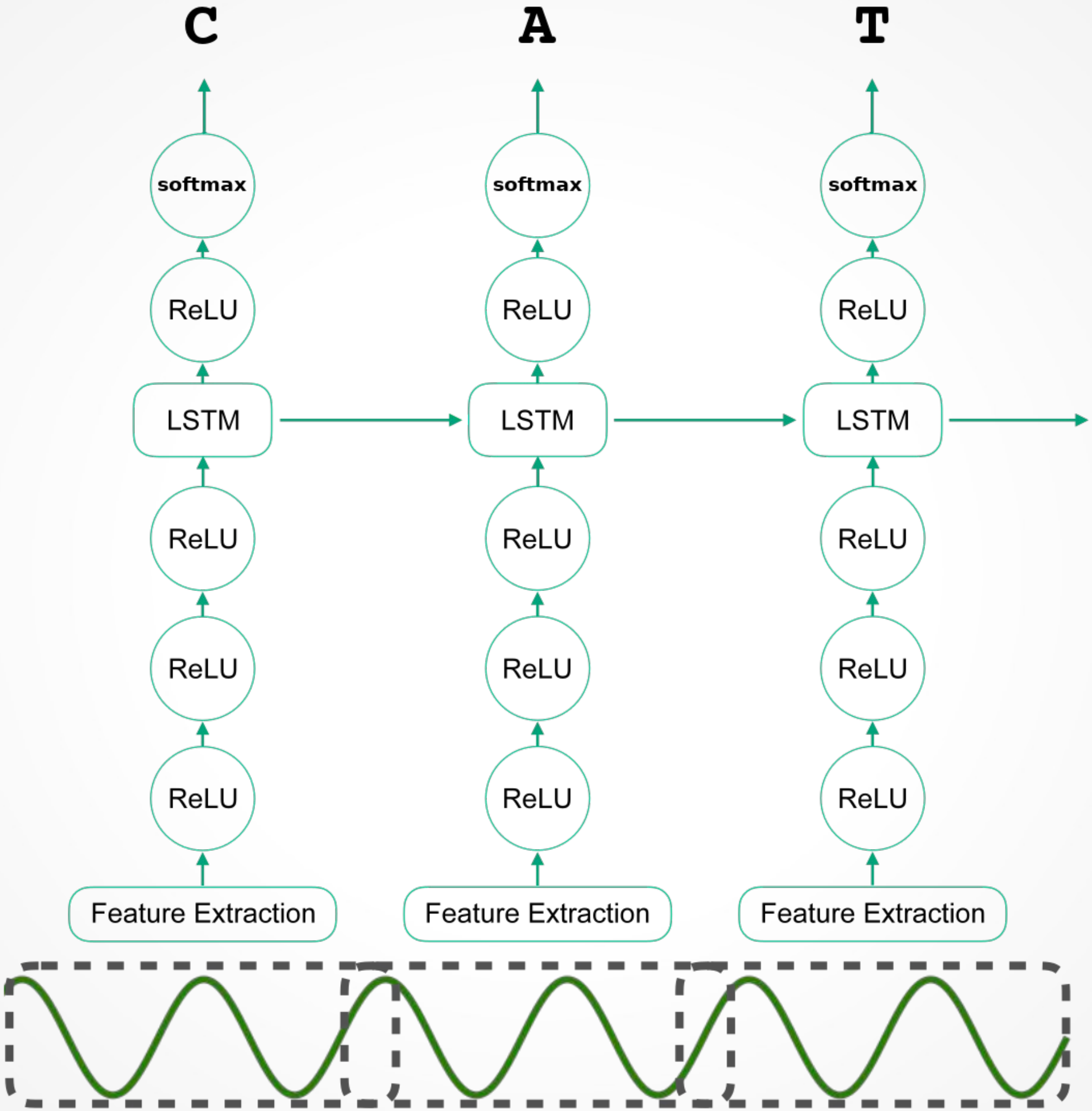
{terrier, small}



Copy-Paste Transfer Studies

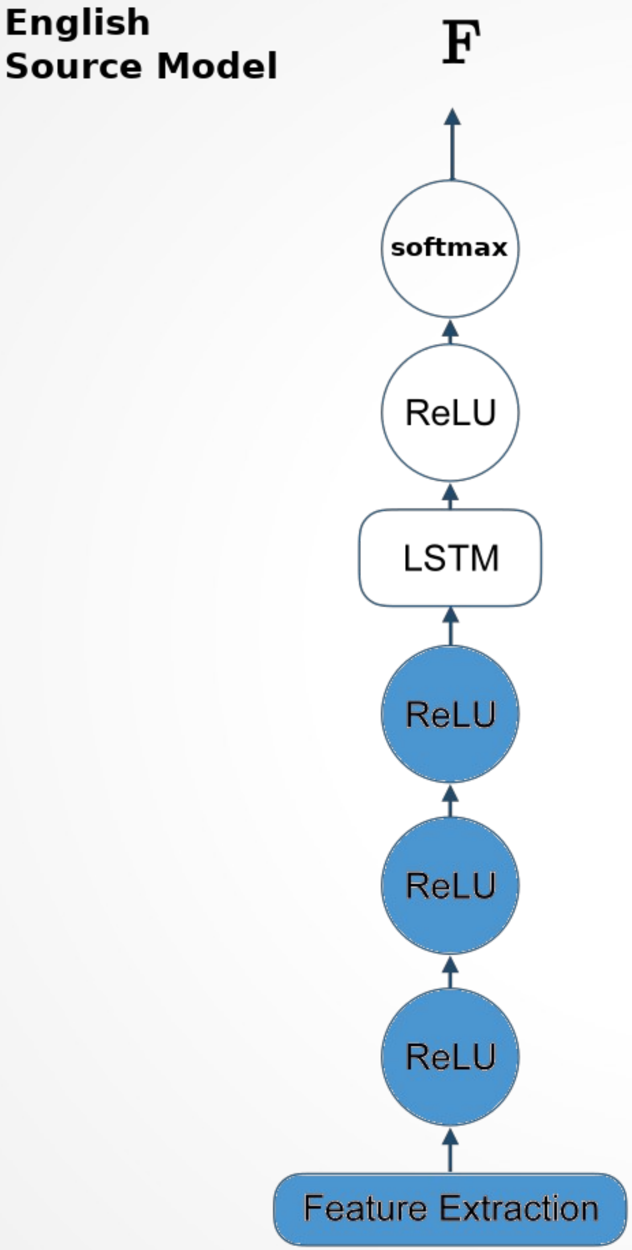
Quick Overview of DeepSpeech

Model Architecture

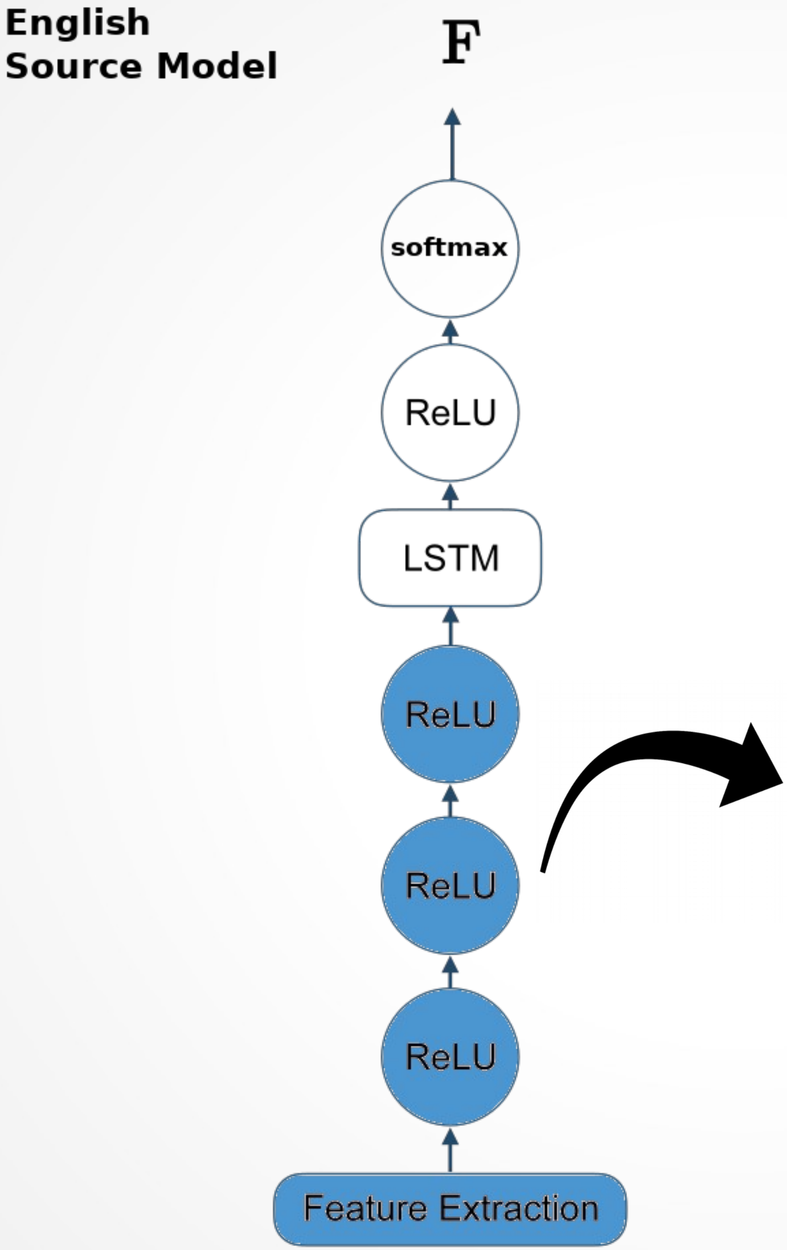


Transfer Experiments on ASR

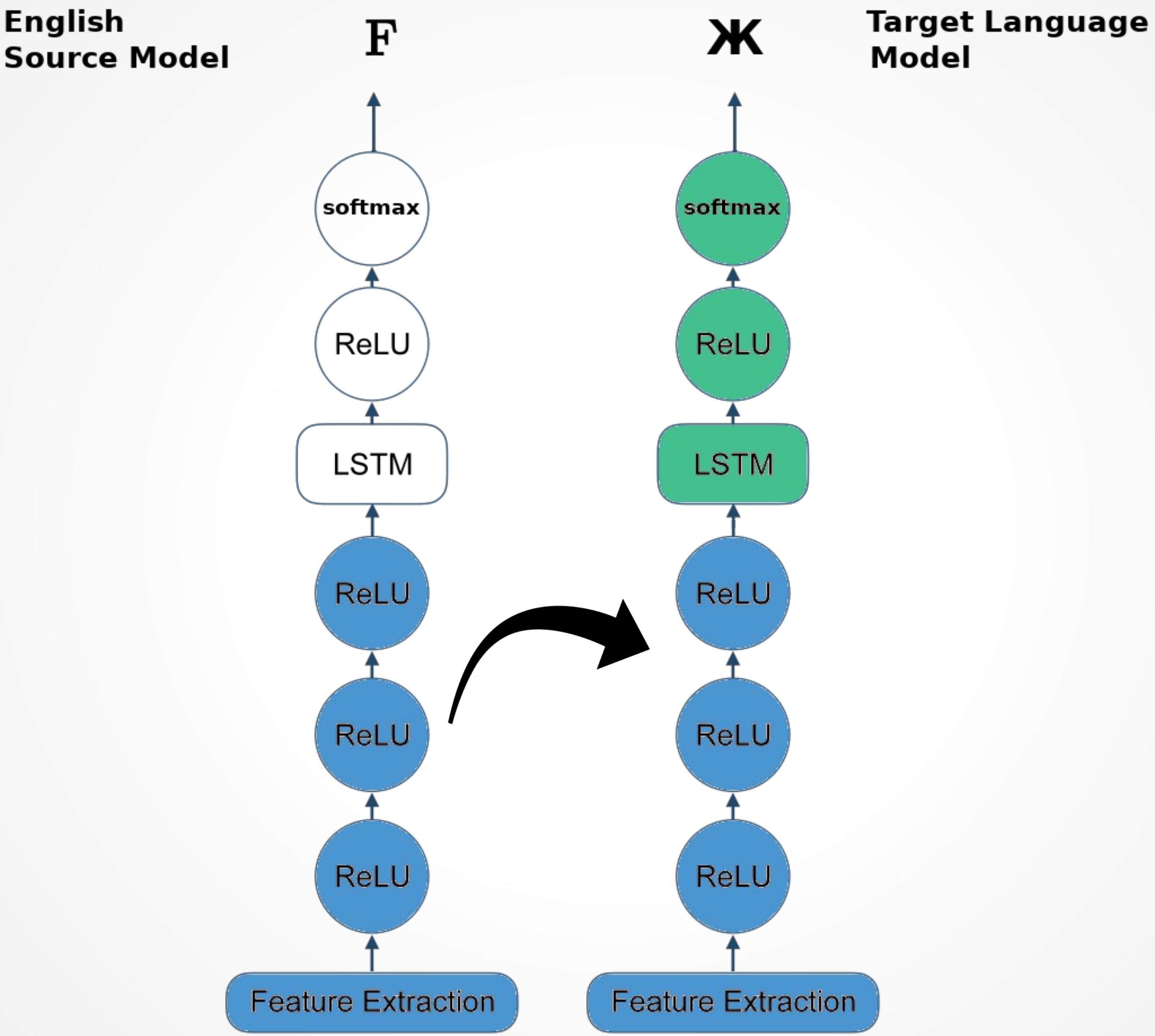
CTC Transfer Experiments



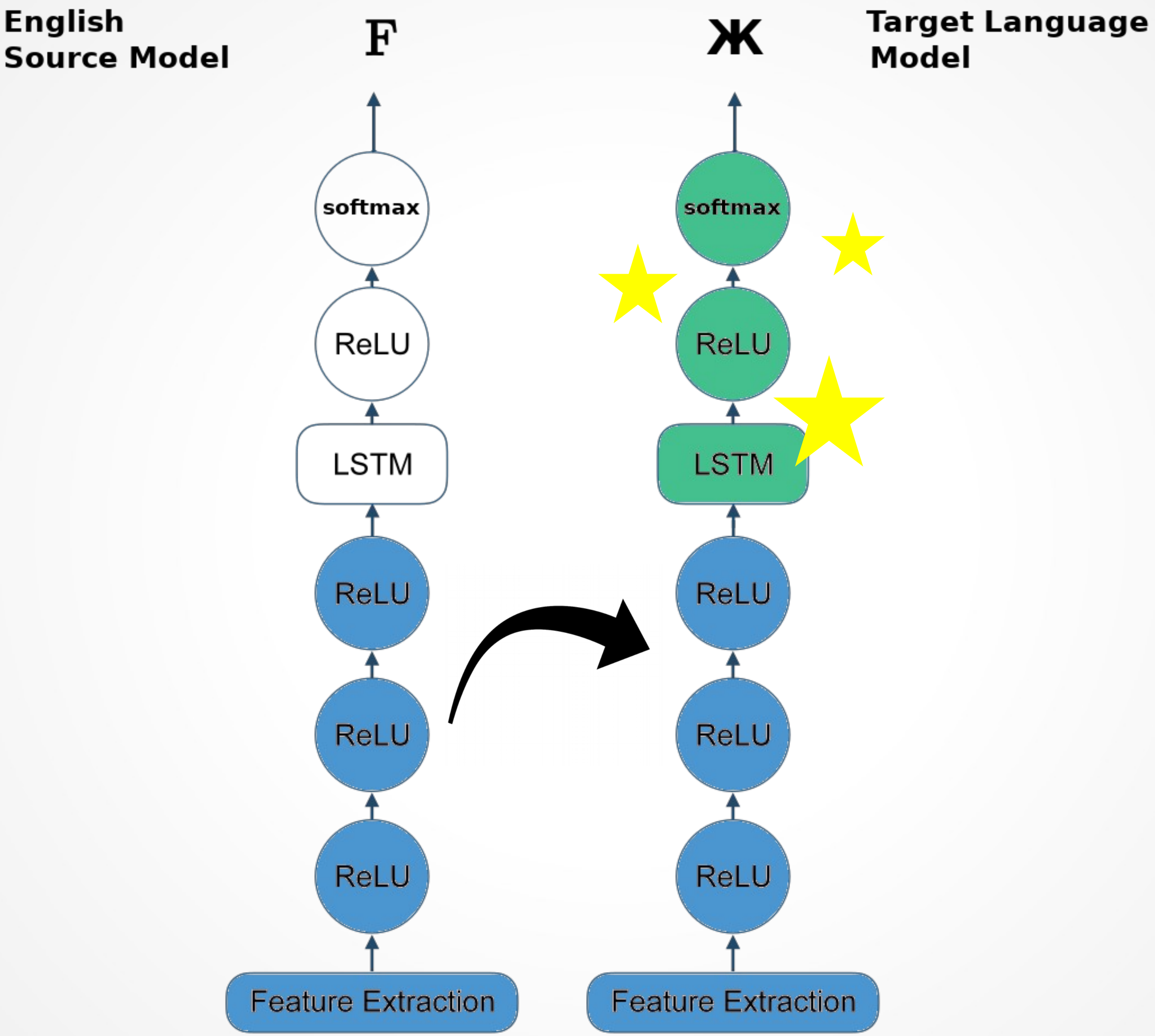
CTC Transfer Experiments



CTC Transfer Experiments



CTC Transfer Experiments



Experimental Design

5 depths for slicing source model

x 2 update scenarios (frozen vs. fine-tuned)

x 12 target languages

TOTAL == 120 experiments

Hyperparameters

Single GPU training

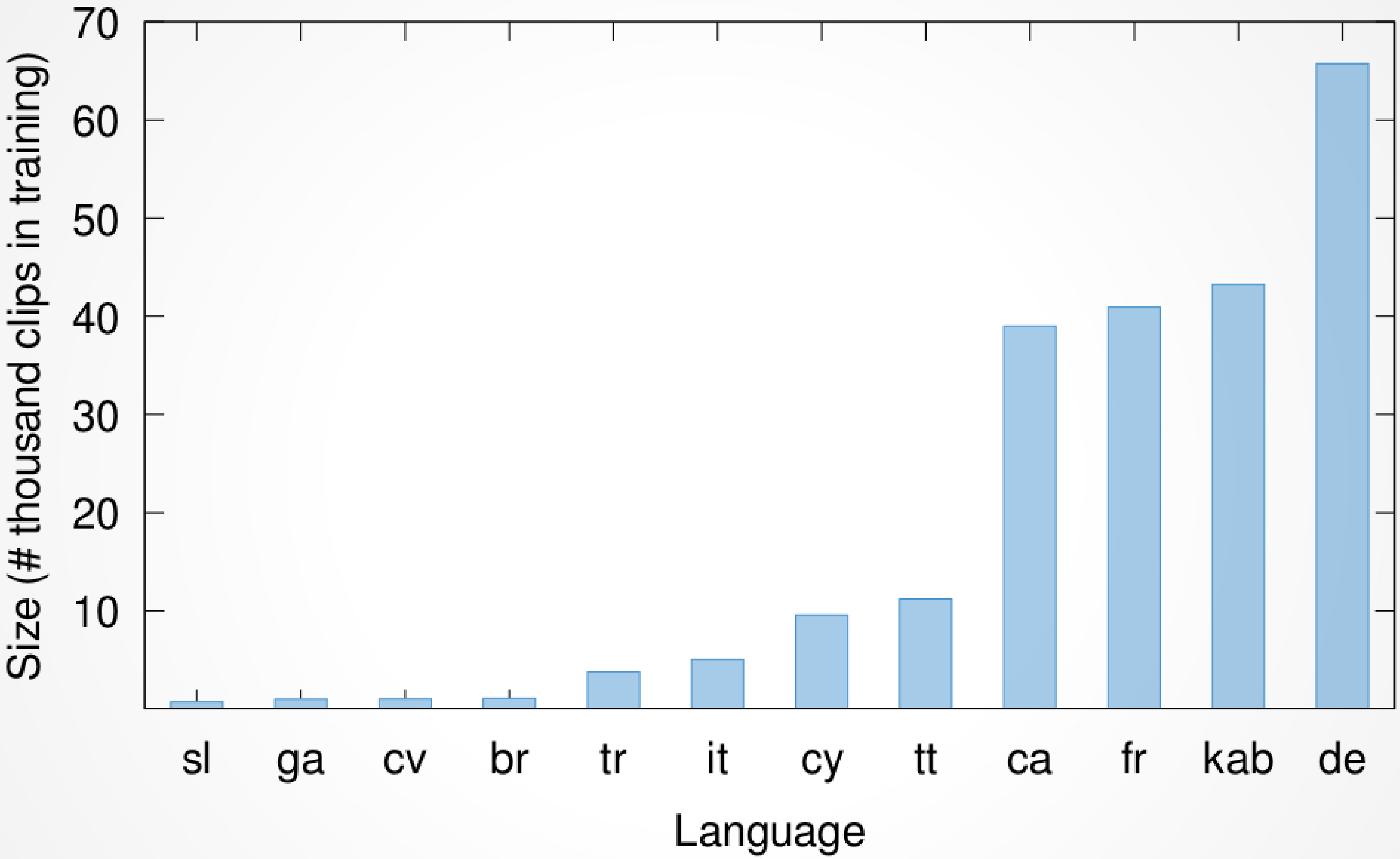
24 train batch, 48 dev batch

20% dropout rate

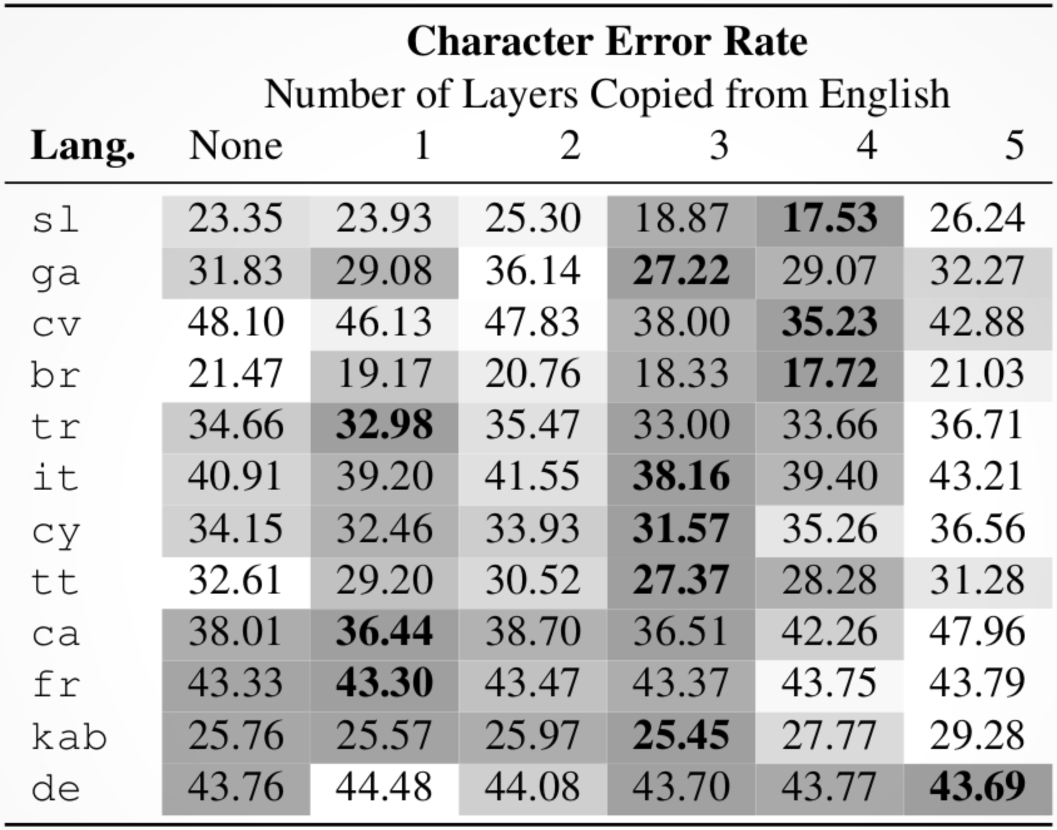
0.0001 learning rate with ADAM

Early stopping based on last 5 steps

Data (Spoken Corpora)



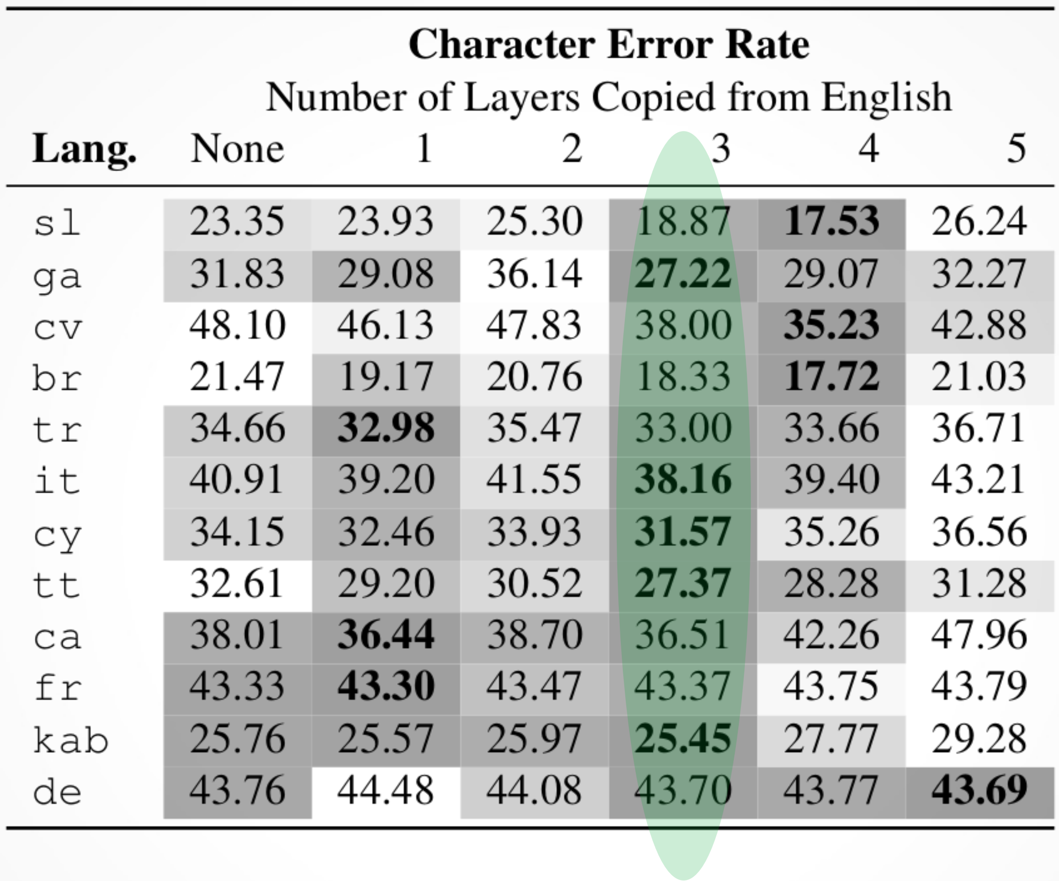
Frozen Transfer Results



	Character Error Rate					
	Number of Layers Copied from English					
Lang.	None	1	2	3	4	5
sl	23.35	23.93	25.30	18.87	**17.53**	26.24
ga	31.83	29.08	36.14	**27.22**	29.07	32.27
cv	48.10	46.13	47.83	38.00	**35.23**	42.88
br	21.47	19.17	20.76	18.33	**17.72**	21.03
tr	34.66	**32.98**	35.47	33.00	33.66	36.71
it	40.91	39.20	41.55	**38.16**	39.40	43.21
cy	34.15	32.46	33.93	**31.57**	35.26	36.56
tt	32.61	29.20	30.52	**27.37**	28.28	31.28
ca	38.01	**36.44**	38.70	36.51	42.26	47.96
fr	43.33	**43.30**	43.47	43.37	43.75	43.79
kab	25.76	25.57	25.97	**25.45**	27.77	29.28
de	43.76	44.48	44.08	43.70	43.77	**43.69**

Table 2. Frozen Transfer Learning Character-error rates (CER)

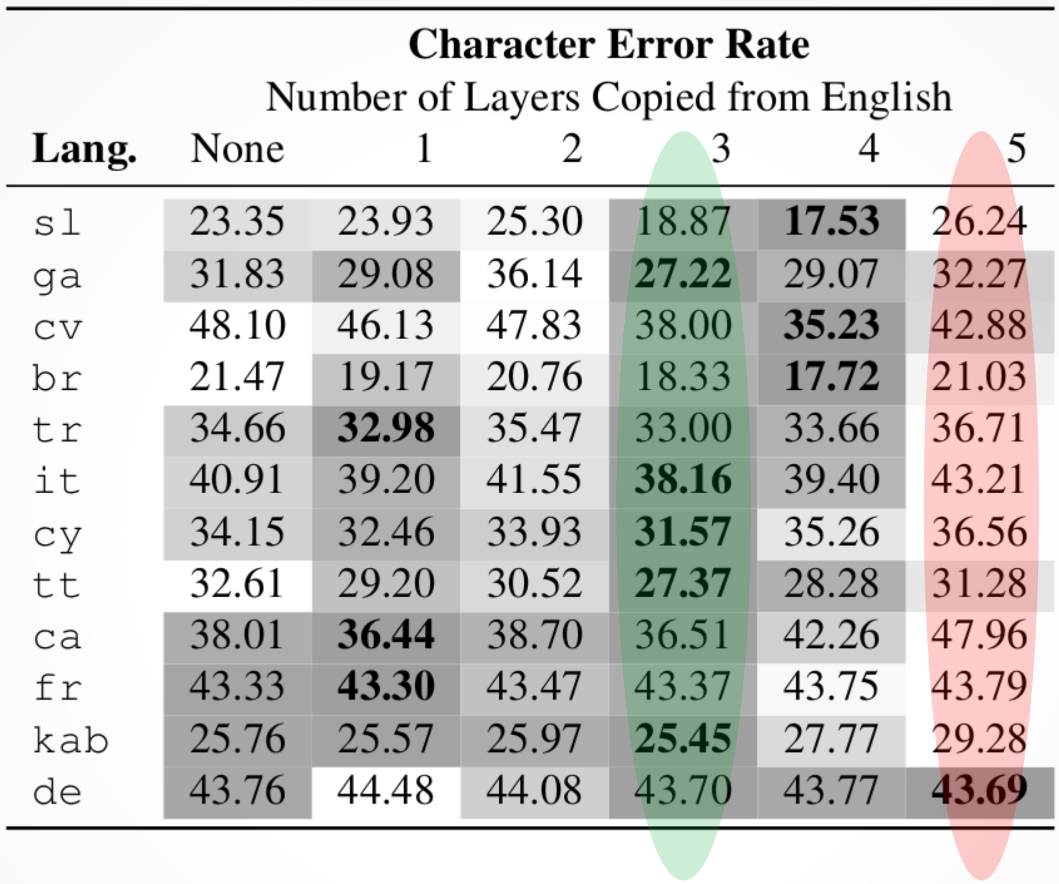
Frozen Transfer Results



	Character Error Rate					
	Number of Layers Copied from English					
Lang.	None	1	2	3	4	5
sl	23.35	23.93	25.30	18.87	**17.53**	26.24
ga	31.83	29.08	36.14	**27.22**	29.07	32.27
cv	48.10	46.13	47.83	38.00	**35.23**	42.88
br	21.47	19.17	20.76	18.33	**17.72**	21.03
tr	34.66	**32.98**	35.47	33.00	33.66	36.71
it	40.91	39.20	41.55	**38.16**	39.40	43.21
cy	34.15	32.46	33.93	**31.57**	35.26	36.56
tt	32.61	29.20	30.52	**27.37**	28.28	31.28
ca	38.01	**36.44**	38.70	36.51	42.26	47.96
fr	43.33	**43.30**	43.47	43.37	43.75	43.79
kab	25.76	25.57	25.97	**25.45**	27.77	29.28
de	43.76	44.48	44.08	43.70	43.77	**43.69**

Table 2. Frozen Transfer Learning Character-error rates (CER)

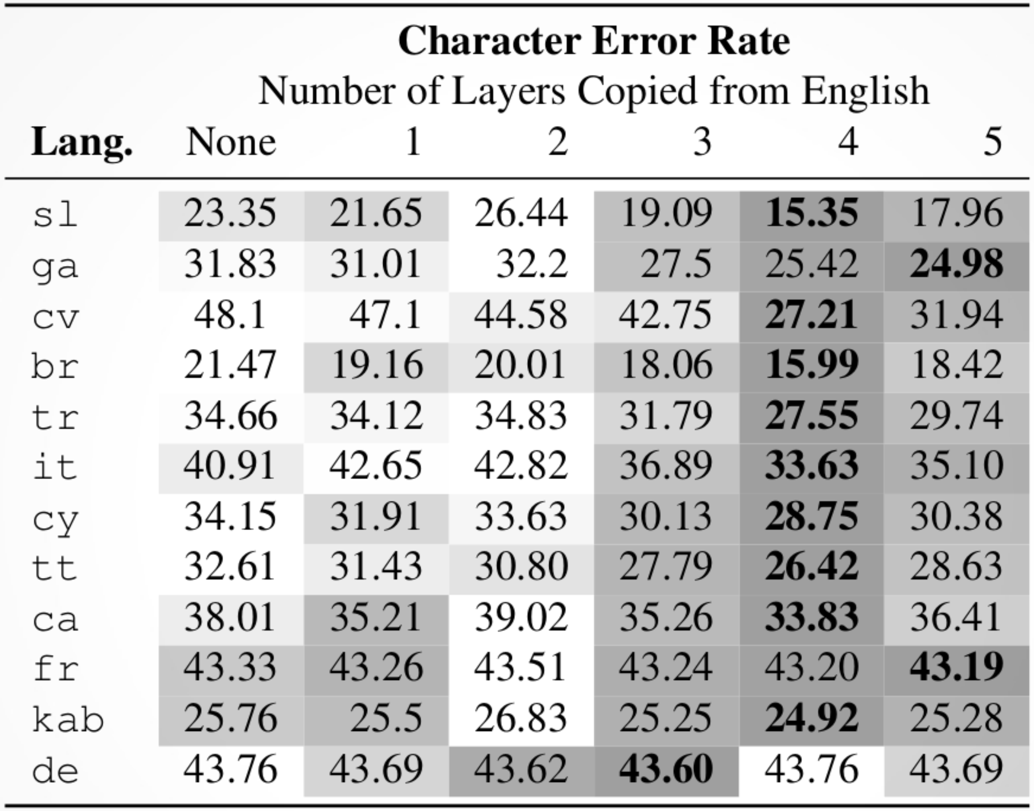
Frozen Transfer Results



	Character Error Rate					
	Number of Layers Copied from English					
Lang.	None	1	2	3	4	5
sl	23.35	23.93	25.30	18.87	**17.53**	26.24
ga	31.83	29.08	36.14	**27.22**	29.07	32.27
cv	48.10	46.13	47.83	38.00	**35.23**	42.88
br	21.47	19.17	20.76	18.33	**17.72**	21.03
tr	34.66	**32.98**	35.47	33.00	33.66	36.71
it	40.91	39.20	41.55	**38.16**	39.40	43.21
cy	34.15	32.46	33.93	**31.57**	35.26	36.56
tt	32.61	29.20	30.52	**27.37**	28.28	31.28
ca	38.01	**36.44**	38.70	36.51	42.26	47.96
fr	43.33	**43.30**	43.47	43.37	43.75	43.79
kab	25.76	25.57	25.97	**25.45**	27.77	29.28
de	43.76	44.48	44.08	43.70	43.77	**43.69**

Table 2. Frozen Transfer Learning Character-error rates (CER)

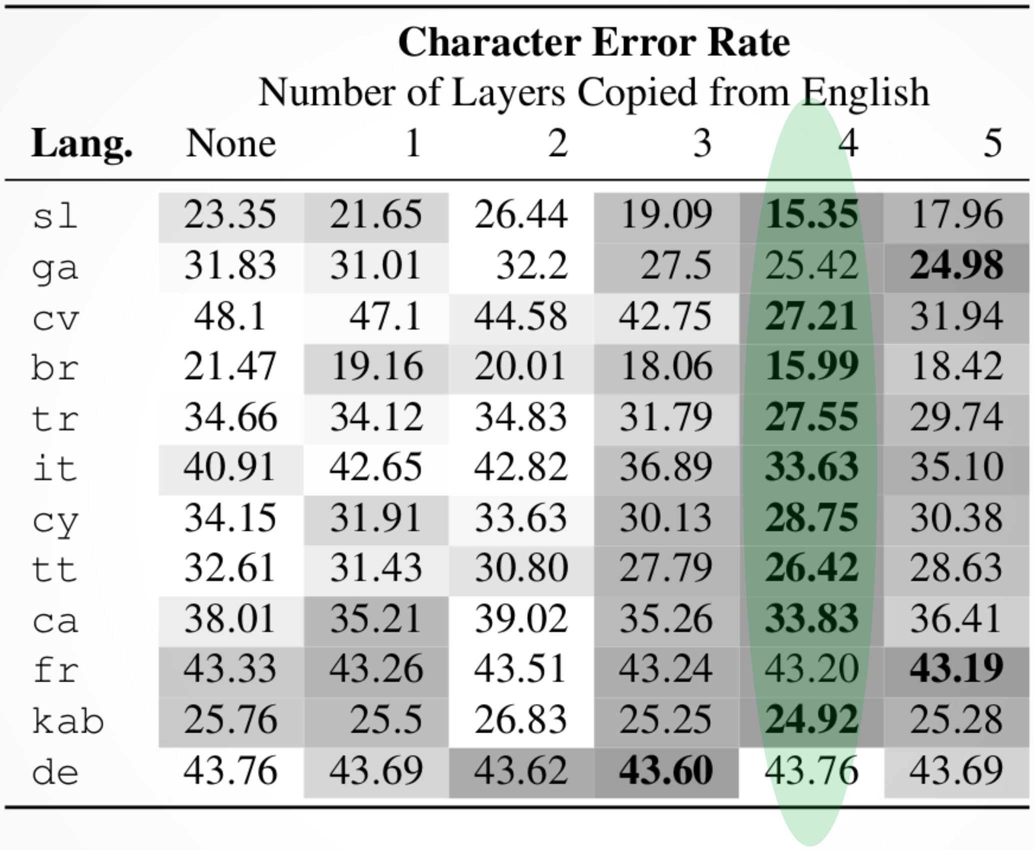
Fine-Tuning Transfer Results



	Character Error Rate					
	Number of Layers Copied from English					
Lang.	None	1	2	3	4	5
sl	23.35	21.65	26.44	19.09	**15.35**	17.96
ga	31.83	31.01	32.2	27.5	25.42	**24.98**
cv	48.1	47.1	44.58	42.75	**27.21**	31.94
br	21.47	19.16	20.01	18.06	**15.99**	18.42
tr	34.66	34.12	34.83	31.79	**27.55**	29.74
it	40.91	42.65	42.82	36.89	**33.63**	35.10
cy	34.15	31.91	33.63	30.13	**28.75**	30.38
tt	32.61	31.43	30.80	27.79	**26.42**	28.63
ca	38.01	35.21	39.02	35.26	**33.83**	36.41
fr	43.33	43.26	43.51	43.24	43.20	**43.19**
kab	25.76	25.5	26.83	25.25	**24.92**	25.28
de	43.76	43.69	43.62	**43.60**	43.76	43.69

Table 3. Fine-Tuned Transfer Learning Character-error rates (CER)

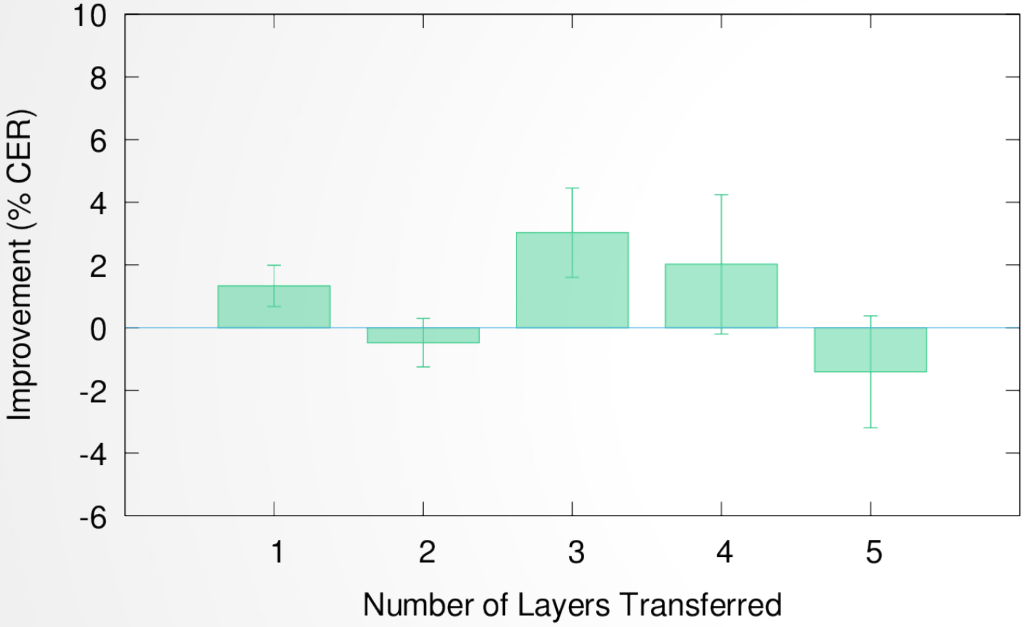
Fine-Tuning Transfer Results

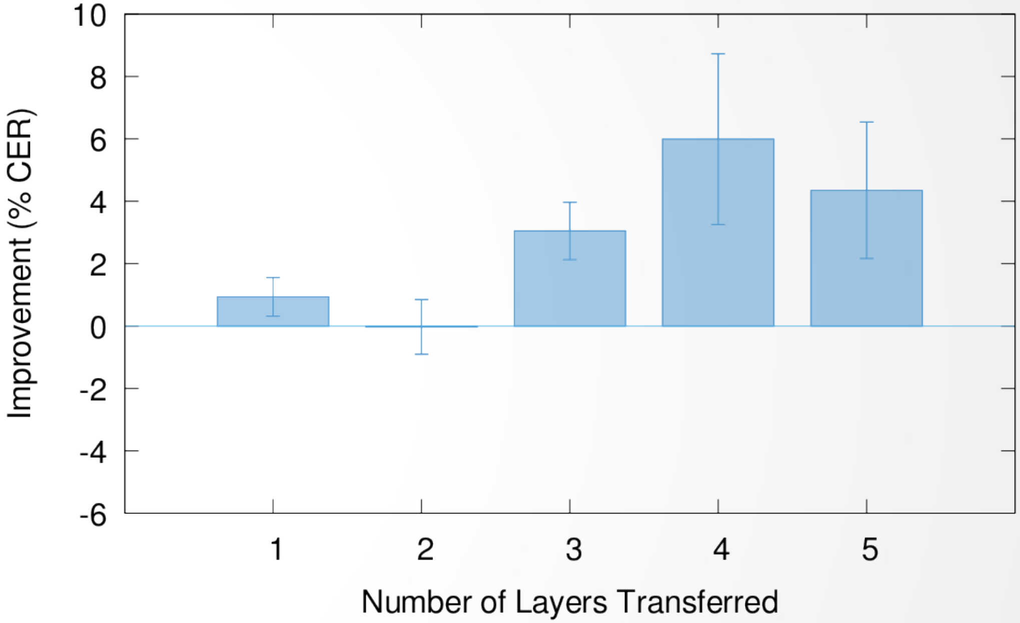


	Character Error Rate					
	Number of Layers Copied from English					
Lang.	None	1	2	3	4	5
sl	23.35	21.65	26.44	19.09	**15.35**	17.96
ga	31.83	31.01	32.2	27.5	25.42	**24.98**
cv	48.1	47.1	44.58	42.75	**27.21**	31.94
br	21.47	19.16	20.01	18.06	**15.99**	18.42
tr	34.66	34.12	34.83	31.79	**27.55**	29.74
it	40.91	42.65	42.82	36.89	**33.63**	35.10
cy	34.15	31.91	33.63	30.13	**28.75**	30.38
tt	32.61	31.43	30.80	27.79	**26.42**	28.63
ca	38.01	35.21	39.02	35.26	**33.83**	36.41
fr	43.33	43.26	43.51	43.24	43.20	**43.19**
kab	25.76	25.5	26.83	25.25	**24.92**	25.28
de	43.76	43.69	43.62	**43.60**	43.76	43.69

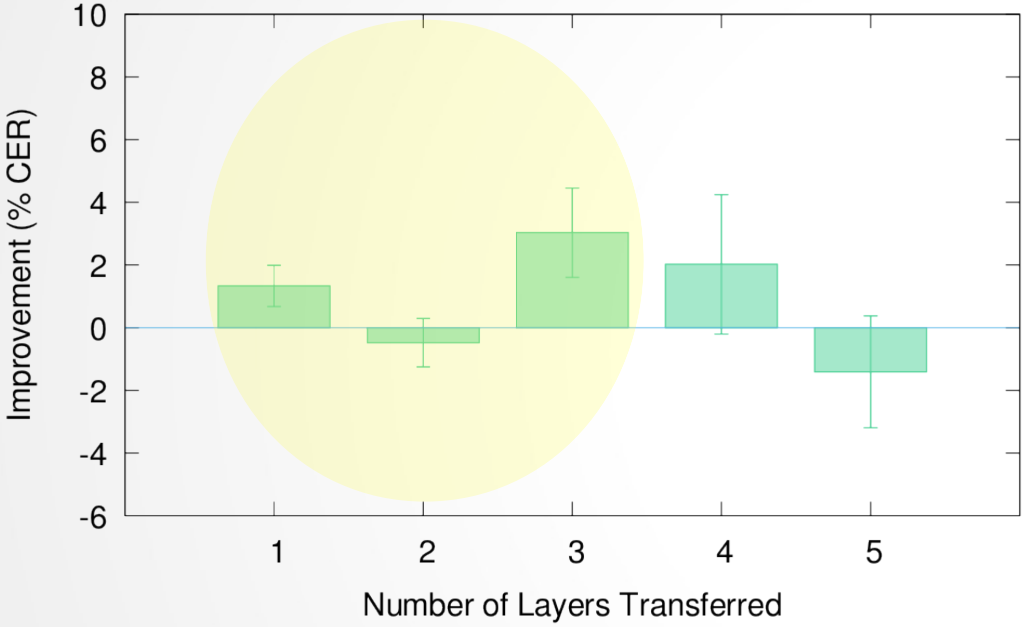
Table 3. Fine-Tuned Transfer Learning Character-error rates (CER)

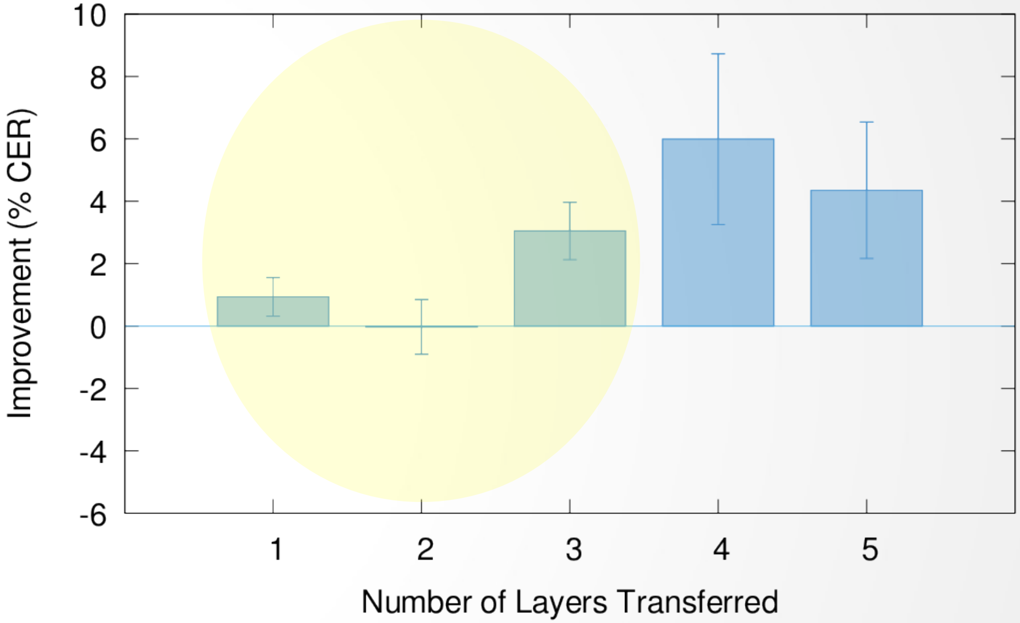
Frozen vs. Fine-Tuned





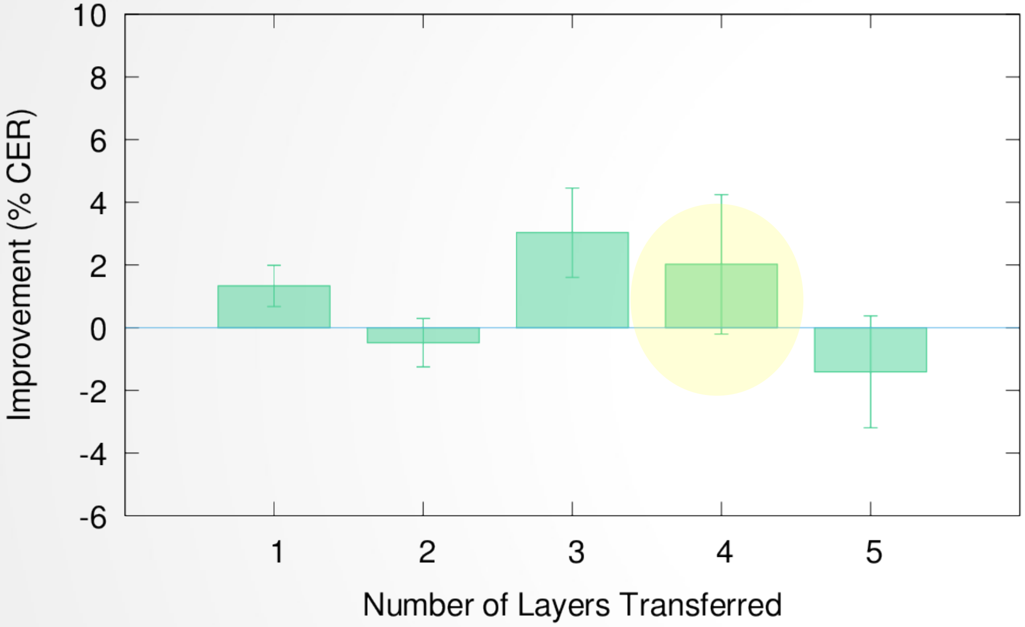
Frozen vs. Fine-Tuned

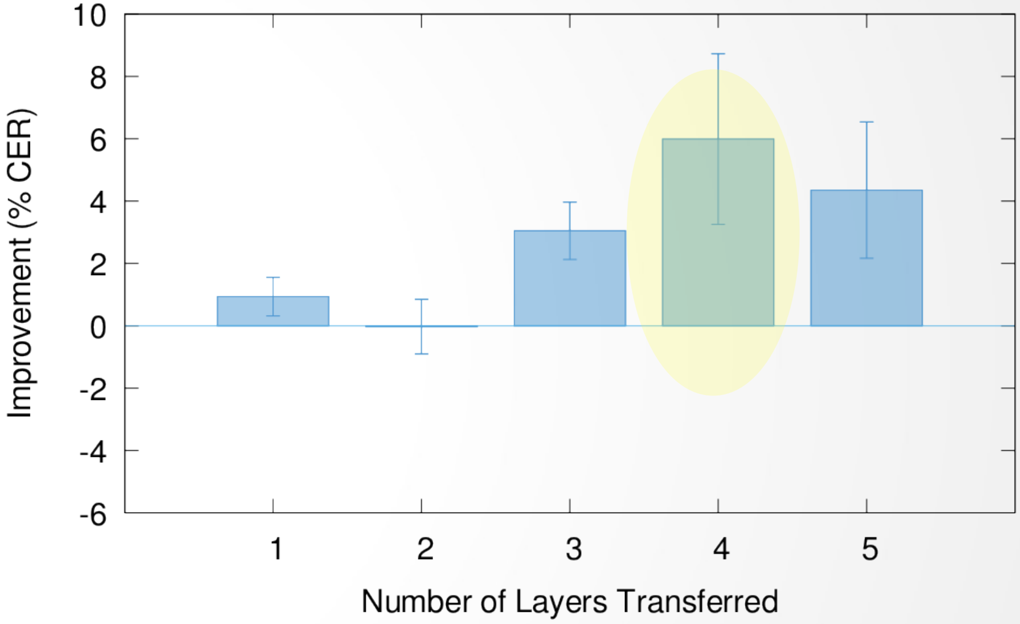




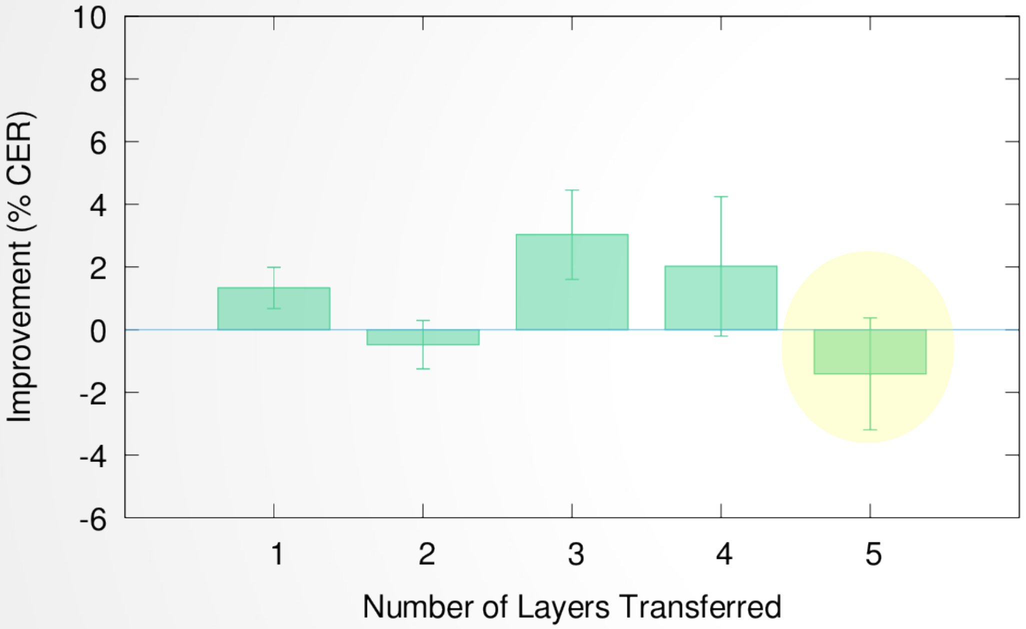
Frozen vs. Fine-Tuned

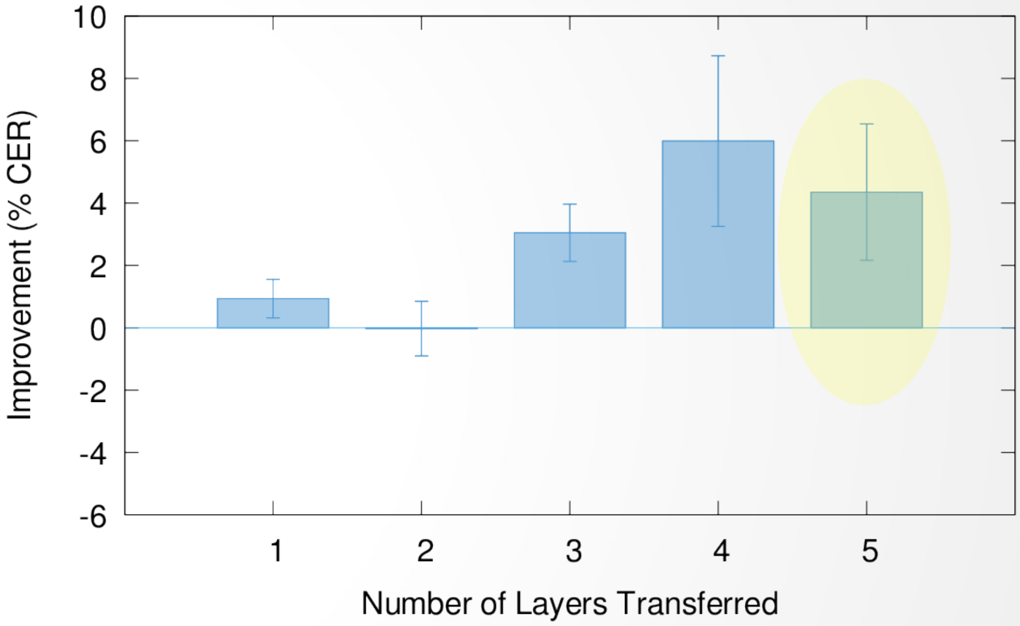
LSTM!





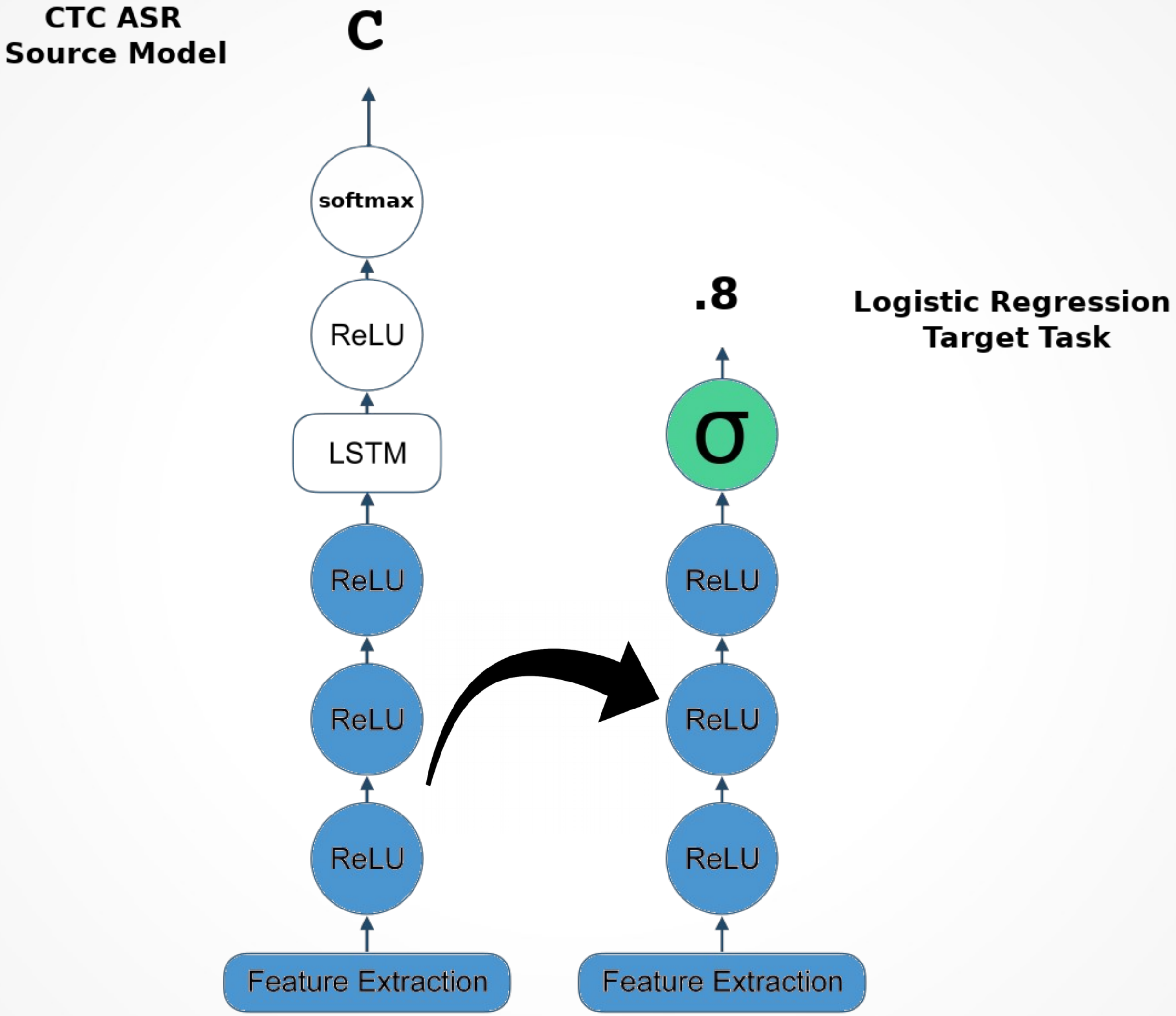
Frozen vs. Fine-Tuned



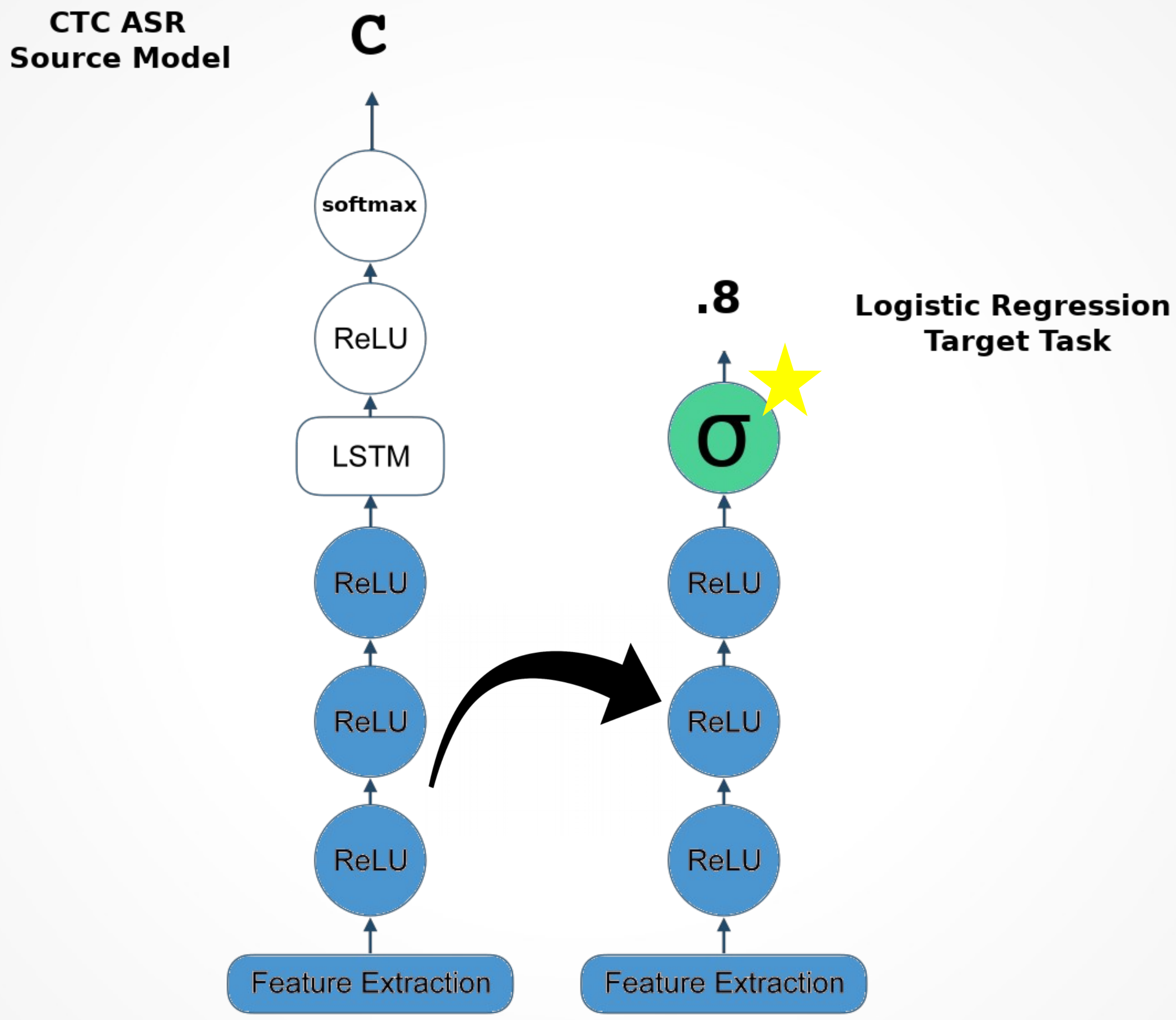


Interpretability Experiments

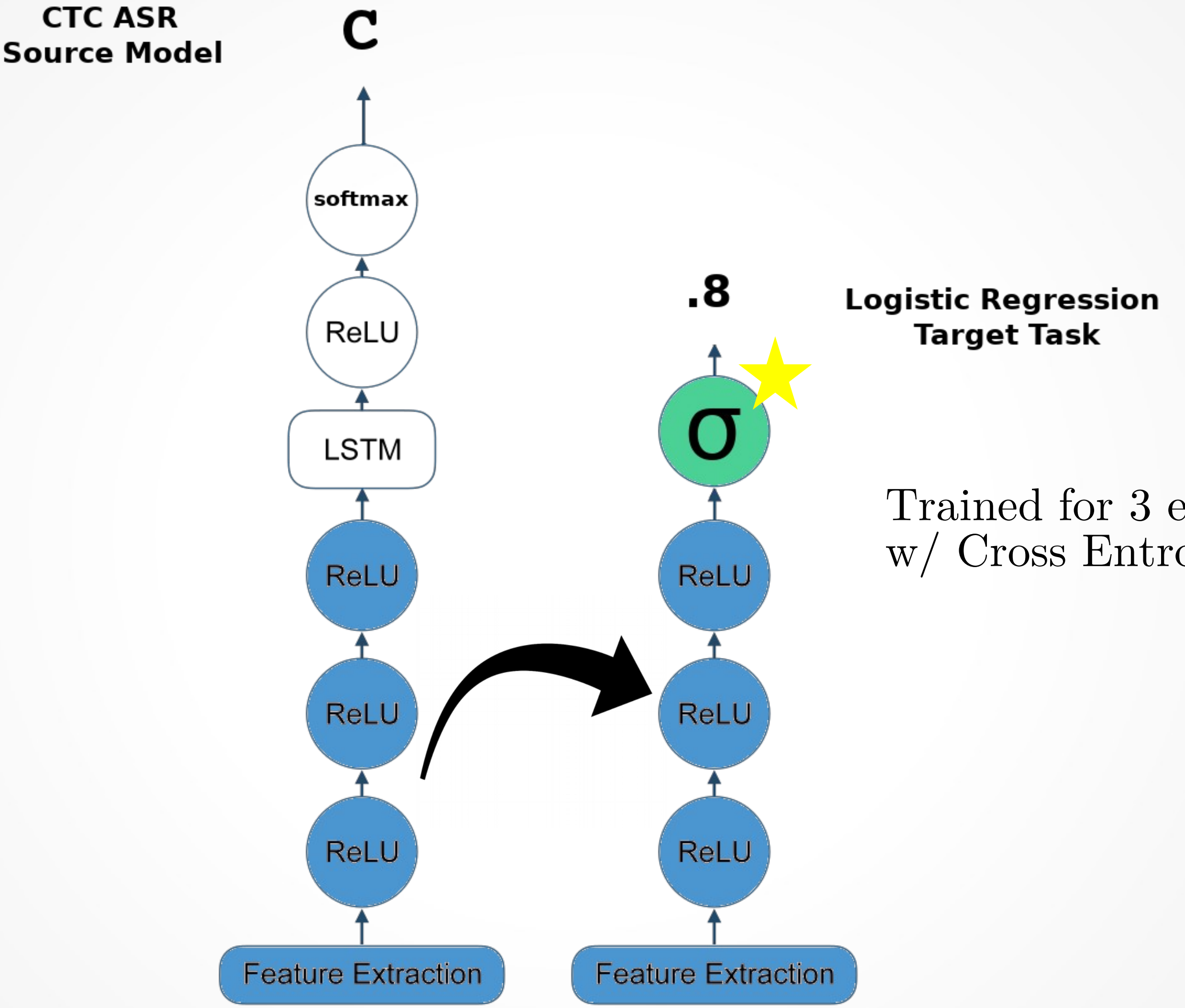
Regression on Embeddings



Regression on Embeddings



Regression on Embeddings



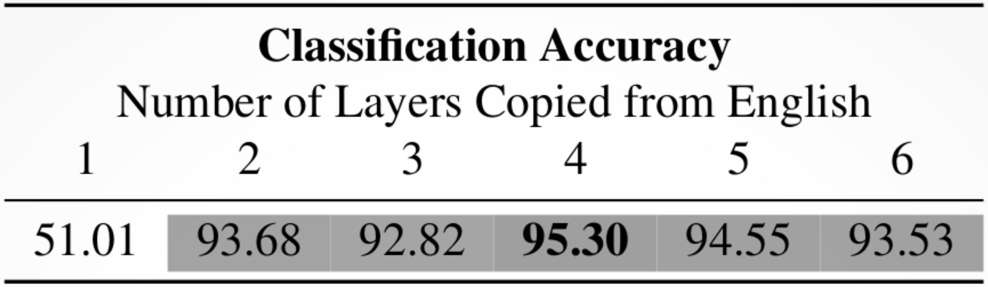
Trained for 3 epochs
w/ Cross Entropy Loss

Regression Results

Speech vs. Noise

- Copied layers, added final FC layer with single output and logistic activation
- 13 languages vs. UrbanSound8k
- 5,005 train clips, 442 test clips per class

Regression Results

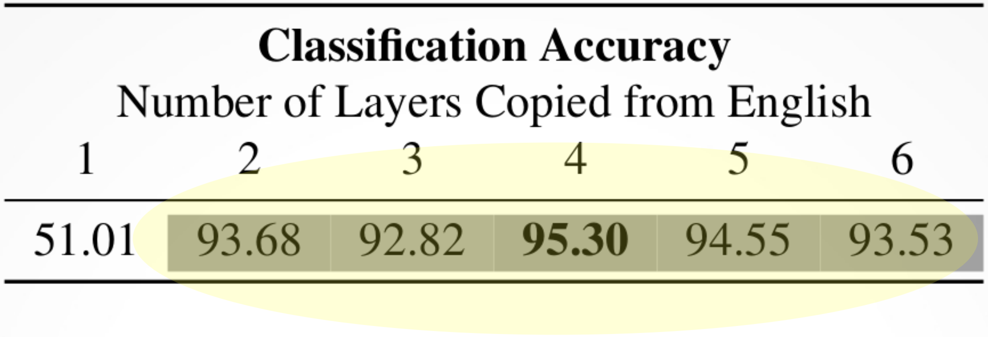


Classification Accuracy					
Number of Layers Copied from English					
1	2	3	4	5	6
51.01	93.68	92.82	**95.30**	94.55	93.53

Table 4. Speech vs. Non-Speech Audio Classification Accuracy

- Copied layers, added final FC layer with single output and logistic activation
- 13 languages vs. UrbanSound8k
- 5,005 train clips, 442 test clips per class

Regression Results



Classification Accuracy					
Number of Layers Copied from English					
1	2	3	4	5	6
51.01	93.68	92.82	**95.30**	94.55	93.53

Table 4. Speech vs. Non-Speech Audio Classification Accuracy

- Copied layers, added final FC layer with single output and logistic activation
- 13 languages vs. UrbanSound8k
- 5,005 train clips, 442 test clips per class

Regression Results

English vs. German

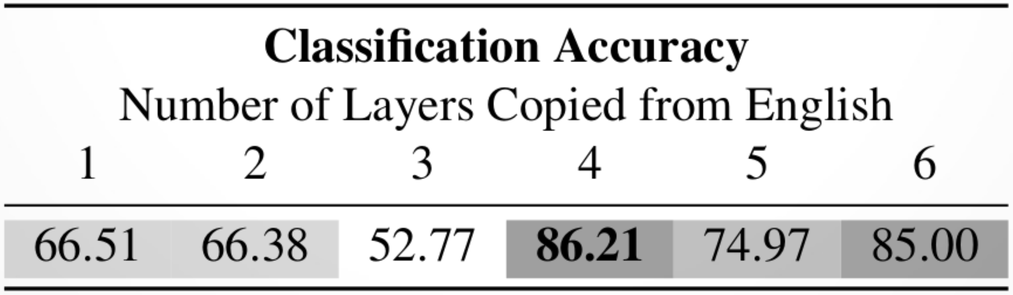
Regression Results

English vs. German

- Copied layers, added final FC layer with single output and logistic activation
- English vs. German
- 5,000 train clips, 500 test clips per class

Regression Results

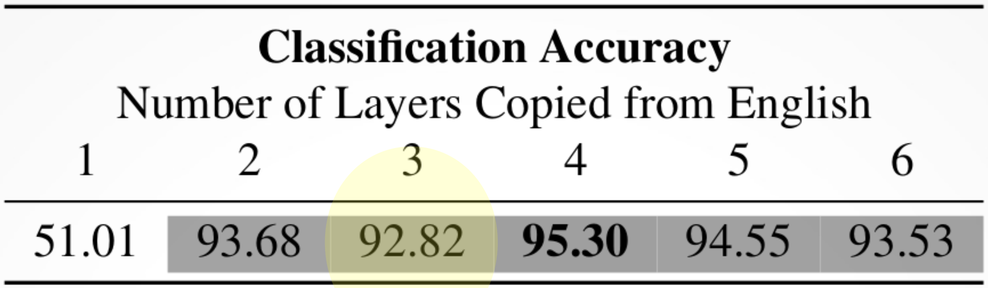
English vs. German



Classification Accuracy					
Number of Layers Copied from English					
1	2	3	4	5	6
66.51	66.38	52.77	**86.21**	74.97	85.00

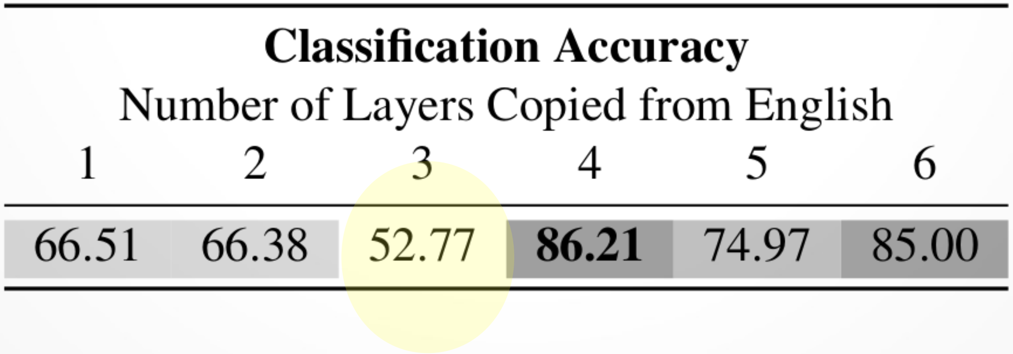
Table 5. English vs. German Audio Classification Accuracy (%)

Regression Results



Classification Accuracy					
Number of Layers Copied from English					
1	2	3	4	5	6
51.01	93.68	92.82	**95.30**	94.55	93.53

Table 4. Speech vs. Non-Speech Audio Classification Accuracy



Classification Accuracy					
Number of Layers Copied from English					
1	2	3	4	5	6
66.51	66.38	52.77	**86.21**	74.97	85.00

Table 5. English vs. German Audio Classification Accuracy (%)

Discussion

Discussion

1) Transfer in ASR

- Fine-tuning always helps

- LSTM transfer is best, but only with fine-tuning

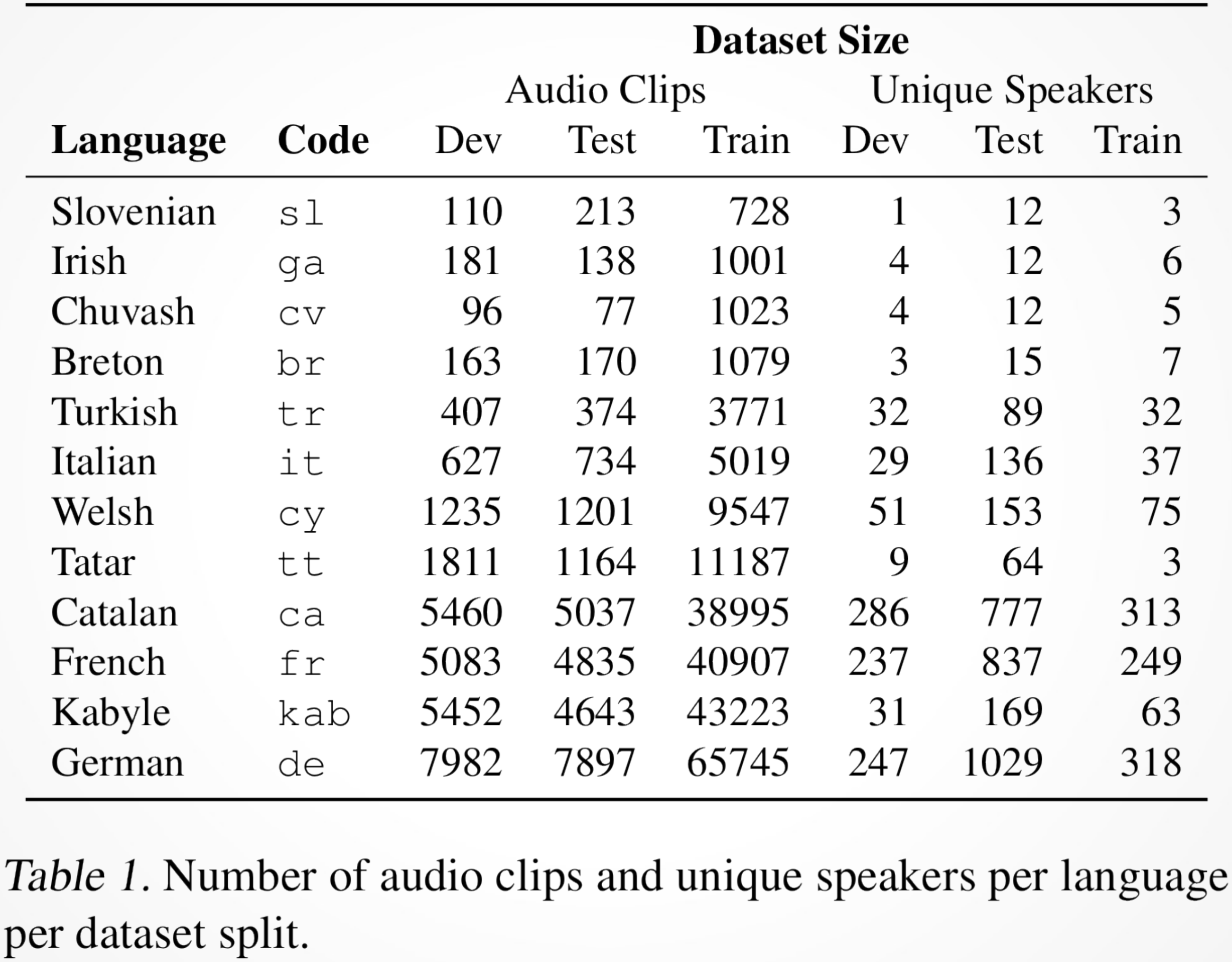
2) Interpretability Studies

- At the third layer, the model has learned general speech, but language-agnostic representations

Thank you for your attention!

APPENDIX B: DeepSpeech

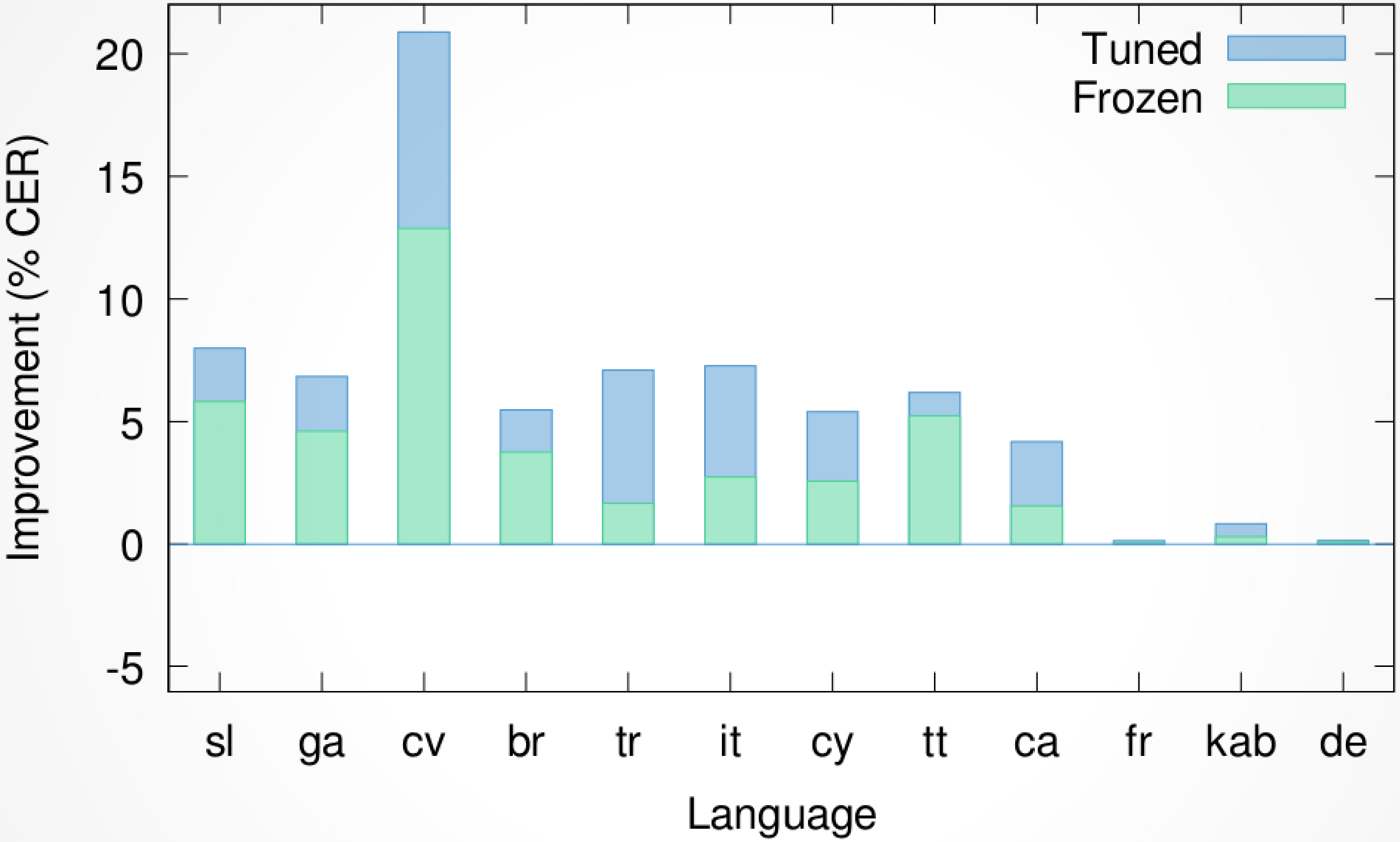
Data Details



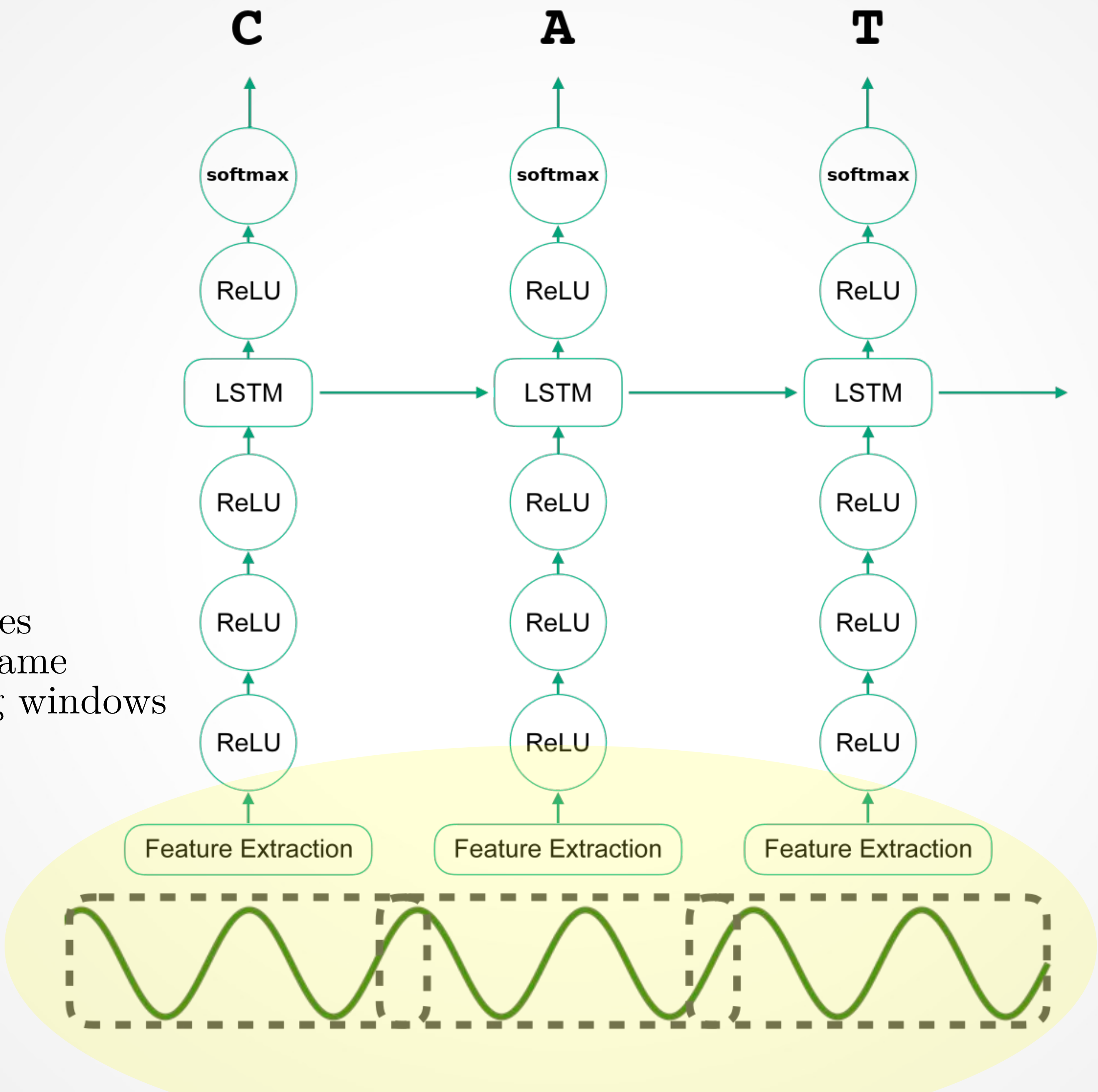
		Dataset Size					
		Audio Clips			Unique Speakers		
Language	**Code**	Dev	Test	Train	Dev	Test	Train
Slovenian	`sl`	110	213	728	1	12	3
Irish	`ga`	181	138	1001	4	12	6
Chuvash	`cv`	96	77	1023	4	12	5
Breton	`br`	163	170	1079	3	15	7
Turkish	`tr`	407	374	3771	32	89	32
Italian	`it`	627	734	5019	29	136	37
Welsh	`cy`	1235	1201	9547	51	153	75
Tatar	`tt`	1811	1164	11187	9	64	3
Catalan	`ca`	5460	5037	38995	286	777	313
French	`fr`	5083	4835	40907	237	837	249
Kabyle	`kab`	5452	4643	43223	31	169	63
German	`de`	7982	7897	65745	247	1029	318

Table 1. Number of audio clips and unique speakers per language per dataset split.

Effect of Data Size

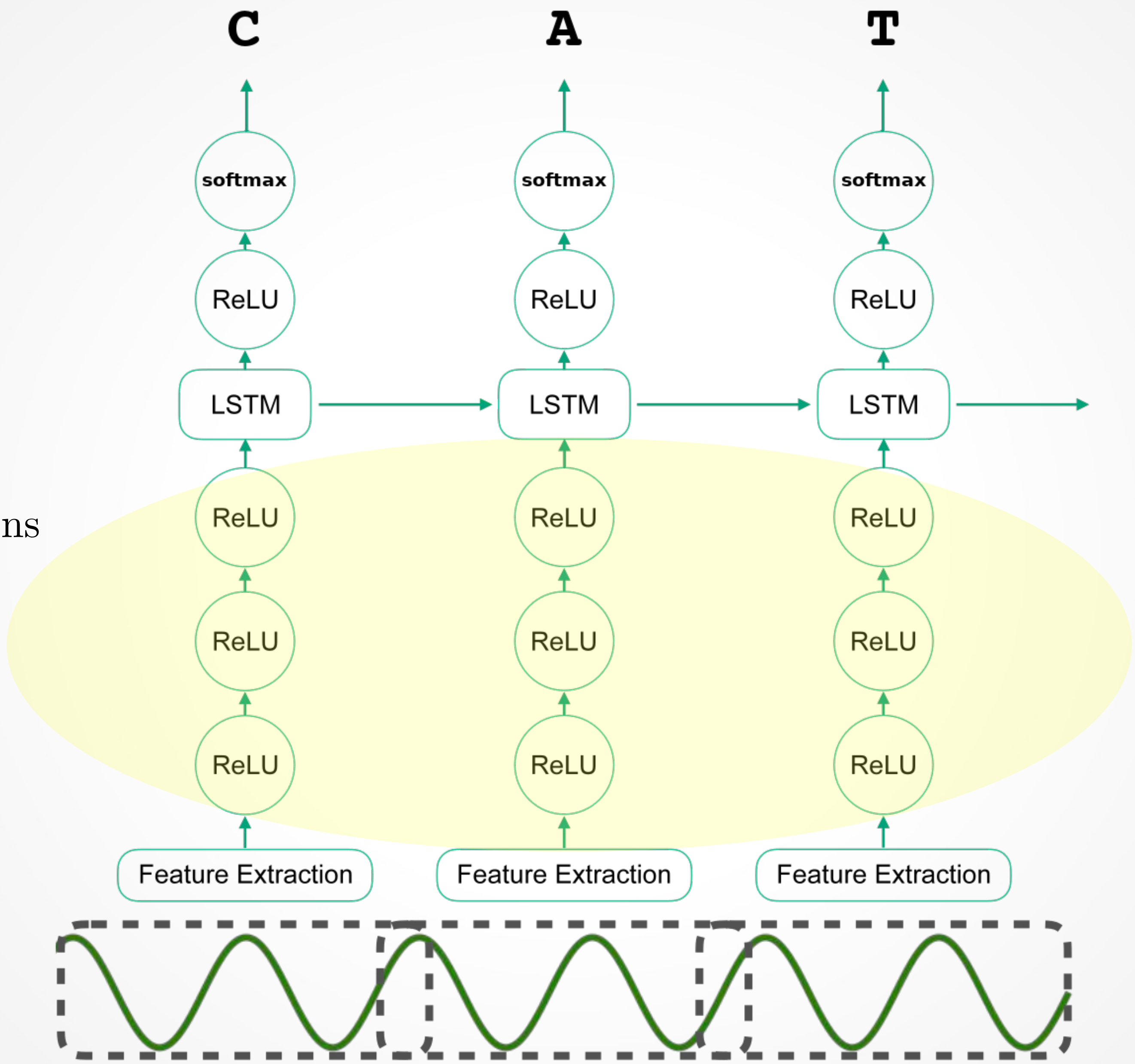


Model Architecture



19 spliced frames
26 MFCCs / frame
32ms Hamming windows
20ms timestep

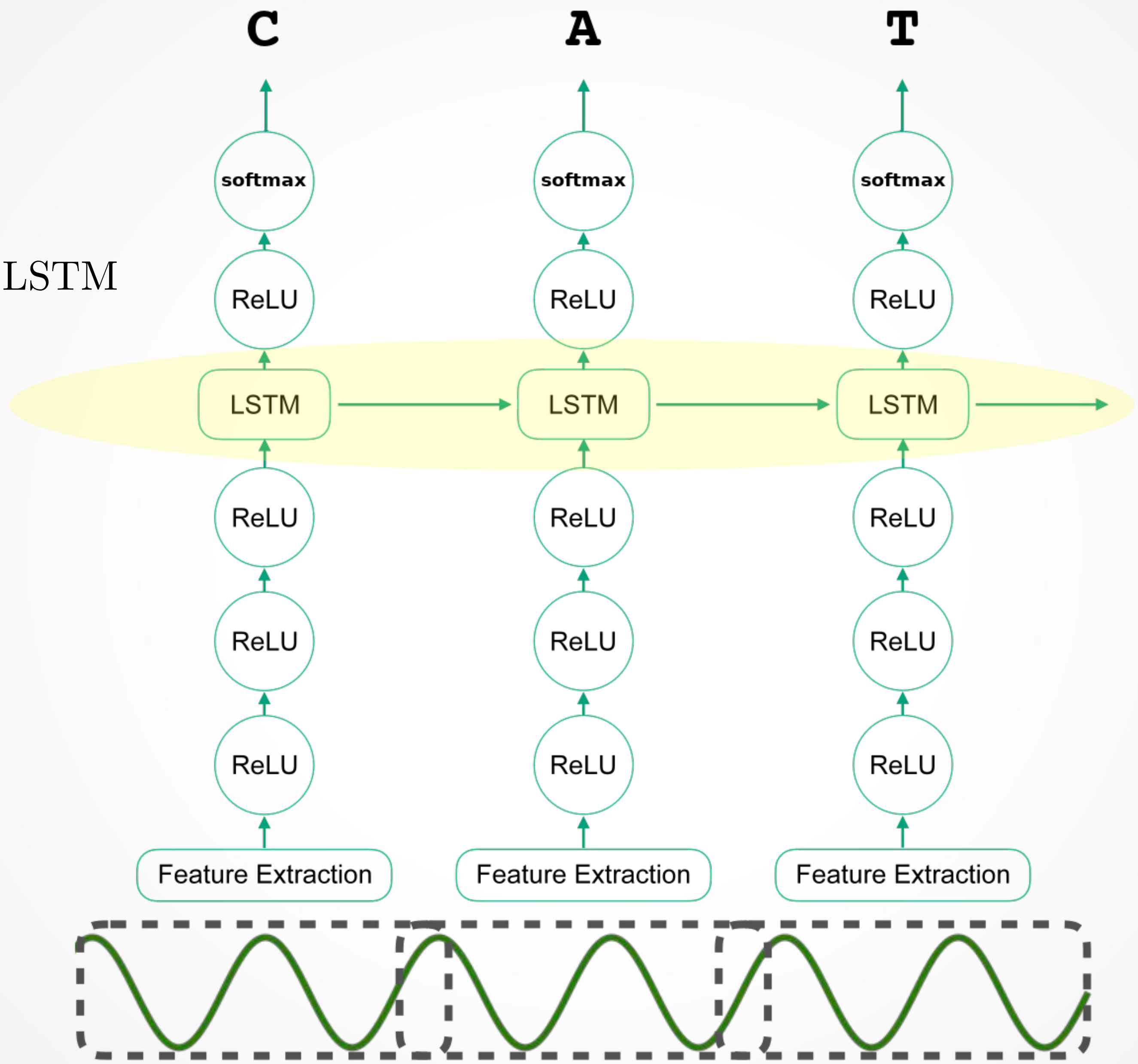
Model Architecture



Fully connected
Feed-Forward
2048 dims
ReLU activiations

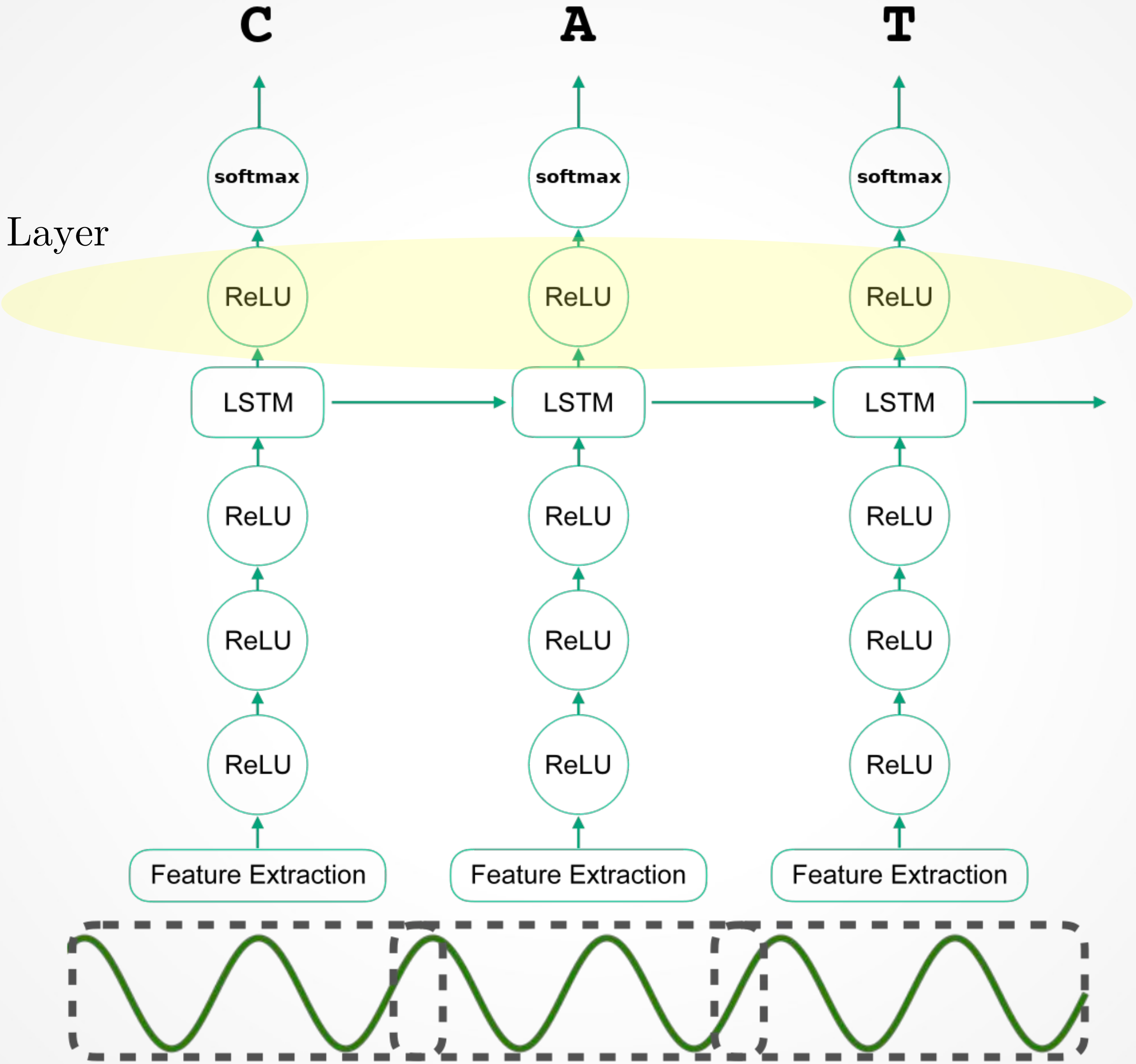
Model Architecture

Unidirectional LSTM
2048 dims



Model Architecture

2048 dim FC Layer

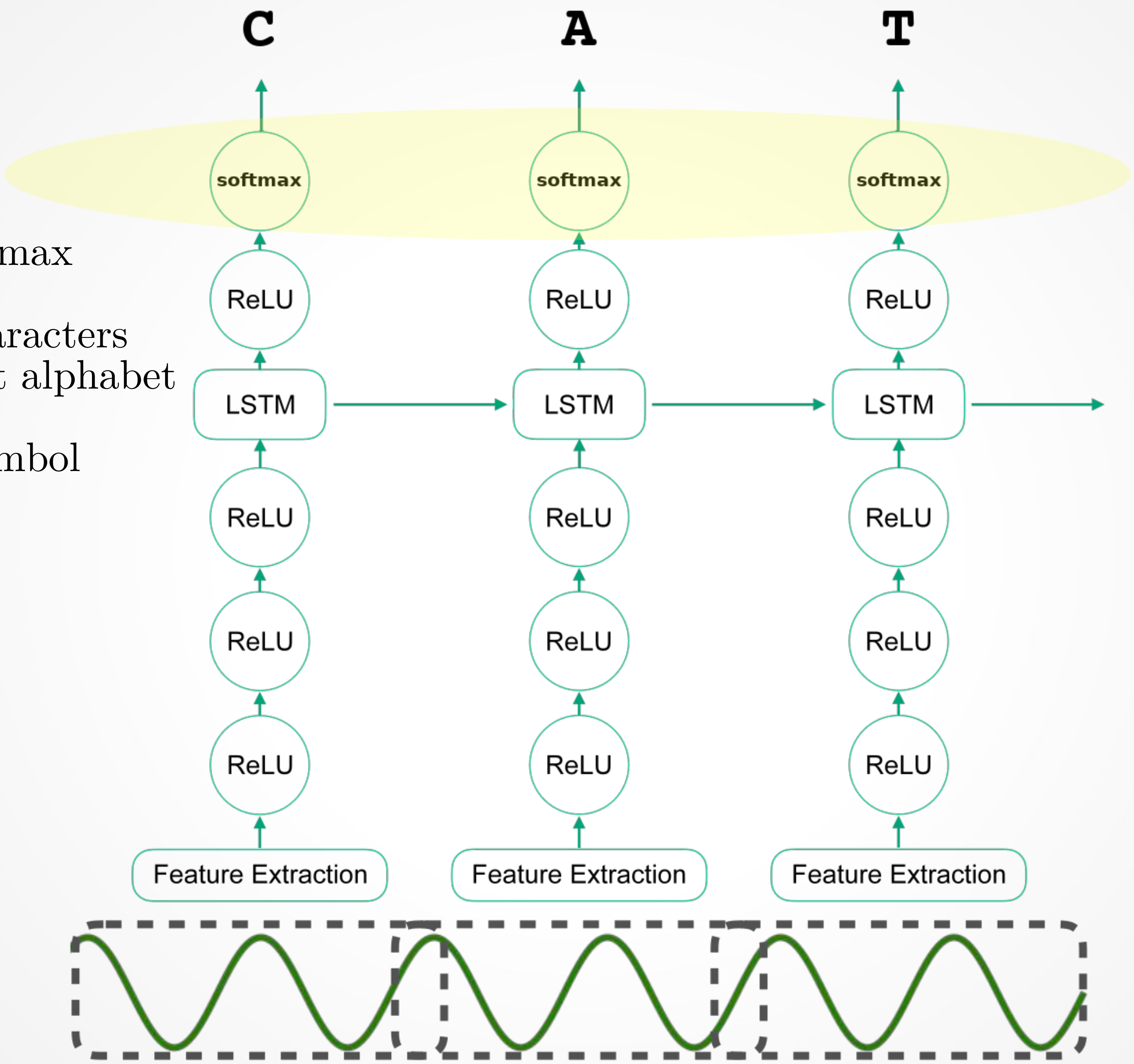


Model Architecture

N+1 dims Softmax

N == num characters in target alphabet

1 == blank symbol


C
A
T
softmax
softmax
softmax
ReLU
ReLU
ReLU
LSTM
LSTM
LSTM
ReLU
ReLU
ReLU
ReLU
ReLU
ReLU
ReLU
ReLU
ReLU
Feature Extraction
Feature Extraction
Feature Extraction