Learning Acoustic Word Embeddings from Sequence-to-Sequence Models

Carnegie Mellon University

Shruti Palaskar

Carnegie Mellon University
Language Technologies Institute

What is this talk about?

How to cram meaning of speech into a vector!?!

But…

“You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!"

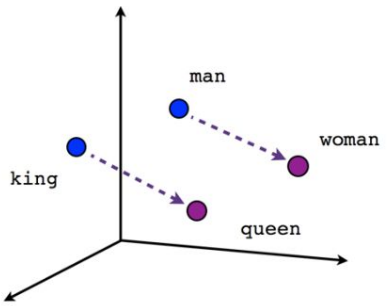
- Raymond Mooney

How to *try to* cram the meaning of a whole sentence into a single vector?

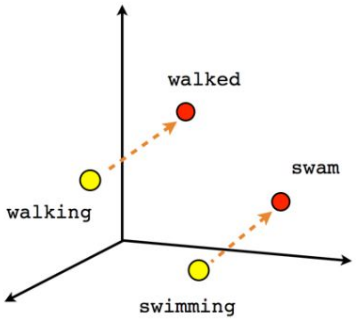
- ELMo, BERT
- word2vec, glove

Text Embeddings

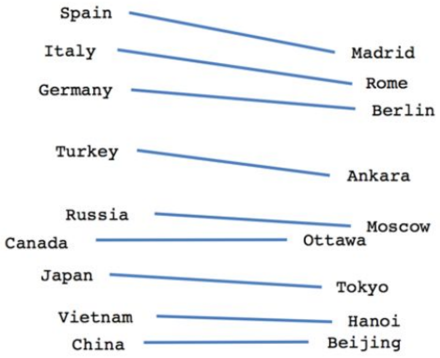
- ➢ Representing written words or sentences as continuous valued fixed dimensional vectors
- ➢ Common representation for various words/sentences/languages
- ➢ Useful as off-the-shelf pre-trained features for other tasks



Male-Female



Verb tense

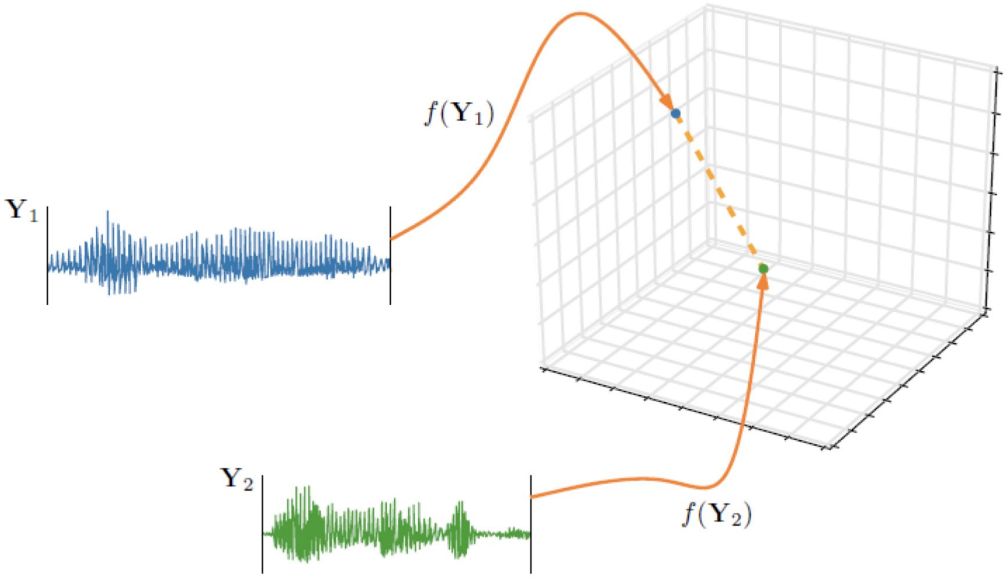


Country-Capital

Acoustic Embeddings

➢ Map speech signal of arbitrary length into a fixed dimensional vector

➢ This speech signal may be for a word or a sentence



[Figure credit: Herman Kamper]

Acoustic Embeddings

- ➢ Represent speech (an inherently continuous signal) into embeddings (fixed dimensional vectors)

- ➢ Speech has many more variations than text like:

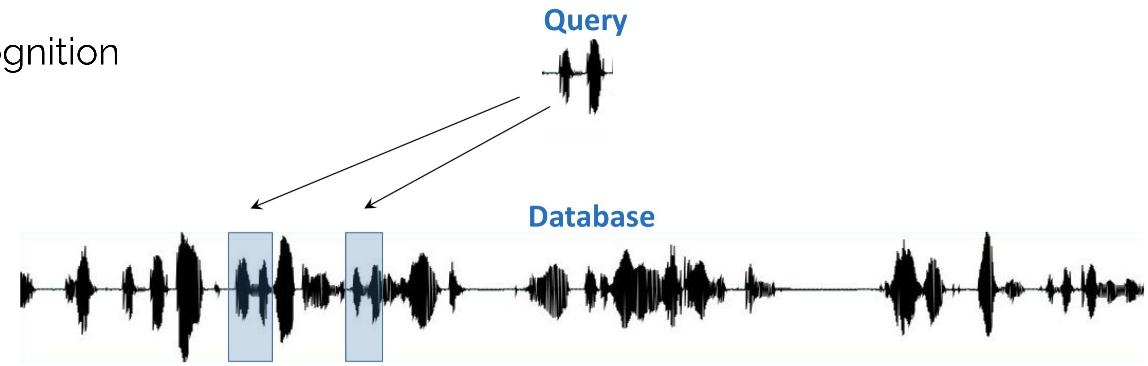
  speaking rate, pronunciation variance, speaker differences,

  acoustic environment, prosody (emotion etc), intonation, ...

- ➢ Can we do the same with speech as text then? Lets see...

Acoustic Embedding: Uses & Applications

➢ Speech Similarity tasks
  - Spoken Language Understanding
  - Whole-word Speech Recognition
  - Spoken Term Discovery
  - Query-by-example



[Figure credit: Herman Kamper]

Acoustic Embedding: Uses & Applications

- Shared representation for speech and other modalities (like text or vision)
  - Easier multimodal interaction for these different modalities
  - Given speech, retrieve text / Given speech retrieve corresponding video!

Speech segment of “CAT”

CAT

Talk Outline

I. Learning Acoustic Word Embeddings
   A. Model: Acoustic-to-Word Speech Recognition
   B. Understanding A2W models
   C. Evaluation
II. Applications of Acoustic Word Embeddings
   A. Spoken Language Understanding
   B. Unsupervised speech recognition and spoken language translation

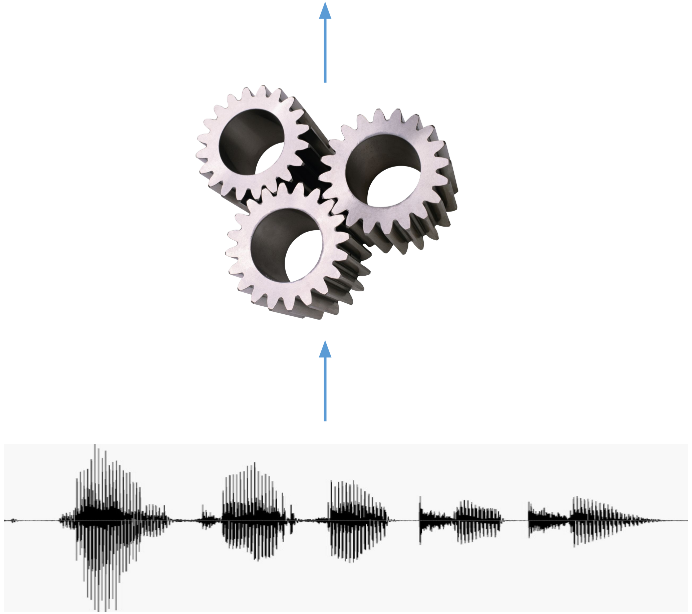
Talk Outline

Acoustic-to-Word Speech Recognition

This Speech Recognizer can *Recognize Speech*

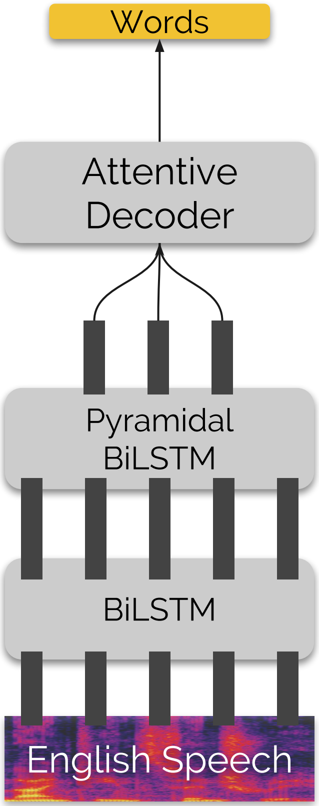
Acoustic-to-Word Speech Recognition

This Speech Recognizer can *Wreck a Nice Beach*



Acoustic-to-Word Speech Recognition

- ➢ Model Probability (Words | Acoustics)
- ➢ These acoustics could be *any* form of representation of speech
- ➢ Sequence-to-Sequence model with attention
- ➢ Around 30,000 words vocabulary
- ➢ Usually 26 character vocabulary (English)
- ➢ No alignment needed like traditional speech recognizers



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

Results

- ➢ Evaluation: Word Error Rate
- ➢ On a standard dataset *Switchboard*

Character models = **15.6%**

Word models = **22.1%**

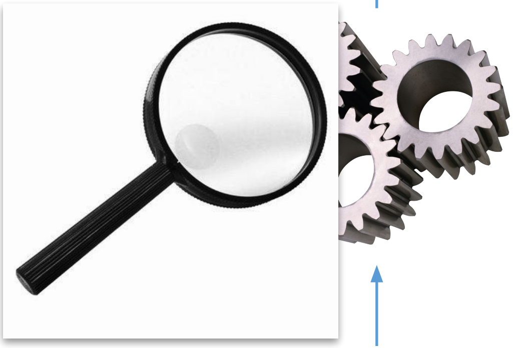
- ➢ But whole words are semantically meaningful units!
- ➢ Can perform non-speech transcription task with speech input!

Palaskar and Metze, "Acoustic-to-Word Recognition with Sequence-to-Sequence Models", 2018

Talk Outline

Understanding Acoustic-to-Word Models

This Speech Recognizer can *Wreck a Nice Beach*

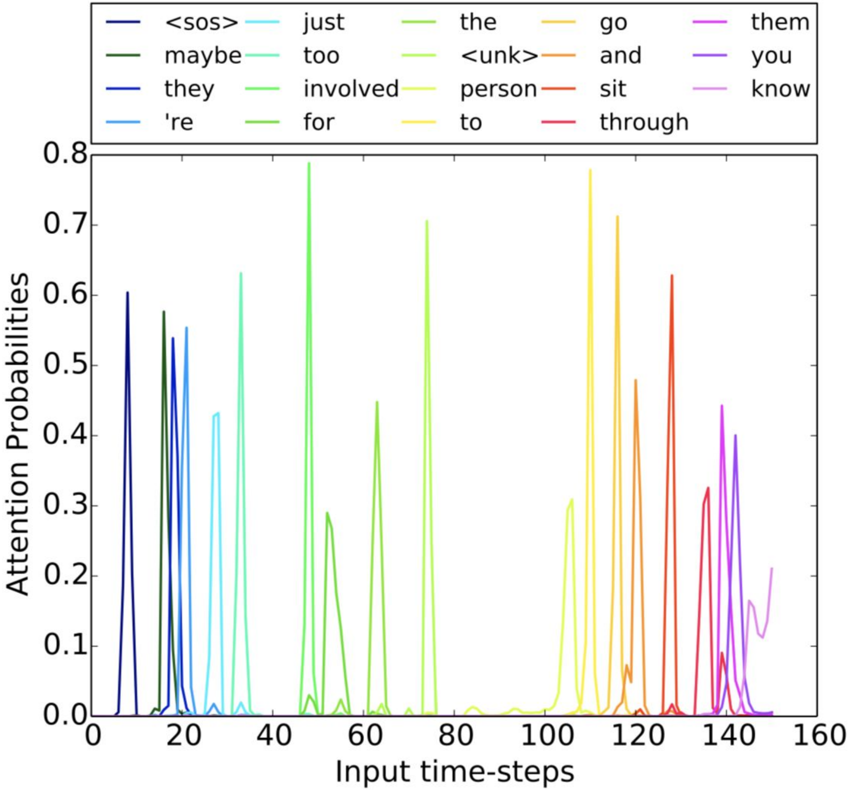


Location-aware Attention

- Attention is a rich source of interpretability and understanding in sequence-to-sequence models
- Specially, input speech and output text are monotonic signals unlike Machine Translation or summarization
- Monotonicity: time-synchronous alignments only
- Can enforcing monotonicity help improve ASR performance? Yes.
  [Chan et al., "Listen, attend and spell", 2015]
- New attention mechanism for sequence-to-sequence based ASR

Analyzing Attention

- ➢ Each color corresponds to a word in the output
- ➢ Highly *localized* attention
- ➢ Peaky distribution
- ➢ Last word attention is non-peaky
- ➢ Time steps 80-100 are silence in speech

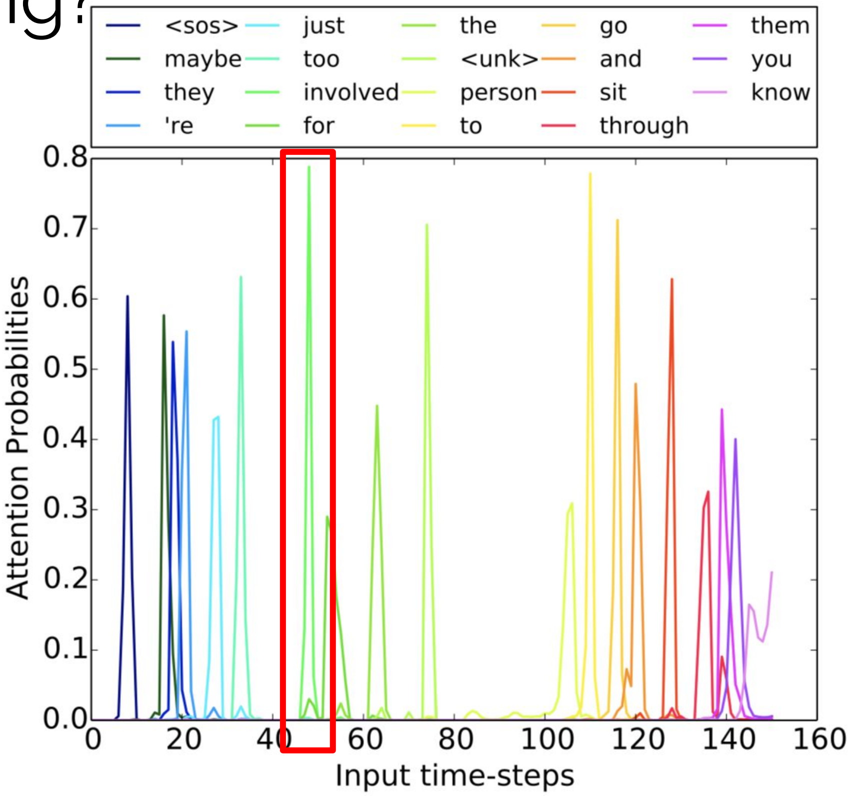


Palaskar and Metze, "Acoustic-to-Word Recognition with Sequence-to-Sequence Models", 2018

What is the model learning?

- Q1. What does it mean that attention is peaky/localized for a word?

- Model focuses on a single input speech frame for every word
- Model localizes word boundaries without supervision

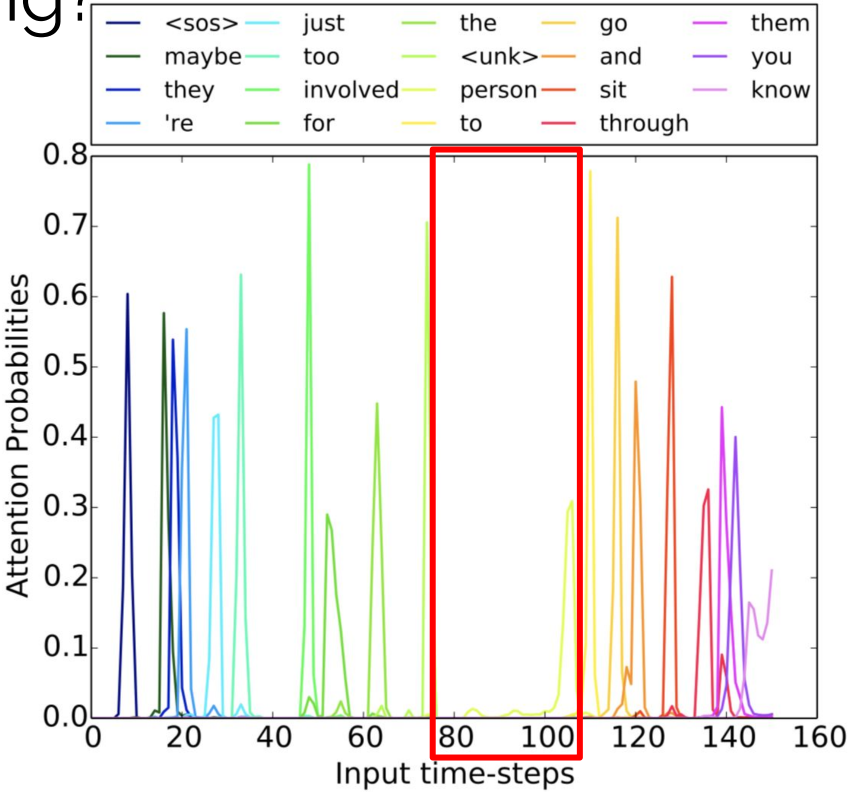


Palaskar and Metze, "Acoustic-to-Word Recognition with Sequence-to-Sequence Models", 2018

What is the model learning?

- Q2. What does it mean that attention is “absent” between timesteps 80-100?

- Model learns to detect speech and non-speech segments without supervision

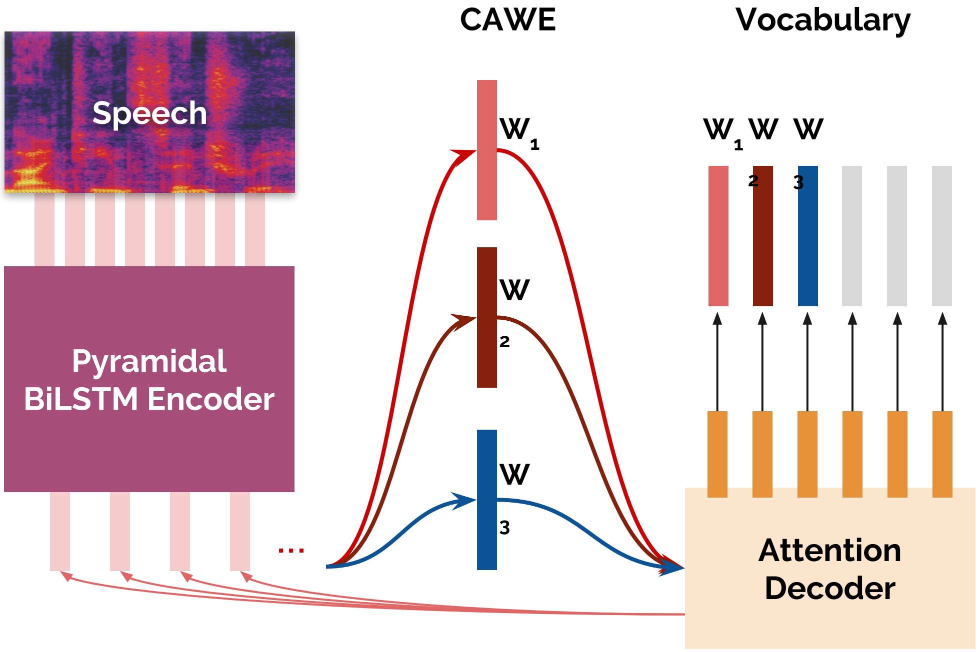


Palaskar and Metze, “Acoustic-to-Word Recognition with Sequence-to-Sequence Models”, 2018

What is the model learning?

- Q3. What does every peak corresponding to a word represent?

- It represents a single fixed-size representation of input speech, or the ***acoustic word embedding***



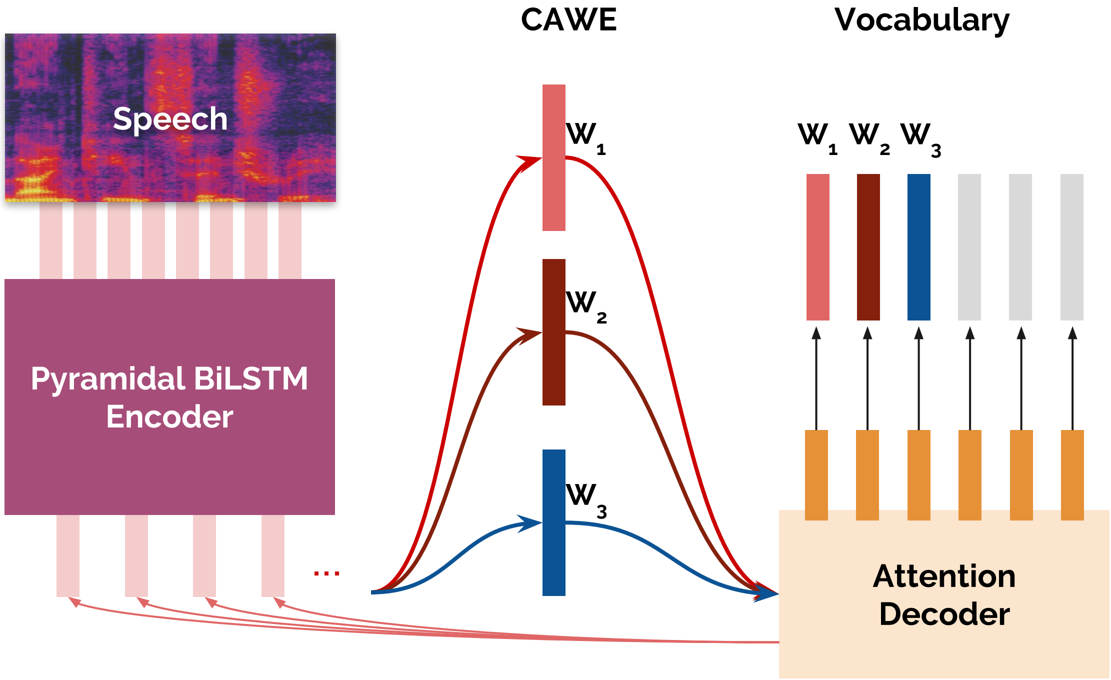
Palaskar and Metze, “Acoustic-to-Word Recognition with Sequence-to-Sequence Models”, 2018

What *all* is the model learning?

1. The model focuses on a single input speech frame for every word
2. It localizes word boundaries in continuous speech without supervision
3. It learns to detect speech and non-speech segments in continuous speech without supervision
4. It represents every output word as a single fixed-size representation of input speech, or the ***acoustic word embedding***

Palaskar and Metze, “Acoustic-to-Word Recognition with Sequence-to-Sequence Models”, 2018

Learning Contextual Acoustic Word Embeddings



- ➢ Learning Acoustic Word Embeddings using Attention
- ➢ Attention distribution helps learn contextual embeddings by applying a soft context of previous and following words in speech

Palaskar*, Raunak* and Metze, "Learned in Speech Recognition: Contextual Acoustic Word Embeddings", 2019

Using Attention to learn CAWE

$$w_i = \frac{\sum_{k \in K} encoder(a_k)}{n(K)} \quad (1)$$

(1) U-AVG: Averaged without attention weights

$$w_i = \frac{\sum_{k \in K} attention(a_k) \cdot encoder(a_k)}{n(K)} \quad (2)$$

(2) CAWE-W: Averaged with attention weights

$$w_i = encoder(a_k) \;\; where \;\; k = \arg\max_{k \in K} attention(a_k) \quad (3)$$

(3) CAWE-M: Arg max of attention weights

➢ Choose based on application

Palaskar*, Raunak* and Metze, "Learned in Speech Recognition: Contextual Acoustic Word Embeddings", 2019

Talk Outline

I. Learning Acoustic Word Embeddings
   A. Model: Acoustic-to-Word Speech Recognition
   B. Understanding A2W models
   C. Evaluation

II. Applications of Acoustic Word Embeddings
   A. Spoken Language Understanding
   B. Unsupervised speech recognition and spoken language translation

Evaluating Acoustic Word Embeddings

- ➢ Standard Sentence Embedding Evaluation Benchmarks
- ➢ There are 17 standard sentence evaluation benchmarks in NLP
- ➢ Most new methods to evaluate sentence embeddings are scored on these methods for fair evaluation
- ➢ We compare CAWE with text-based **word2vec** embeddings learned on the transcripts
- ➢ A2W models trained on **Switchboard (conversational)** and **How2 (planned but free speech, outdoors, distance microphone)**

Palaskar*, Raunak* and Metze, "Learned in Speech Recognition: Contextual Acoustic Word Embeddings", 2019

SentEval

- Standard Sentence Embedding Evaluation Benchmarks
- Fixed datasets on Sentence Textual Similarity, classification (movie reviews, product reviews etc), entailment, sentiment analysis, question type etc.
- **Human annotated** similarity scores present for this dataset
- Proposed **word** embeddings are plugged for all words in a **sentence (1)**
- Similarly, baseline **word** embeddings are plugged in for all words in a **sentence (2)**
- Correlation or Classification scores are computed with these two **sentence** embeddings

https://github.com/facebookresearch/SentEval

Comparing CAWE methods

	Switchboard			How2		
Dataset	U-AVG	CAWE-W	CAWE-M	U-AVG	CAWE-W	CAWE-M
STS 2012	0.3230	0.3281	**0.3561**	0.3255	0.3271	**0.3648**
STS 2013	0.1252	0.1344	**0.1969**	0.2070	0.2071	**0.2716**
STS 2014	0.3358	0.3389	**0.3888**	0.3375	0.3426	**0.3940**
STS 2015	0.3854	0.3881	**0.4275**	0.3852	0.3843	**0.4173**
STS 2016	0.2998	0.2974	**0.3833**	0.3248	**0.3271**	0.3159
STS B	0.3667	0.3510	**0.4010**	0.3343	0.3440	**0.4000**
SICK-R	0.5640	0.5800	**0.6006**	0.5800	0.6060	**0.6440**
MR	63.86	63.75	**64.69**	63.46	63.19	**63.64**
MRPC	**70.67**	69.45	69.80	68.29	67.83	**70.61**
CR	71.42	72.13	**72.93**	**74.12**	73.99	73.03
SUBJ	**82.45**	82.22	81.19	81.48	**81.88**	81.01
MPQA	**73.76**	73.28	73.75	**74.21**	74.18	73.53
SST	66.45	**66.61**	65.02	63.43	63.43	**65.13**
SST-FG	32.81	32.04	**33.53**	31.95	**32.35**	32.03
TREC	63.80	62.40	**67.60**	**66.60**	66.00	60.60
SICK-E	**74.20**	73.41	74.06	75.14	75.34	**75.97**

CAWE-M always performs better in STS tasks

CAWE-W more generalizable but noisy

U-AVG noisiest

Palaskar*, Raunak* and Metze, "Learned in Speech Recognition: Contextual Acoustic Word Embeddings", 2019

Comparing CAWE with word2vec

	Switchboard			How2		
Dataset	CAWE-M	CBOW	Concat	CAWE-M	CBOW	Concat
STS 2012	0.3561	**0.3639**	0.3470	0.3648	0.3688	**0.3790**
STS 2013	0.1969	0.1960	**0.2010**	**0.2716**	0.2524	0.2675
STS 2014	**0.3888**	0.3745	0.3795	0.3940	**0.3973**	0.3971
STS 2015	0.4275	0.4459	**0.4481**	0.4173	**0.4781**	0.4710
STS 2016	**0.3833**	0.3471	0.3651	0.3159	**0.4023**	0.3388
STS B	0.401	**0.4100**	0.3995	0.4000	**0.4720**	0.4487
SICK-R	0.6006	0.6170	**0.6228**	0.6440	0.6550	**0.6945**
MR	64.69	66.24	**66.89**	63.64	66.03	**66.89**
MRPC	**69.80**	68.99	68.00	**70.61**	69.68	68.52
CR	72.93	74.49	**75.39**	73.03	**74.89**	74.84
SUBJ	81.19	**84.62**	84.59	81.01	84.75	**85.04**
MPQA	73.75	**76.44**	75.36	73.53	75.56	**75.60**
SST	65.02	68.37	**68.97**	65.13	67.66	**68.20**
SST-FG	33.53	34.71	**35.79**	32.08	33.62	**33.67**
TREC	67.60	69.80	**71.40**	60.60	**68.40**	67.40
SICK-E	74.06	75.02	**76.19**	75.97	76.29	**78.14**

CAWE performs competitively with word2vec

Improvement in concatenation shows both embeddings contribute unique features

Gains more prominent in SWBD as it is conversational while How2 is planned

Palaskar*, Raunak* and Metze, "Learned in Speech Recognition: Contextual Acoustic Word Embeddings", 2019

Talk Outline

I. Learning Acoustic Word Embeddings
   A. Model: Acoustic-to-Word Speech Recognition
   B. Understanding A2W models
   C. Evaluation
II. Applications of Acoustic Word Embeddings
   A. Spoken Language Understanding
   B. Unsupervised speech recognition and spoken language translation

Spoken Language Understanding

- Speech-based downstream task other than transcription
- ATIS dataset of flight queries with intent, domain, and named entities
- Widely used corpus for SLU
- Classification Task: Given query identify intent, domain and named entities
- Prior work used transcription of speech rather than audio input for this task [Mesnil et al. 2013]
- Performance in this task will help validate use of CAWE

Using CAWE for Spoken Language Understanding

	F1 Score		
	CAWE-M	CAWE-W	CBOW
RNN	91.49	91.67	91.82
GRU	93.25	93.56	93.11

- ➢ Two simple models:
  RNN and GRU
- ➢ F1 score for classification on CAWE-M, CAWE-W and CBOW
- ➢ CAWE performs competitively with text embeddings highlighting its utility
- ➢ Can be used as off-the-shelf embeddings for other speech-based tasks when trained on larger data

Palaskar*, Raunak* and Metze, "Learned in Speech Recognition: Contextual Acoustic Word Embeddings", 2019

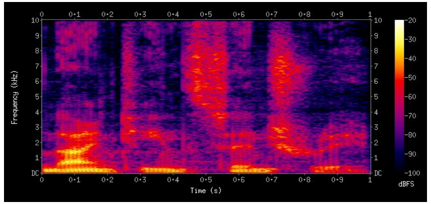
Talk Outline

I. Learning Acoustic Word Embeddings
   A. Model: Acoustic-to-Word Speech Recognition
   B. Understanding A2W models
   C. Evaluation
II. Applications of Acoustic Word Embeddings
   A. Spoken Language Understanding
   B. Unsupervised speech recognition and spoken language translation

Multimodal applications: example dataset

Video

Speech



Video

Transcript (en)

Once I have my jack stand there on the rear axle, go ahead and release the hydraulic pressure...

Translation (pt)

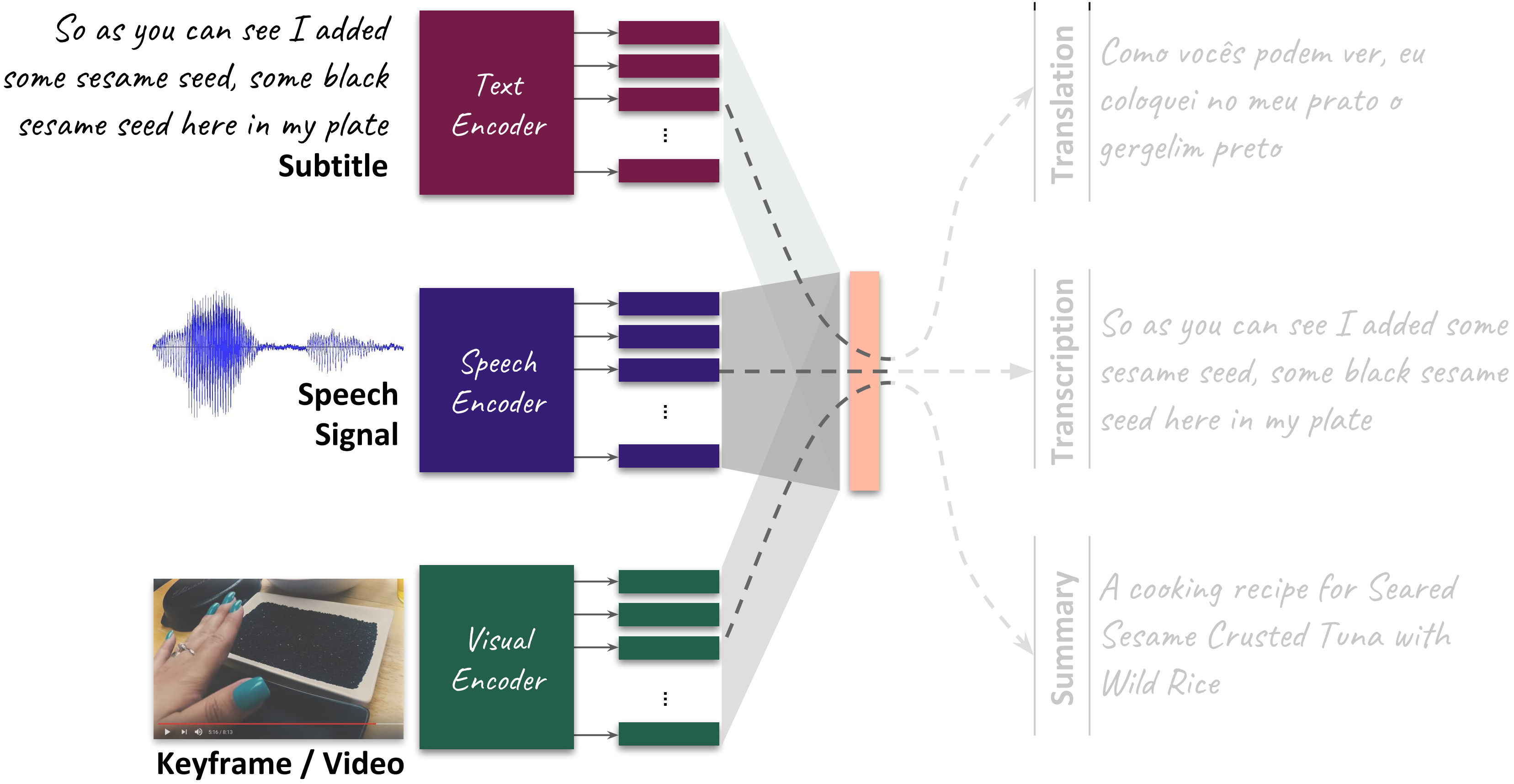
Quando eu tiver meu macaco parado no eixo traseiro, vá em frente e libere a pressão hidráulica...

Summary (en)

Changing flat tires doesn't have to be done with car jacks. Learn how to use an automotive hydraulic lift...

Sanabria et al., “How2: A Large Scale dataset for Multimodal Language Understanding”, 2018

The big picture



Learning Multimodal Embeddings

I. Each is different but all views share similar information

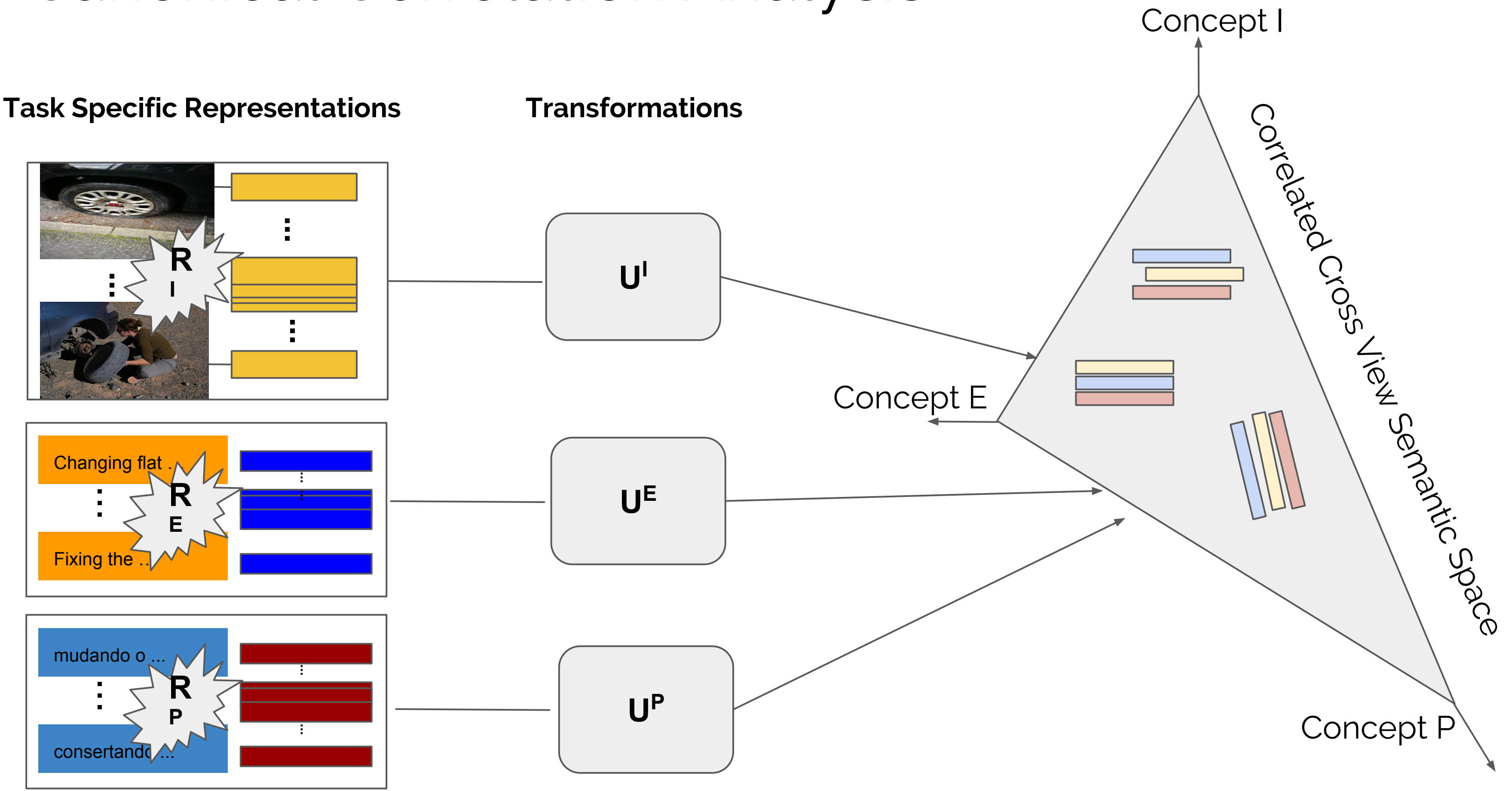
II. Visual, Auditory and Language views are aligned

III. Views in the same modality v/s Views in multiple modalities

IV. Unit level representations v/s Sequence Level Representations

Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019

Canonical Correlation Analysis



CCA in a Nutshell

Pairs of points: $(X, Y) \sim \mathcal{D}_{X,Y}$

View 1 View 2

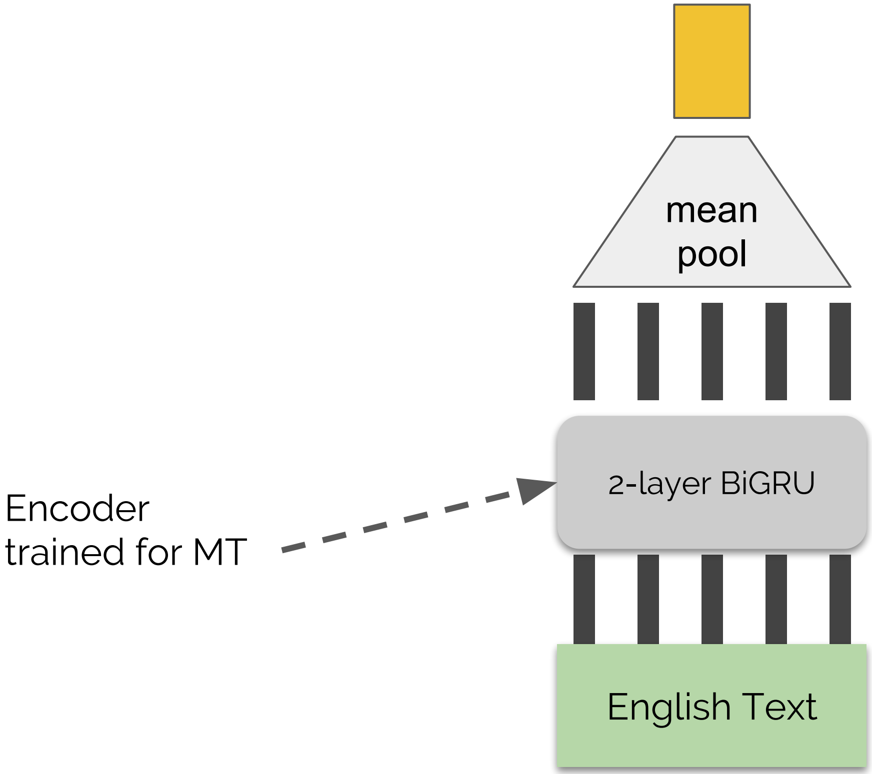
"A man in an orange hat staring at something."

Find transformations $\mathbf{u} \in \mathbb{R}^{d_x}, \mathbf{v} \in \mathbb{R}^{d_y}$

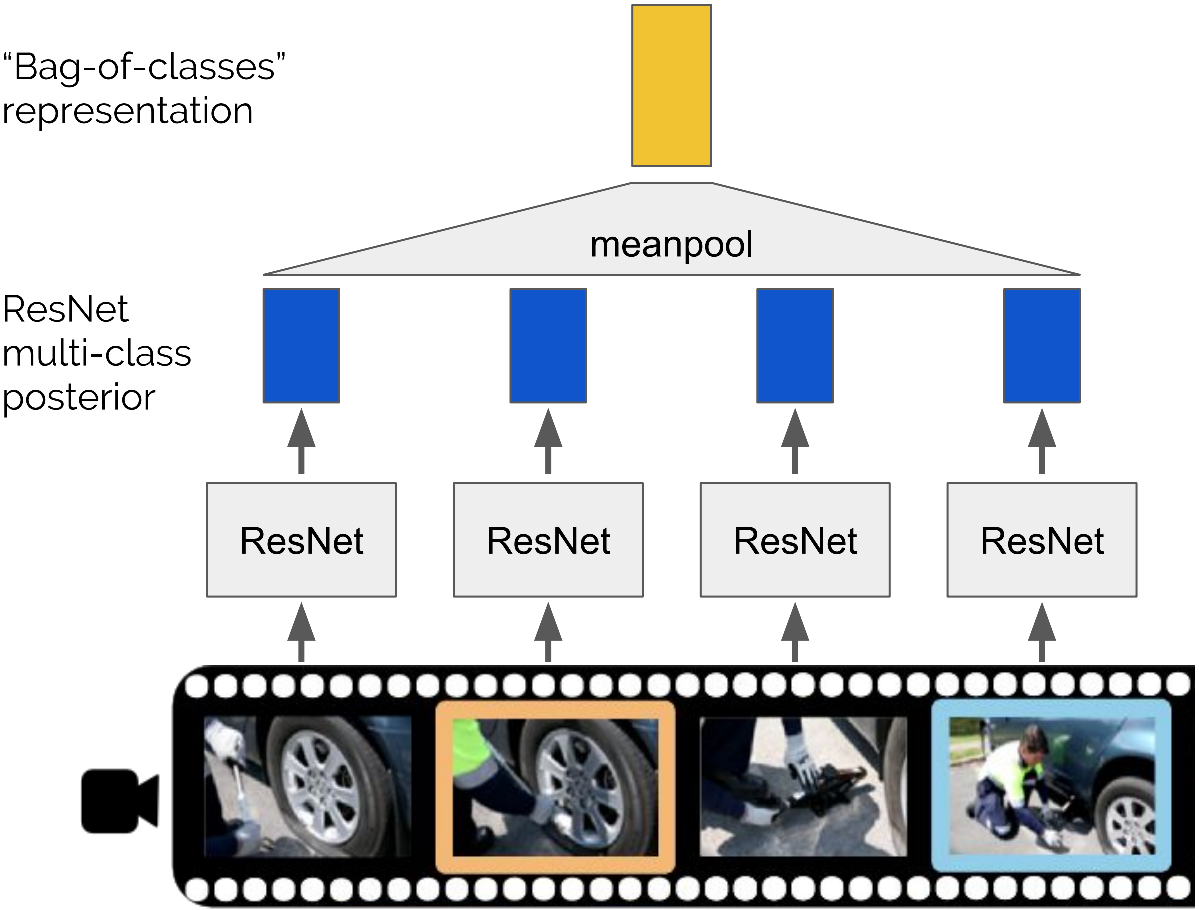
to maximize $\text{correlation}(\mathbf{u}^T f_\theta(X), \mathbf{v}^T g_\phi(Y))$

Hotelling, 1936; Wang et al., 2016

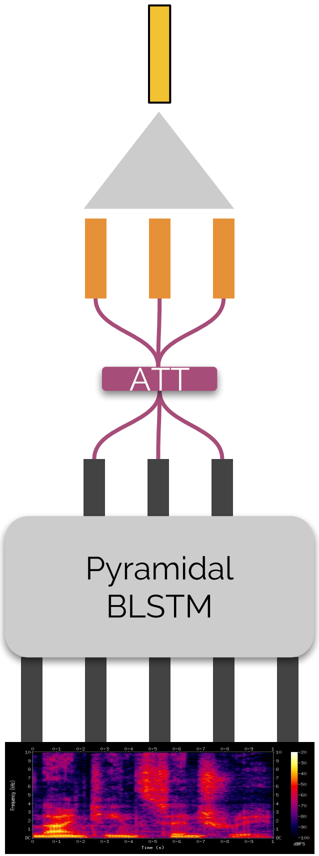
Text Representations - Sentences



Video Representations



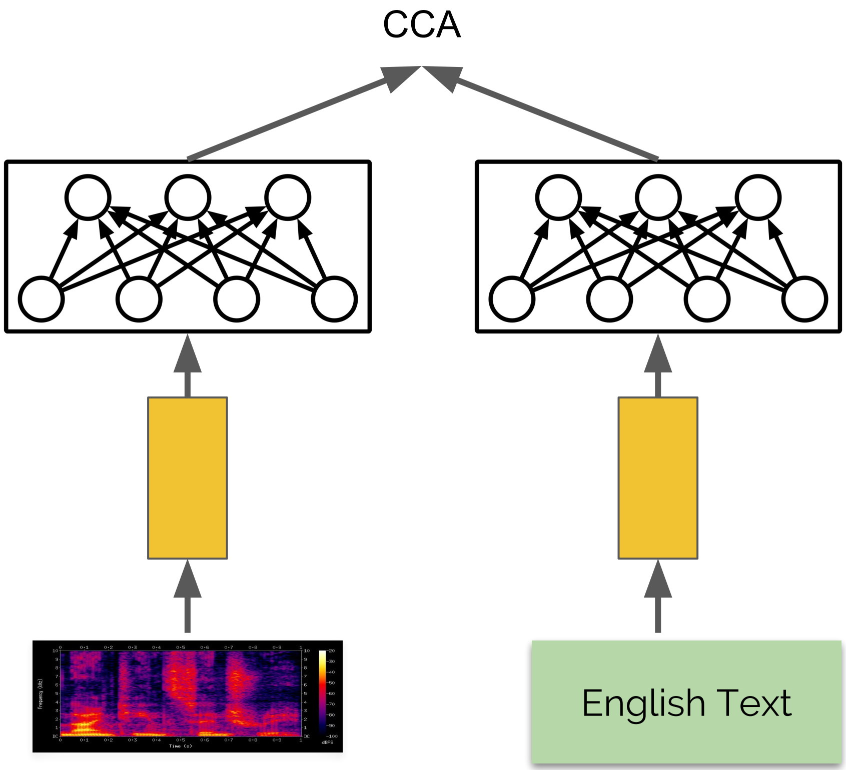
Speech Representations - Sentences [CAWE]



- bag-of-audio-words
- Meanpool
- CAWE

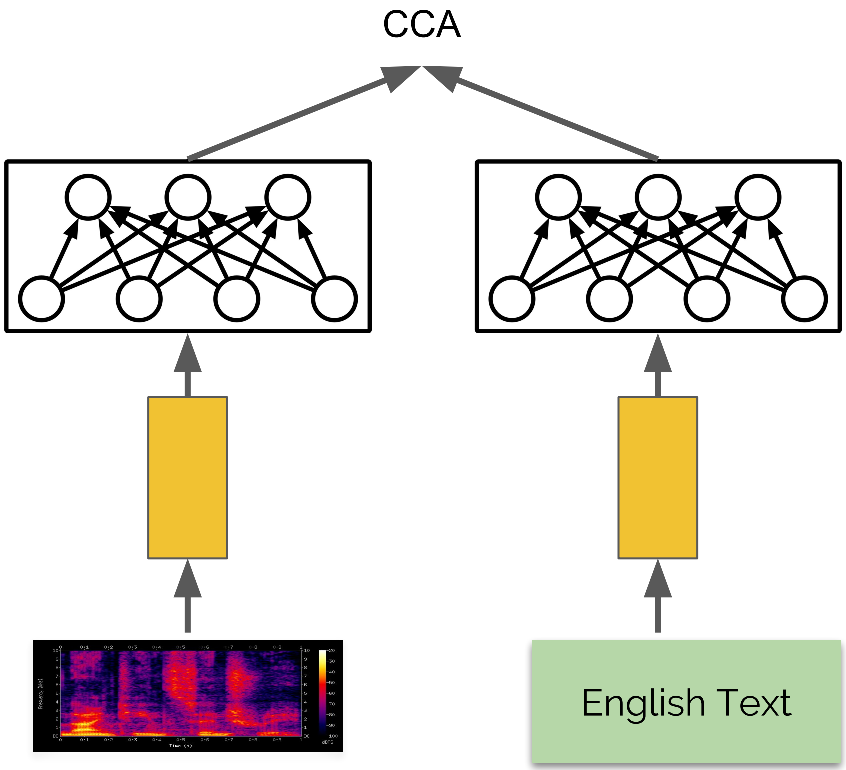
Palaskar and Metze, “Acoustic-to-Word Recognition with Sequence-to-Sequence Models”, 2018

Speech and Text Representations



Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., “Learning from Multiview Correlations in Open-Domain Videos”, 2019

Retrieve Text Given Speech

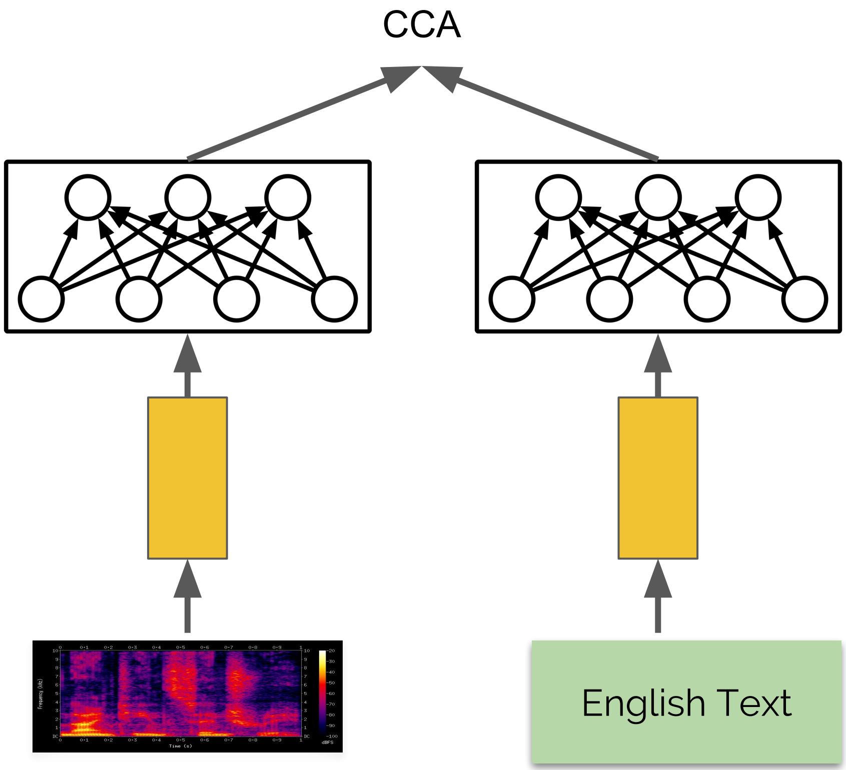


Recall@10
over Test set

Linear CCA	Deep CCA
96.9%	**90.1%**

Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019

Retrieve Speech Given Text

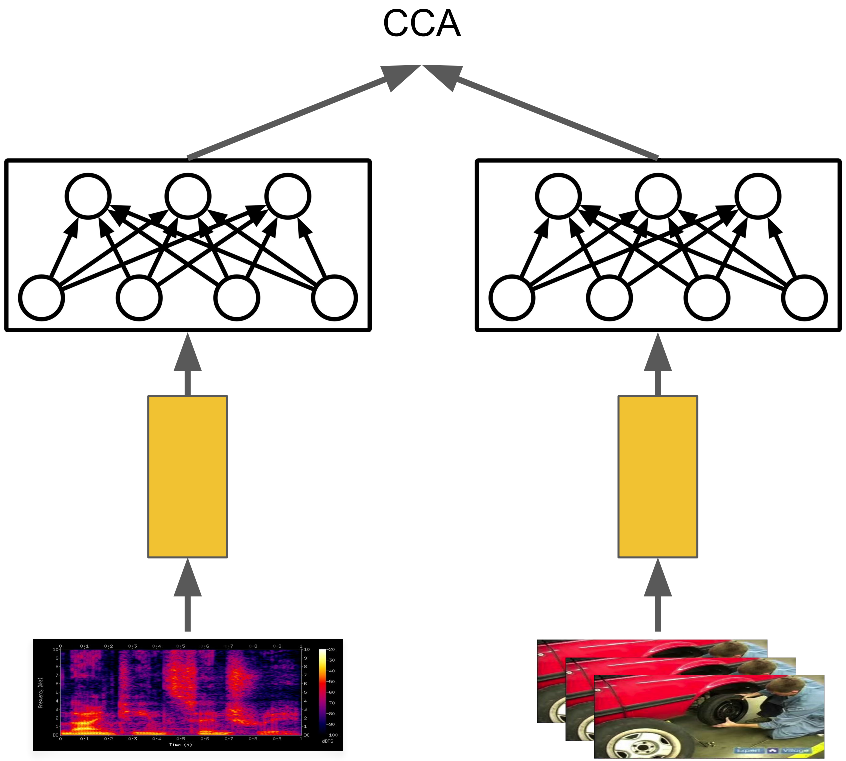


Recall@10
over Test set

Linear CCA	Deep CCA
96.1%	**89.7%**

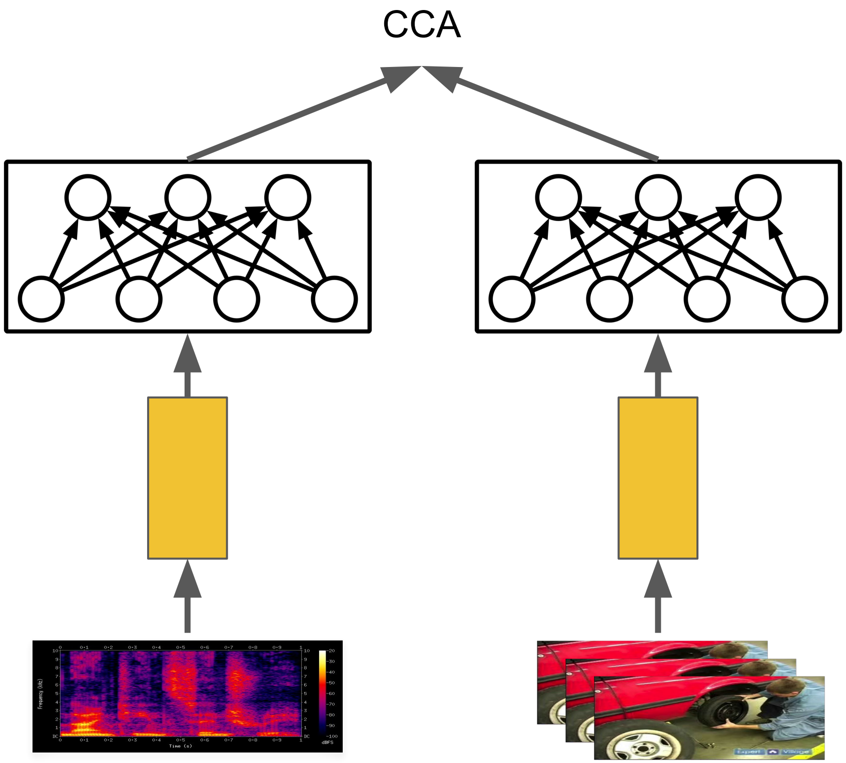
Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019

Speech and Video Representations



Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019

Retrieve Video Given Speech

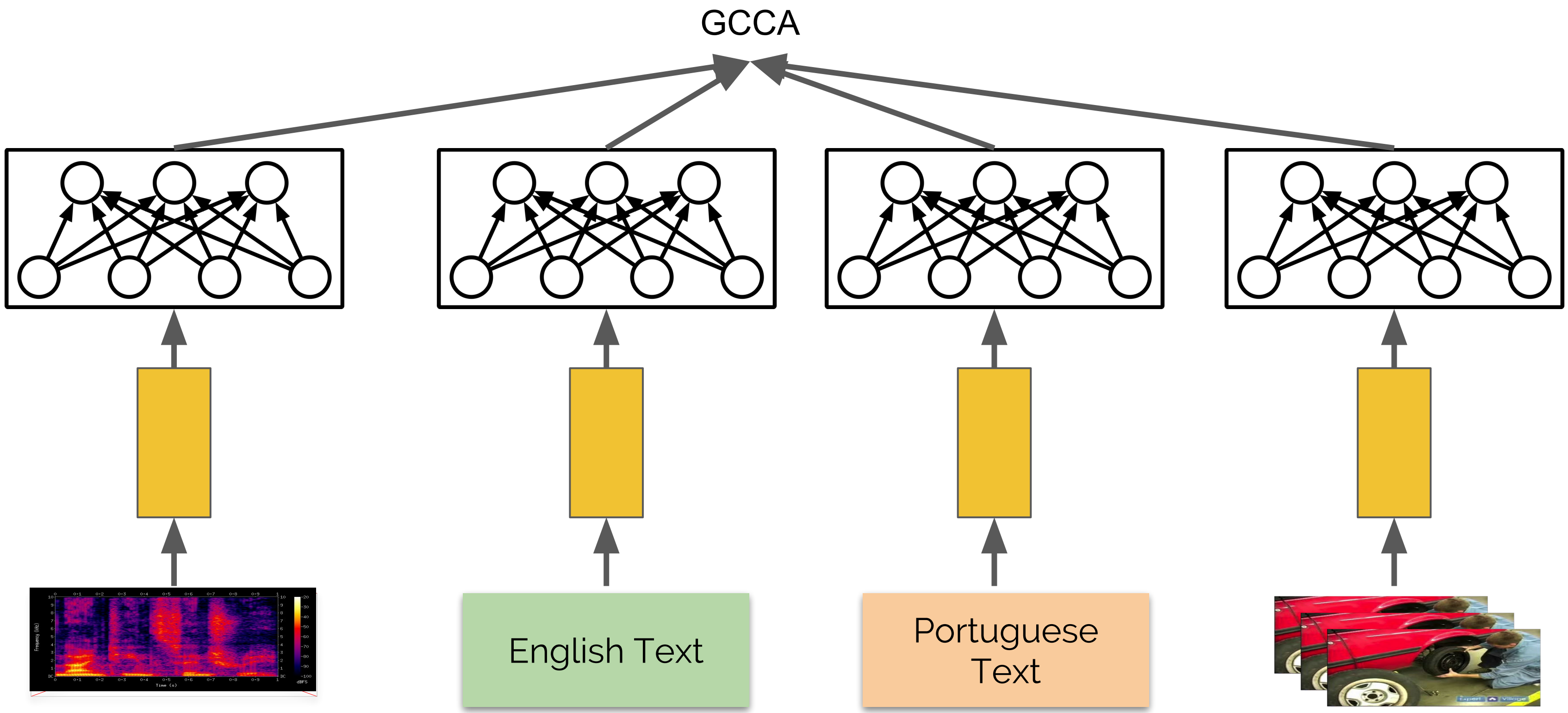


Recall@10
over Test set

Linear CCA	Deep CCA
0.5%	1.8%

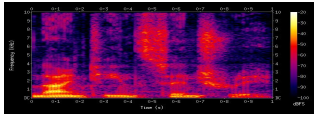
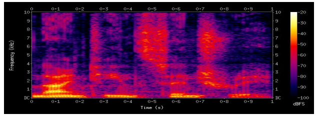
Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., “Learning from Multiview Correlations in Open-Domain Videos”, 2019

Speech, Text and Video Representations



Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019

Retrieval: Speech, Text (En & Pt) and Video on Test Set

Recall@10		English Text	Portuguese Text	
	-	85.4	70.7	1.0
English Text	85.4	-	98.4	0.9
Portuguese Text	71.0	98.3	-	1.1
	1.1	1.1	0.9	-

Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019

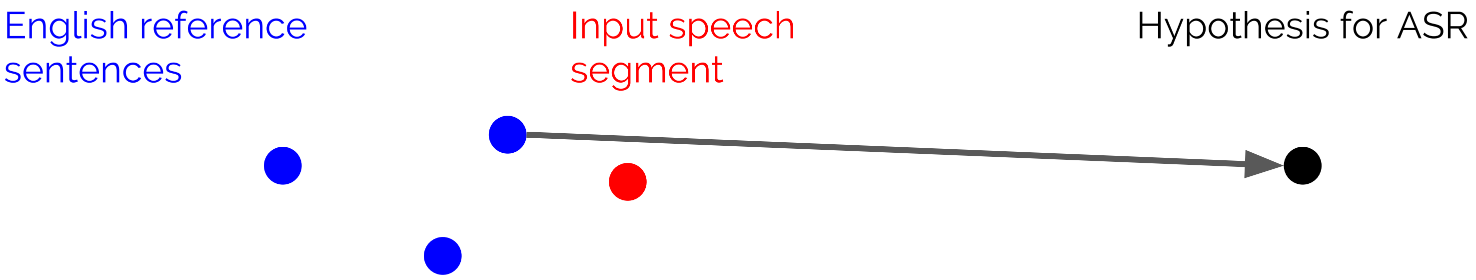
Retrieve Text Given Speech - Comparison

Model	Recall@10
Speech & En Text	90.1%
Speech, En Text, Pt Text & Video	85.4%

Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., “Learning from Multiview Correlations in Open-Domain Videos”, 2019

Retrieval for ASR

Given a Speech segment from the test set, retrieve the closest English sentence in a reference set.



Reference set	WER ↓
S2S Model	24.2 %
Train	134 %
Train + Test	27.4 %

Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019

Retrieval for SLT

Given a Speech segment from the test set, retrieve the closest Portuguese sentence in a reference set.

Portuguese reference sentences

Input speech segment

Hypothesis for Spoken Language Translation

Reference set	BLEU ↑
S2S Model	27.9
Train	0.2
Train + Test	19.8

Holzenberger*, Palaskar*, Madhyastha, Metze and Arora., "Learning from Multiview Correlations in Open-Domain Videos", 2019

To conclude

Main Takeaways

1. Possible to learn pre-trained acoustic word embeddings similar to text (bert, elmo) and vision (alexnet, vggnet)
2. These embeddings perform well with text based embeddings and capture complimentary information than text embeddings
3. Can perform non-transcription tasks with speech inputs: spoken language understanding
4. Can learn shared global multimodal embedding spaces to perform unsupervised ASR, SLT etc

Main Takeaways

1. Possible to learn pre-trained acoustic word embeddings similar to text (bert, elmo) and vision (alexnet, vggnet)
2. AWE performs competitively with word2vec and capture complimentary information than text embeddings
3. Can perform non-transcription tasks with speech inputs: spoken language understanding
4. Can learn shared global multimodal embedding spaces to perform unsupervised ASR, SLT etc

Main Takeaways

1. Possible to learn pre-trained acoustic word embeddings similar to text (bert, elmo) and vision (alexnet, vggnet)
2. These embeddings perform well with text based embeddings and capture complimentary information than text embeddings
3. Can perform non-transcription tasks with speech inputs: spoken language understanding
4. Can learn shared global multimodal embedding spaces to perform unsupervised ASR, SLT etc

Main Takeaways

1. Possible to learn pre-trained acoustic word embeddings similar to text (bert, elmo) and vision (alexnet, vggnet)
2. These embeddings perform well with text based embeddings and capture complimentary information than text embeddings
3. Can perform non-transcription tasks with speech inputs: spoken language understanding
4. Can learn shared global multimodal embedding spaces to perform unsupervised ASR, SLT etc

Thank you!

Questions?

spalaska@cs.cmu.edu