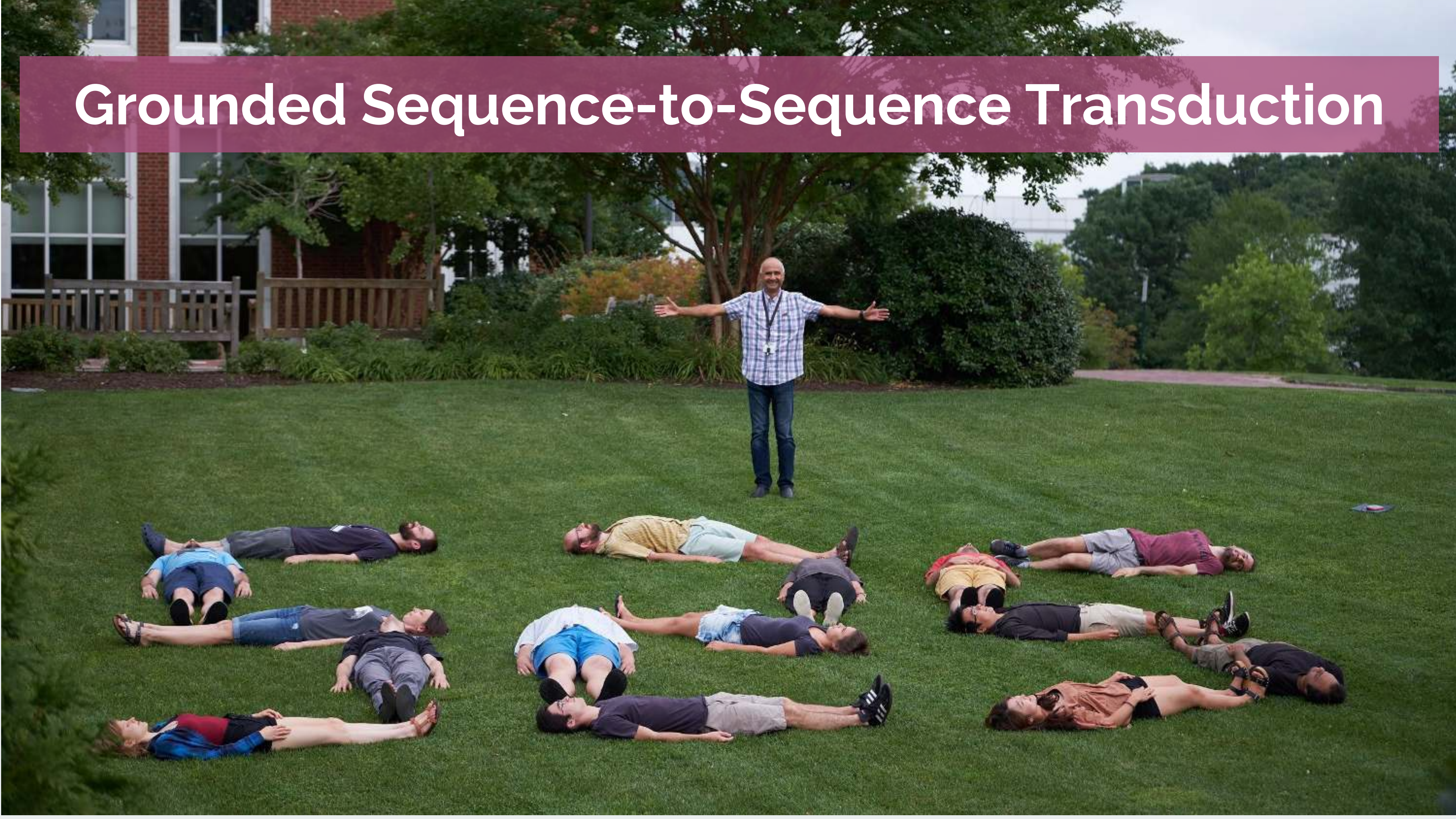
Grounded Sequence-to-Sequence Transduction



Team

Undergraduate Students



Alissa Ostapenko - WPI

Karl Mulligan - Rutgers

Sun Jae (Jasmine) Lee - UPenn

Graduate Students






Jindrich Libovicky - Charles

Ramon Sanabria - CMU

Shruti Palaskar - CMU

Nils Holzenberger - JHU

Amanda Duarte - UPC

Ozan Caglayan - Le Mans

Senior Researchers






Lucia Specia - Sheffield

Florian Metze - CMU

Loïc Barrault - Le Mans

Des Elliott - Edinburgh / Copenhagen

Josiah Wang - Sheffield

Pranava Madhyastha - Sheffield

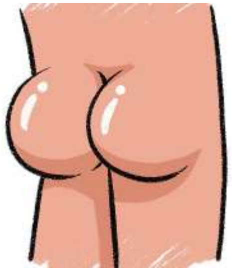
Remotely


Spandana Gella - Edinburgh

Chiraag Lala - Sheffield

Motivation

Understanding language is hard



Motivation

Humans interact with the world in **multimodal** ways. **Language** understanding & generation is not an exception

- **Multimodality** in computational models
  - Richer context modelling
  - Grounding of language

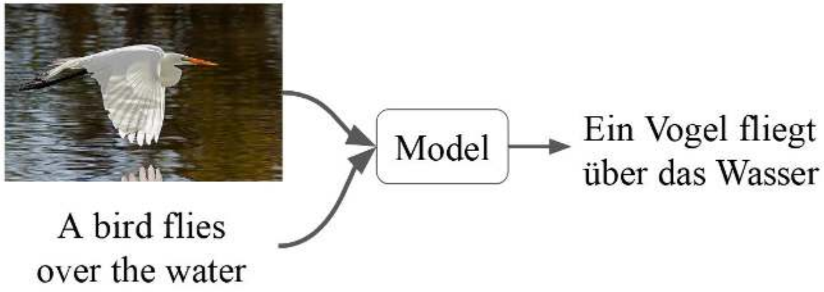
- True for a wide range of NL **tasks**

- **Sequence-to-sequence** NN is a convenient approach

Previous to JSALT...

Multimodality useful for MT

#	Raw	z	System
1	77.8	0.665	LIUMCVC_MNMT_C
2	74.1	0.552	UvA-TiCC_IMAGINATION_U
3	70.3	0.437	NICT_NMTrerank_C
	68.1	0.325	CUNI_NeuralMonkeyTextualMT_U
	68.1	0.311	DCU-ADAPT_MultiMT_C
	65.1	0.196	LIUMCVC_NMT_C
	60.6	0.136	CUNI_NeuralMonkeyMultimodalMT_U
	59.7	0.08	UvA-TiCC_IMAGINATION_C
	55.9	-0.049	CUNI_NeuralMonkeyMultimodalMT_C
	54.4	-0.091	OREGONSTATE_2NeuralTranslation_C
	54.2	-0.108	CUNI_NeuralMonkeyTextualMT_C
	53.3	-0.144	OREGONSTATE_1NeuralTranslation_C
	49.4	-0.266	SHEF_ShefClassProj_C
	46.6	-0.37	SHEF_ShefClassInitDec_C
15	39.0	-0.615	Baseline (text-only NMT)
	36.6	-0.674	AFRL-OHIOSTATE_MULTIMODAL_U



Multimodal
Text

(Elliott et al., 2017)

Previous to JSALT...

Multimodality useful for ASR

- 90h of **how-to** video data
- Object and place features
- Word Error Rates:
    - 23.4% with DNN/HMM + WFST (baseline)
    - 22.3% with AM adaptation
    - 22.6% with LM adaptation (RNNLM)
    - **21.5% with AM+LM** n-best rescoring
- Improvements make sense intuitively
    - Higher for acoustically hard videos

(Gupta et al., 2017; Palaskar et al., 2018)

Previous to JSALT...

Promising results, but...

- 'Easy', small data (for MT)
- Limited types of **modalities**: static visual information
- Limited number of **tasks**
- **Representations** not shared across tasks
- Not clear **where improvements** are coming from

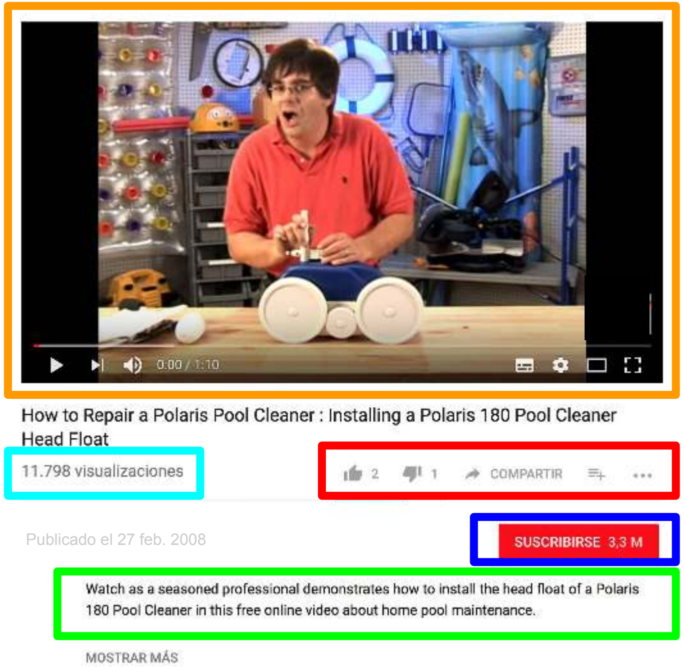
JSALT goals

More **data**, more **modalities**, more **tasks**

Better **models**, better **representations**, better **understanding**

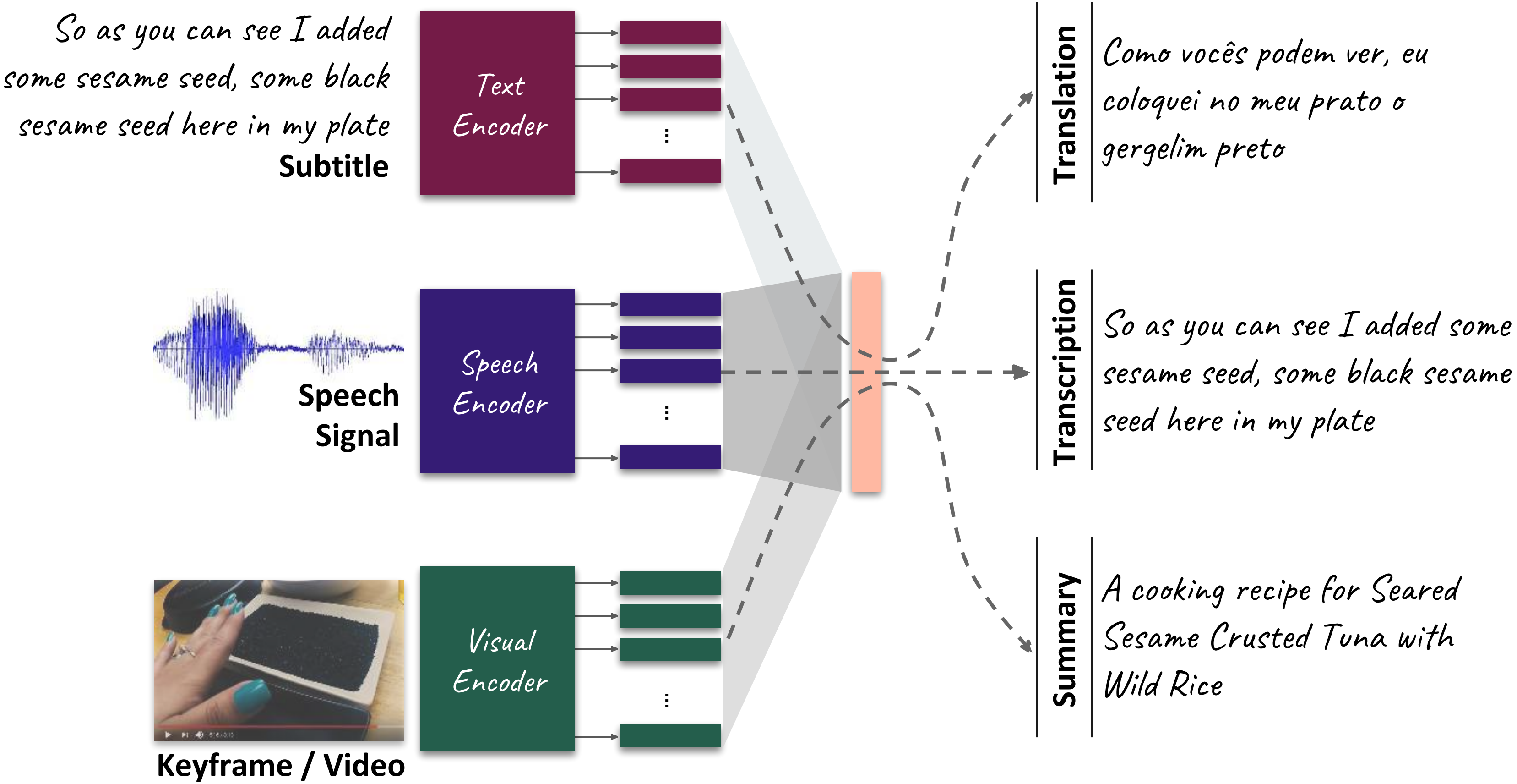
Dataset

- 2000h of **how-to** videos (Yu et al., 2014)
  - 300h for MT, 480h for ASR (as of today)
  - Shared splits, held-out data
- Ground truth captions
- Metadata
  - Number of likes / dislikes
  - Visualizations
  - Uploader, Date
  - Tags
- Video descriptions (“summaries”)
  - 80K descriptions for 2000h
- Very different topics
  - Cooking, fixing things, playing instruments, etc.
- 300,000 segments translated into Portuguese



Dataset - example

The big picture



Groups

- Automatic Speech Recognition and Spoken Language Translation
- Text Summarization
- Region-specific Machine Translation
- Multiview Learning
- Multitask Learning

Highlights

- ASR & SLT:
  - Multi-task learning approaches that improve both tasks
  - One-to-many model generalizes better than many-to-one model

- Summarization:
  - Models that successfully generate teasers for videos
  - Multimodal models using action features that outperform text models

- Region-specific MMT:
  - Supervised attention that successfully grounds words to image regions
  - Models for explicit grounding and its integration into MT

Highlights

- Multi-view learning:
  - Implementation and exploration of DGCCA models
  - High cross-view retrieval scores; exploration of integration in MT & ASR
- Multi-task learning:
  - Single framework for multi-task learning over multiple inputs & outputs
  - New models: Shared Recurrent Space and Mutual Projection Networks
- How-to dataset & evaluation methods:
  - Same dataset used for a number of diverse & challenging tasks
  - Established best practices and common framework for these tasks

Highlights

~13K lines of code added

nmtpytorch

- New data loaders for audio, video, arbitrary feature vectors
- Layers:
  - Auxiliary feature integration into RNN encoder & decoder
  - Hierarchical attention, coattention, supervised attention
  - Video encoder & video decoder
  - Sequence convolutions
  - Latent Recurrent Space Layer, ...
- New models: ASR, SLT, MMT, MPN, ...
- Multi-tasking
  - Scheduling
  - One-to-many, many-to-one, many-to-many

Schedule

- 1:30 - 1:45: Introduction
- 1:45 - 2:10: ASR/SLT
- 2:10 - 2:35: Teaser generation
- 2:35 - 3:00: Region-specific MT

- 3:00 - 3:15: Break

- 3:15 - 3:40: Multiview learning
- 3:40 - 4:05: Multitask learning
- 4:05 - 4:10: Take home messages

? After each section

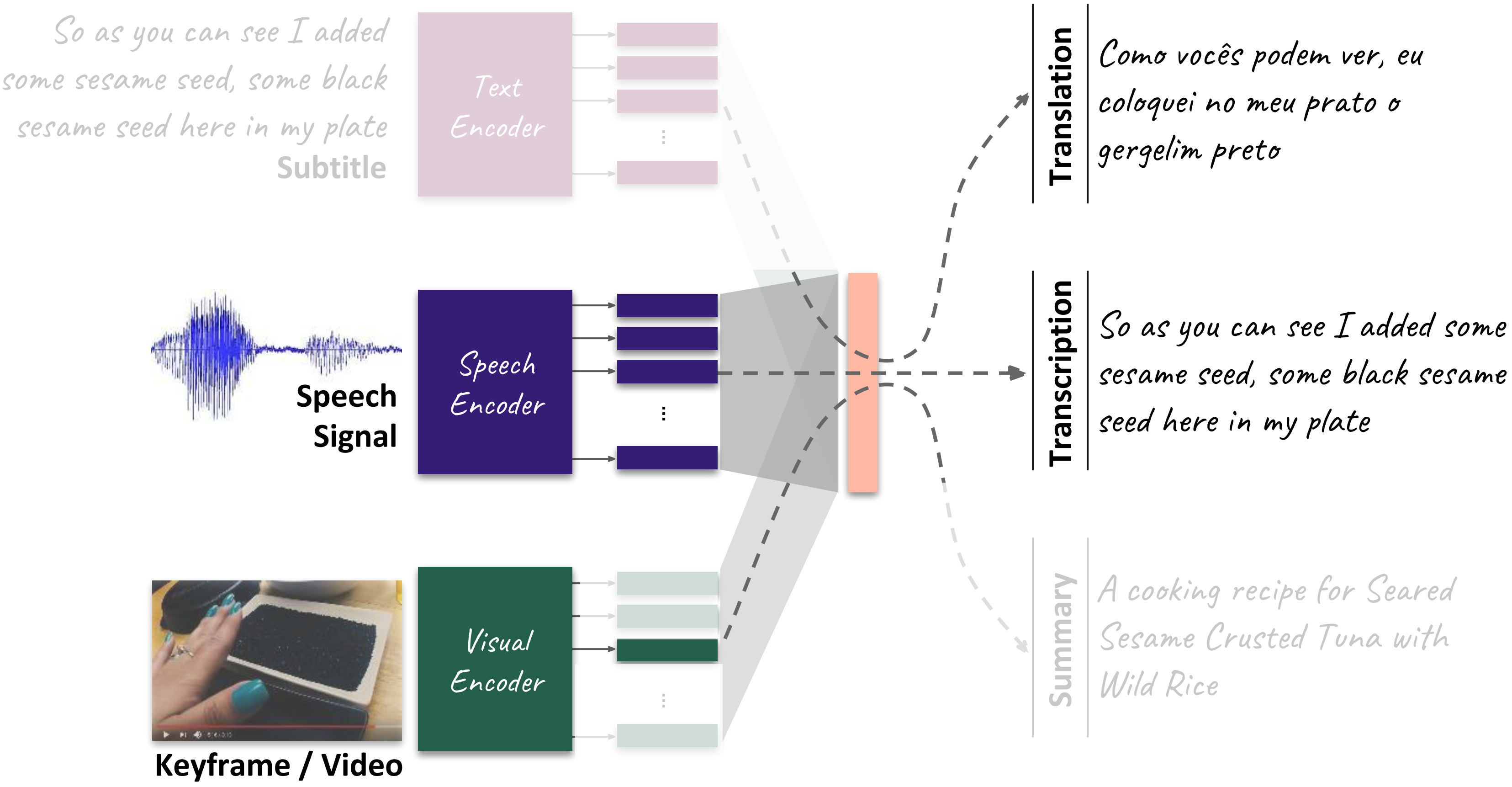
Automatic Speech Recognition
Spoken Language Translation





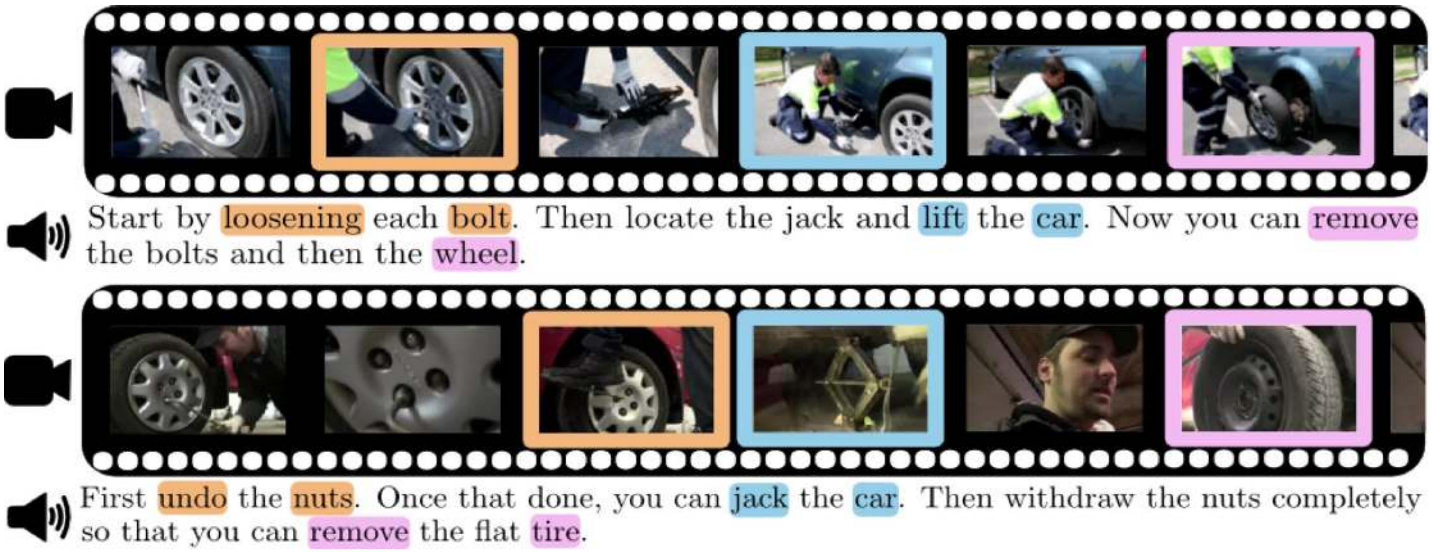
Florian, Jindrich, Ozan, Ramon, Shruti

The big picture



Motivation

- In how-to videos, speech and visuals are often highly correlated
  - Earlier work suggests that gains can be obtained by fusing
- S2S models provide an elegant framework (no separate AM / LM)



From [Alayrac et al., 2016]

Related & Previous Results

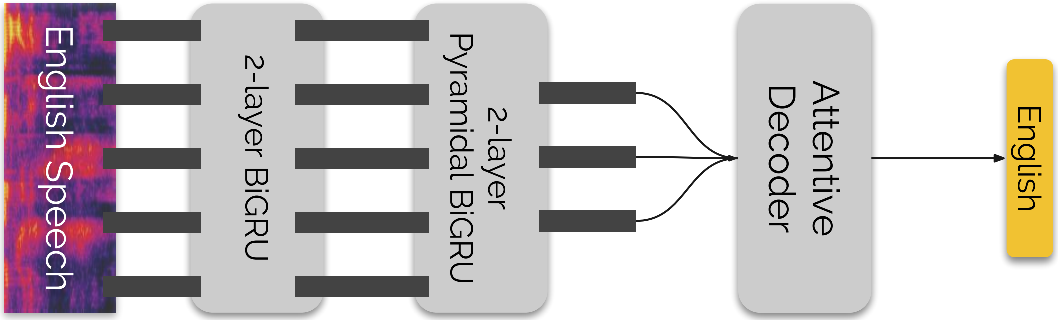
- Have seen improvements in the past (on `devtest`)
  - 23.4% → 21.5% WER - HMM / GMM using LM rescoring on **90h**
  - 15.2% → 14.1% TER - CTC on **480h**
  - 89 → 74 PPL - NNLM on **480h**

- Introduced new **300h** training set
  - Compatible with S2S machine translation experiments
  - 5K SentencePiece token vocab for EN and PT

- Baselines on 300h (on `cv05`)
  - 19.6% WER - ESPNet Character S2S (TER=11.8%)
  - 23.6% WER - ESPNet Word S2S (preliminary)
  - 23.0% WER - `nmtpy` Word baseline (Small -- 4.3M params)
  - 19.6% WER - `nmtpy` Word Baseline (Medium -- 13.7M params, ~ESPNet)

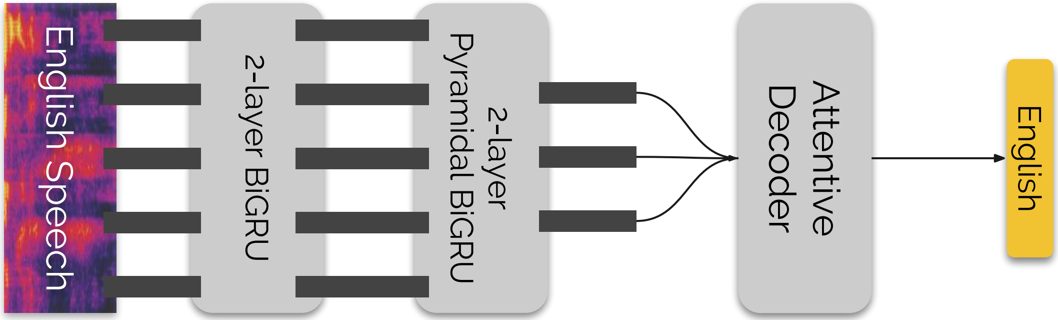
Automatic Speech Recognition

S2S ASR Baseline



- 4-Layer BiGRU Encoder (200D)
- 200D Embeddings
- 2-Layer Conditional GRU Decoder
- MLP Attention
- Dropout (p=0.4)

S2S ASR Baseline



- 4-Layer BiGRU Encoder (200D)
- 200D Embeddings
- 2-Layer Conditional GRU Decoder
- MLP Attention
- Dropout (p=0.4)

	# of Params	Tokens	cv05 WER	dev5 WER
ASR	4.3M	SentPiece-5K	23.0	24.0
ASR w/ 6-layer BiLSTMp encoder	13.7M	SentPiece-5K	19.6	21.1
ESPNet 6-layer BiLSTMp encoder	-	Char	19.6	19.8

We use a small ASR for faster experimental turnaround time.

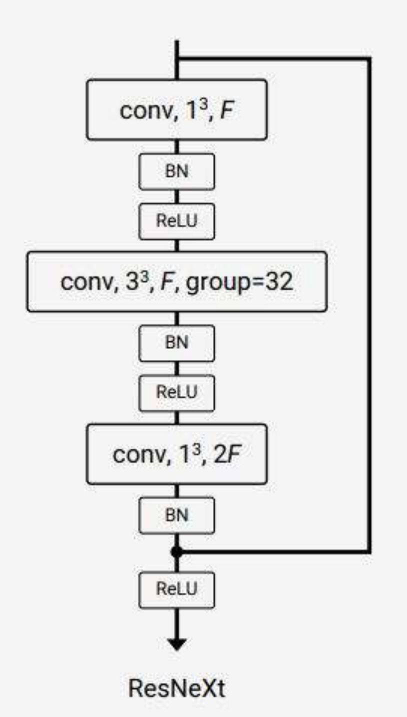
Multimodal ASR

Multimodal ASR: Motivations

- “Speech and visual are often highly correlated”
- Can we improve the decoder LM by providing visual context?
  - Action-level **global** visual features

- Can we benefit from multimodal attention?
  - Let the model learn when to pay attention to multiple modalities
  - Action-level **temporal** visual features

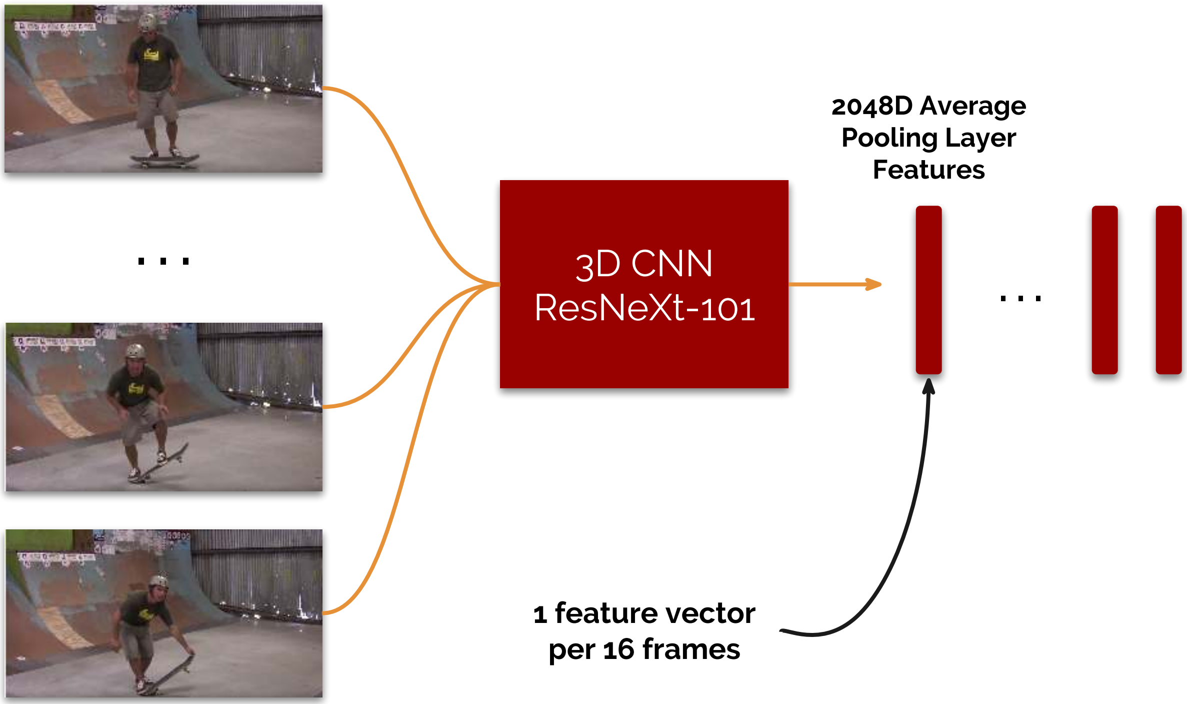
Action-level Video Features [Hara et al., 2018]



Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?

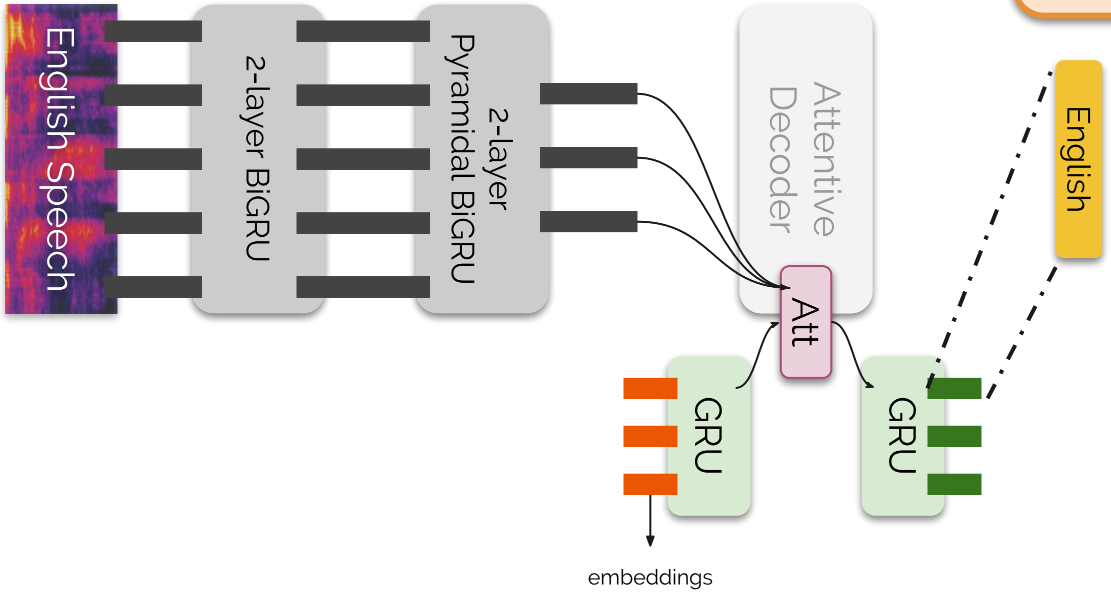
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan
{kensho.hara, hirokatsu.kataoka, yu.satou}@aist.go.jp

Action-level Video Features [Hara et al., 2018]



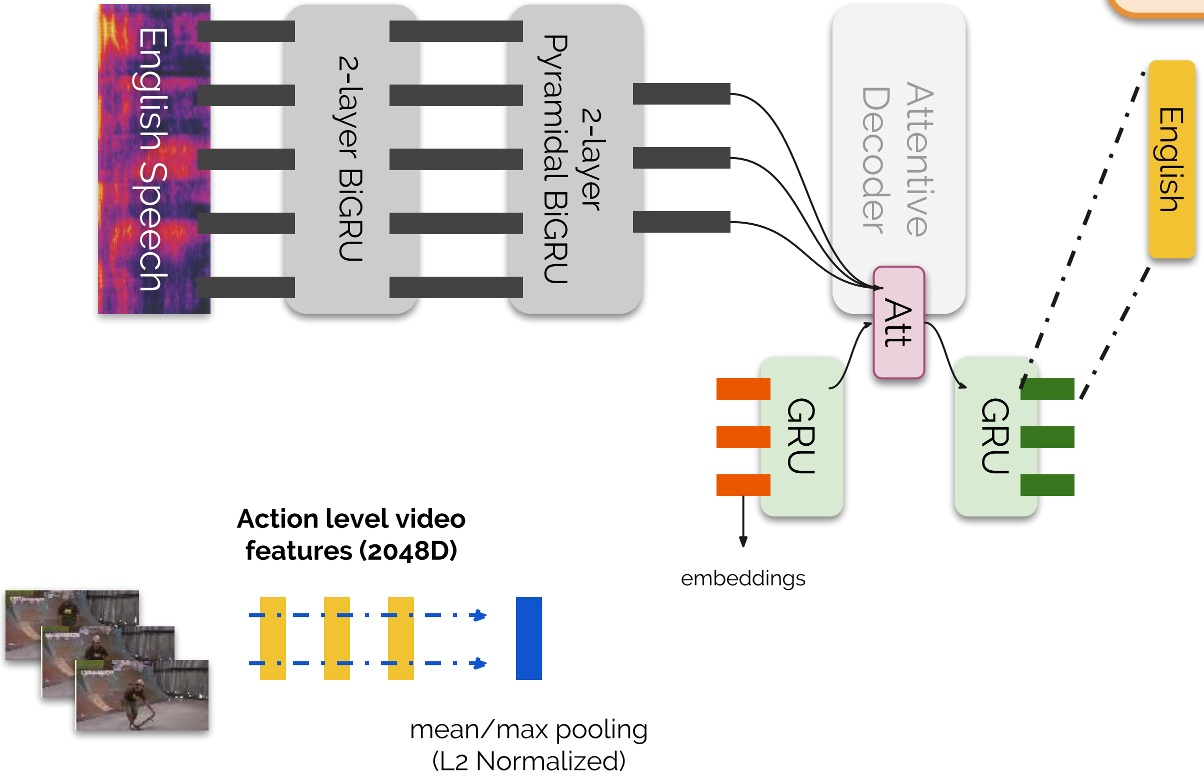
Integration of Features

Motivation: Can we improve decoder LM by visual grounding?



Integration of Features

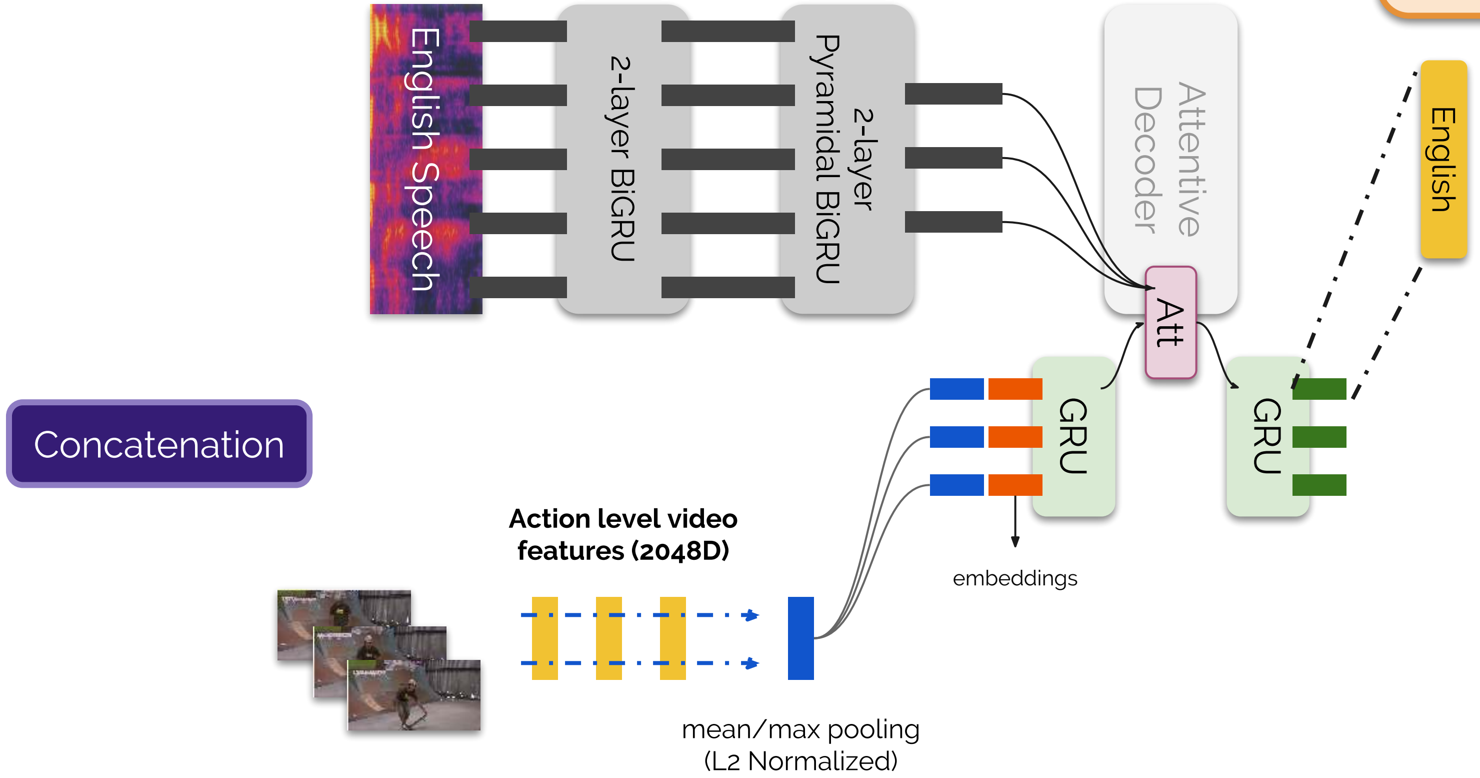
Motivation: Can we improve decoder LM by visual grounding?



[Caglayan et al. 2017]

Integration of Features

Motivation: Can we improve decoder LM by visual grounding?

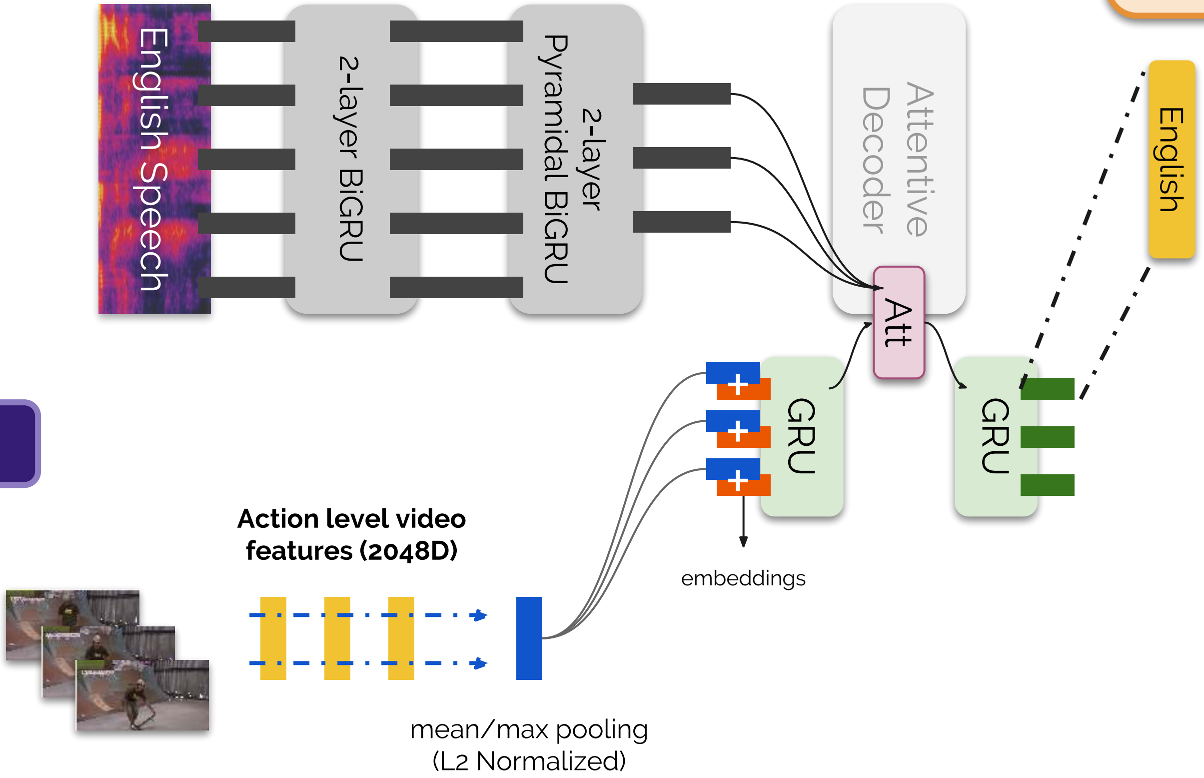


[Caglayan et al. 2017]

Integration of Features

Motivation: Can we improve decoder LM by visual grounding?

Additive

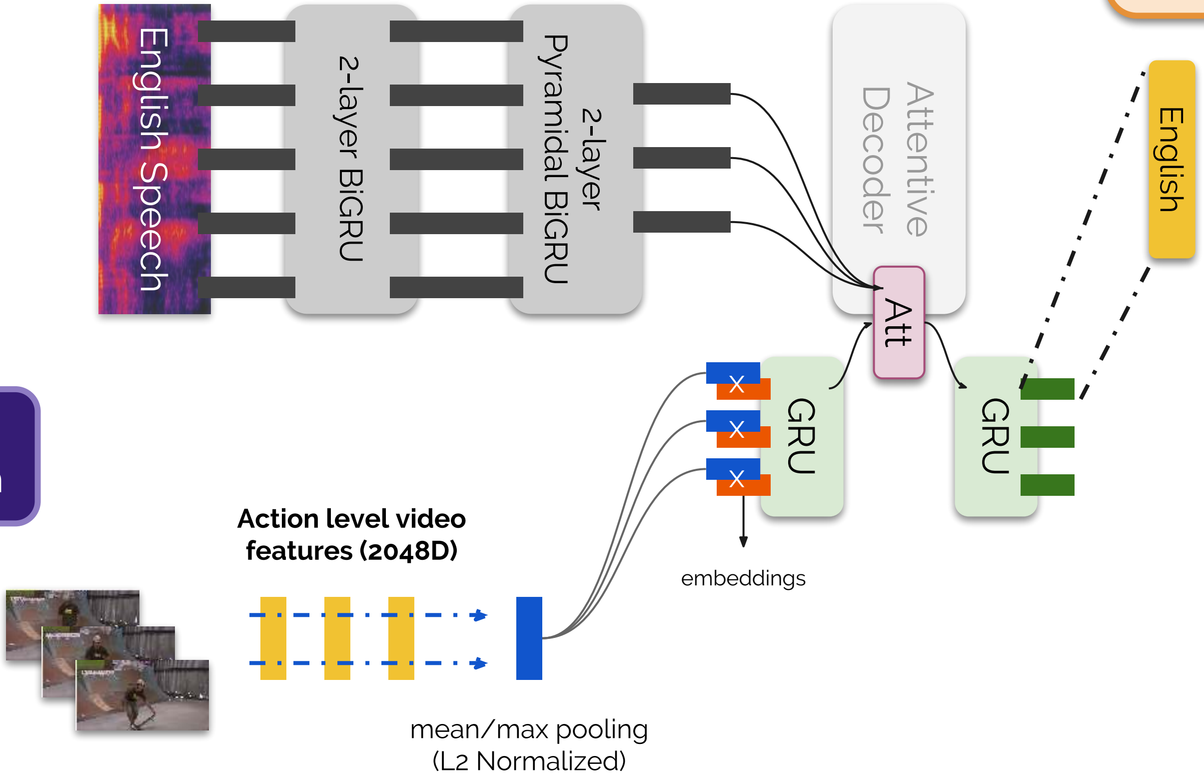


[Caglayan et al. 2017]

Integration of Features

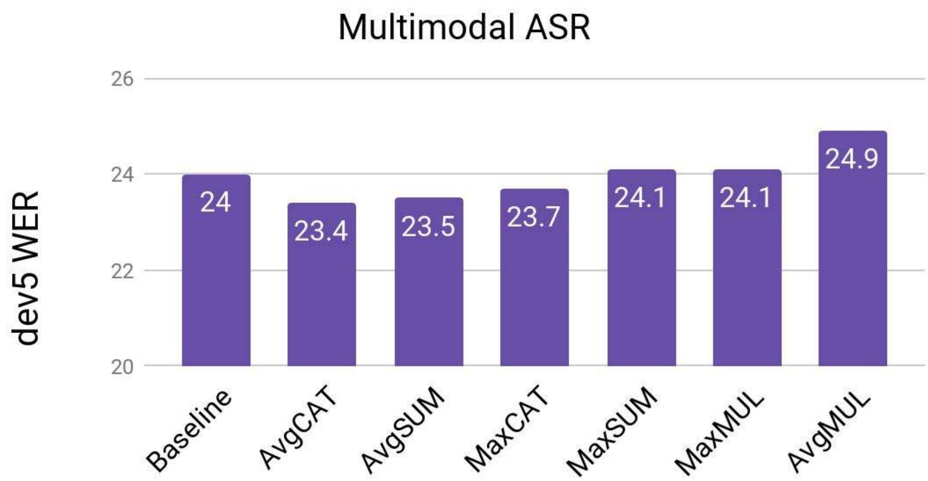
Motivation: Can we improve decoder LM by visual grounding?

Elementwise Multiplication



[Caglayan et al. 2017]

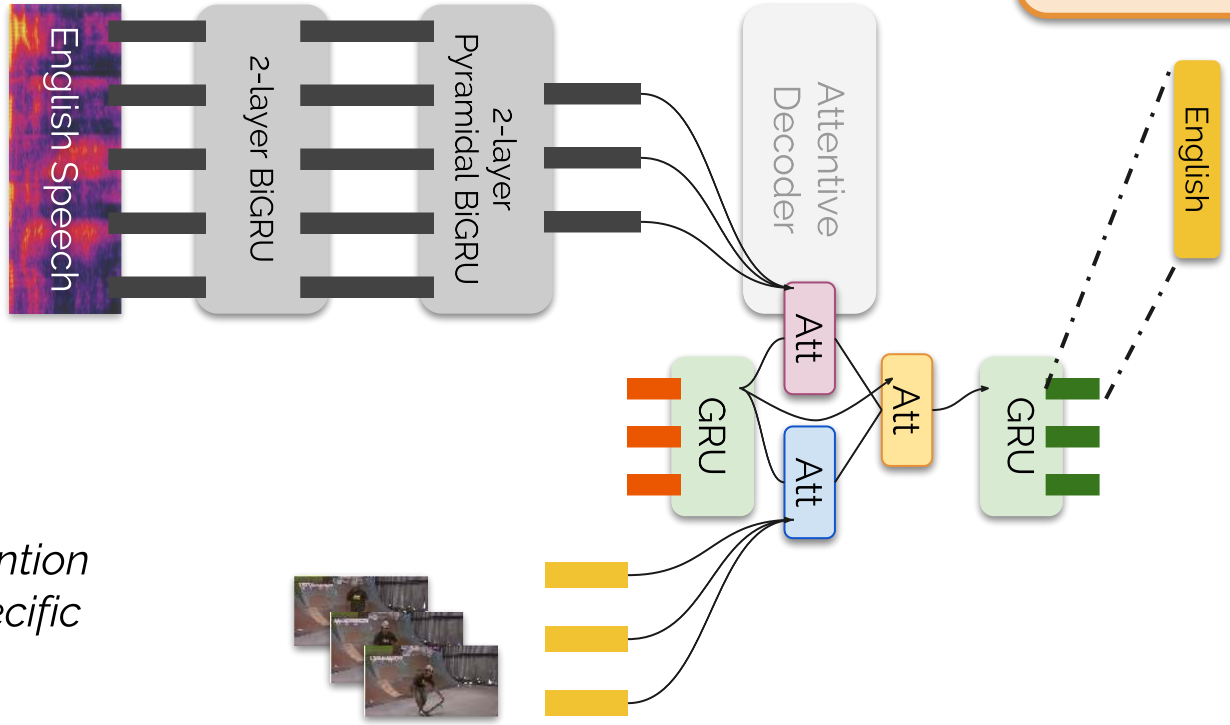
Decoder-side Interaction



- Previous work
  - LM benefits from visual adaptation in terms of PPL [Gupta et al., 2018]
  - Visual features improve acoustic modeling in HMM [Miao & Metze, 2016]
- Hard to conclude for S2S models
  - Need to experiment with bigger models and different features
  - Encoder-side adaptation should be re-explored for 300h

Hierarchical Attention

Motivation: Can we benefit from selective multimodal attention?

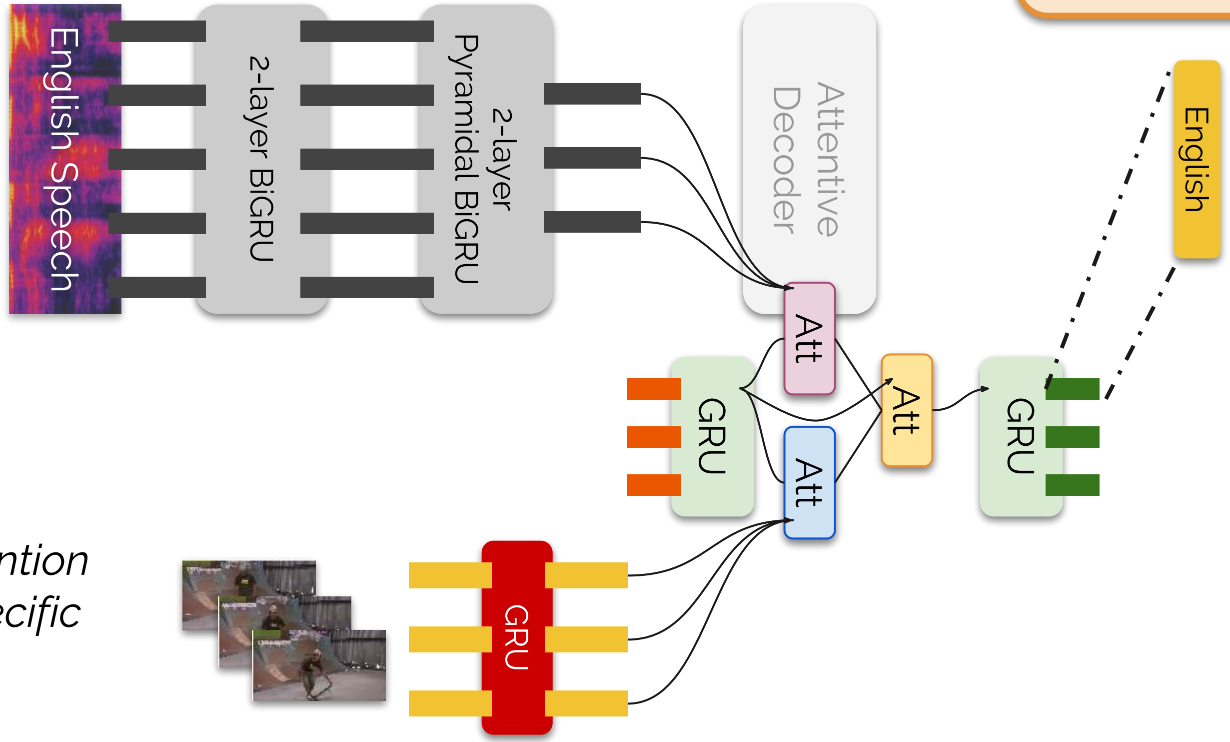


Another layer of attention to fuse modality-specific contexts.

[Libovický et al. 2017]

Hierarchical Attention + ActionGRU

Motivation: Can we benefit from selective multimodal attention?

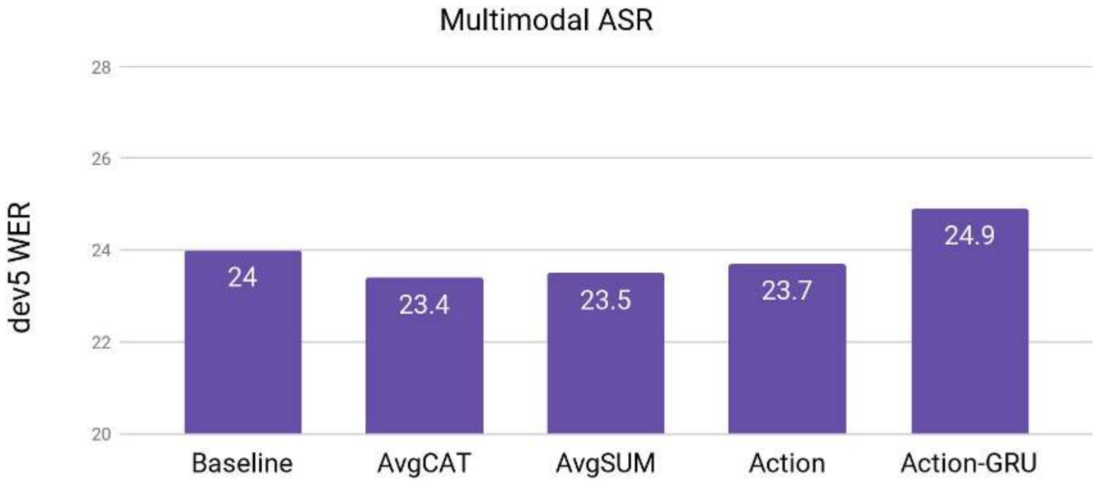


Action level video features (2048D)

Another layer of attention to fuse modality-specific contexts.

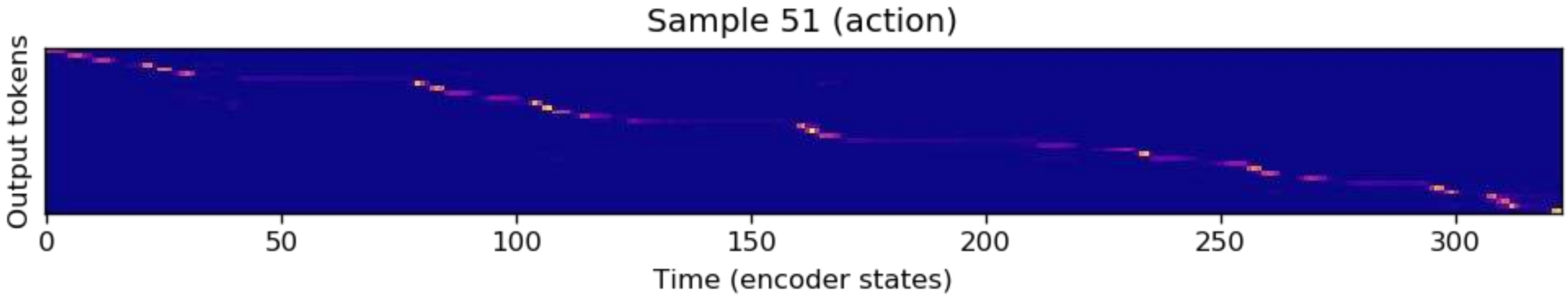
[Libovický et al. 2017]

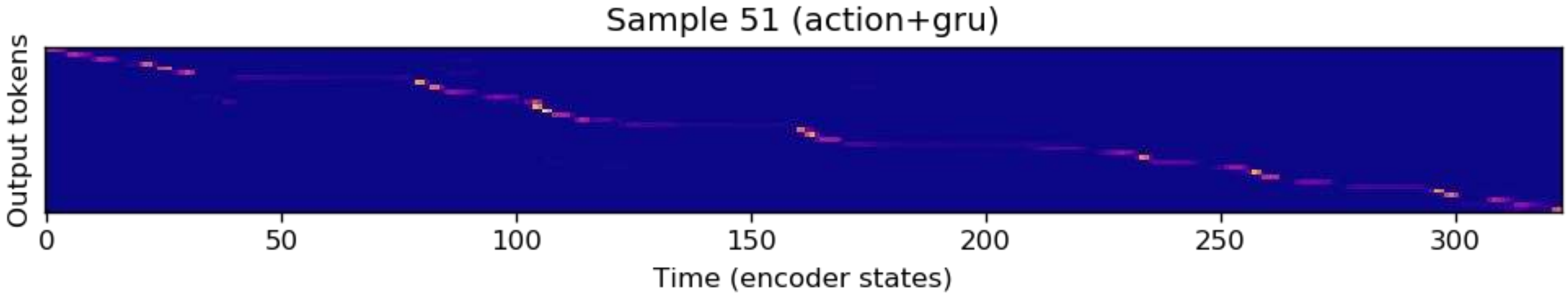
Hierarchical Attention



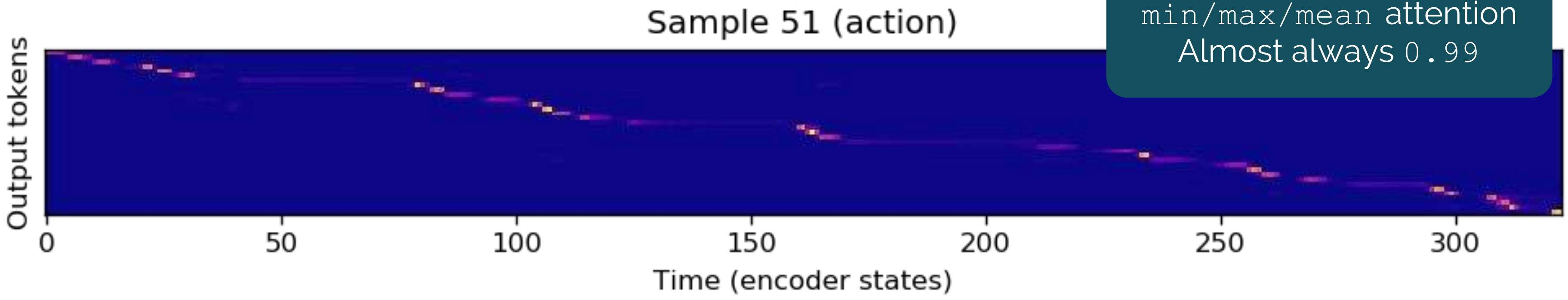
- `AvgCAT/AvgSUM/Action` are comparable: needs further exploration
- Encoding temporal action features with an RNN hurts WER
    - Reason → the model shifts attention

Hierarchical Attention: Example #1

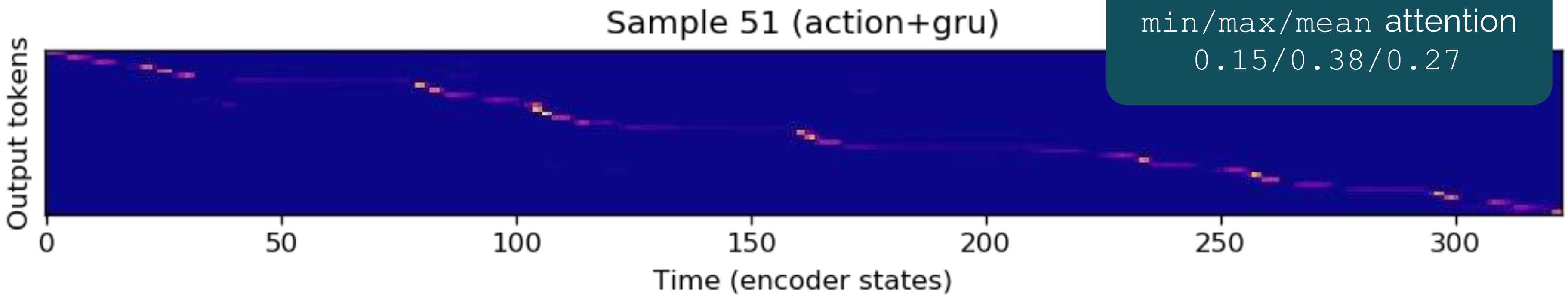




Hierarchical Attention: Example #1

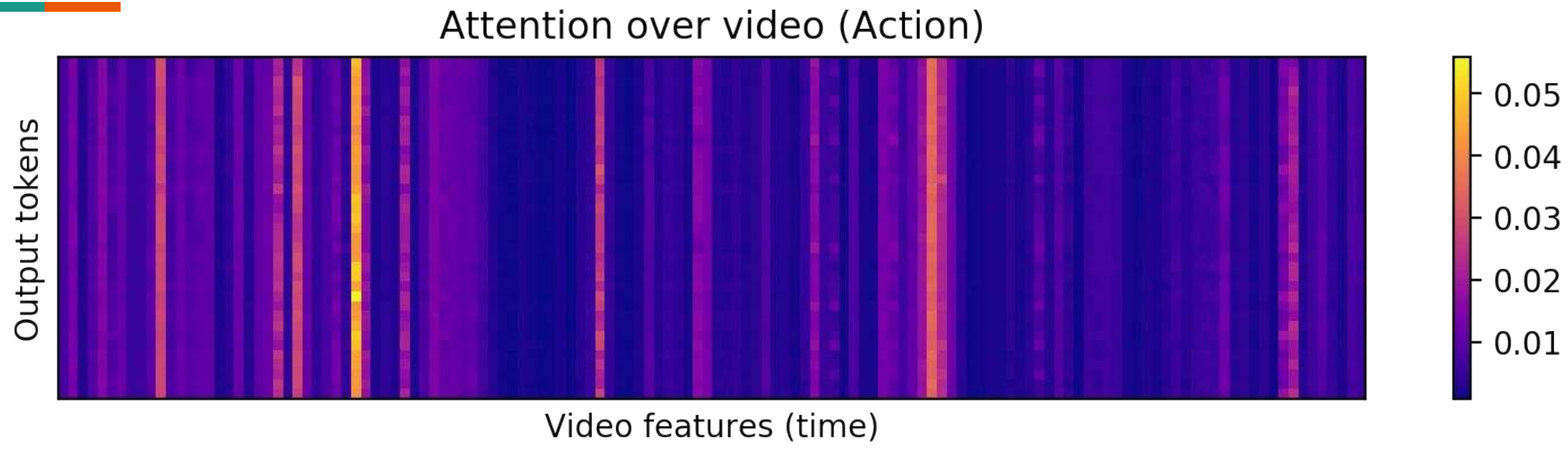


`min/max/mean` attention
Almost always `0.99`

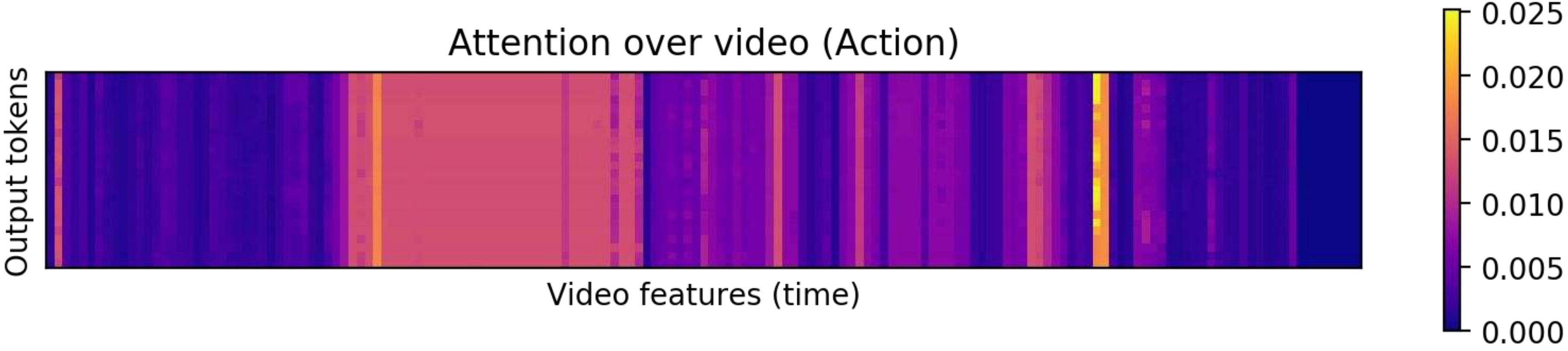


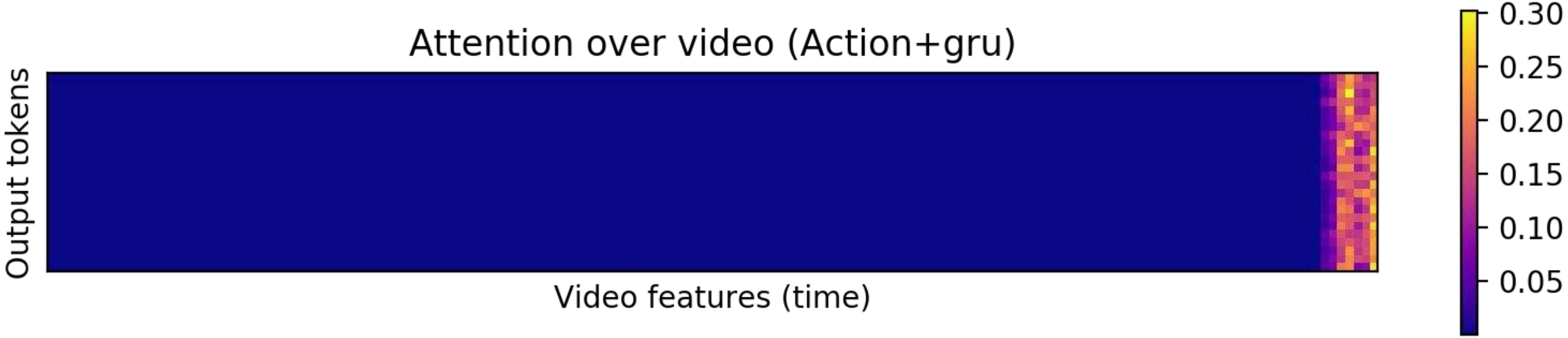
`min/max/mean` attention
`0.15/0.38/0.27`

Hierarchical Attention: Example #1



Hierarchical Attention: Example #2





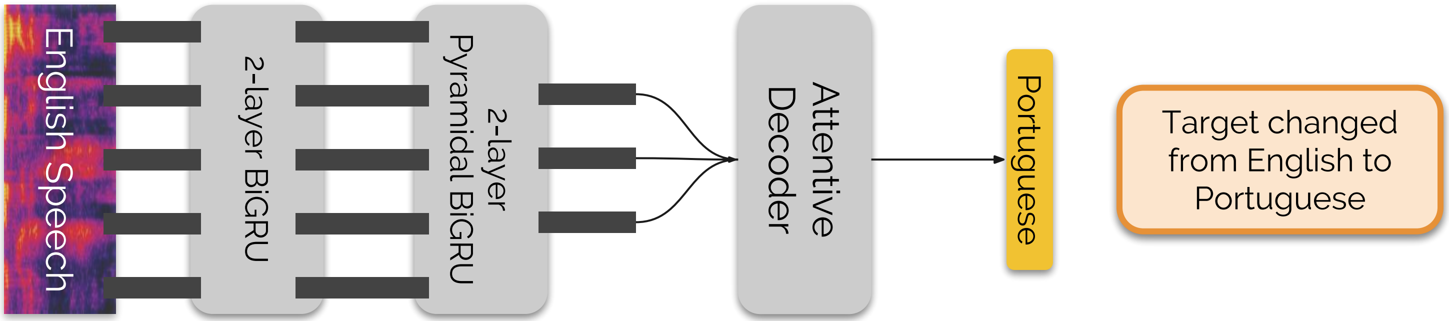
Spoken Language Translation

Spoken Language Translation (SLT)

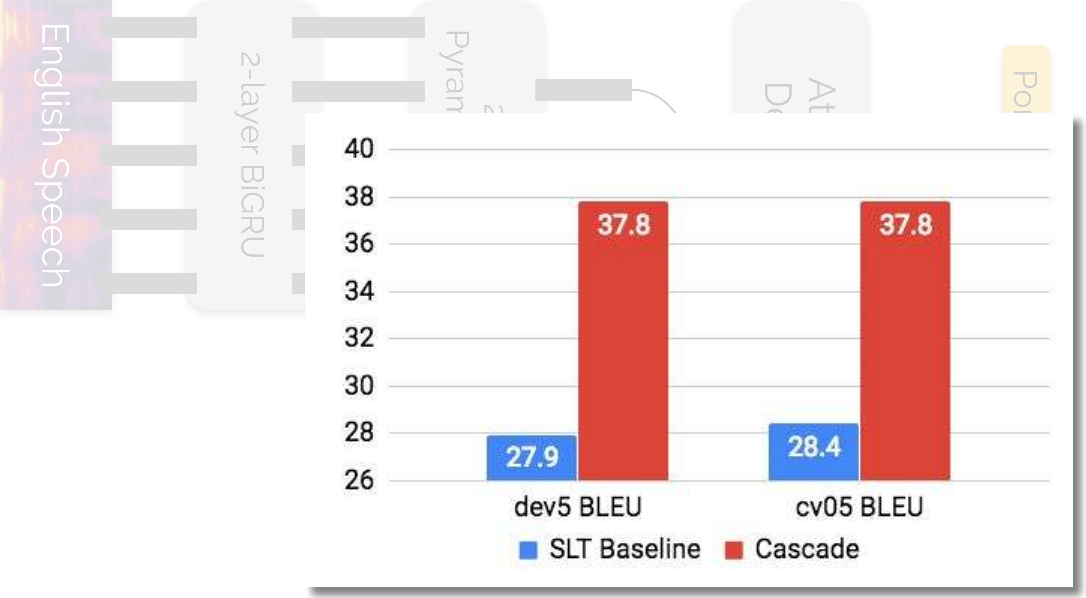
- We have access to English speech, English Text and Portuguese Text
  - Can we improve ASR? `En Speech → En Text`
  - Can we improve SLT? `En Speech → Pt Text`
  - Can we improve MT? `En Text → Pt Text`

- Multi-task Learning
  - Many-to-one
  - One-to-many
  - Hierarchical (auxiliary supervision)

From ASR to SLT



From ASR to SLT

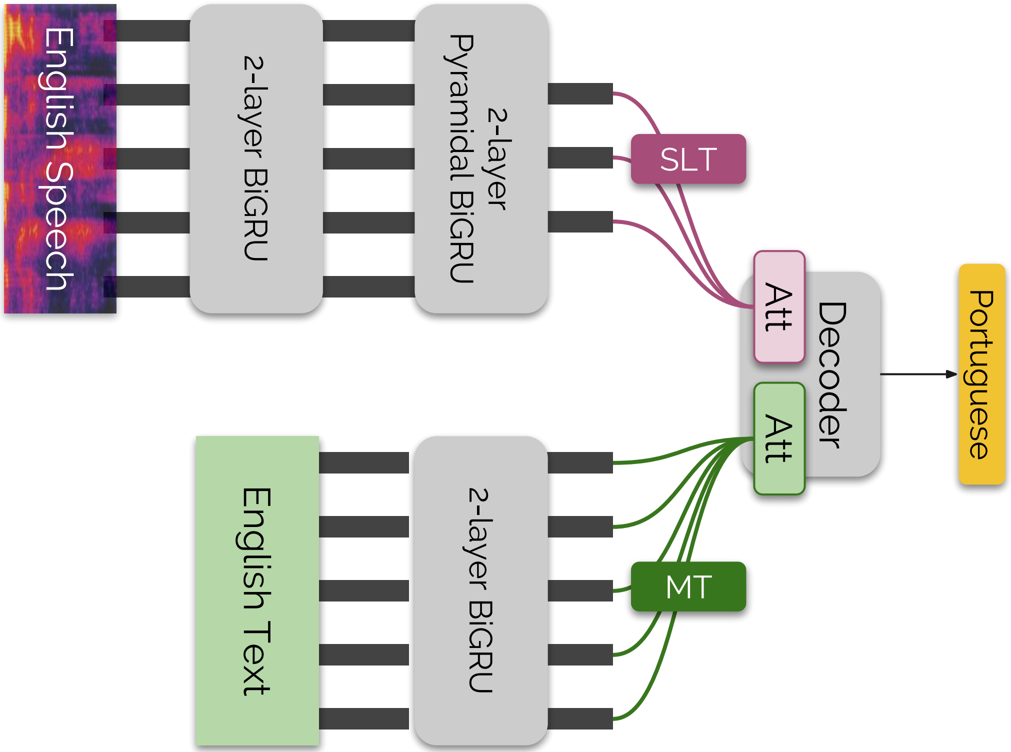


Target changed from English to Portuguese

Almost ~10 BLEU difference

Multi-task Learning “Many-to-One (MTO)”

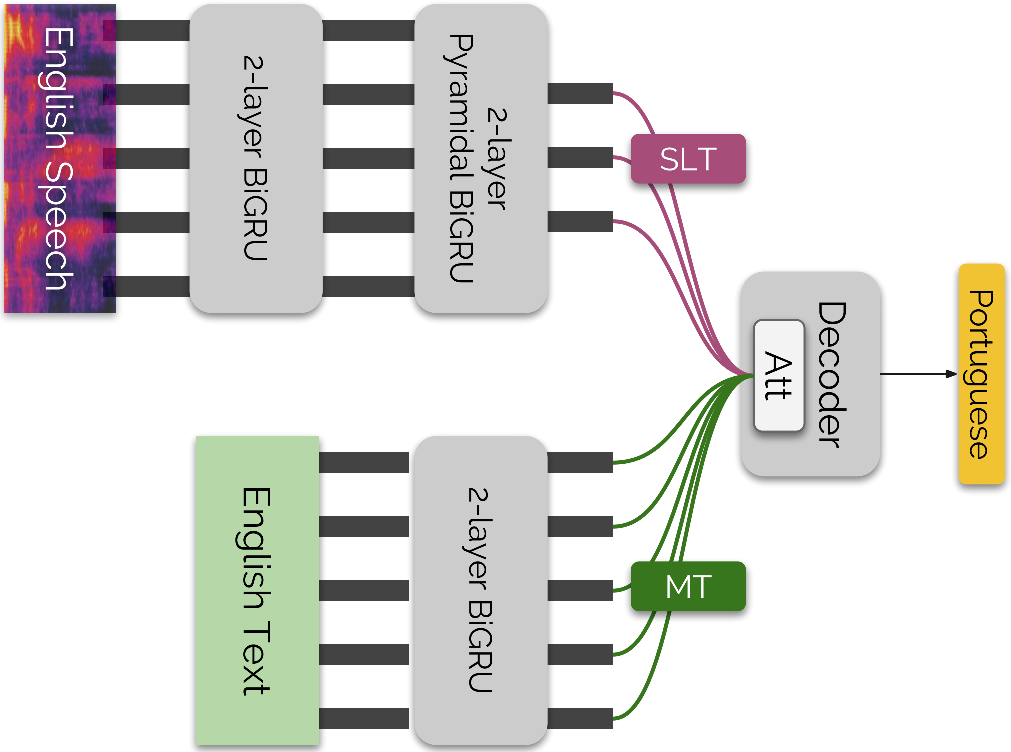
Can we improve SLT/MT?



- **Motivation:** Generalized decoder
- Modality-specific encoders/batches
- Multiplexed training
  - Alternating encoders
  - Sample TASK with p=0.7
- Shared decoder
  - **Separate attention**
  - Shared attention

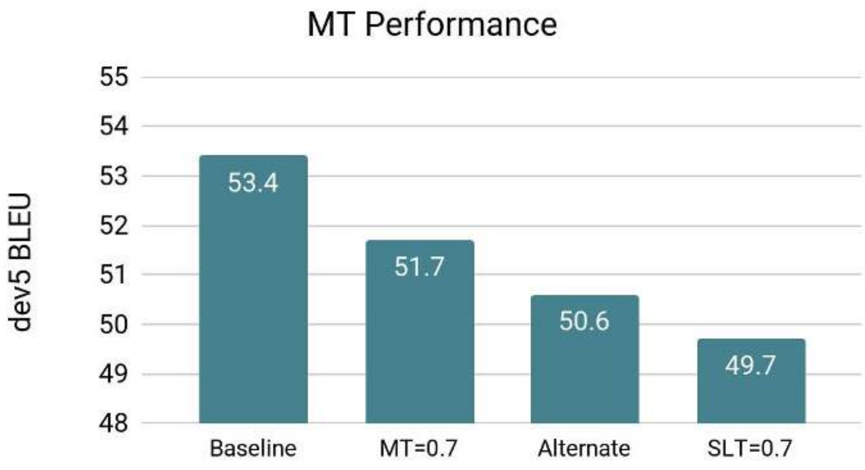
Multi-task Learning “Many-to-One (MTO)”

Can we improve SLT/MT?



- **Motivation:** Generalized decoder
- Modality-specific encoders/batches
- Multiplexed training
  - Alternating encoders
  - Sample TASK with p=0.7
- Shared decoder
  - Separate attention
  - **Shared attention**

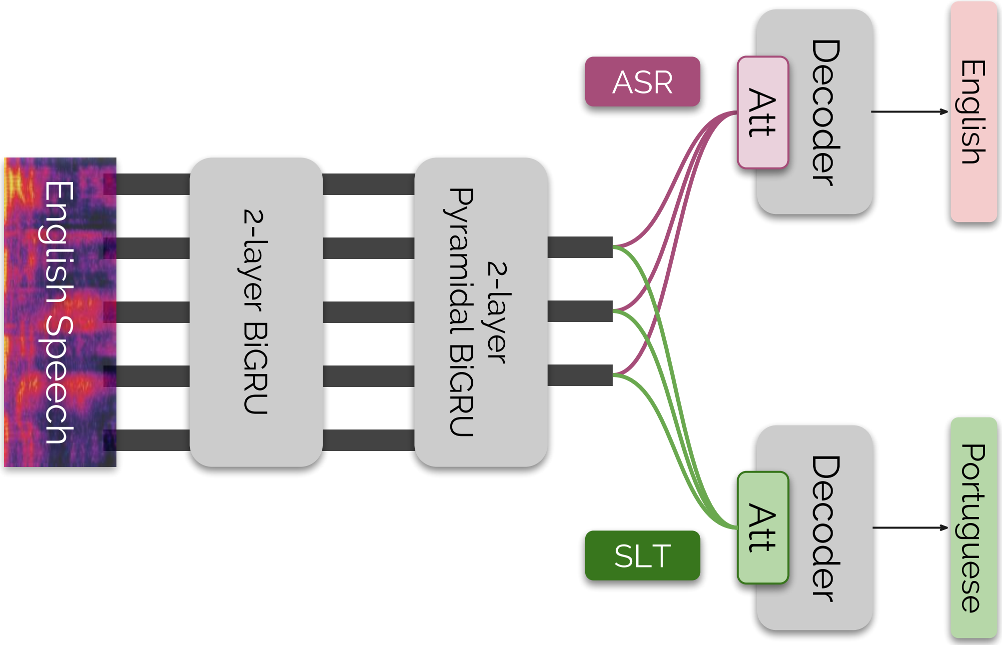
Many-to-one: Speech & EN → PT



- **SLT** benefits from **MT** even with alternating policy
- **MT** does not benefit from **SLT**

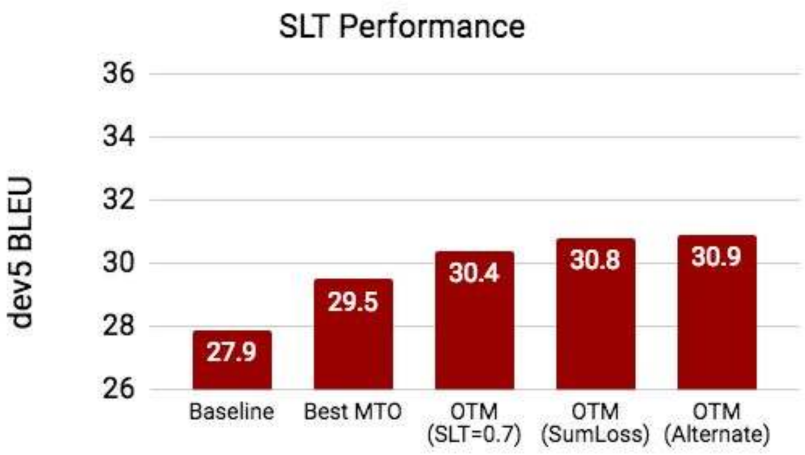
One-to-many: Speech → EN & PT

Can we improve SLT/ASR?



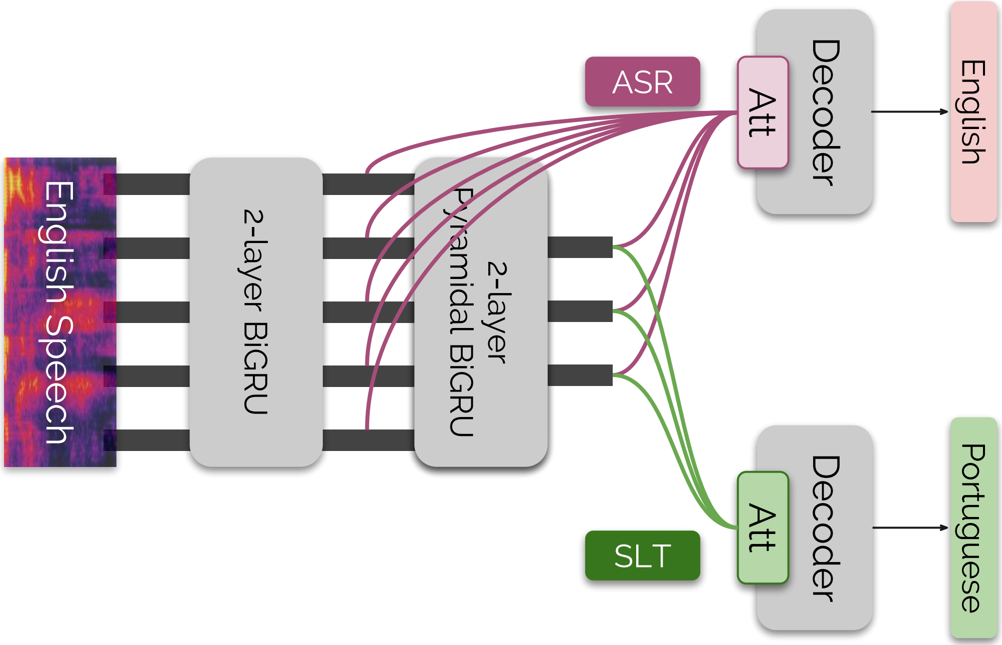
- **Motivation:** Generalized encoder
- Task-specific decoders
- In addition to scheduling:
  - Sum-of-losses model

One-to-many: Speech → EN & PT



- **OTM** clearly better than **MTO**
- `SumLoss` and `Alternate` better than `SLT=0.7`
  - No need to schedule for **OTM**
  - **Alternate** → 3 BLEU and 1.5 WER improvements

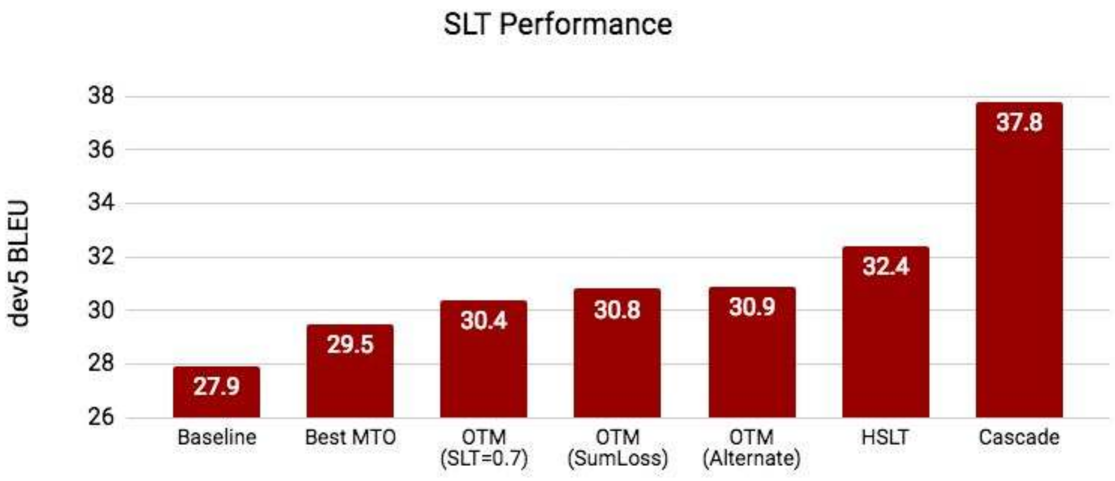
Hierarchical SLT (HSLT)



One-to-Many architecture with sum of losses

- **Motivation:** Ground the intermediate representation of the encoder with ASR supervision

One-to-Many vs HSLT



- **HSLT** even better than **OTM** for SLT
- ASR performance of HSLT very bad

Multimodal ASR and SLT Conclusions

- Multimodal ASR
  - Decoder side improvements consistent with MNMT [Caglayan et al., 2017]
  - Further exploration: Temporal smoothing of visual features
  - More analysis in later parts of the talk

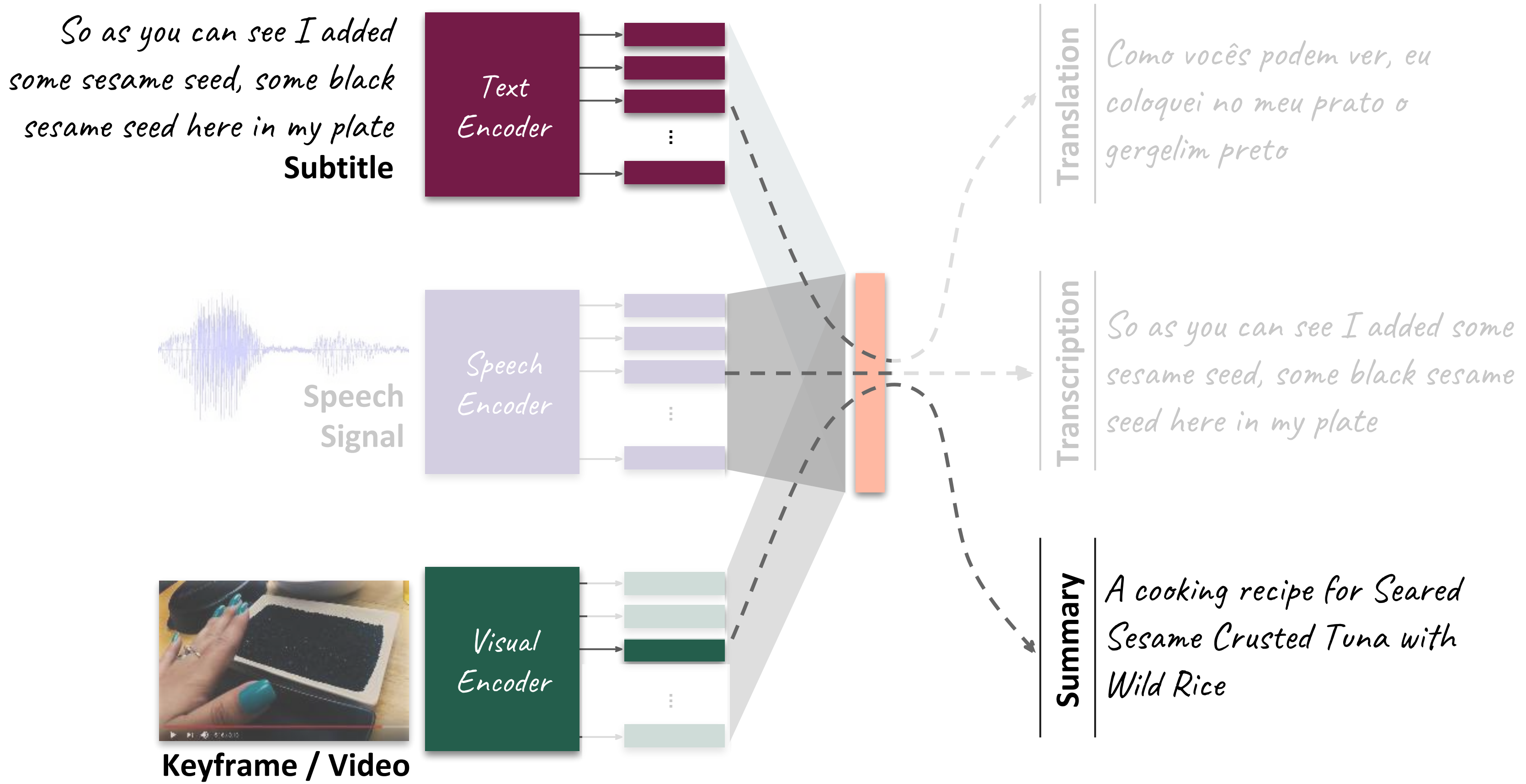
- Spoken Language Translation
  - Mutual benefits between SLT and ASR tasks
  - One-to-Many (OTM) better than Many-to-One (MTO)
  - Hierarchical SLT performs best, closing gap to “Cascade”

Summarization (“Teaser Generation”)

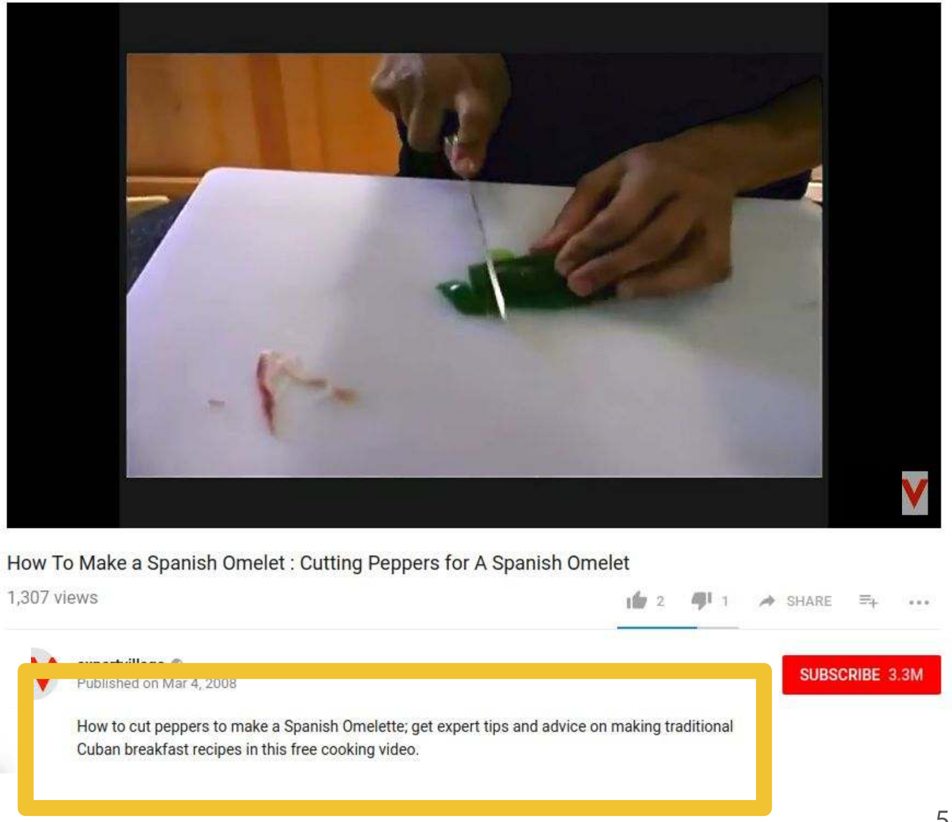
Florian, Jasmine, Jindrich, Shruti, Spandana

The big picture

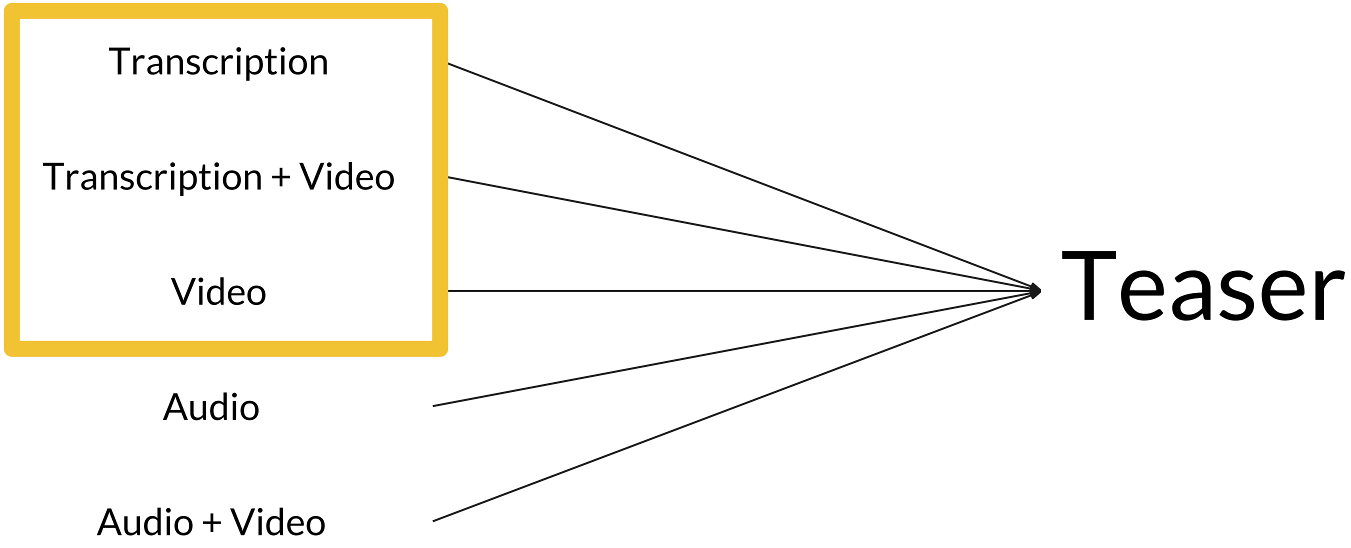


Teaser Generation

- Summarization
  - Present subset of information in a more compact form (maybe across modalities)
- "Description" field
  - 2-3 sentences of meta data: template based, uploader provides
  - "Informative" and abstractive summary of a how-to video
  - Should generate interest of a potential viewer



General Experimental Setup



Used 2000h of data: 74k videos for training, and 5k for validation/ test
(keeping original dev/ test/ heldout sets intact)

Spanish Omelet

~1.5 minutes of audio and video

"Teaser" (33 words on avg)

```
how to cut peppers to make a spanish
omelette ; get expert tips and advice on
making cuban breakfast recipes in this free
cooking video .
```

Transcript (290 words on avg)

```
on behalf of expert village my name is lizbeth muller and today we are going to show you how to make spanish
omelet . i 'm going to dice a little bit of peppers here . i 'm not going to use a lot , i 'm going to use very
very little . a little bit more then this maybe . you can use red peppers if you like to get a little bit color
in your omelet . some people do and some people do n't . but i find that some of the people that are mexicans
who are friends of mine that have a mexican she like to put red peppers and green peppers and yellow peppers in
hers and with a lot of onions . that is the way they make there spanish omelets that is what she says . i loved
it , it actually tasted really good . you are going to take the onion also and dice it really small . you do n't
want big chunks of onion in there cause it is just pops out of the omelet . so we are going to dice the up also
very very small . so we have small pieces of onions and peppers ready to go .
```

Dataset statistics

Most frequent words in transcript

41812	,	5627	have
41125	.	5035	with
33193	the	5022	are
30993	to	5007	just
25738	you	4555	be
25348	and	4459	for
19516	a	4294	want
15838	it	4078	up
14457	that	3860	if
13966	of	3805	'm
12594	is	3621	or
11573	i	3586	here
9731	going	3572	like
9652	in	3487	one
9384	we	3475	as
8698	your	3465	now
8491	this	3324	there
8185	's	3278	they
7873	so	3259	what
6877	on	3148	go
6571	're	2956	then
6347	do	2933	get

Most frequent words in teasers

4806	.	579	your
3806	a	387	clip
3799	in	369	when
3058	this	360	get
2922	**free**	349	-
2883	the	339	more
2876	to	328	that
2832	**video**	327	you
2264	and	307	lesson
1948	**learn**	298	are
1779	from	285	by
1720	on	273	's
1639	with	268	make
1460	**how**	262	be
1321	**tips**	257	can
1220	,	242	do
1117	for	232	music
1036	of	225	or
756	**expert**	221	it
675	an	218	use
654	**about**	217	out
634	is	214	as

Evaluation Metrics (1)

Reference

```
a ukulele is a cousin instrument to the guitar with four strings
played in folk music . learn about ukulele anatomy from a musician
in this free guitar video .
```

Hypothesis

```
the banjo 's ukulele has many different types of guitar . learn
more about the banjo string and guitar with tips from a guitar
instructor in this free music lesson video .
```

Evaluation Metrics (2)

Catchphrases in teasers

```
3799 in
3058 this
2922 free
2832 video
1948 learn
1460 how
1321 tips
 756 expert
```

>=500 times

- **Rouge-L**
  - Standard summarization evaluation metric
  - F-score over longest common subsequence
    → captures structural coherence

- **Content word F-score** (using Meteor code)
  - No crossover penalty (Gamma)
  - Zero weight to function words (Delta)
  - Equal weight to Precision and Recall (Alpha)

ROUGE-L

Reference

a ukulele is a cousin instrument to **the guitar** with four strings played in folk music **. learn about** ukulele anatomy **from a** musician **in this free** guitar **video .**

Hypothesis

the banjo 's ukulele has many different types of **guitar . learn** more **about** the banjo string and guitar with tips **from a** guitar instructor **in this free** music lesson **video .**

Reference length = 30
Hypothesis length = 32
Common subsequence length = 12

Recall = $^{12}/_{30}$ = .40

Precision = $^{12}/_{32}$ = .38

F1 score = .39

Content word F-score

Reference

~~a~~ ukulele ~~is a~~ cousin instrument ~~to the~~ guitar ~~with~~ four strings
played ~~in~~ folk music ~~.~~ **~~learn~~** ~~about~~ ukulele anatomy ~~from a~~ musician
~~in~~ ~~this~~ **~~free~~** guitar **~~video~~** ~~.~~

Hypothesis

~~the~~ banjo ~~'s~~ ukulele ~~has many~~ different types ~~of~~ guitar ~~.~~ **~~learn~~**
~~more~~ ~~about~~ ~~the~~ banjo string ~~and~~ guitar ~~with~~ **~~tips~~** ~~from~~ ~~a~~ guitar
instructor ~~in~~ ~~this~~ **~~free~~** music lesson **~~video~~** ~~.~~

Reference content words = 13
Hypothesis content words = 12
Matching words = 4

Recall = $^4/_{13} = .31$

Precision = $^4/_{12} = .33$

F1 score = .32

Evaluation Metrics

Catchphrases in teasers

```
3799 in
3058 this
2922 free
2832 video
1948 learn
1460 how
1321 tips
 756 expert
```

>=500 times

- **Rouge-L**
  - Standard summarization evaluation metric
  - F-score over longest common subsequence → captures structural coherence
  - Prefers style over content
- **Content word F-score** (using Meteor code)
  - No crossover penalty (Gamma)
  - Zero weight to function words (Delta)
  - Equal weight to Precision and Recall (Alpha)
  - Ignores fluency

Rule-based Baseline

- Rule based extractive summary - 1 most informative sentence
  - Sentence contains "`how to`"
  - The predicate is "`learn`", "`tell`", "`show`", "`discuss`", "`explain`"
  - Second sentence in the transcript

```
on behalf of expert village my name is
lizbeth muller and today we are going to show
you how to make spanish omelet .
```

Rouge-L	Content F1
16.4	18.8

Random Baseline

- Train a language model on the teasers and sample from the model
- Nice text, correct style, nonsense content

```
learn tips on how to play the bass drum beat
variation on the guitar in this free video
clip on music theory and guitar lesson.
```

Rouge-L	Content F1
27.5	8.3

S2S models: Vocabulary

- S2S model with attention
- Vocabulary matters

```
how to add tomatoes to a spanish omelette ;
get expert tips and advice on making
traditional cuban breakfast recipes in this
free cooking video .
```

	Rouge-L	Content F1
BPE 10k	45.1	35.5
BPE 20k	46.5	37.8
Tokens 20k	53.9	47.4
Tokens 30k	53.5	46.3

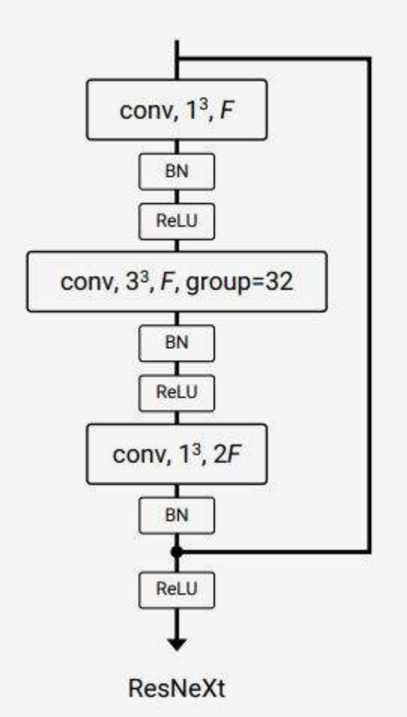
Almost no proper names, no place for BPE to show off

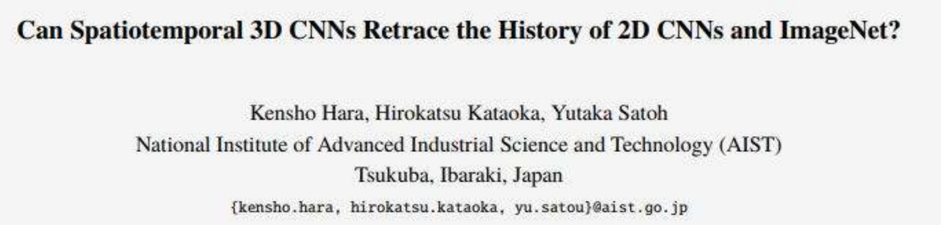
No gain from from larger vocabulary, just trains slowly

Do we need the complete transcript?

	Rouge-L	Content F1
No input = Language model	27.5	8.3
Extracted sentence (itself 18.8 F1 points)	46.6	36.0
First 200 tokens	40.3	27.5
Complete transcript (up to 650 tokens)	53.9	47.4

Action Recognition Features

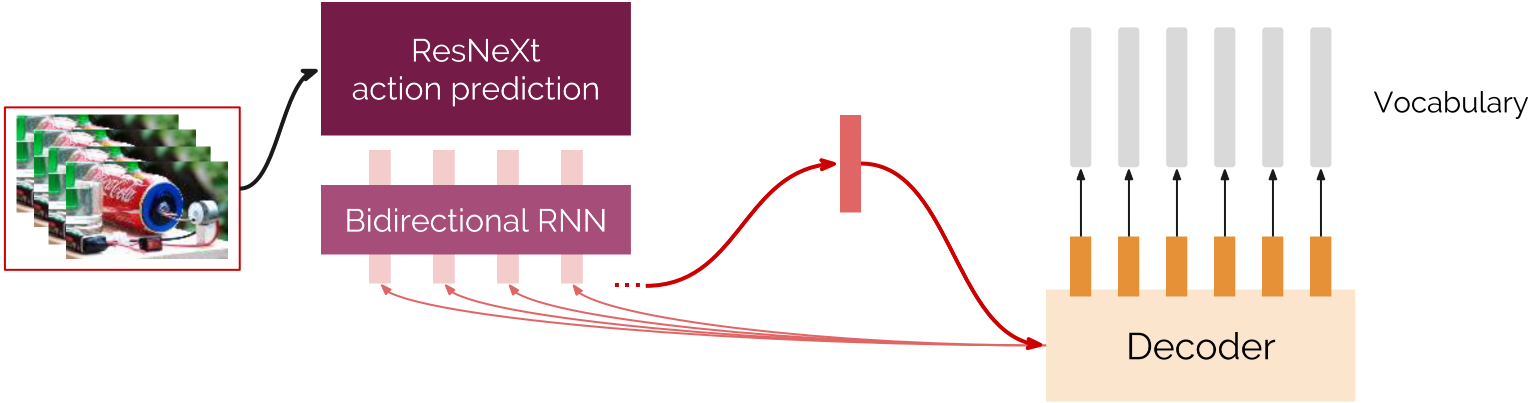




Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?

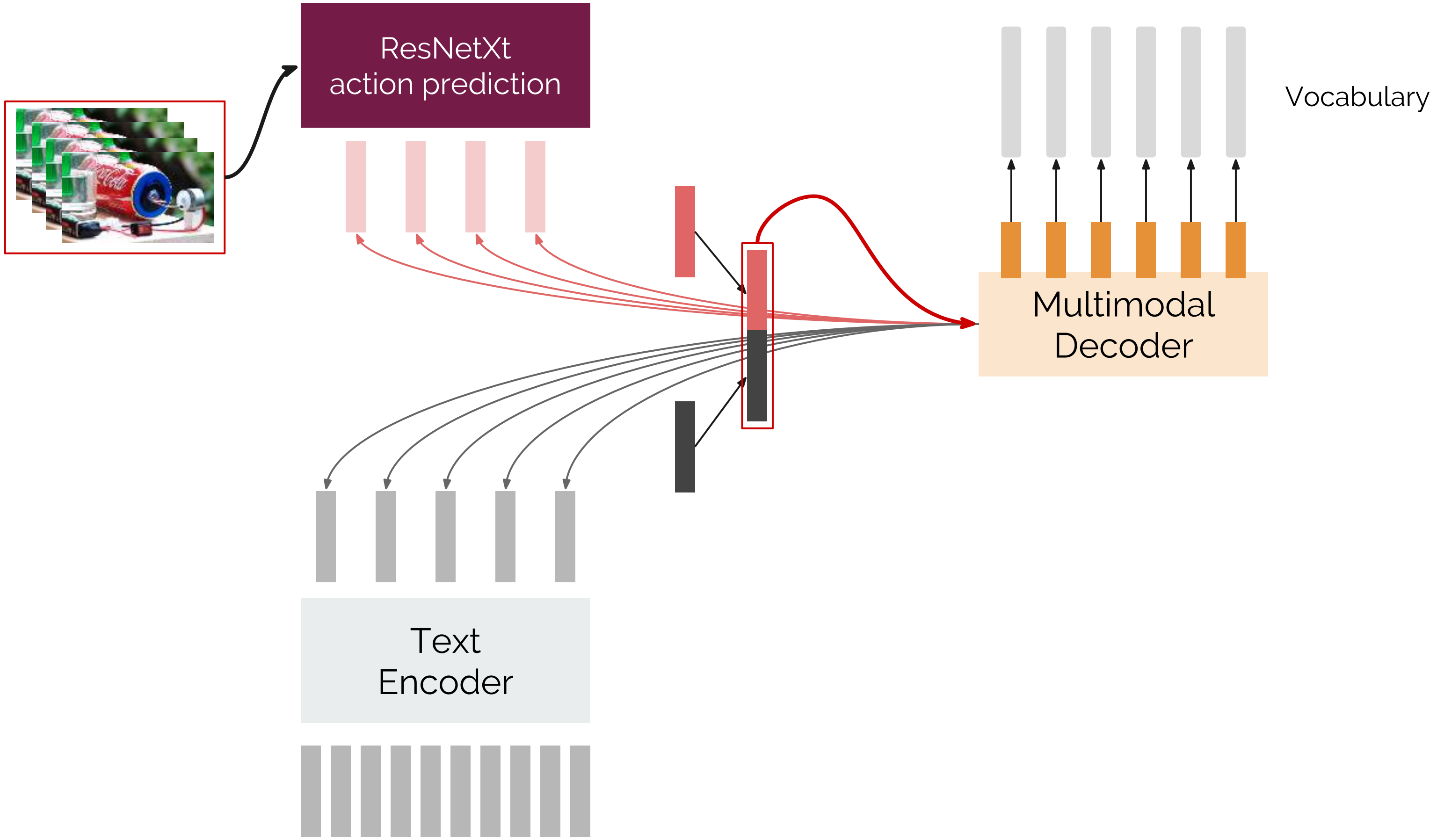
Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan
{kensho.hara, hirokatsu.kataoka, yu.satou}@aist.go.jp

Video Features as Input

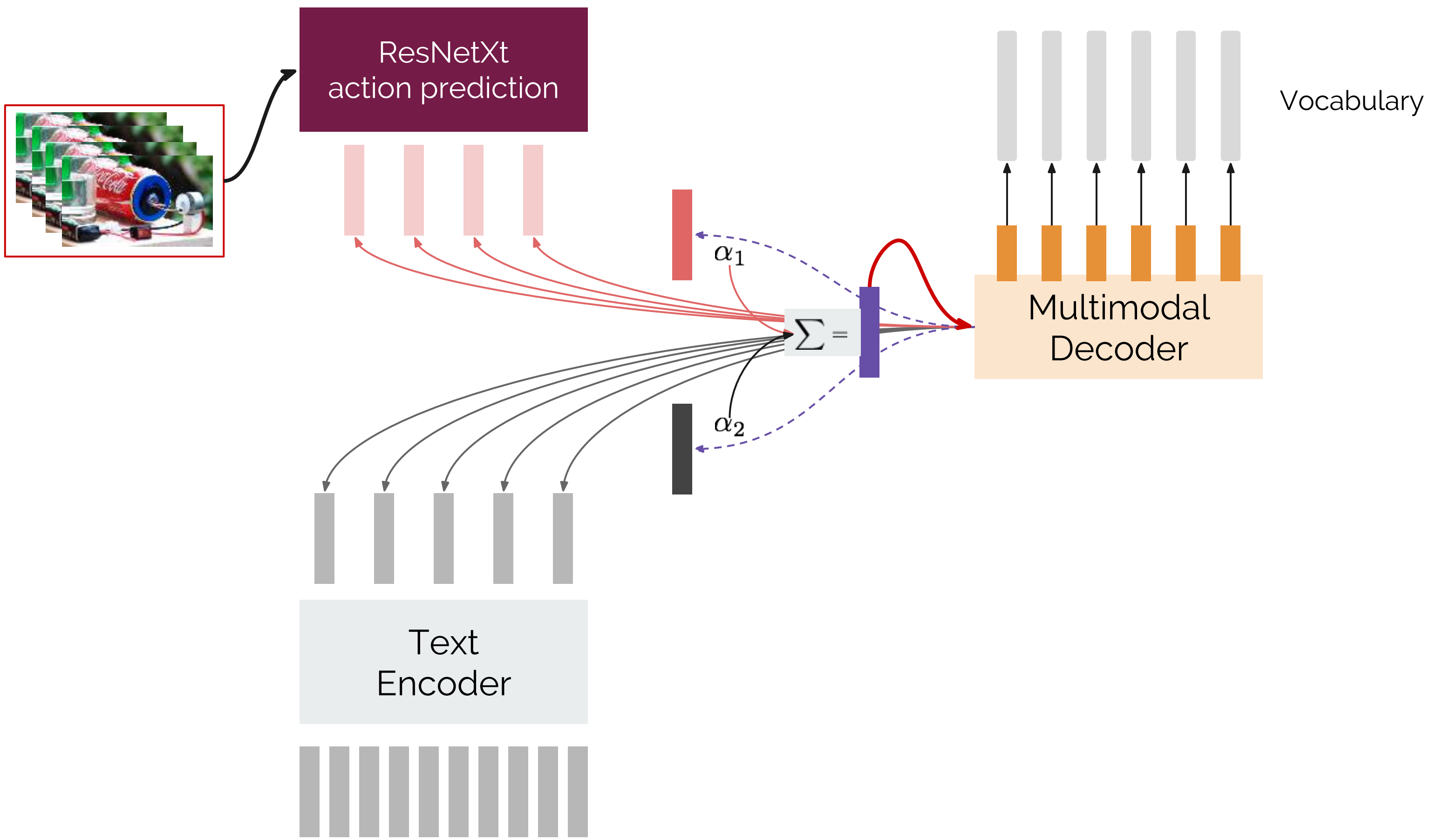


	Rouge-L	Content F1
Text-only input	53.9	47.4
Features only	38.5	24.8
Features + RNN	46.3	34.9

Multi-modal Attention



Hierarchical Multi-modal Attention

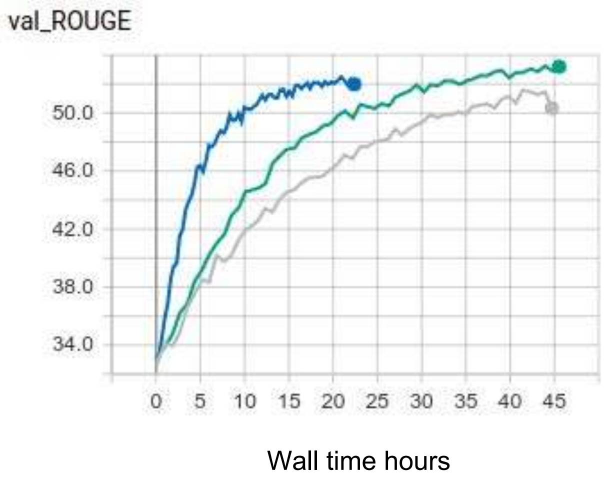


Results of Attention Combination

- Modest improvements when we combine text and video

	Rouge-L	Content F1
Text-only input	53.9	47.4
Context vector concatenation	51.0	44.4
Hierarchical attention	54.9	48.9

Slow to converge

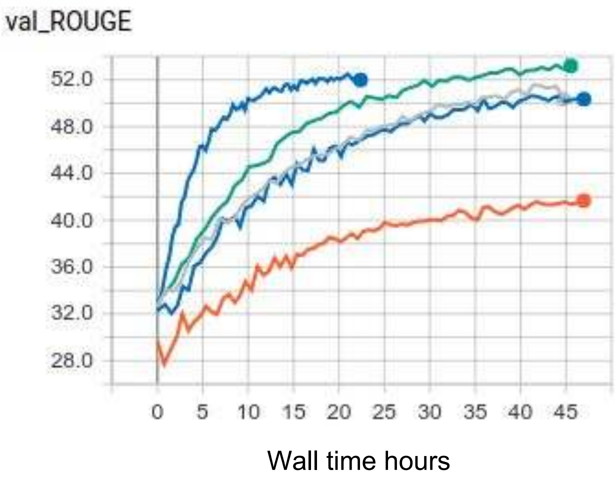


Results of Attention Combination

- Modest improvements when we combine text and video
- RNN over action features does not seem to help

	Rouge-L	Content F1
Text-only input	53.9	47.4
Context vector concatenation	51.0	44.4
+ RNN over actions	42.2	30.3
Hierarchical attention	54.9	48.9
+ RNN over actions	53.4	46.8

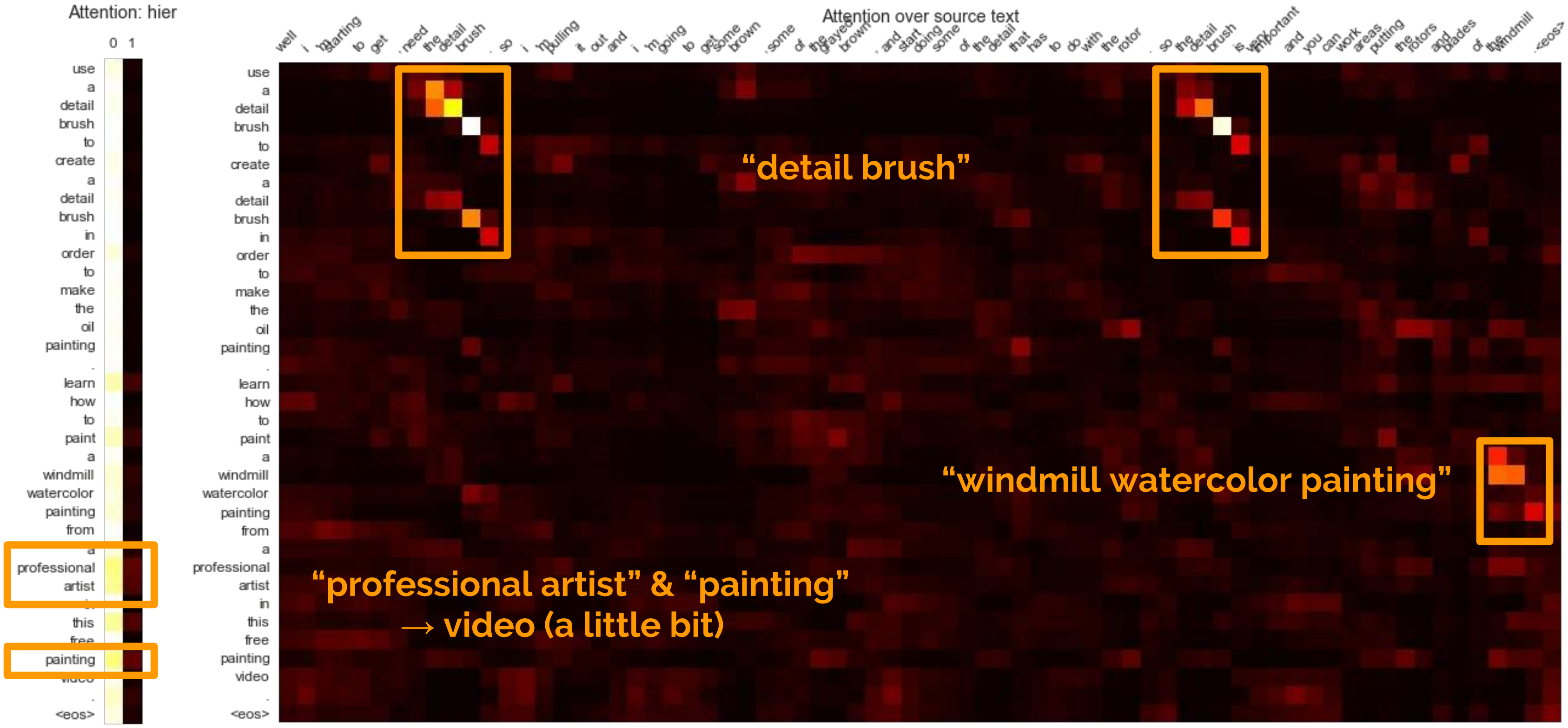
Slow to converge



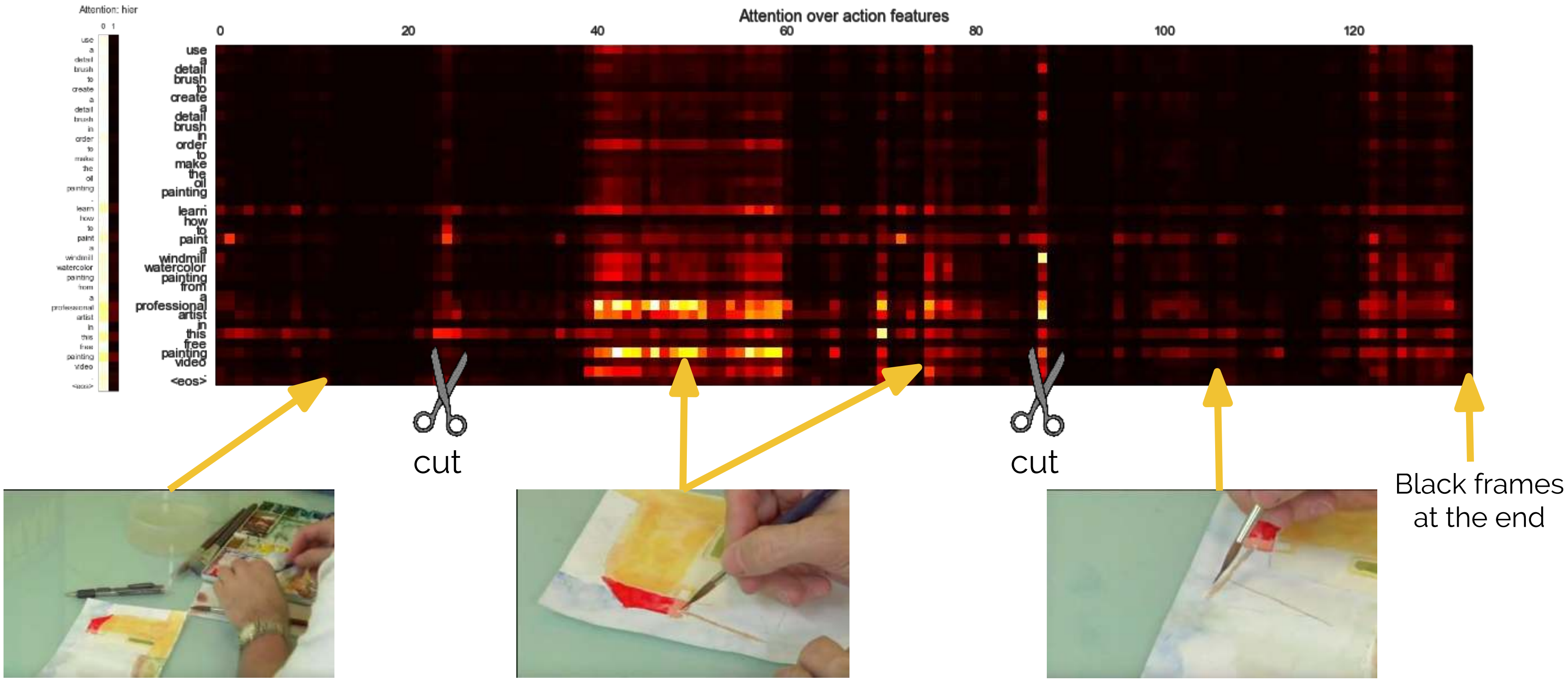
Overview of the Result

	Rouge-L	Content F1
Language model	27.5	8.3
Extractive rules	16.4	18.8
S2S from extractive rules	46.6	36.0
Text-only input	53.9	47.4
Action features	38.5	24.8
Action features + RNN	46.3	34.9
Text + action features w/o RNN	54.9	48.9
Text + action features w/ RNN	53.4	46.8

Attention over the Transcriptions



Attention over the Video Features



Talking and preparing the brush

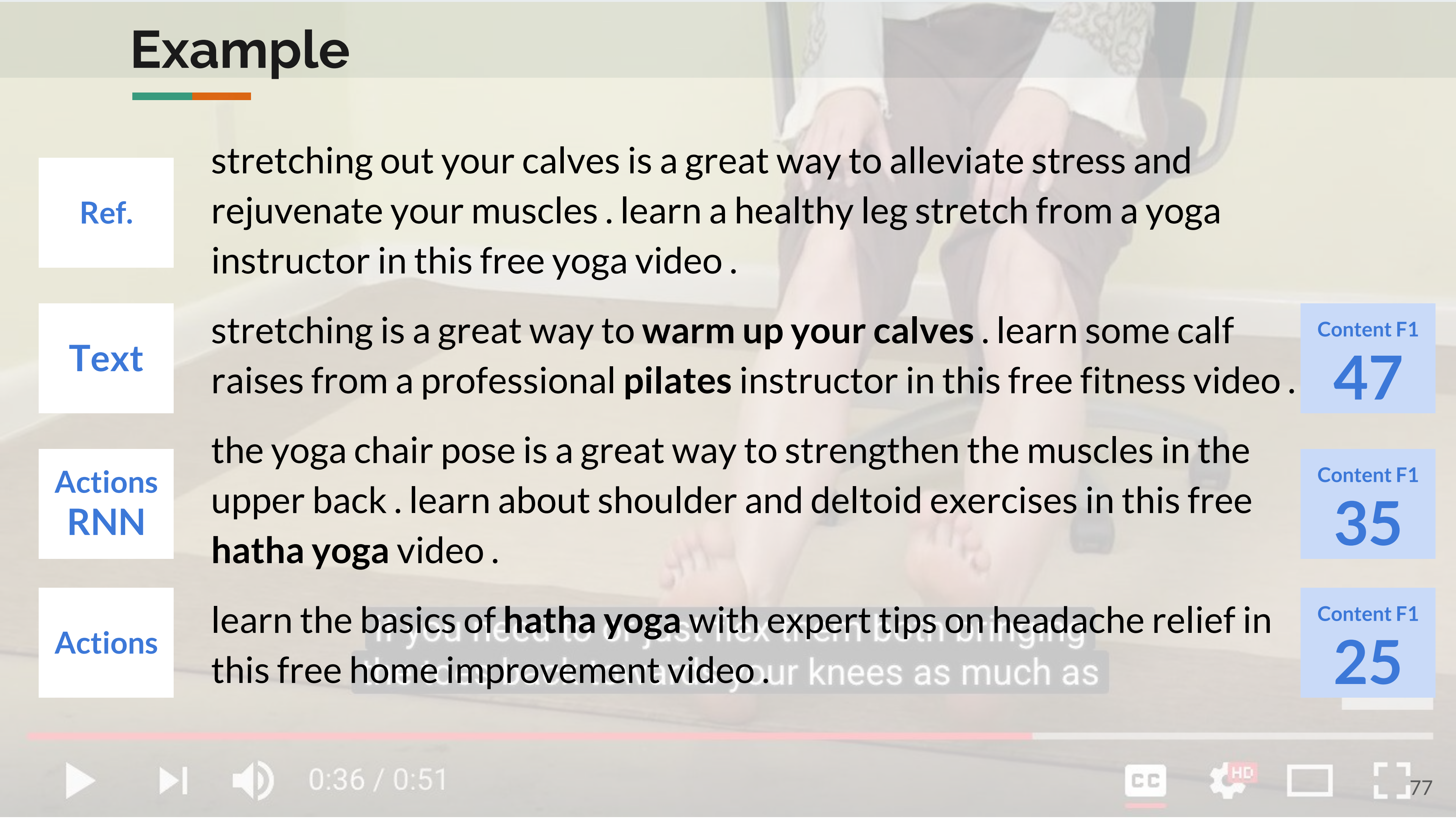
Close-up of brushstrokes **w/ hand**

Close-up of brushstrokes **no hand**

Example

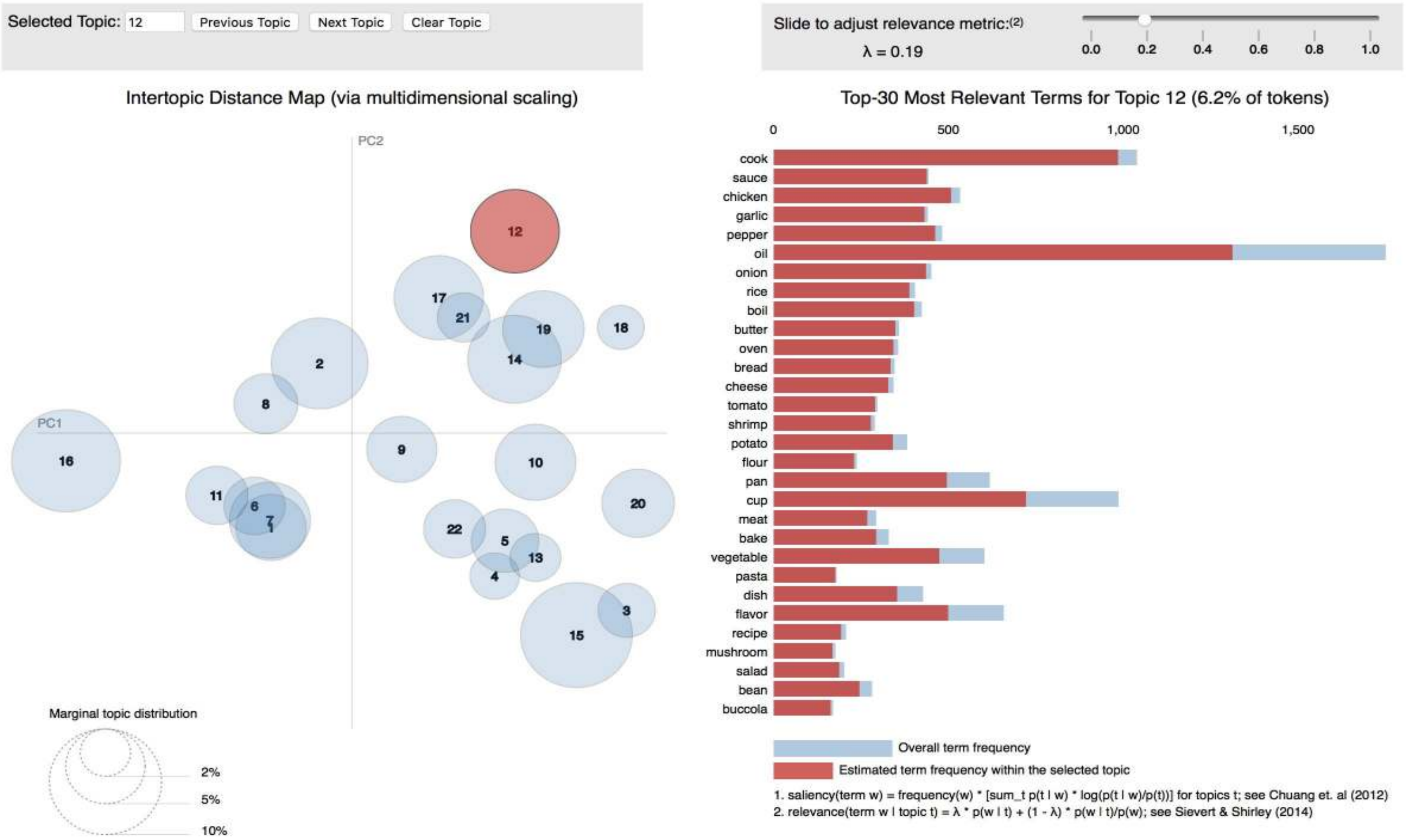
Ref.	**partial dentures** come in both plastic and metal versions . examine different types of **partial dentures** with information from a dentist in this free oral hygiene video .	
Text	**partial dentures** will help to prevent dentures . learn about **partial dentures** from a dentist in this free oral hygiene video .	Content F1 47
Actions RNN	do n't leave a home drug test . learn about **vacuum cleaners** with expert tips from a dentist in this free oral hygiene video .	Content F1 35
Actions	in order to make an nail art design , get expert tips and advice on housecleaning in this free video series that will teach you everything you need to know to make your own ceviche in this free video .	Content F1 25

Example



Ref.	stretching out your calves is a great way to alleviate stress and rejuvenate your muscles . learn a healthy leg stretch from a yoga instructor in this free yoga video .	
Text	stretching is a great way to **warm up your calves** . learn some calf raises from a professional **pilates** instructor in this free fitness video .	Content F1 47
Actions RNN	the yoga chair pose is a great way to strengthen the muscles in the upper back . learn about shoulder and deltoid exercises in this free **hatha yoga** video .	Content F1 35
Actions	learn the basics of **hatha yoga** with expert tips on headache relief in this free home improvement video .	Content F1 25

Topics in How-To Videos (LDA on Transcripts)



Use of Topics

- What if we take the teaser from the next neighbor video in topic space?
    - wearing a bra is almost universal in western countries , but did you ever wonder why ? learn about why women wear bras and what function they serve in this free women 's fashion video .
    - do n't wrinkle you suit right after ironing it ! learn how to hang a jacket while ironing a men 's suit in this free clothing care video from a wardrobe professional .

- This performs similarly to our rule-based baseline!
- Worse in content F1 than all S2S models.

Rouge-L	Content F1
31.8	17.9

Ongoing Work

- Treat context vector like visual feature - use for adaptation
  - General framework for adaptation of S2S models

- Multi-document summarization
  - Create captions for multiple videos together - this would be really useful
  - A bit slow to train (2000h …), but running now using multi-task encoders (two)
  - Need to think about evaluation some more (currently: ROUGE=52.1 vs 53.0)
  - Form of data augmentation?

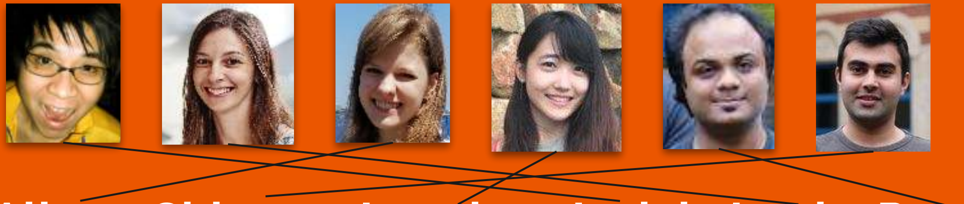
- Discriminative summarization
  - See three videos at the same time: two similar, one different
  - Explain (e.g. generate text) how one is different from the other(s)
  - Use ranking loss for discrimination

Summarization Conclusion

- It works! Kind of. Still looking at ...
  - Multi-document summarization
  - End-to-end summarization from speech
  - Multi-modal summarization with temporal structure and/ or object & scene features

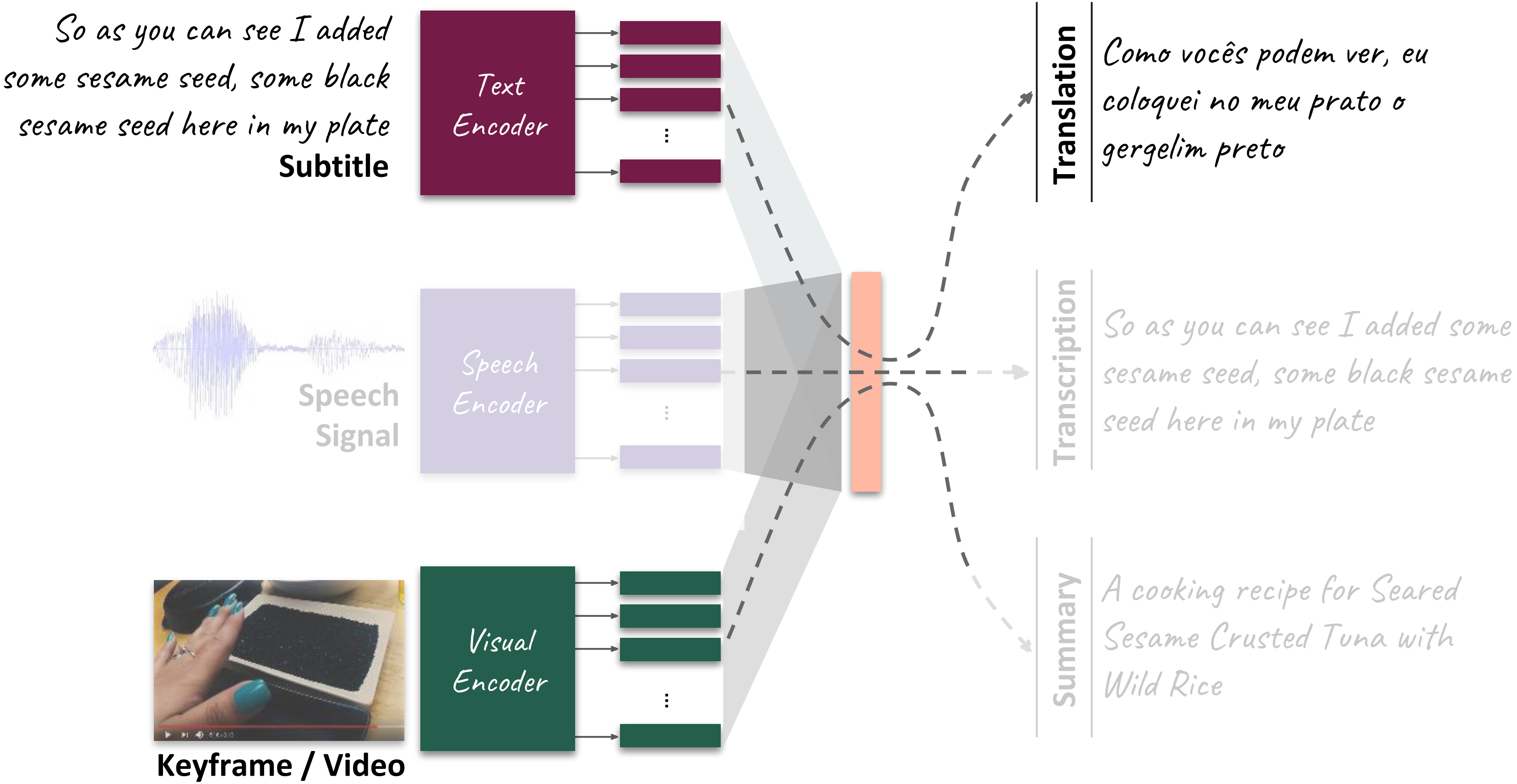
- *Text-generated descriptions* are generative, pretty detailed and often repeats certain key phrases.
- *Action-feature generated* text is boiler-plate but accurate, *Act-RNN text* is more diverse and more self-consistent.
- Need to tie in with representation learning and investigate portability

Region-specific Machine Translation



Alissa, Chiraag, Jasmine, Josiah, Lucia, Pranava

The big picture



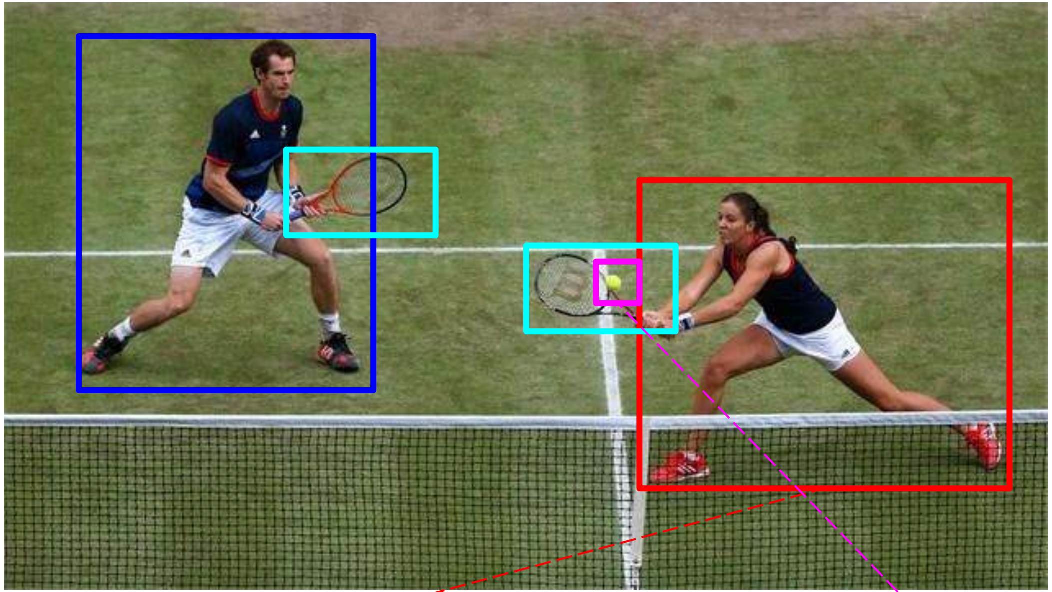
Q: Can region-specific multimodal MT improve translation quality?

Grounding Machine Translation

The player on the right has just hit the ball

~~O jogador~~ à direita acaba de acertar a bola

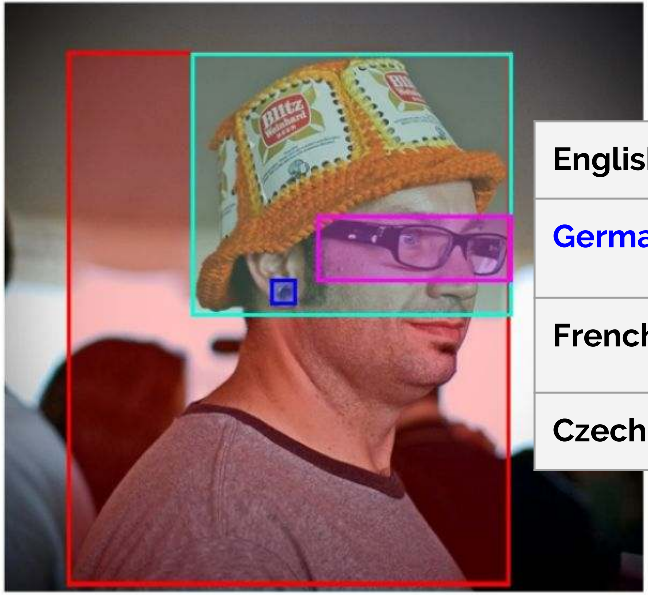
Grounding Machine Translation to Image Regions



The player on the right has just hit the ball

A jogadora à direita acaba de acertar a bola

Dataset: Multi30K + Flickr30k Entities



A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

English	A man in an orange hat staring at something.
German	Ein Mann mit einem orangefarbenen Hut, der etwas anstarrt.
French	Un homme avec un chapeau orange regardant quelque chose.
Czech	Muž v oranžovém klobouku na něco zírá.

30K (image, sentence) pairs per language

Region-specific Grounded MT

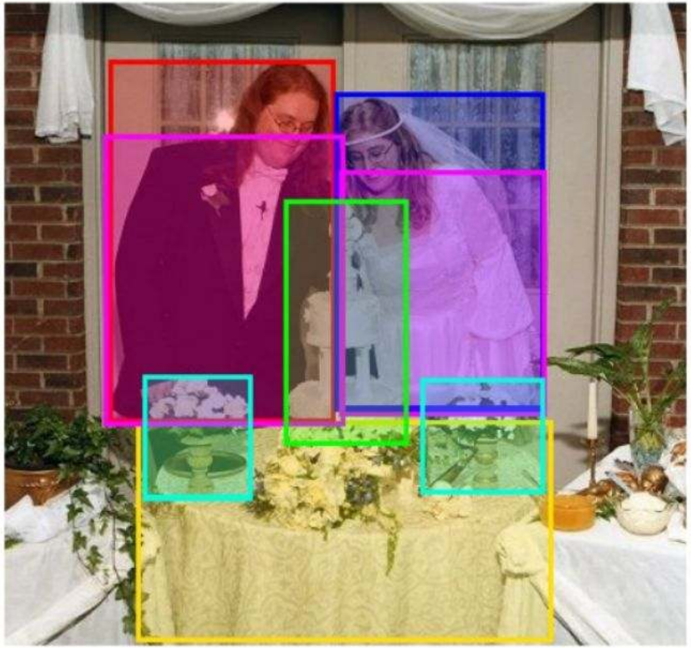
Step 1	Step 2	Step 3	Step 4
Obtain **image regions**	Represent image regions	Devise algorithms to **learn associations** between **visual** and **text** information	Use grounded representation to **guide MT**

Step 1: Obtaining Image Regions

Step 1	Step 2	Step 3	Step 4
Obtain **image regions**	Represent image regions	Devise algorithms to **learn associations** between **visual** and **text** information	Use grounded representation to **guide MT**

Step 1: Obtaining Image Regions

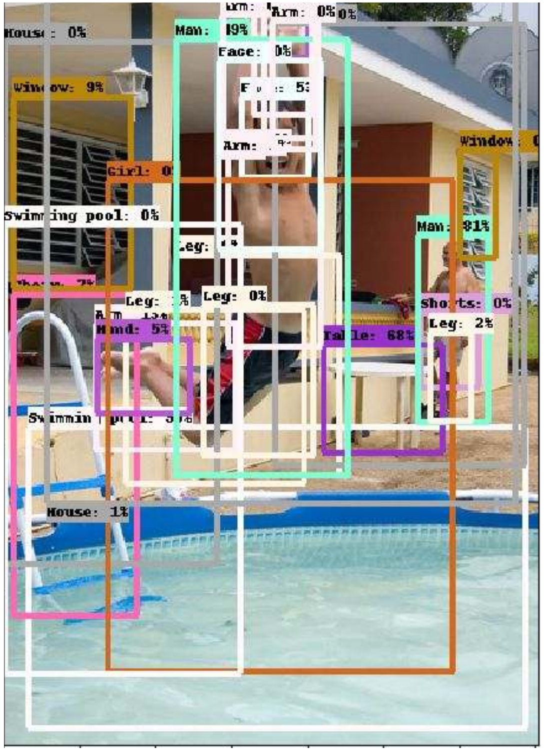
- Oracle regions (Flickr30k Entities)

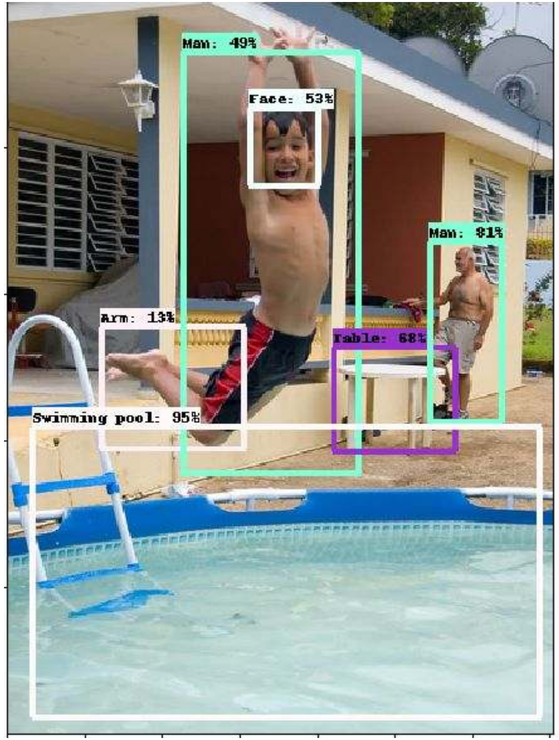


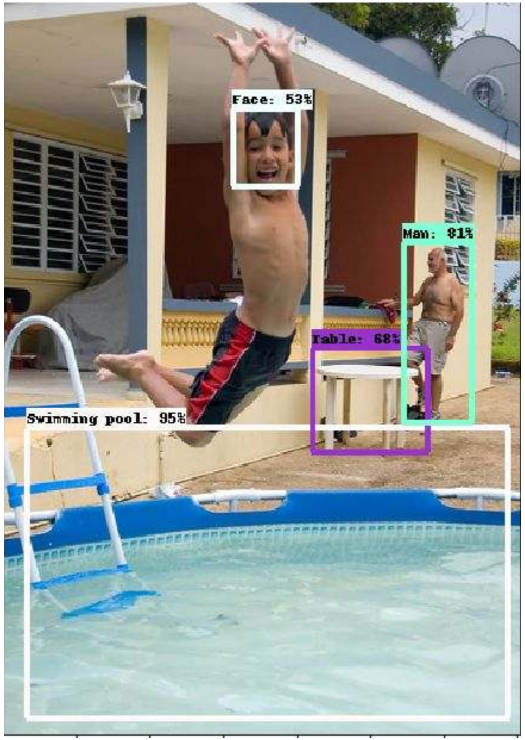
A **bride** and **groom** are standing in front of **their wedding cake** at **their reception**.
A **bride** and **groom** smile as **they** view **their wedding cake** at **a reception**.

Step 1: Obtaining Image Regions

- Output of a detector (545 categories -- Open Images)





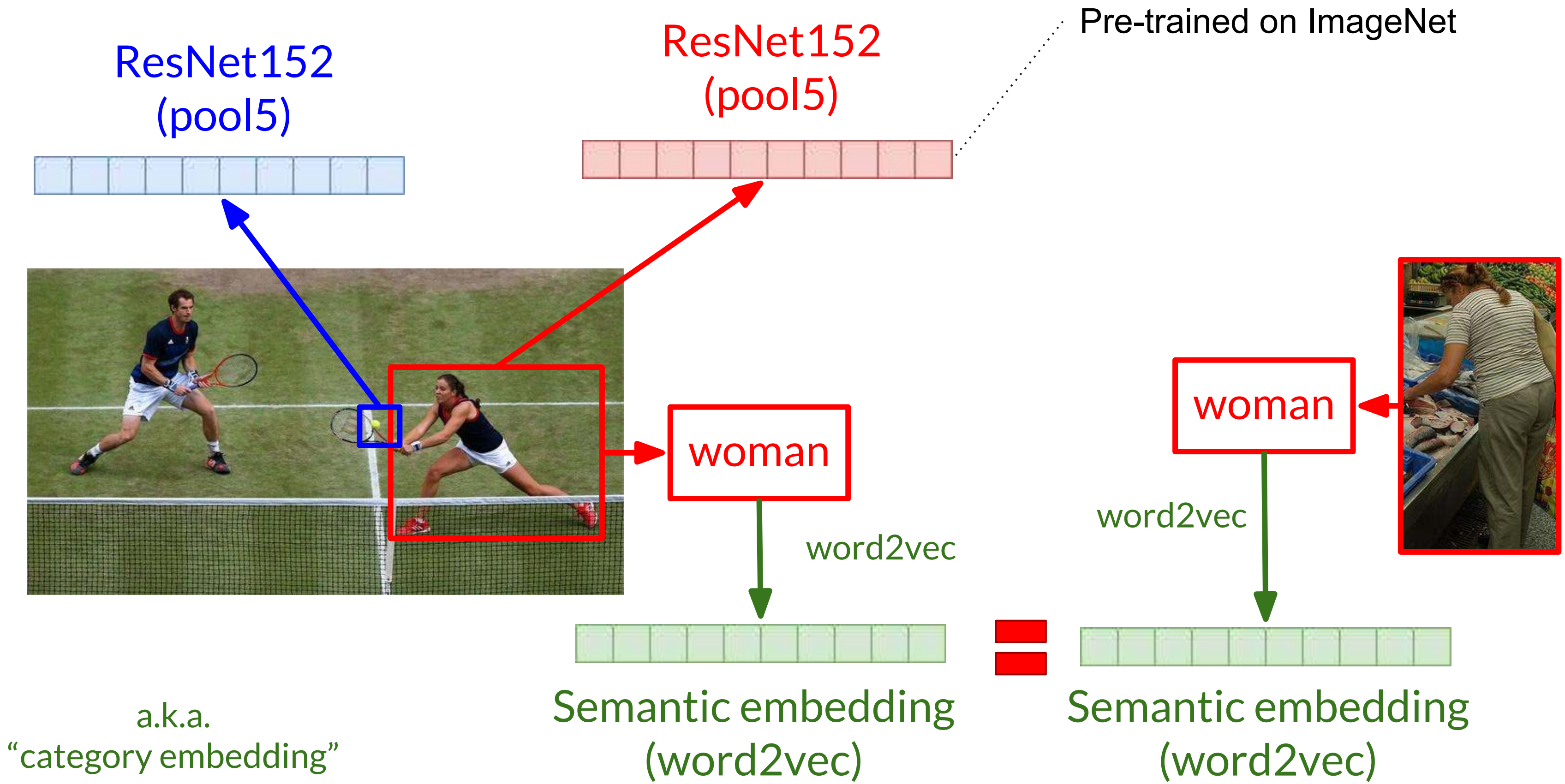


Step 1: Obtaining Image Regions

Step 2: Representing Image Regions

Step 1	Step 2	Step 3	Step 4
Obtain **image regions**	Represent image regions	Devise algorithms to **learn associations** between **visual** and **text** information	Use grounded representation to **guide MT**

Step 2: Representing Image Regions



Grounding Regions and MT

Implicit

Alignment and MT jointly

Explicit

Alignment, then MT

Grounding Regions and MT

Implicit

Alignment and MT jointly

Explicit

Alignment, then MT

Steps 3 & 4: Joint Alignment and MT

Step 1

Step 2

Step 3

Step 4

Obtain **image regions**

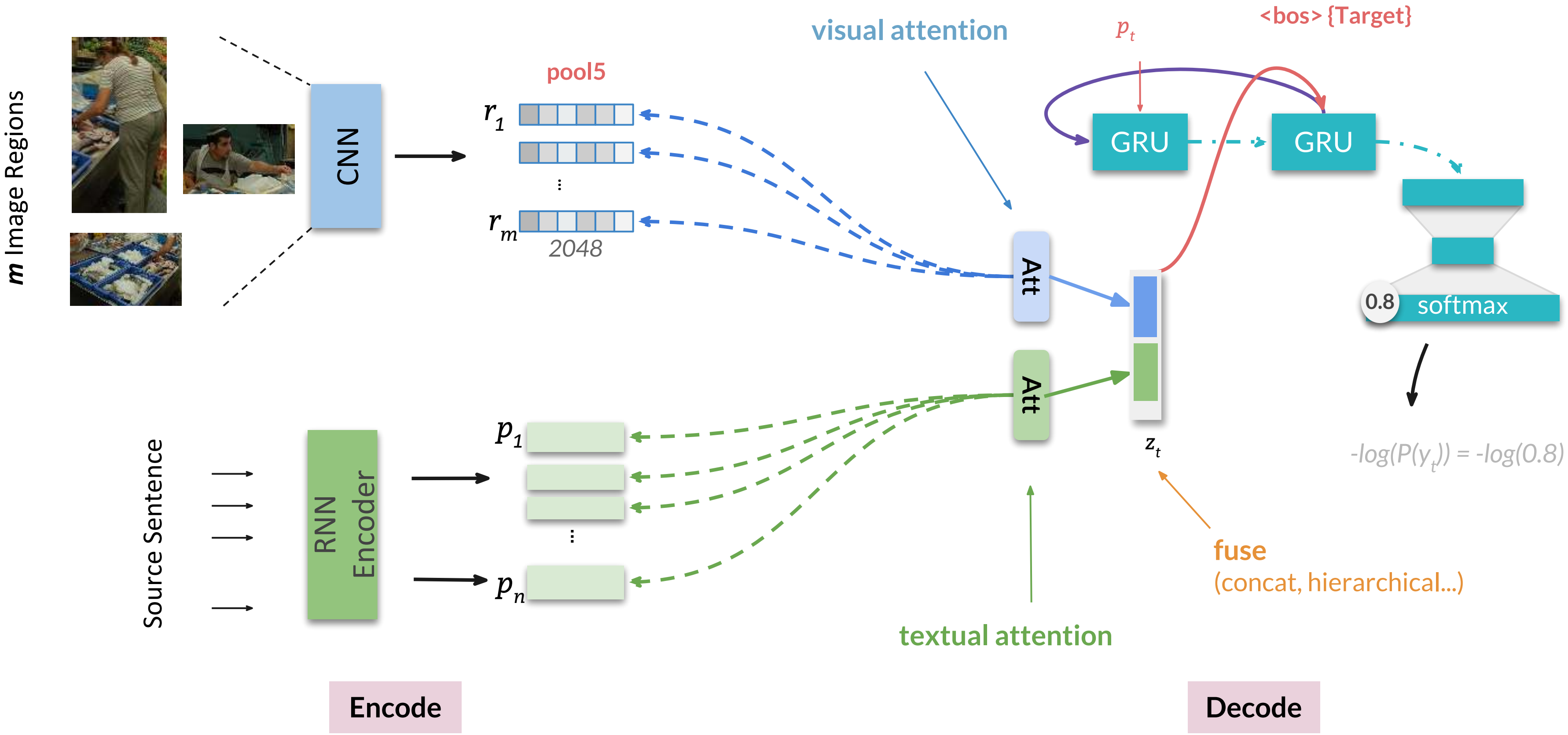
Represent image regions

Devise algorithms to **learn associations** between **visual** and **text** information

Use grounded representation to **guide MT**

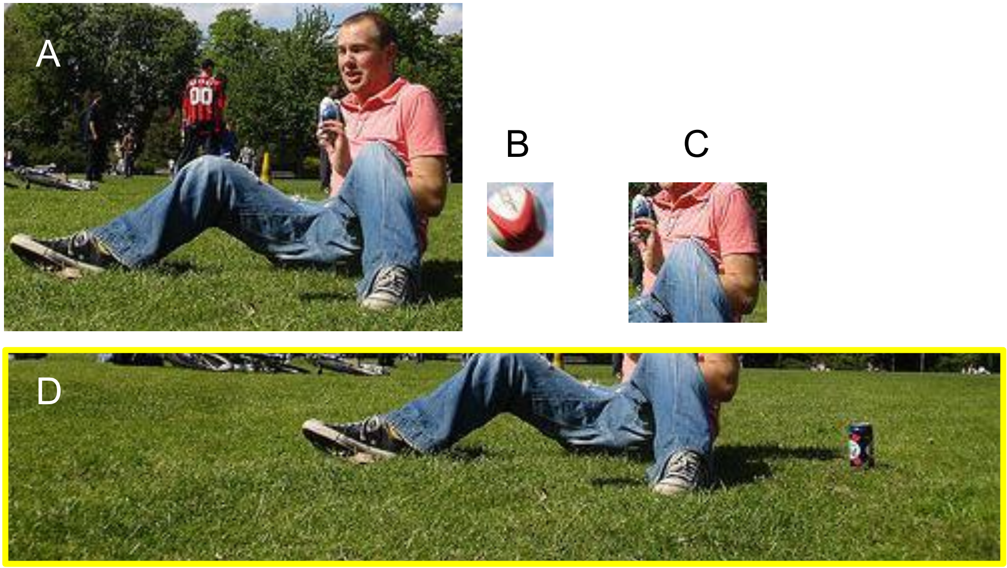
End-to-end joint alignment and MT

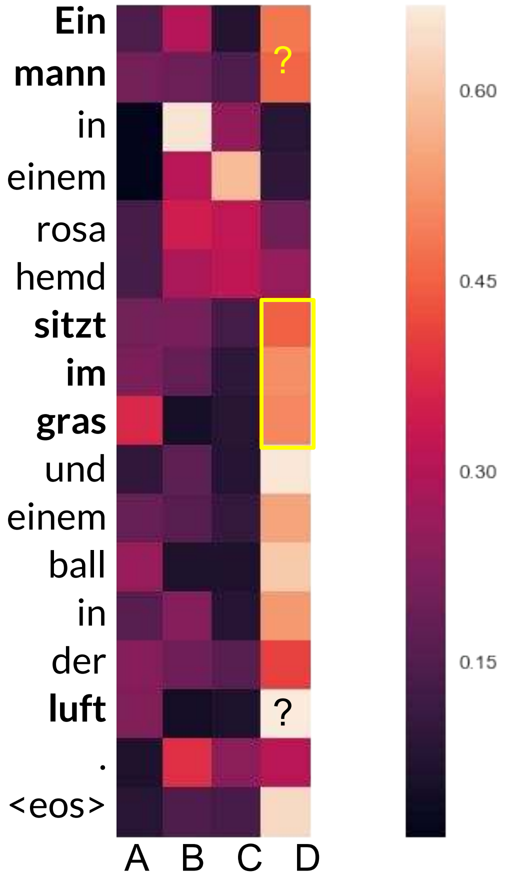
Standard Decoder Attention



Fusion: concat

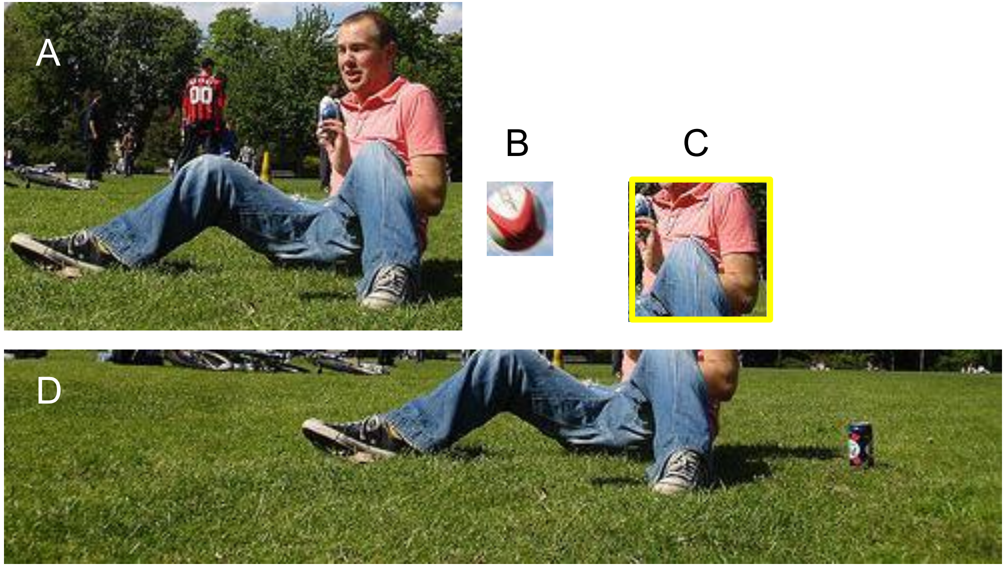
S: A man in a pink shirt is sitting in the grass and a ball is in the air.

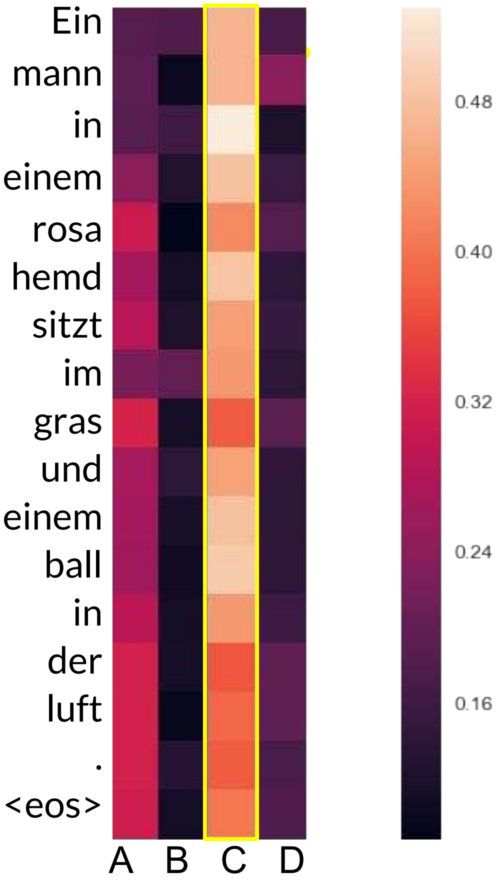




Fusion: hierarchical

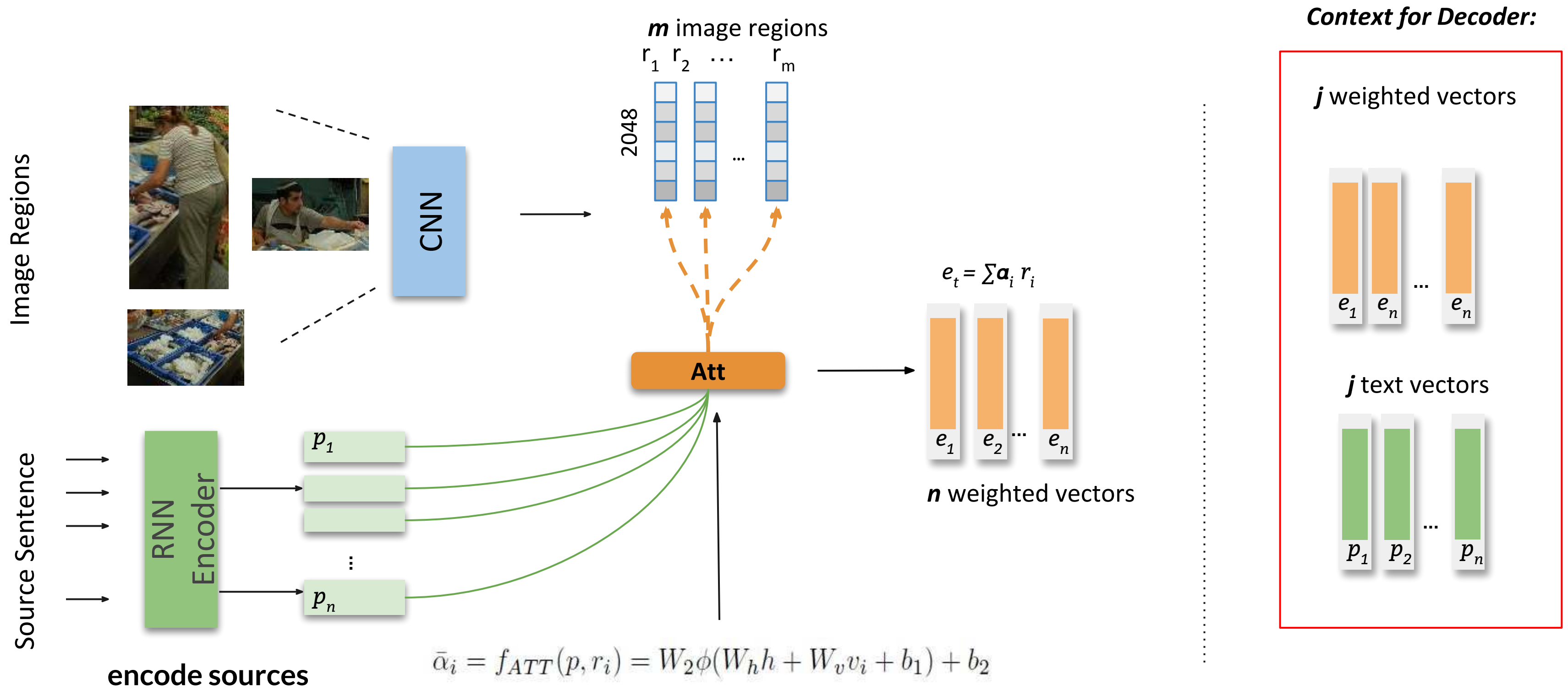
S: A man in a pink shirt is sitting in the grass and a ball is in the air.





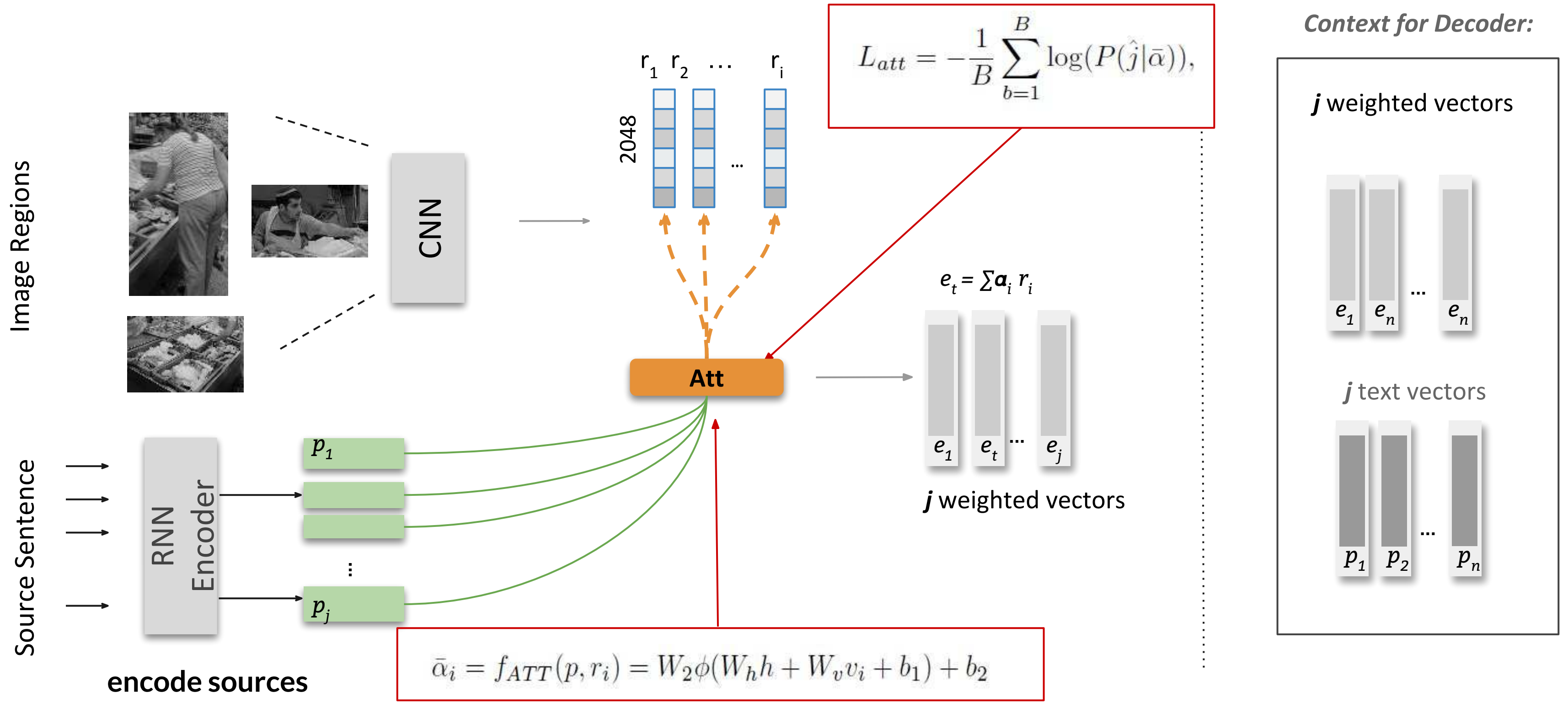
Encoder Attention Model

Idea: Ground the images in the ***source***



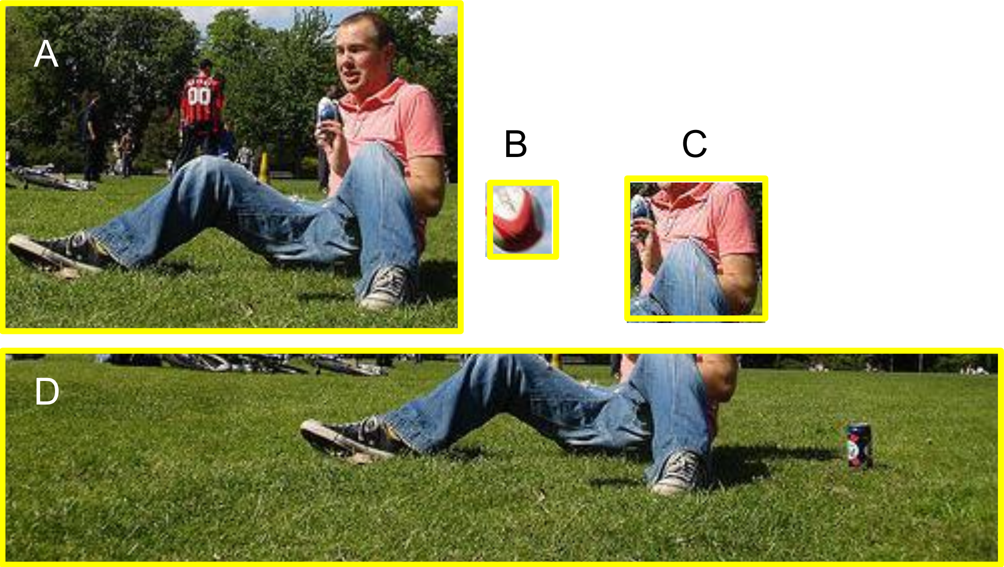
Supervised Encoder Attention Model

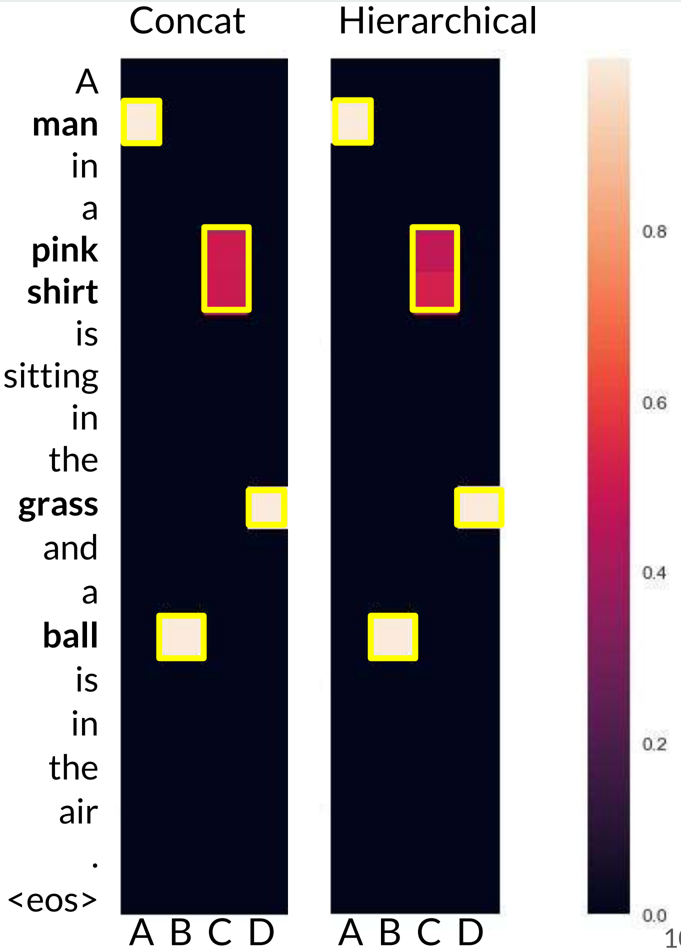
Given gold word-region alignments, add an auxiliary loss to main MT loss



Fusion: concat, hierarchical

Alignments are much clearer! Even though metrics don't improve...





Grounding Regions and MT

Implicit

Alignment and MT jointly

Explicit

Alignment, then MT

Step 3: Explicit Alignment

Step 1 | Step 2 | **Step 3** | Step 4

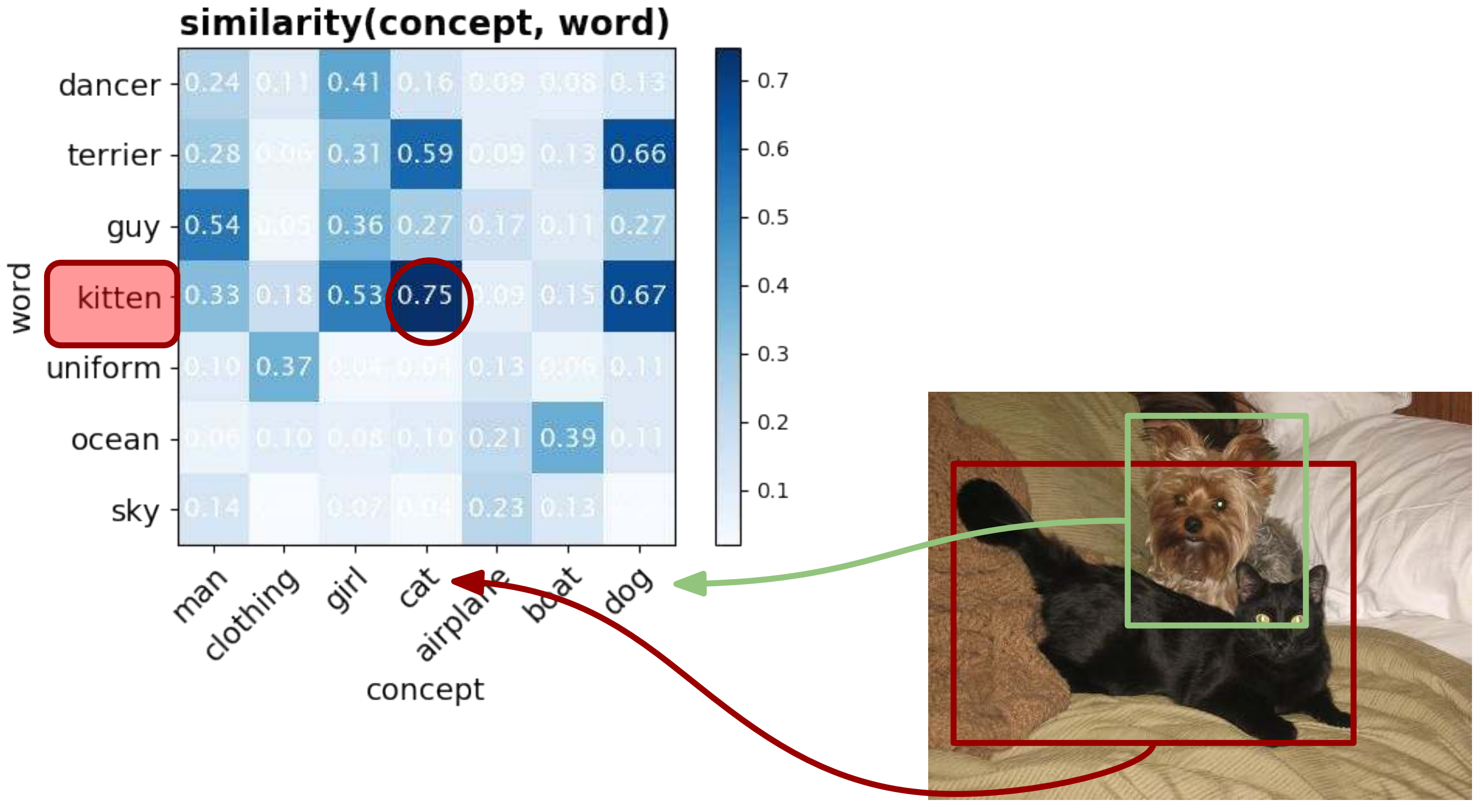
Obtain **image regions**

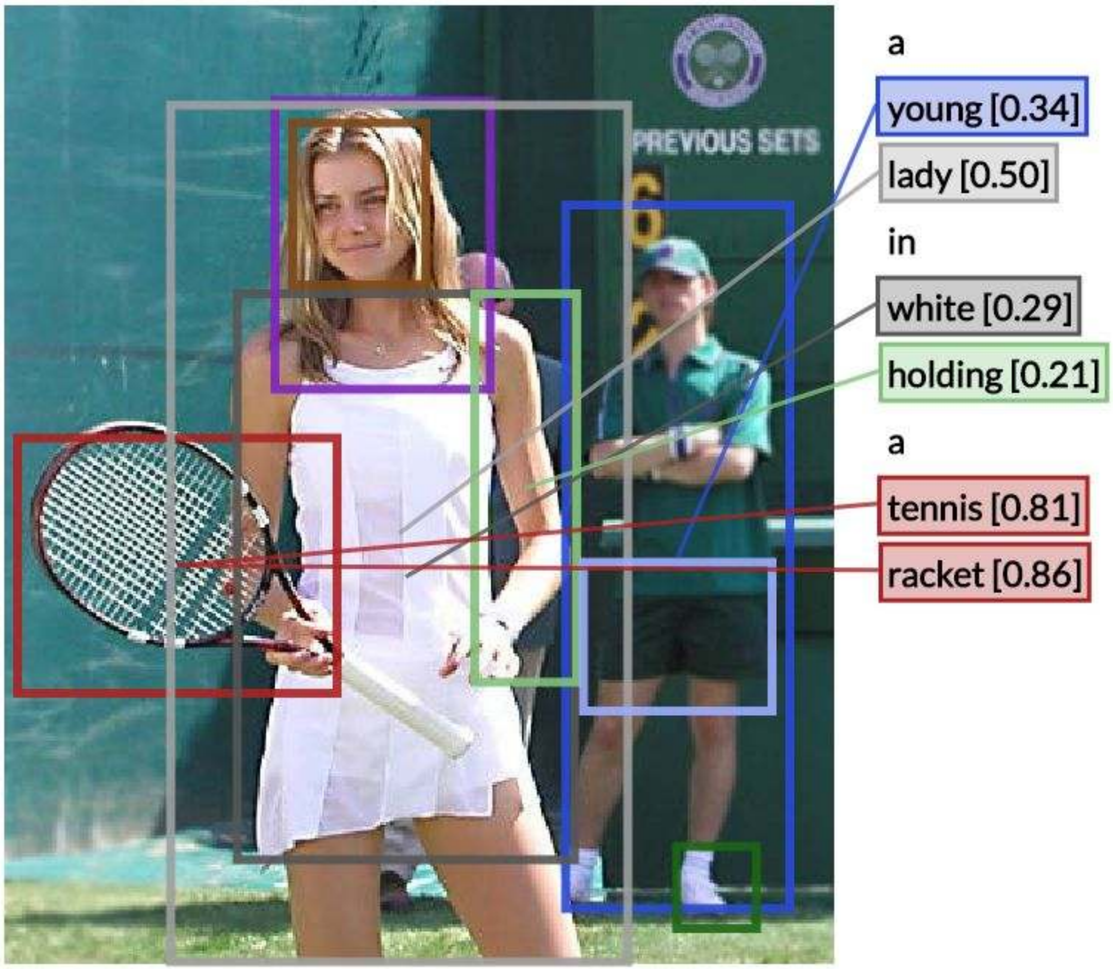
Represent image regions

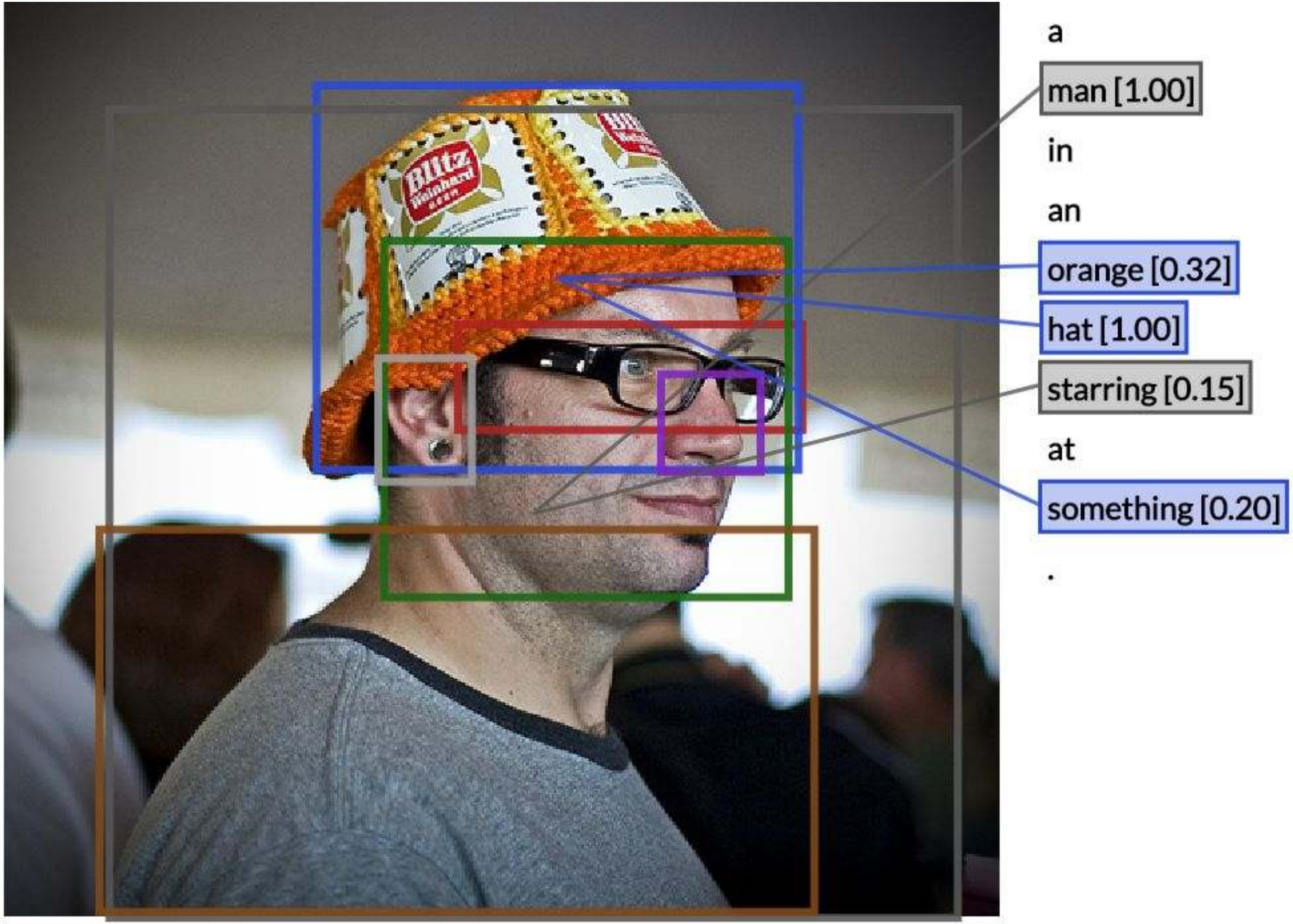
Devise algorithms to **learn associations** between **visual** and **text** information

Use grounded representation to **guide MT**

Alignments Learnt Explicitly




PREVIOUS SETS
a
young [0.34]
lady [0.50]
in
white [0.29]
holding [0.21]
a
tennis [0.81]
racket [0.86]


a
man [1.00]
in
an
orange [0.32]
hat [1.00]
starring [0.15]
at
something [0.20]
.

Step 4: Using Explicitly Learnt Alignments for MT

Step 1

Step 2

Step 3

Step 4

Obtain **image regions**

Represent image regions

Devise algorithms to **learn associations** between **visual** and **text** information

Use grounded representation to **guide MT**

Idea

- Further specify source words with respective image region visual info

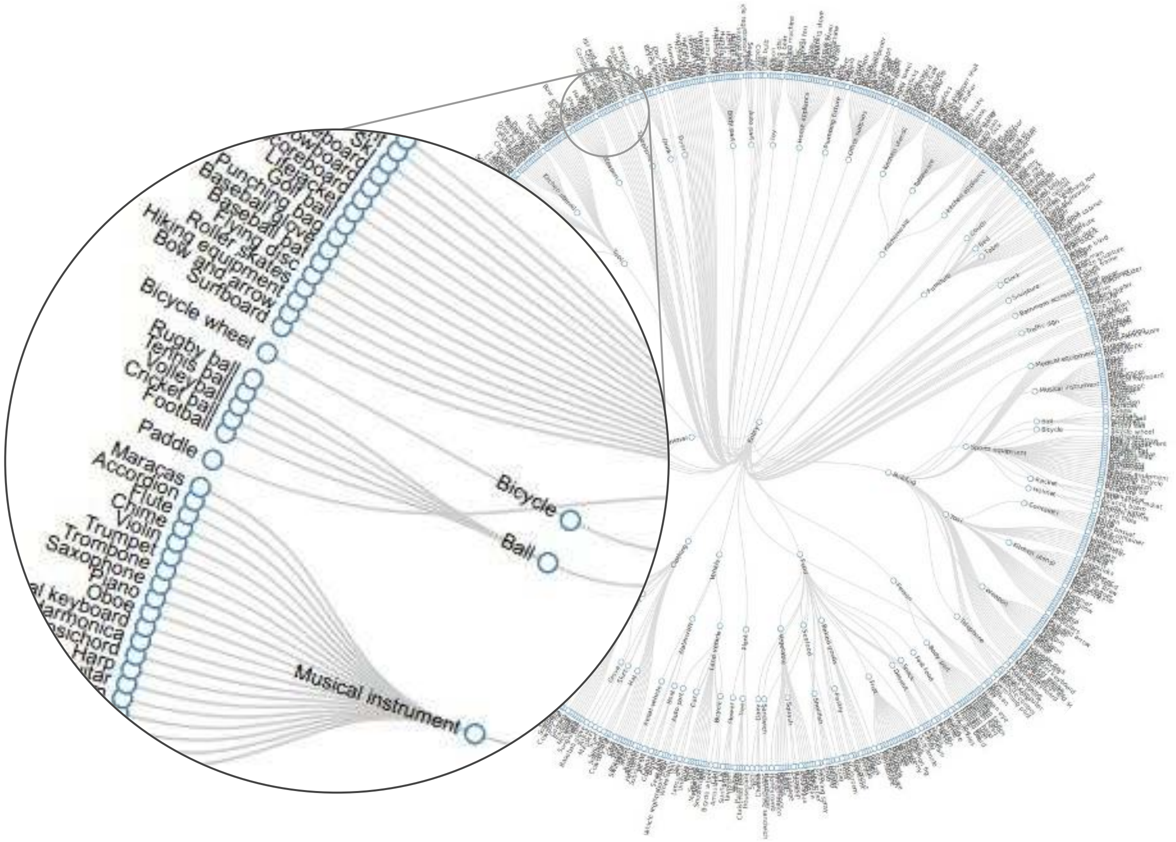
Category:
clothing

The man in yellow pants is raising his arms

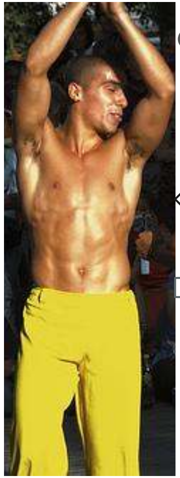
Categories from Image Regions

- Oracle (8)
  - People
  - Clothing
  - Scene
  - Animals
  - Vehicles
  - Instruments
  - Body parts
  - Other

- Predicted (545)



Categories from Image Regions



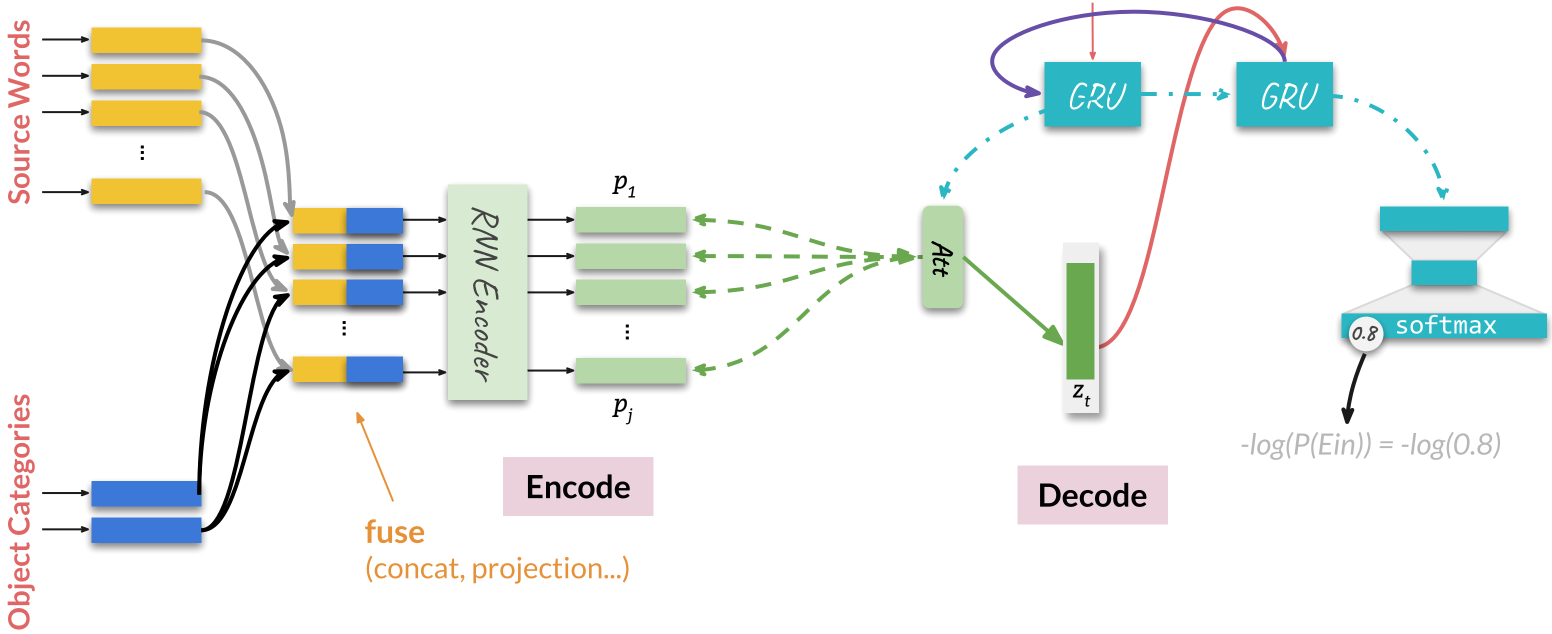
- Take category of image region to be the category of head noun of corresponding text phrase

Sentence:	The	man	in	yellow	pants	is	raising	his	arms
Categories:		people			clothing				body part

- For any other word, set category to "empty"

Sentence:	The	man	in	yellow	pants	is	raising	his	arms
Categories:	empty	people	empty	empty	clothing	empty	empty	empty	body part

Categories from Image Regions



Examples (En-De)

Gold	Baseline	With Categories
five people in **winter jackets** and helmets stand in the snow .	five people in **winter jackets** and helmets stand in the snow.	five people in **winter clothes** and with their helmets standing in the snow.
a man is standing by a group of **video games** in a bar .	a man is standing next to a group of **students** in a bar.	a man is standing in a bar next to a group of **video games.**

Noun Drop

- "Drop" head nouns in source sentences, but keep category information

The man sat in the rain .

The <DEL> sat in the <DEL> .

- In the absence of words, can visual information can guide model to generate better translations?

Sentence Drop

- In training, “drop” 20% of source sentences, but keep category information

The man sat in the rain .

<DEL> <DEL> <DEL> <DEL> <DEL> <DEL> <DEL>

- In the absence of sentences, can visual information guide model to generate better translations?

Sentence Drop Examples (En-De)

Gold	Baseline	With Categories
a group of Asian boys is waiting for meat to **be grilled**. a **boston terrier** is running on lush green grass in front of a white fence.	a group of Asian boys is waiting for meat to **be grilled**. a **boston cook** runs in front of a white fence on green grass and runs over green grass.	a group of Asian boys is waiting for meat to **be photographed**. a **boston shepherd dog** runs in front of a white fence on a green meadow.

Drop Results

Noun Drop

	Features	en-de	en-fr	en-cs
Text-only	-	31.28	49.81	25.77
Explicit alignment	Cat. embeddings	30.31	49.65	25.12

Sentence Drop

	Features	en-de	en-fr	en-cs
Text-only	-	35.35	57.84	26.71
Explicit alignment	Cat. embeddings	**36.29**	**58.64**	**30.14**

General results

Results (test2016)

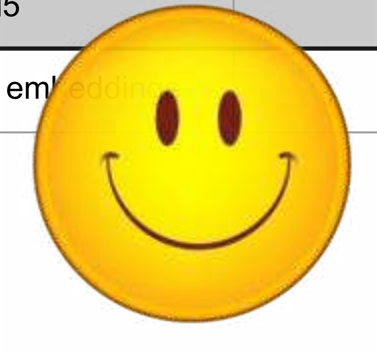
METEOR	Features	en-de	en-fr	en-cs
Text-only (no image)	-	57.35	75.16	29.35
Decoder init. (full image)	Pool5	56.97	74.82	29.04
Attention over regions (decoder)	Pool5	56.77	74.74	28.86
Attention over regions (decoder)	Cat. em[illegible]	56.48	73.65	28.42
Encoder attention over regions	Pool5	57.30	75.36	**30.48**
Encoder attention over regions	Cat. embeddings	57.29	**75.97**	**30.78**
Supervised attention over regions	Pool5	56.34	75.07	**30.19**
Supervised attention over regions	Cat. embeddings	56.64	75.56	**30.39**
Explicit alignment - projection	Cat. embeddings	57.39	75.25	**30.64**
Explicit alignment - concatenation	Cat. embeddings	57.44	75.47	**30.77**

Results - lexical ambiguity (test2016)

ACCURACY	Features	en-de	en-fr	en-cs
Text-only (no image)	-	37.00	53.62	10.44
Decoder init. (full image)	Pool5	37.53	53.31	**13.65**
Attention over regions (decoder)	Pool5	37.82	53.62	10.84
Attention over regions (decoder)	Cat. embeddings	37.76	52.31	**14.46**
Encoder attention over regions	Pool5	**38.06**	**55.16**	**12.45**
Encoder attention over regions	Cat. embeddings	37.94	54.24	**14.06**
Supervised attention over regions	Pool5	37.47	53.39	**13.25**
Supervised attention over regions	Cat. embeddings	36.89	54.08	**14.06**
Explicit alignment - projection	Cat. embeddings	**38.41**	54.08	**13.65**
Explicit alignment - concatenation	Cat. embeddings	38.06	53.78	**12.85**

Results - lexical ambiguity accuracy (test2018)

ACCURACY	Features	en-de	en-fr	en-cs
Text-only (no image)	-	44.14	43.06	-
Decoder init. (full image)	Pool5	46.85	43.06	-
Attention over regions (decoder)	Cat. embedding	**48.65**	**45.83**	-



Results - human eval

- Proportion of times each system is better (meaning preservation)

	Features	en-de		en-fr		en-cs	
Text-only (no image)	-	22%		32%		20%	
Multimodal	Pool5	[illegible]	= 78%	37% +	= 68%	34% +	= 80%
	Cat. embed	[illegible]		32%		46%	

- Text-only system is more fluent but has less correct content words

Conclusions and Future Work

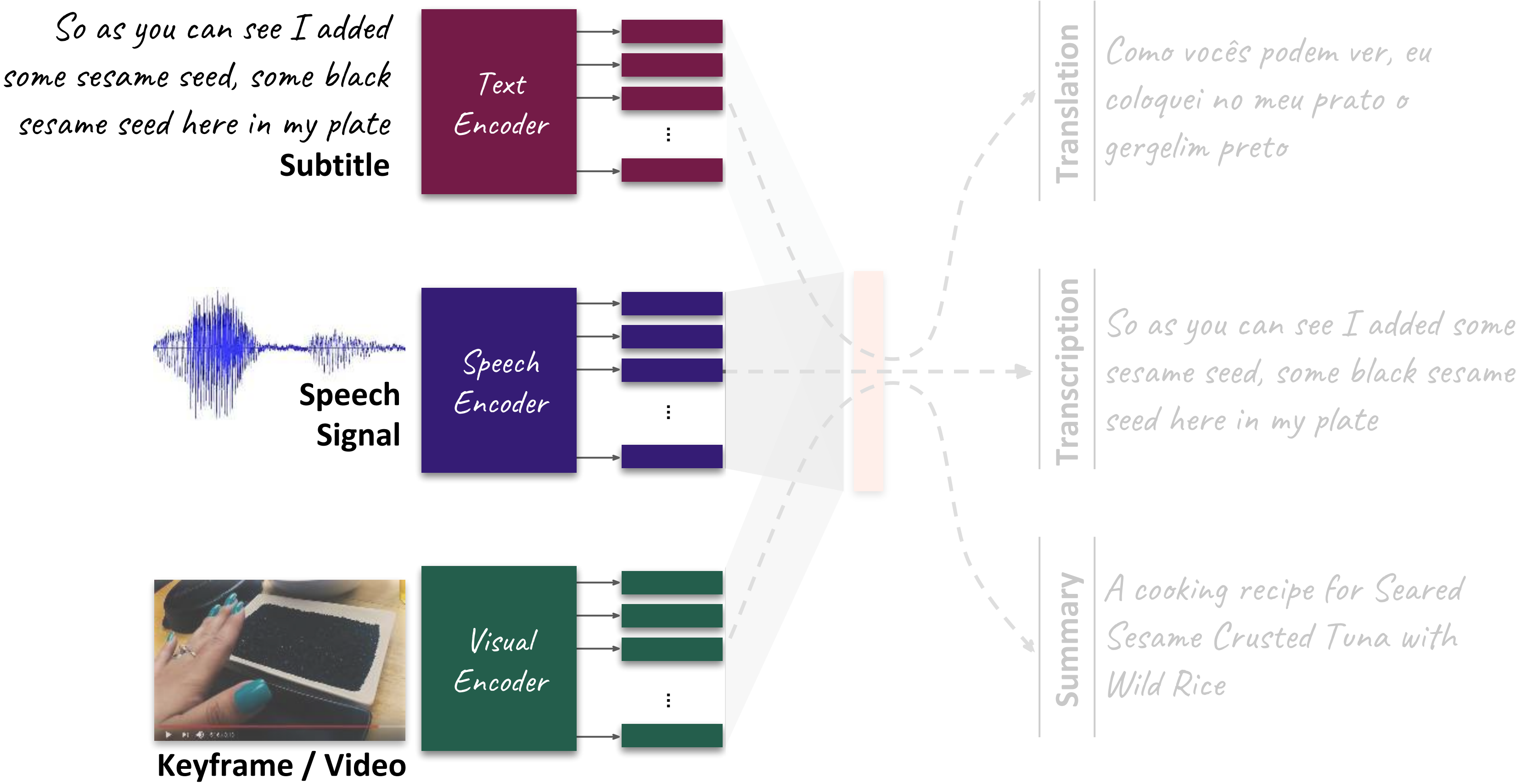
- **Text-only** vs **region-specific**
  - Region-specific always better
- **Oracle** vs **predicted** regions and alignment
  - Predictions do not degrade performance substantially
- Representations: **pool5** vs **category embeddings**
  - Similar but category embeddings more interpretable
- **Meteor/BLEU** are not indicative of performance variations
  - Lexical ambiguity evaluation: more indicative but only subset of words
  - Human evaluation: much more telling
- **Future**: more human **eval**, better use of explicit & implicit **alignments**

Multiview Learning



Nils, Pranava, Shruti

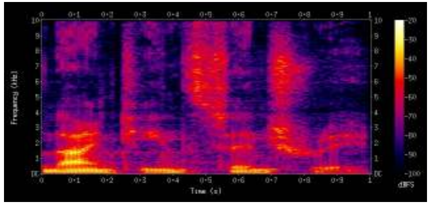
The big picture



A look at our Dataset

How-to 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)

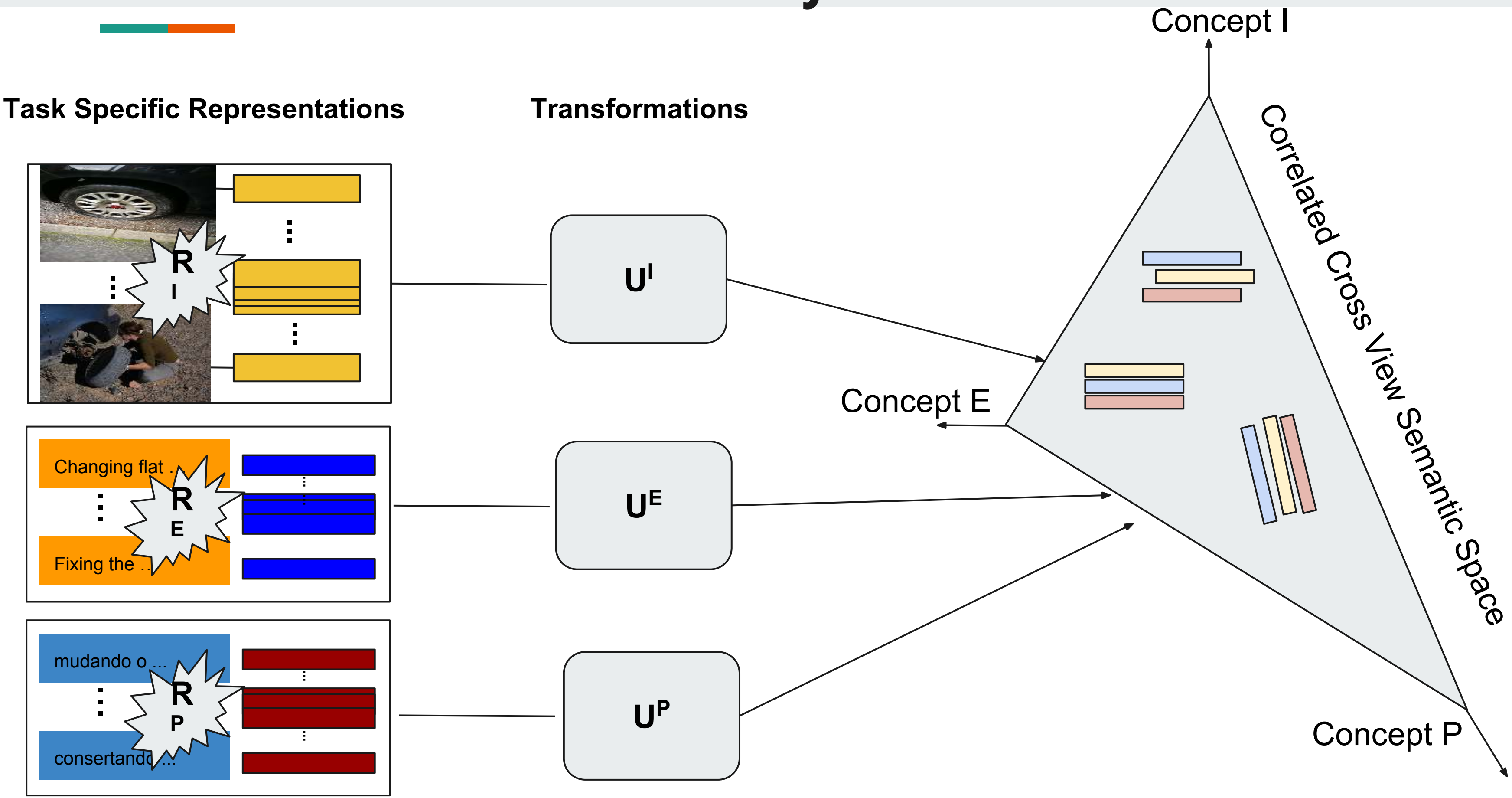
Changing flat tires doesn't have to be done with car jacks. Learn how to use an automotive hydraulic lift...

Q: What could explicit representation learning give us?

Learning from Multiple Views

- Each is different but all views share similar information
- Visual, Auditory and Language views are aligned
- Views in the same modality v/s Views in multiple modalities
- Unit level representations v/s Sequence Level Representations

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

Hotelling, 1936; Wang et al., 2016

CCA in a Nutshell

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

View 1 View 2

"...mething."

Elegant closed form solution

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

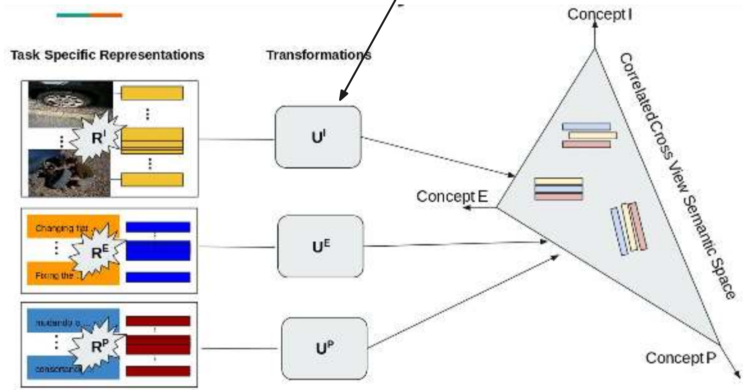
to maximize $\text{correlation}(\mathbf{u}^T f_\theta(X), \mathbf{v}^T g_\phi(Y))$

Hotelling, 1936; Wang et al., 2016

CCA: Extentions

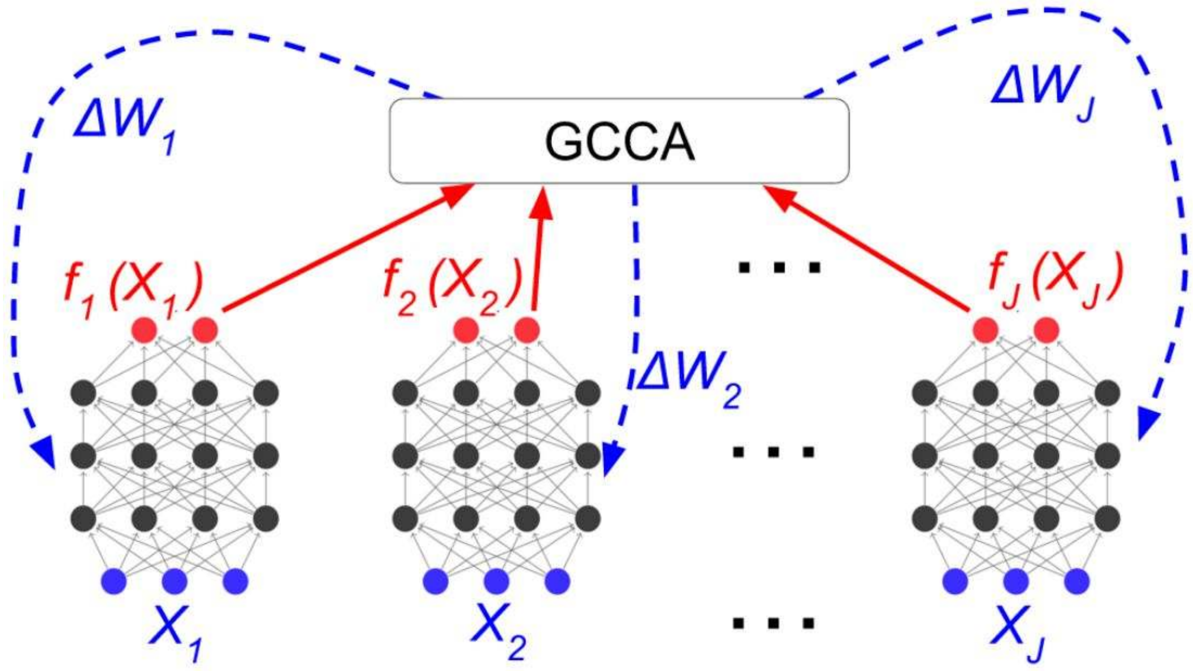
- Extending from two views to multiple views

$$\underset{U_j \in \mathbb{R}^{d_j \times r}, G \in \mathbb{R}^{r \times N}}{\text{minimize}} \sum_{j=1}^{J} \|G - U_j^\top X_j\|_F^2$$



CCA: Extensions

- Deep Generalized CCA: At the bleeding edge!

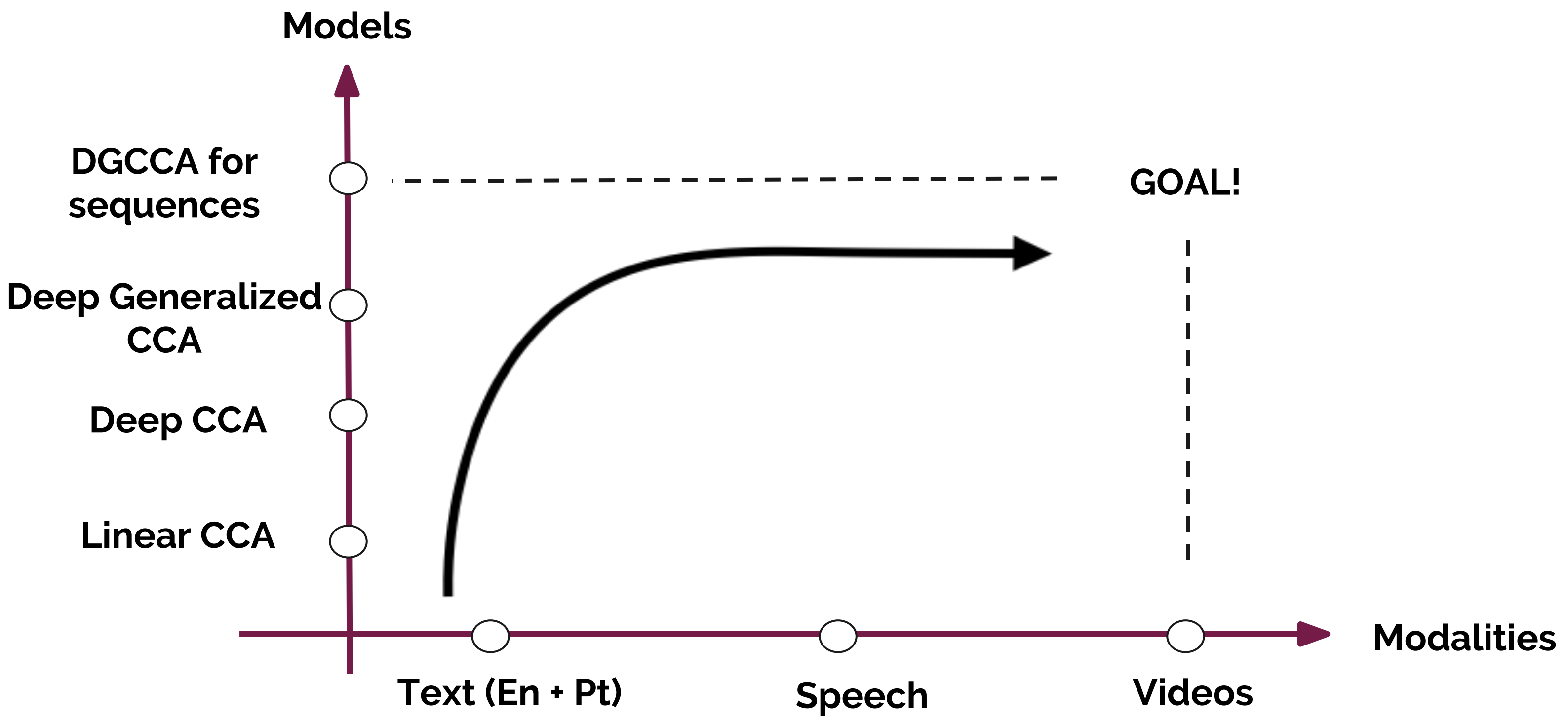


Benton et al., 2017

Salient Properties

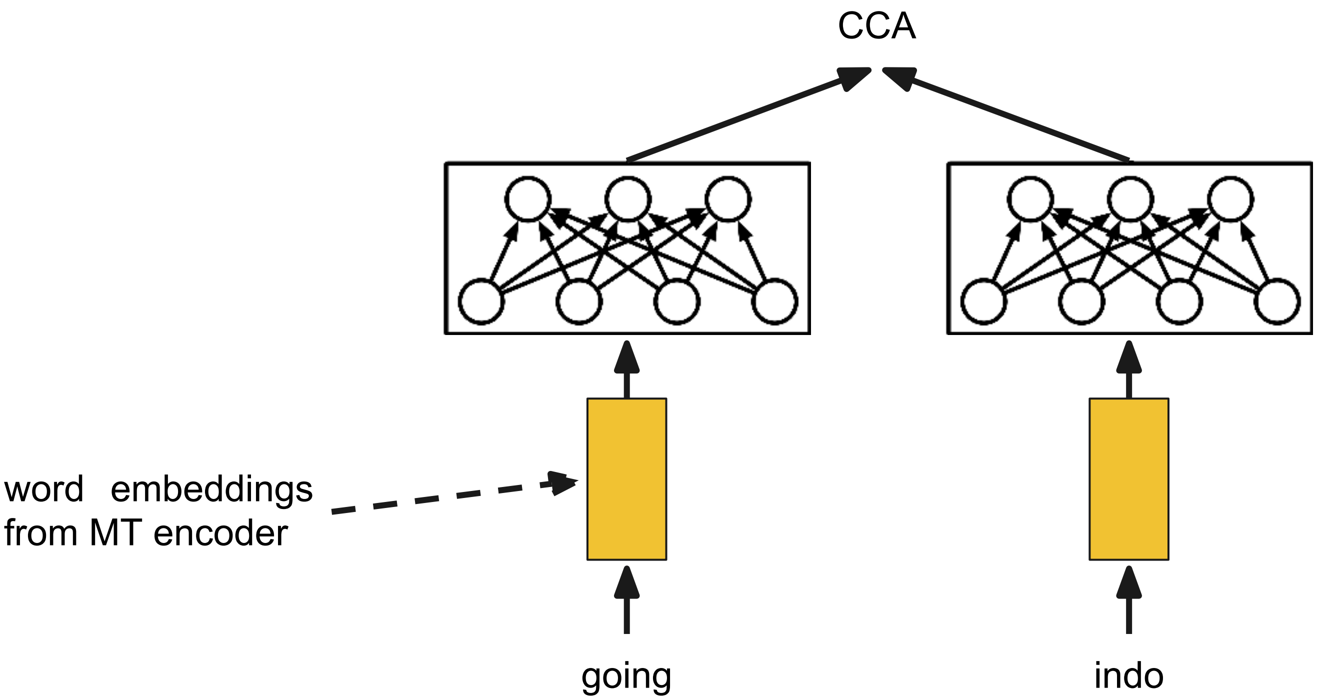
- (DG)CCA helps us obtain maximally correlated information that is consistent with each view
- Gives us a handle on the amount of variance shared
- Grounds information consistent with other view(s)
- It also helps in denoising and maximizing mutually relevant information

Our Goal



Models
DGCCA for sequences
Deep Generalized CCA
Deep CCA
Linear CCA
GOAL!
Modalities
Text (En + Pt)
Speech
Videos

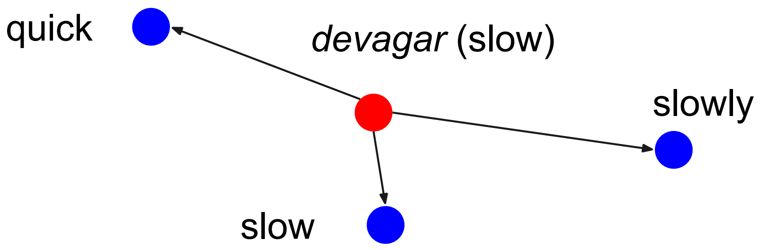
Text Representations - Words



Text Representations - Words

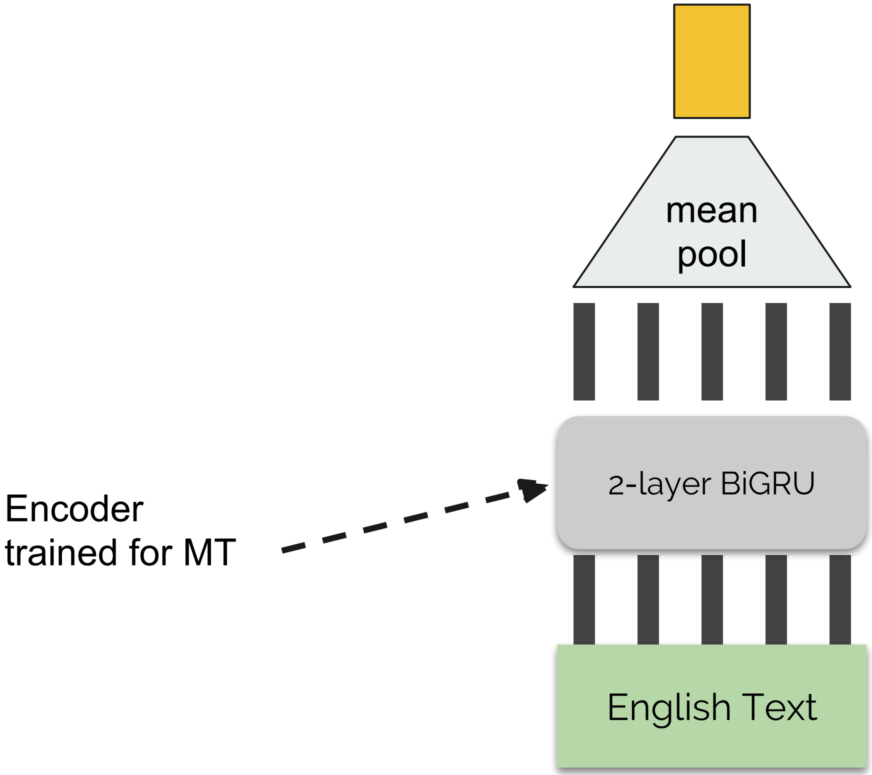
Recall@10

Frequency-based retrieval	Linear CCA	Deep CCA
0.1%	20.7%	13.8%

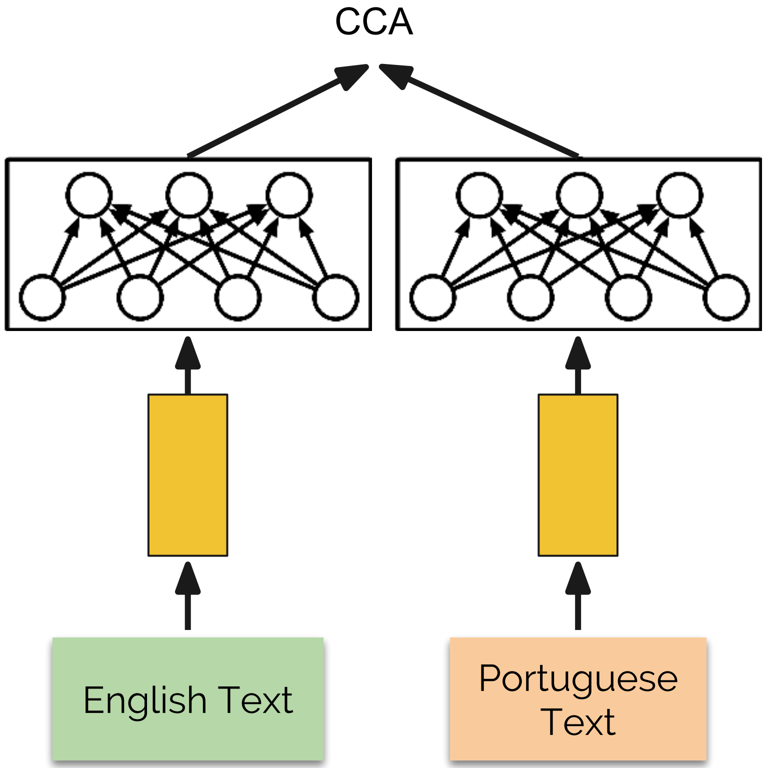


Nearest neighbors before CCA	**After CCA**
os (the)	os (the)
1. trinkets 2. sells 3. wins	1. the 2. your 3. their
devagar (slow)	devagar (slow)
1. hotel 2. tetra 3. dispute	1. tightly 2. slowly 3. totally

Text Representations - Sentences



Text Representations - Sentences



Recall@10 over test set

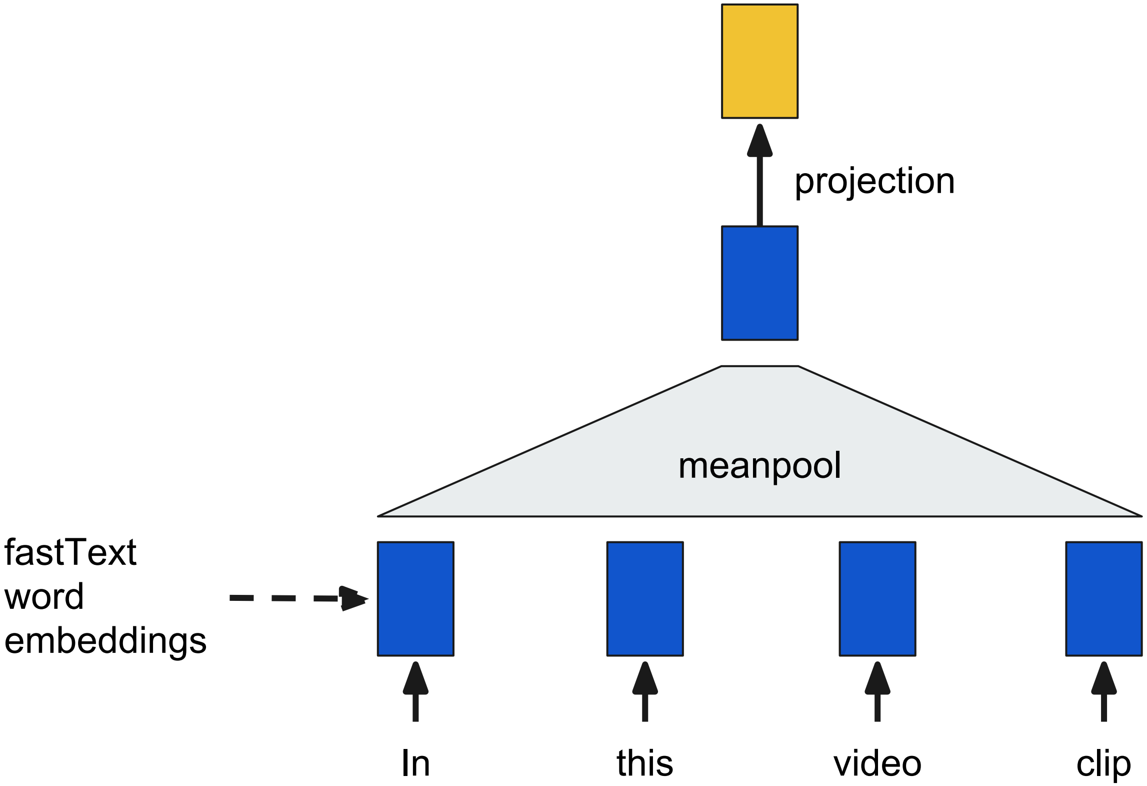
Linear CCA

81.4%

Deep CCA

95.0%

Text Representations - Sentences



Recall@10 over test set

Linear CCA

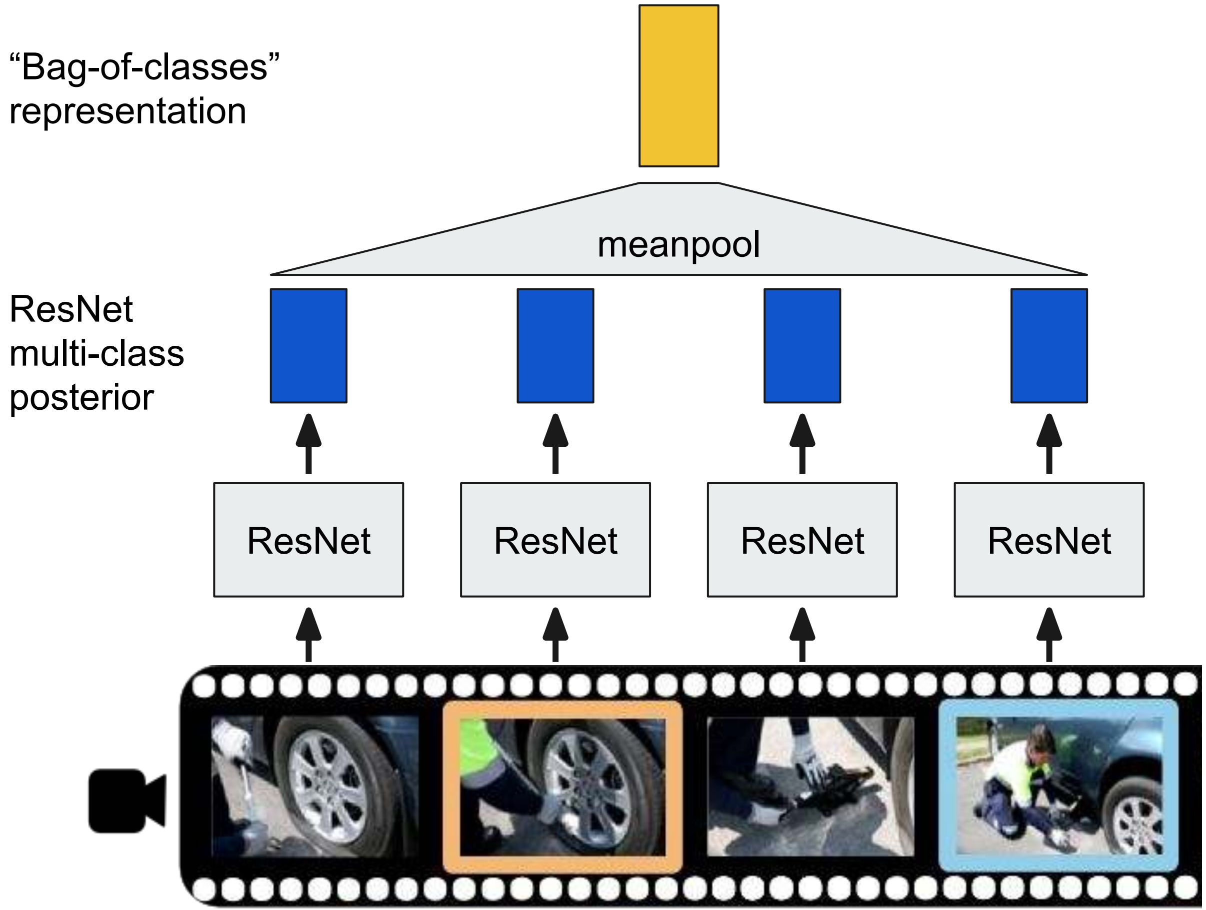
97.0%

Deep CCA

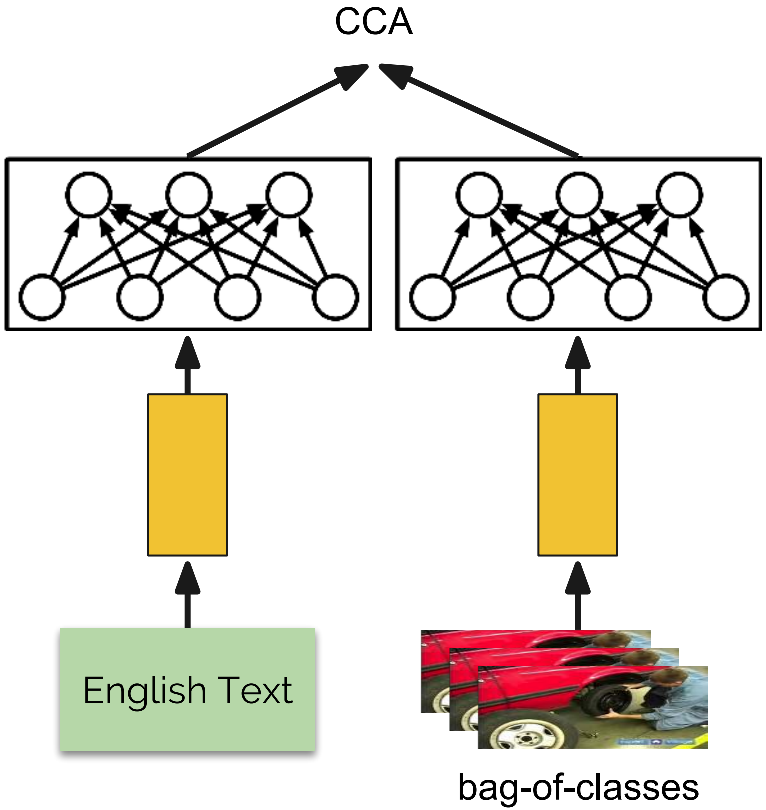
96.2%

Arora et al., 2017.

Video Representations



Text and Video Representations - Sentences



Recall@10
over test set

Linear CCA

0.8%

Deep CCA

1.6%

Text Representations - Summary

Recall@10	Portuguese Words	Portuguese Sentences (MT)	Portuguese Sentences (FT)	
English Words	21.2	-	-	-
English Sentences (MT)	-	95.0	-	1.6
English Sentences (FT)	-	-	97.0	-

Text Representations - Summary

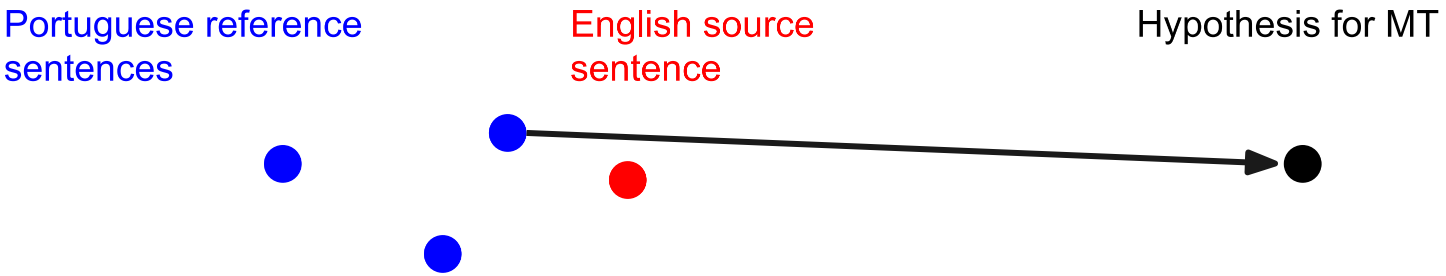
Recall@10	Portuguese Words	Portuguese Sentences (MT)	Portuguese Sentences (FT)	
English Words	21.2	-	-	-
English Sentences (MT)	-	95.0	-	1.6
English Sentences (FT)	-	-	97.0	-

Text Representations - Summary

Recall@10	Portuguese Words	Portuguese Sentences	Portuguese Sentences (FT)	
English Words	21.2	-	-	-
English Sentences (MT)	-	95.0	-	1.6
English Sentences (FT)	-	-	97.0	-

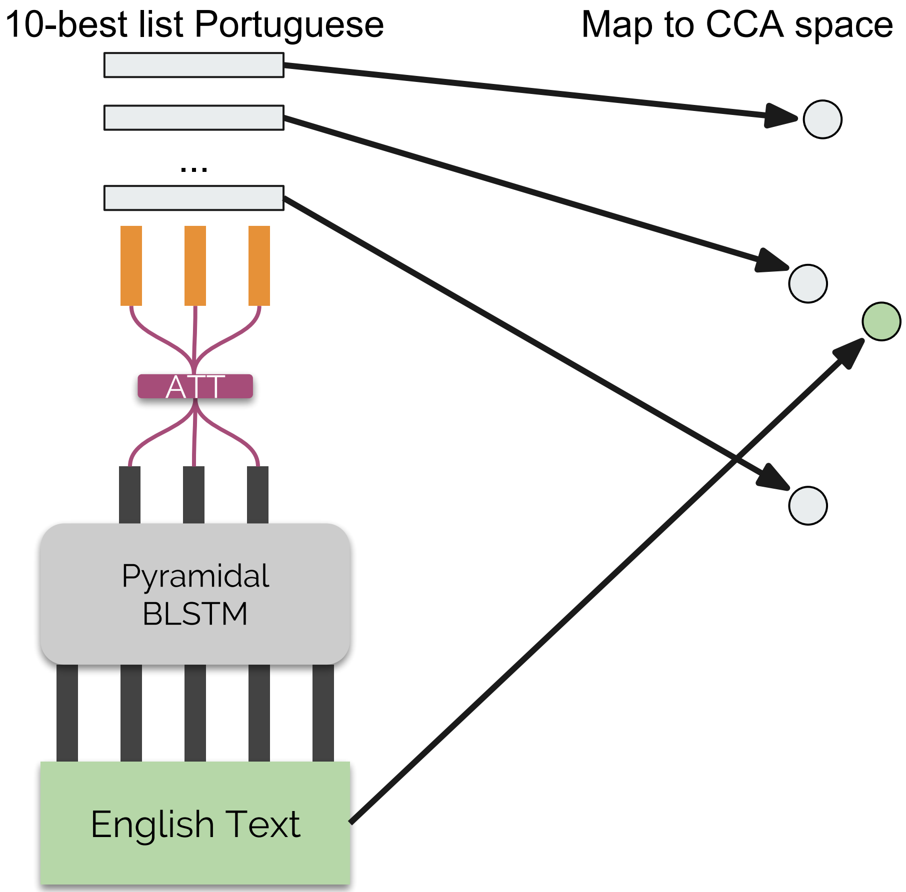
Retrieval for MT

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

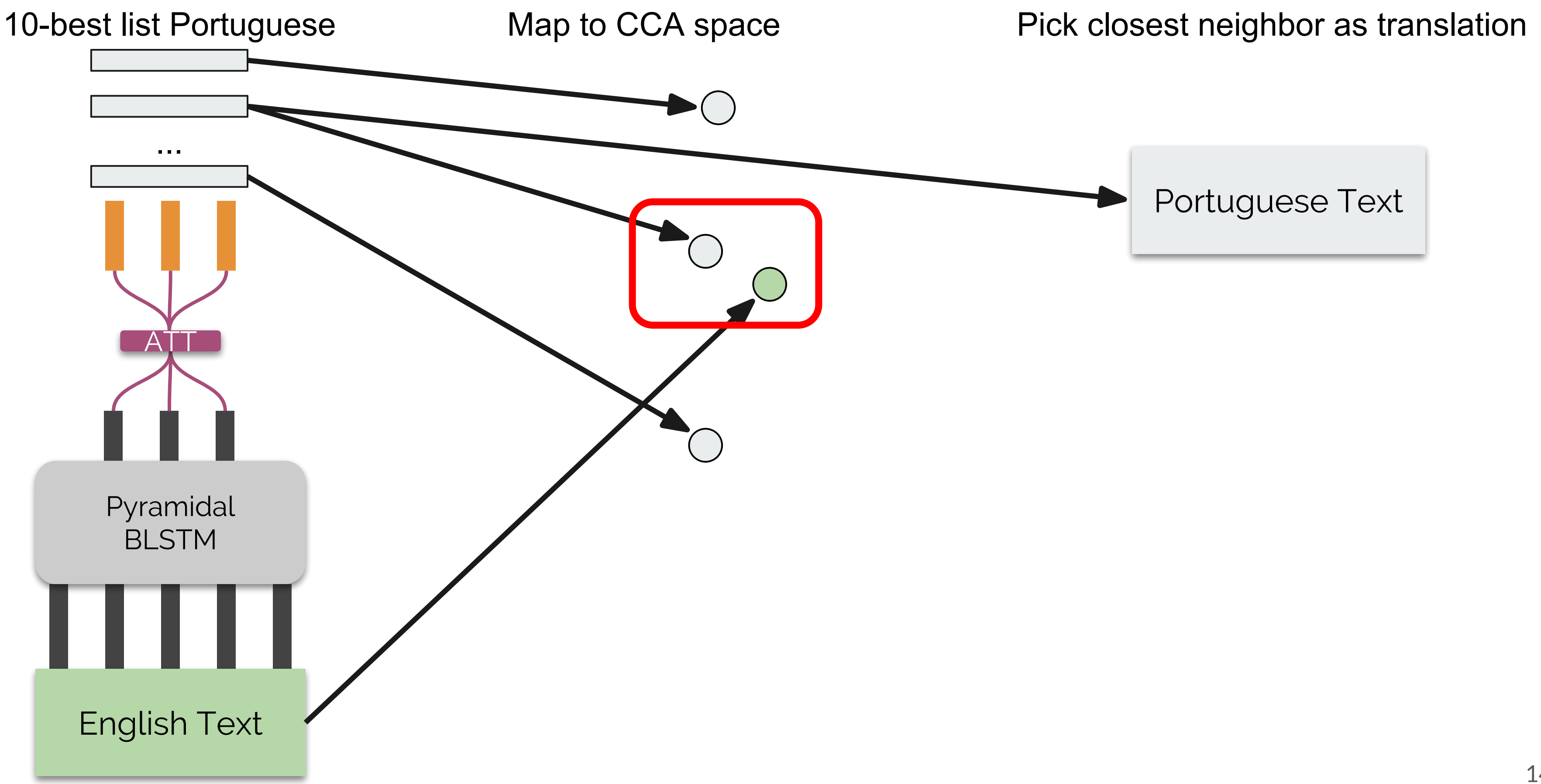


Reference set	BLEU (top 1 retrieval)	BLEU (random pick)
train	5.2	0.4
train + test	80.7	0.4

Re-ranking in MT



Re-ranking in MT

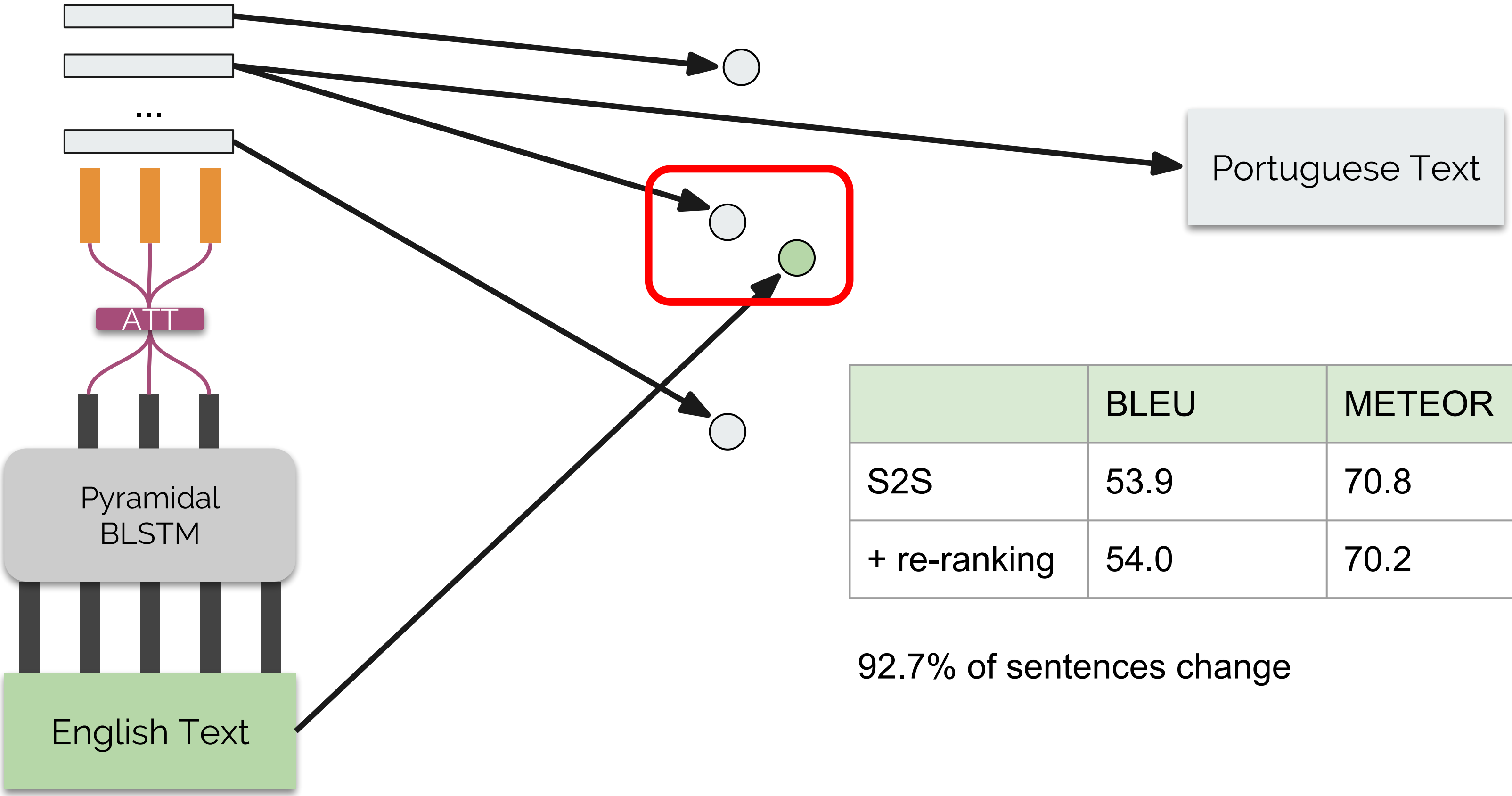


Re-ranking in MT

10-best list Portuguese

Map to CCA space

Pick closest neighbor as translation

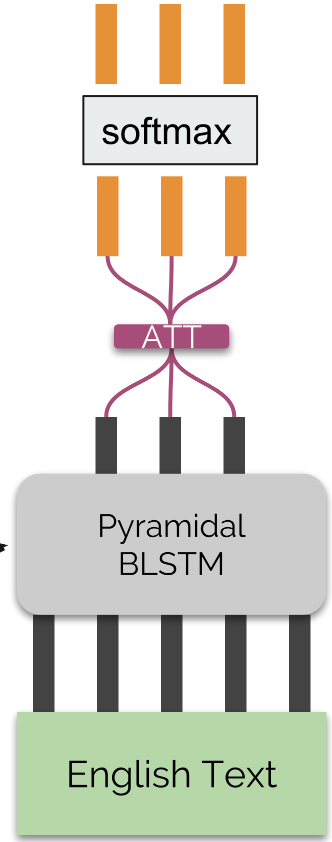


	BLEU	METEOR
S2S	53.9	70.8
+ re-ranking	54.0	70.2

92.7% of sentences change

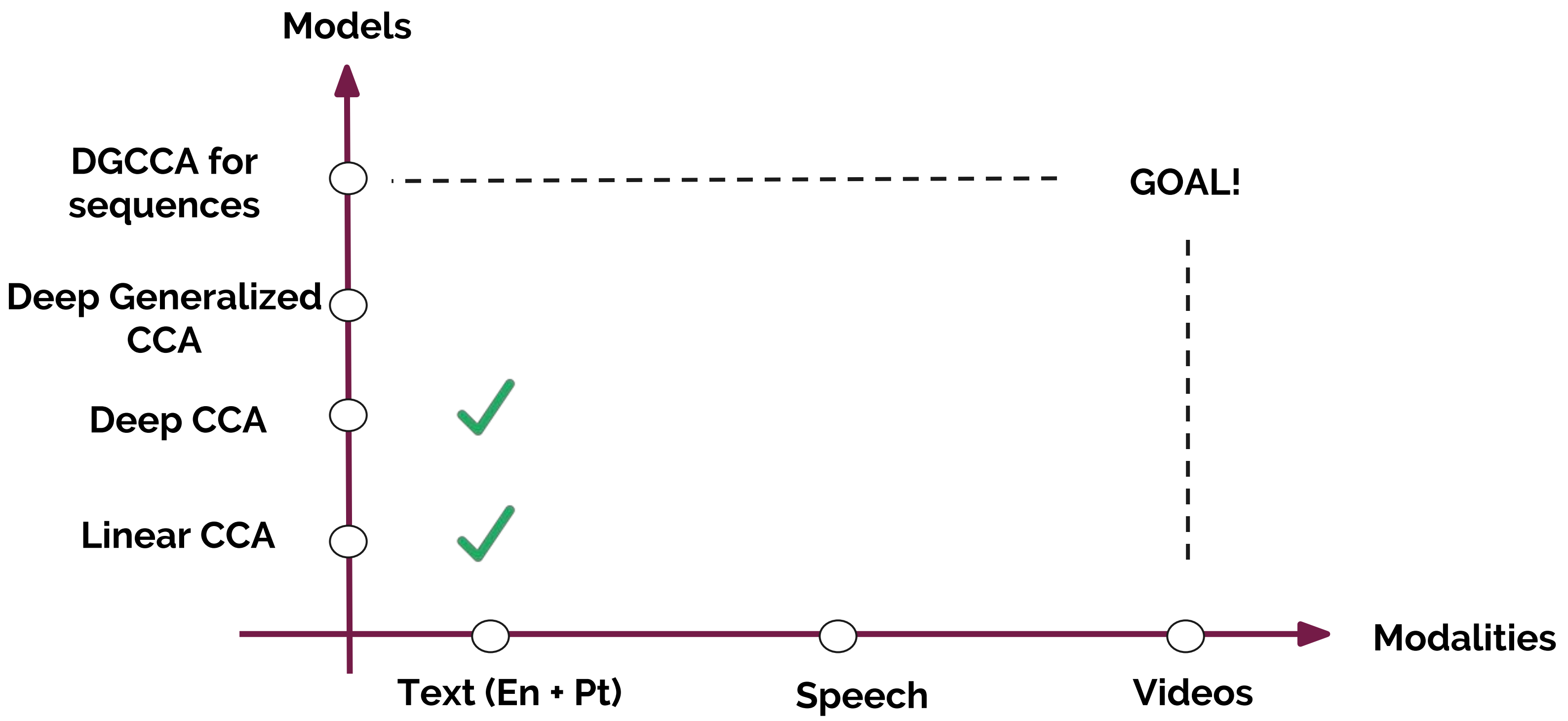
Integration in MT

Inject word embeddings as side information

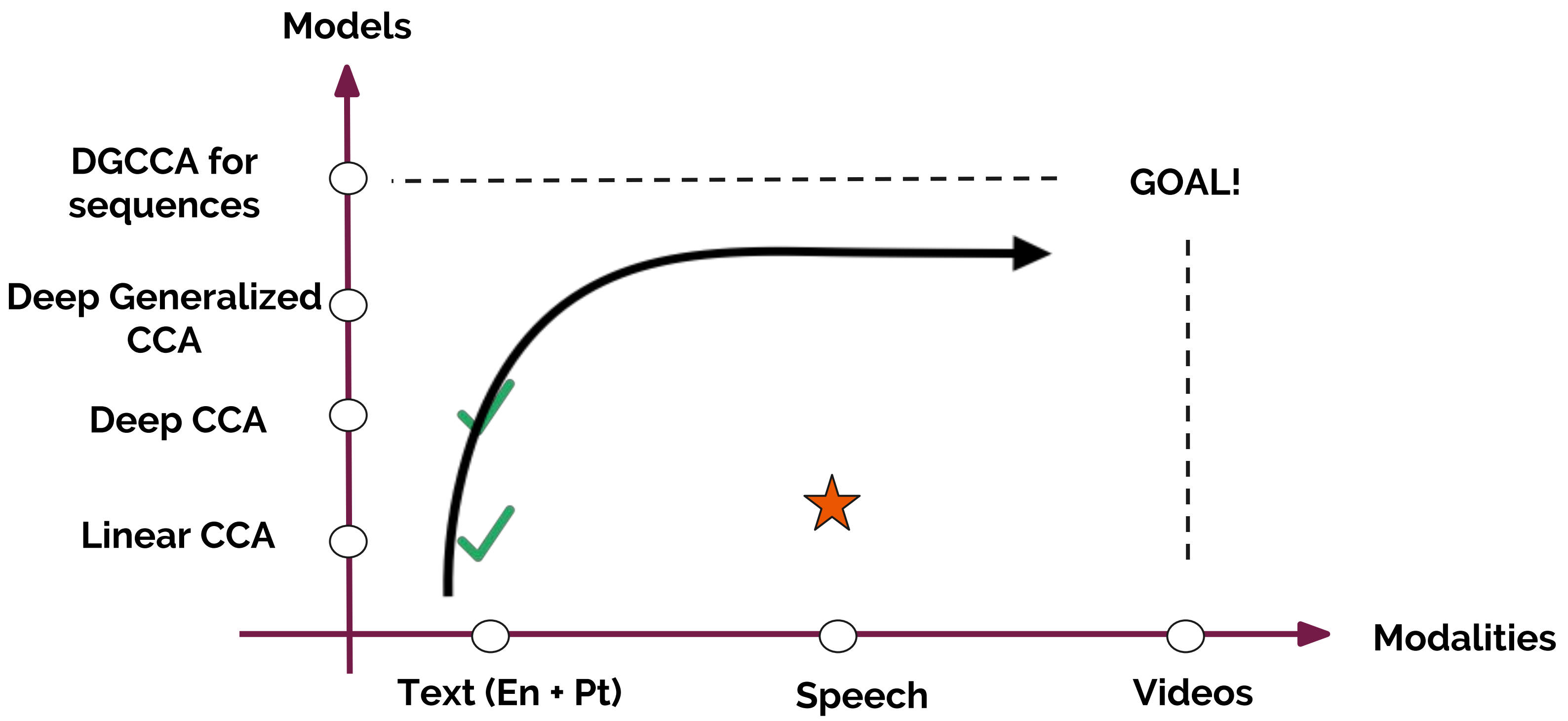


	BLEU	METEOR
S2S	57.3	73.0
+ word embeddings CCA	56.0	72.6
+ word embeddings DCCA	57.1	73.1

Recap: Our Goal

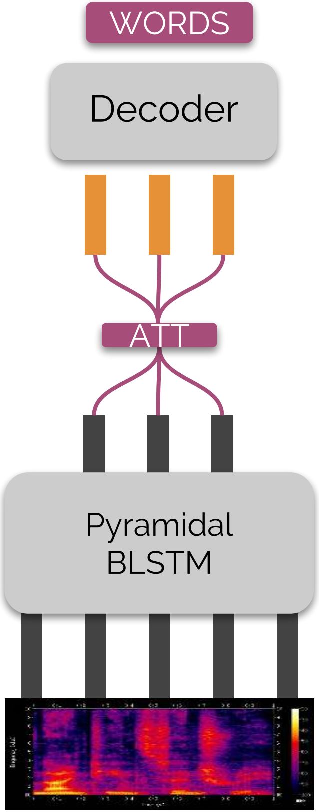


Recap: Our Goal



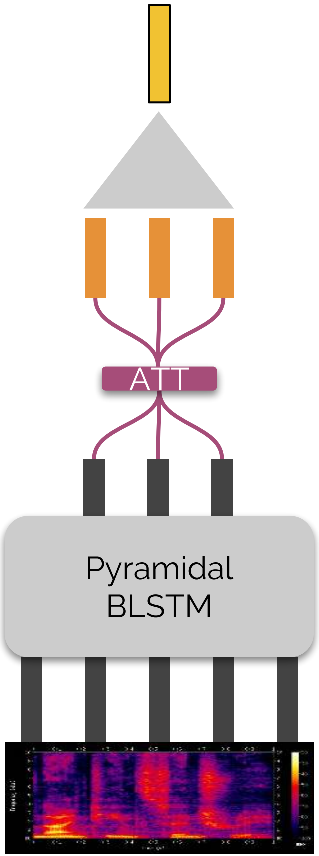
Speech Representations - S2S Model

- Char-based ASR model has a scale mismatch with NMT (words)
- End-to-End Word-based Speech Recognition Model



Palaskar et al. 2018.

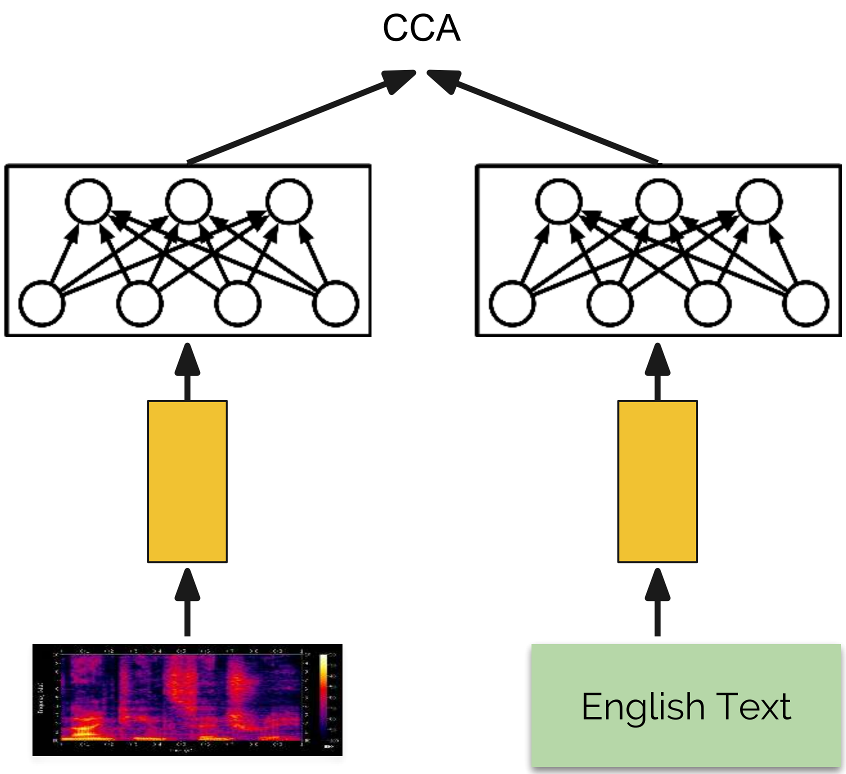
Speech Representations - Sentences



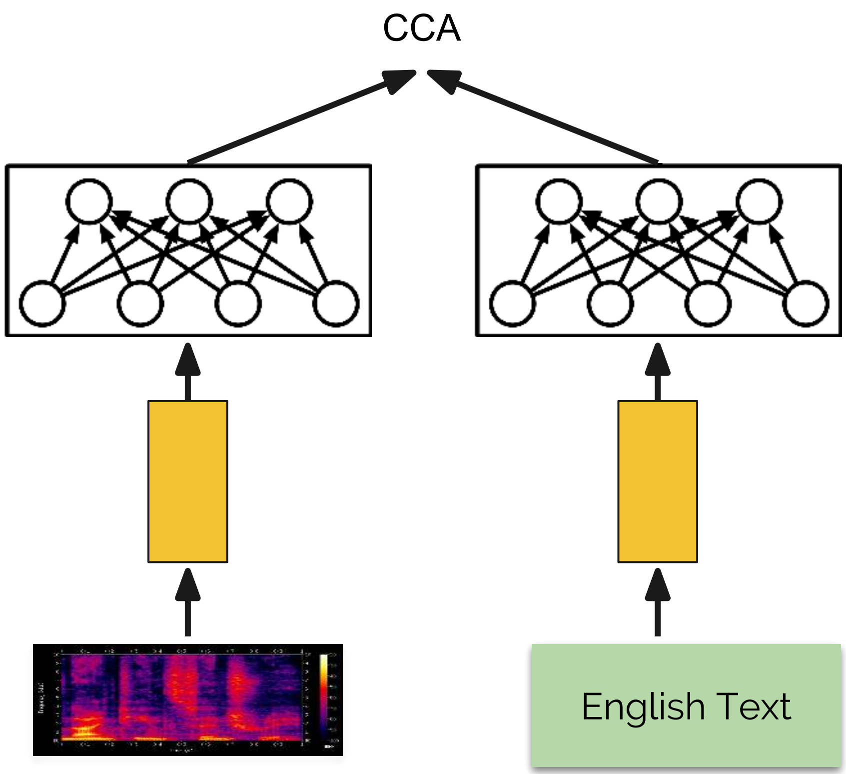
- bag-of-audio-words
- Meanpool
- Context Vector

Palaskar et al. 2018

Speech and Text Representations



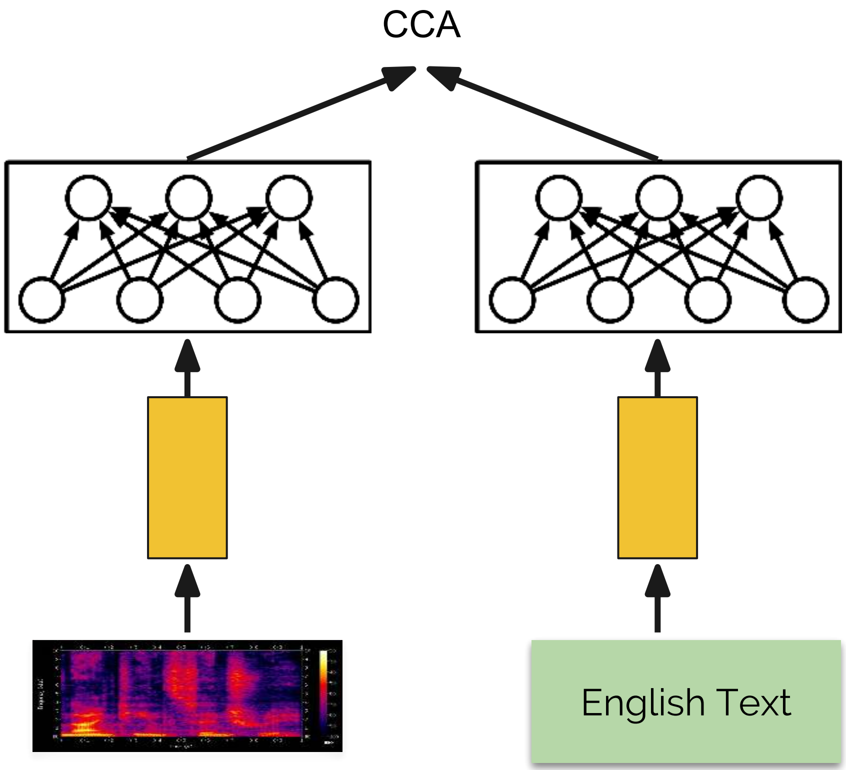
Retrieve Text Given Speech



Recall@10
over Test set

Linear CCA	Deep CCA
96.9%	90.1%

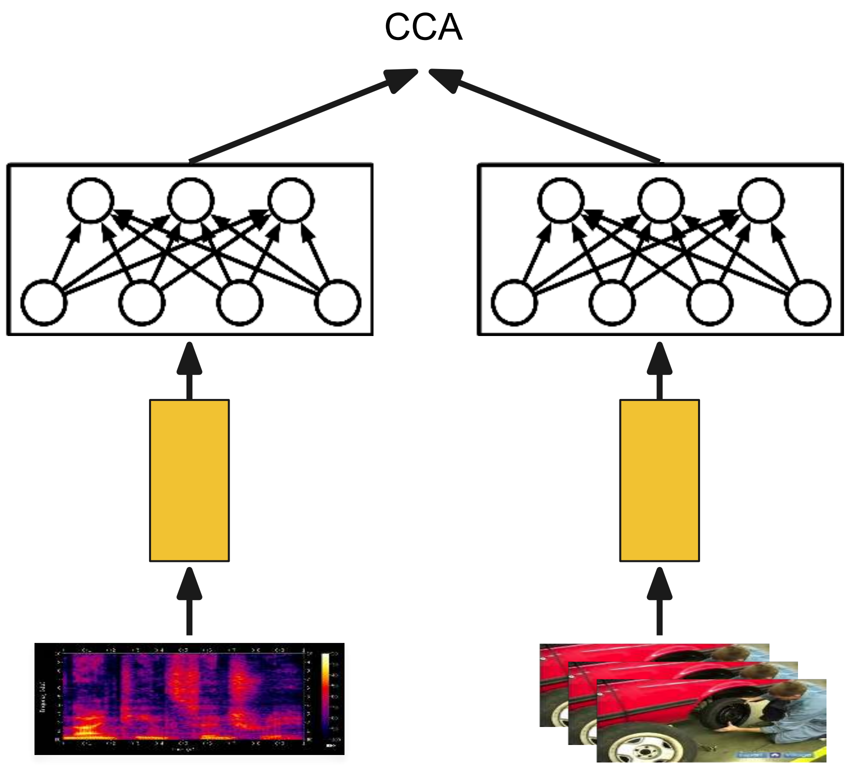
Retrieve Speech Given Text



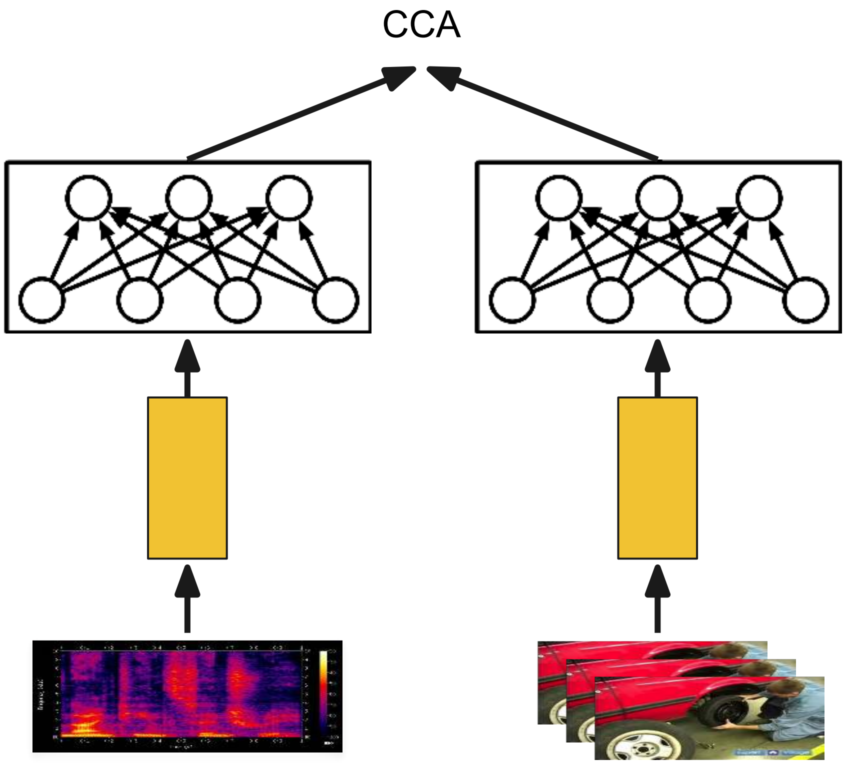
Recall@10
over Test set

Linear CCA	Deep CCA
96.1%	89.7%

Speech and Video Representations



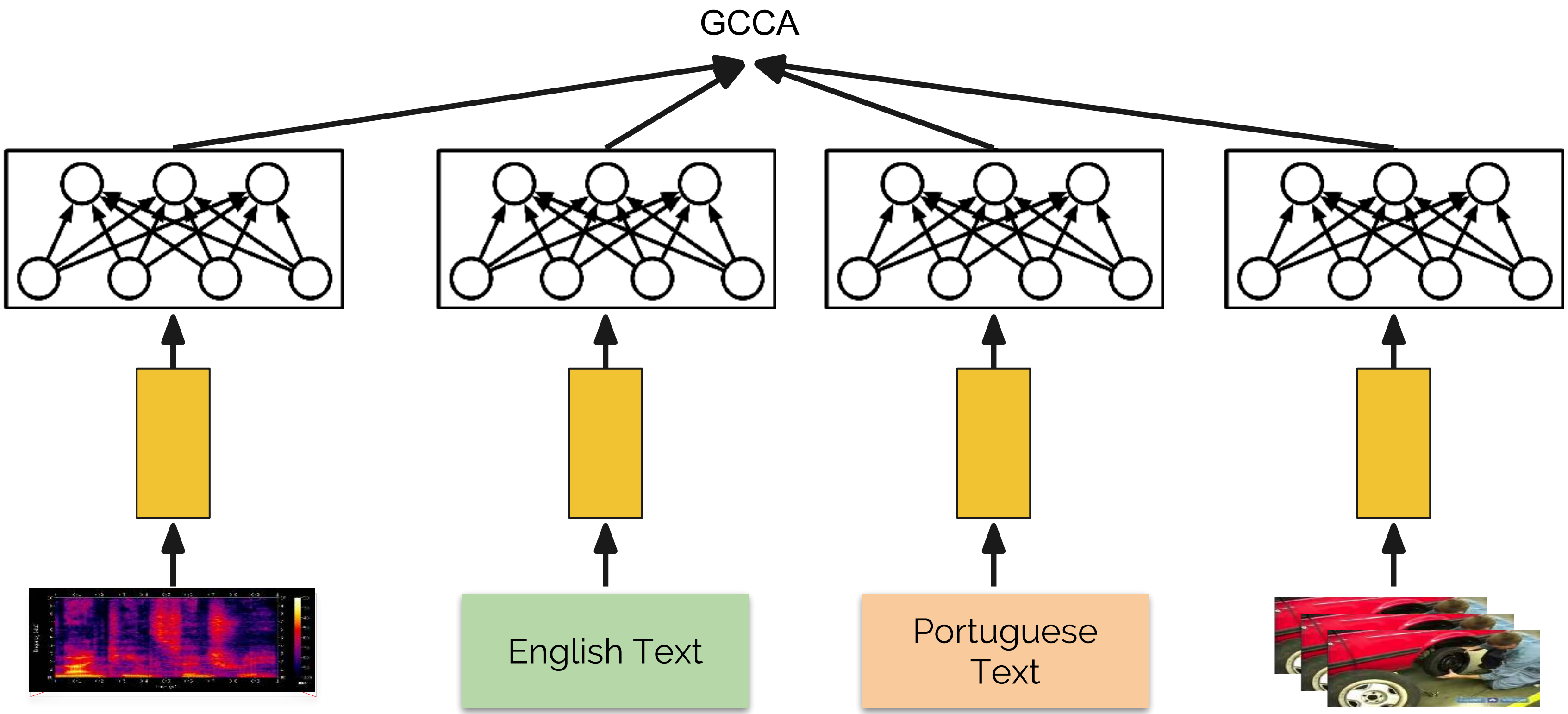
Retrieve Video Given Speech



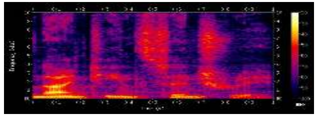
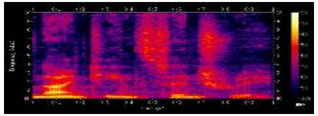
Recall@10
over Test set

Linear CCA	Deep CCA
0.5%	1.8%

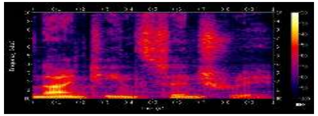
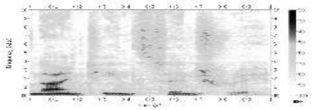
Speech, Text and Video Representations



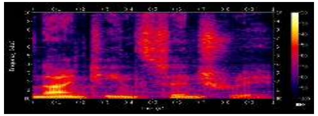
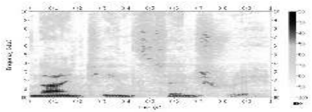
Retrieval: Speech, Text (En & Pt) and Video on Test Set

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	-

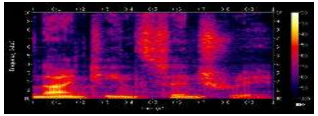
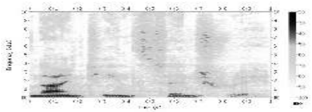
Retrieval: Speech, Text (En & Pt) and Video on Test Set

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	-

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	-

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

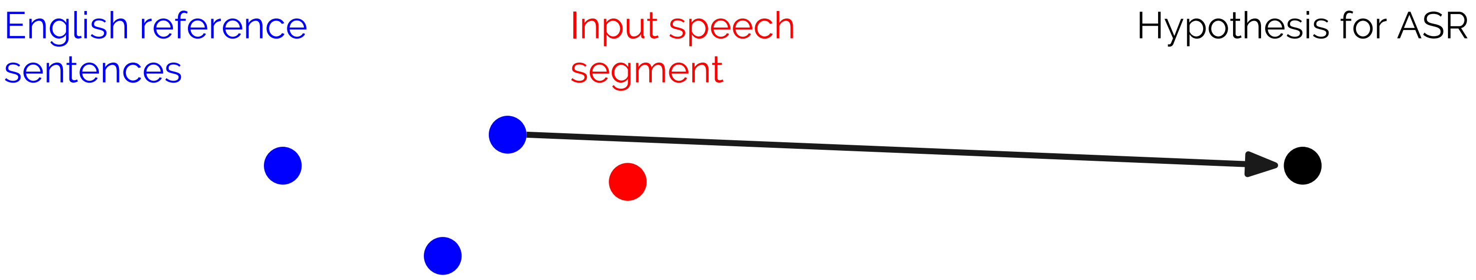
Recall@10		English Text	Portuguese Text	
	-	85.4	70.7	1.0
English Text	85.4	-	98.4	0.9
Portuguese Text	71.0	98.3	-	1.1
	1.1	1.1	0.9	-

Retrieve Text Given Speech - Comparison

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

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 %

Retrieve Pt Text Given Speech - Comparison

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

Portuguese reference sentences

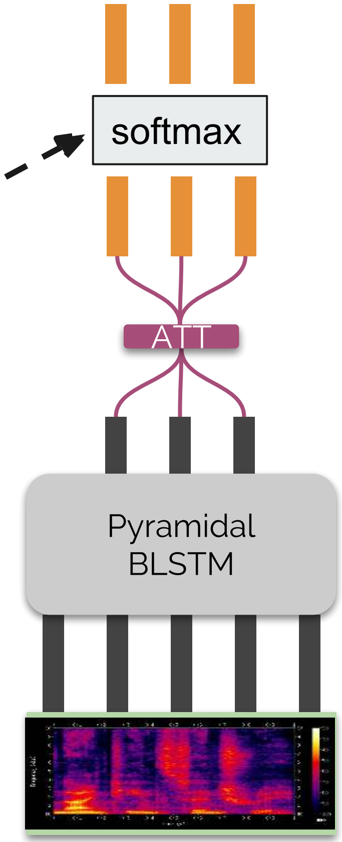
Input speech segment

Hypothesis for Spoken Language Translation

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

Speech Representations - Integration in ASR

Learned CCA representation

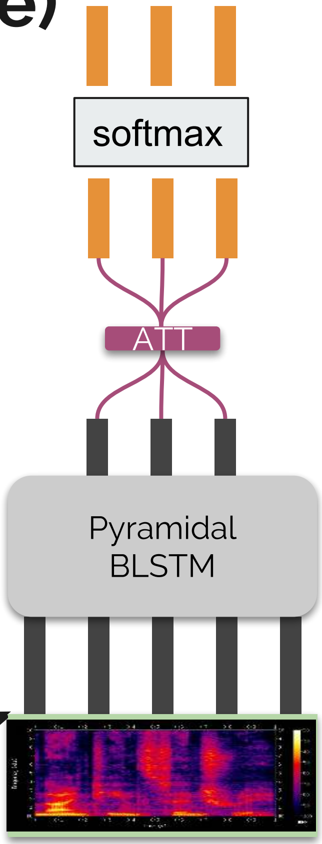


Word Based ASR model
Vocabulary: 19k words

	WER (%) ↓
S2S Model	24.2
+ CCA projections	25.3

Substitutions ↑ 7%

Speech Representations - Integration in ASR (Encoder-side)



Learned CCA representation

Word Based ASR model
Vocabulary: 19k words

	WER (%) ↓
S2S Model	24.2
+ CCA projections	27.3

Substitutions ↑ 14%

Deletions ↑ 11%

Insertions ↑ 11%

Conclusion

- Implementation and exploration of DGCCA models

- CCA can learn strong representations with high cross-view retrieval scores (even with a simple, closed form linear version)

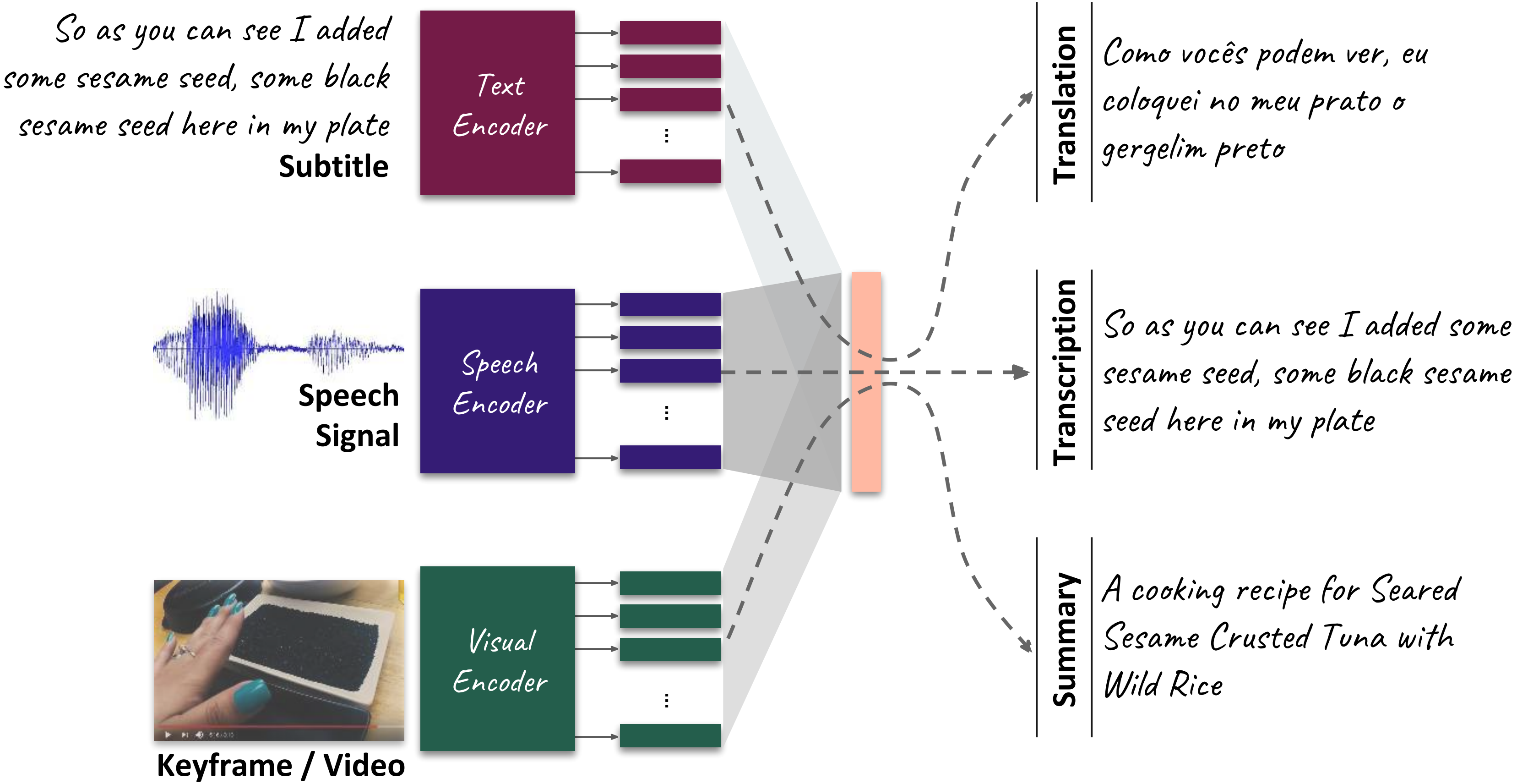
- Exploration of integration into task-specific models

Multitask learning

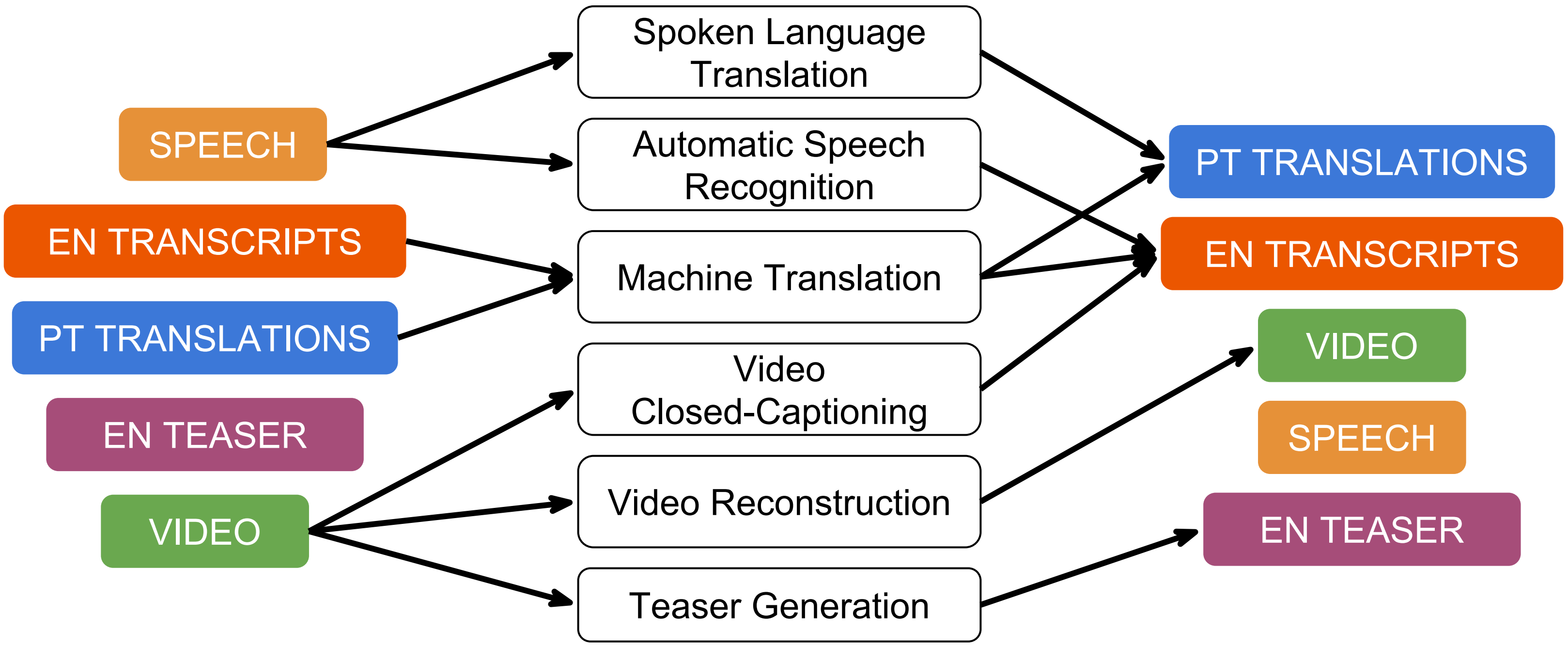
   

Amanda, Desmond, Loïc, Karl

The big picture

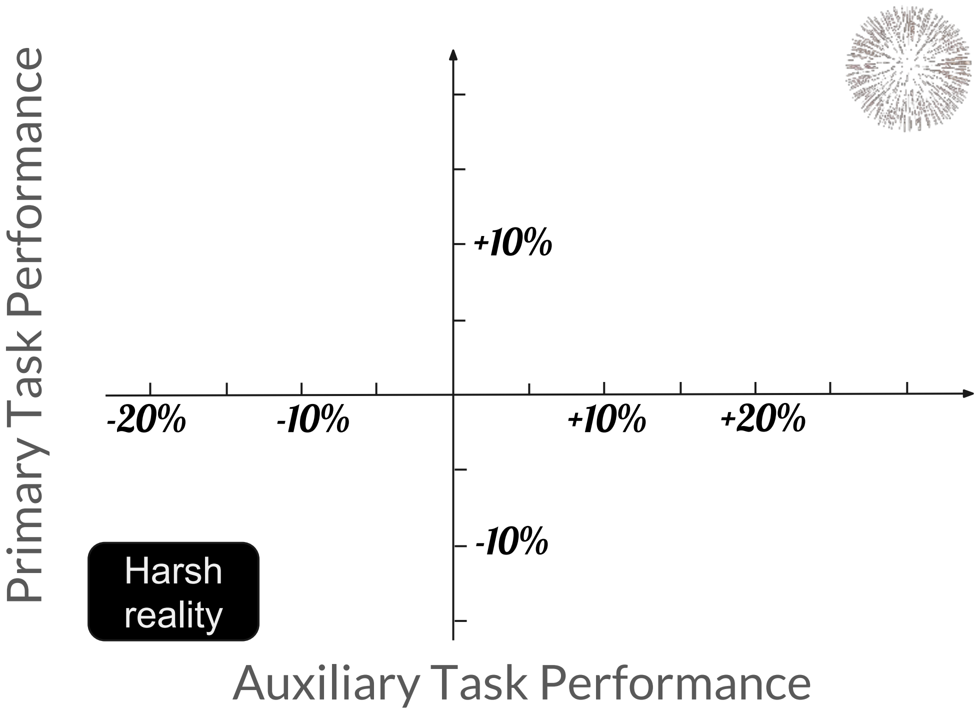


Our big picture



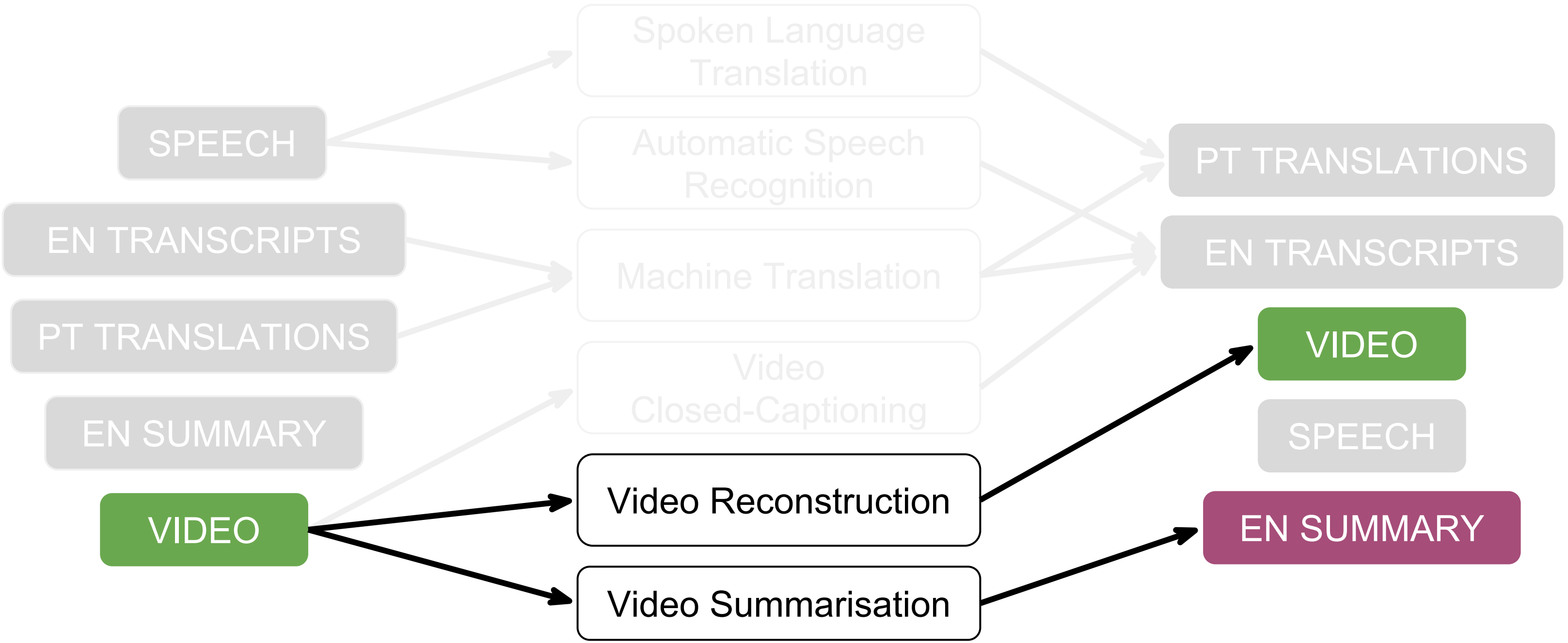
Q: *How* and *when* is it useful to learn a shared representation between different modalities?

Defining useful Multitask Learning



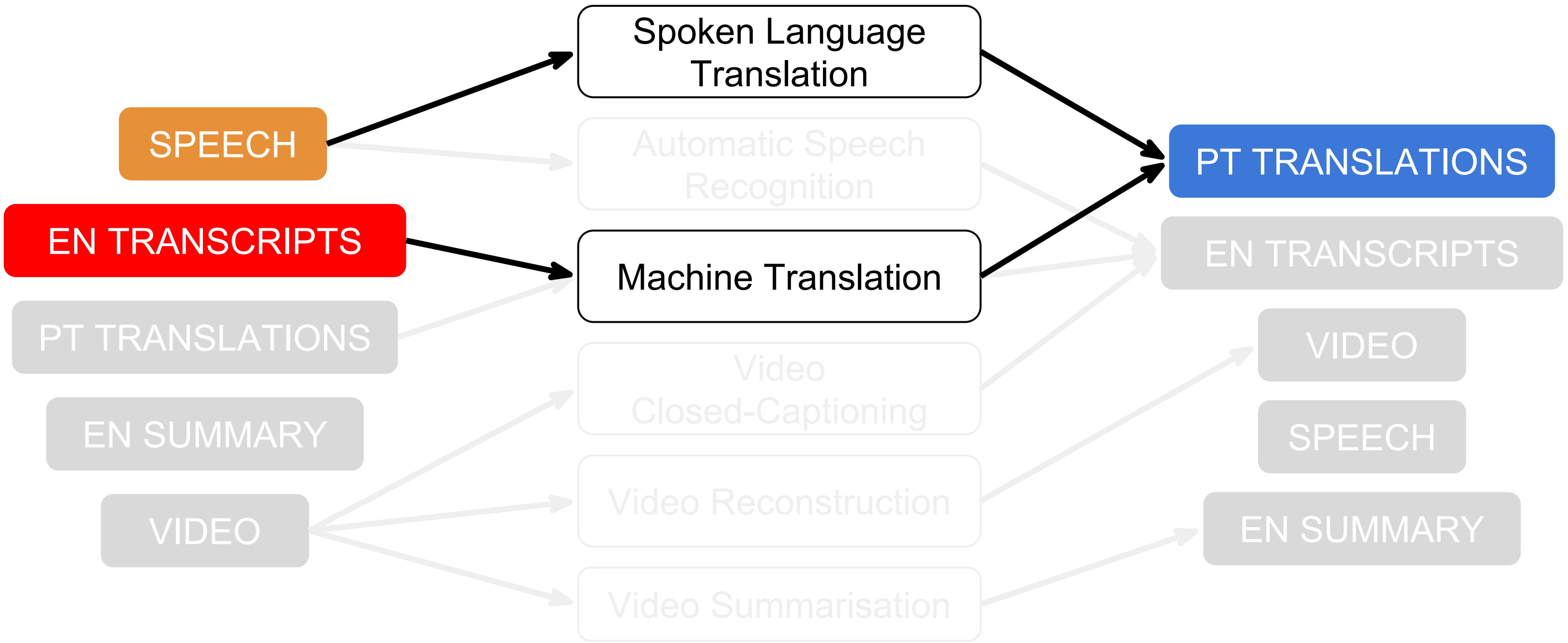
When: Shared Encoder

Video Reconstruction + Teaser Generation



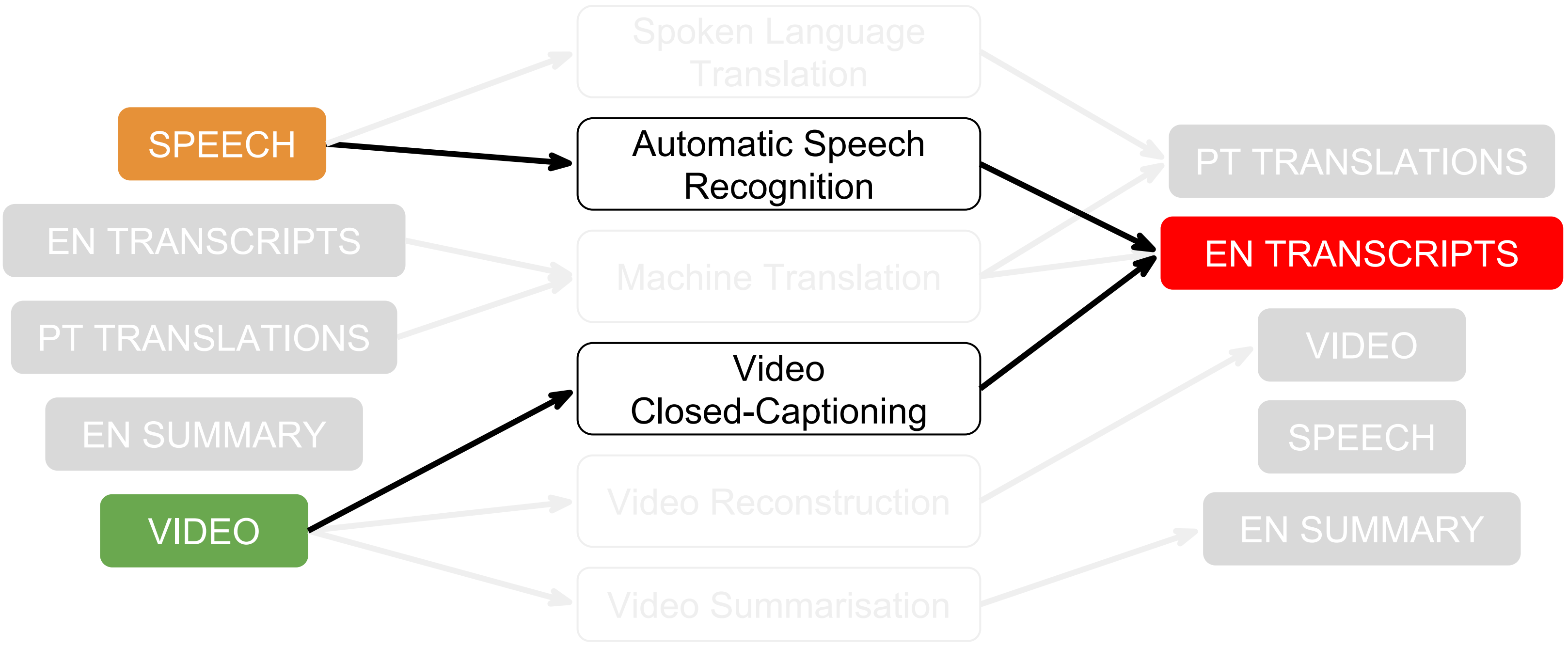
When: Shared Decoder

Spoken Language Translation + Machine Translation



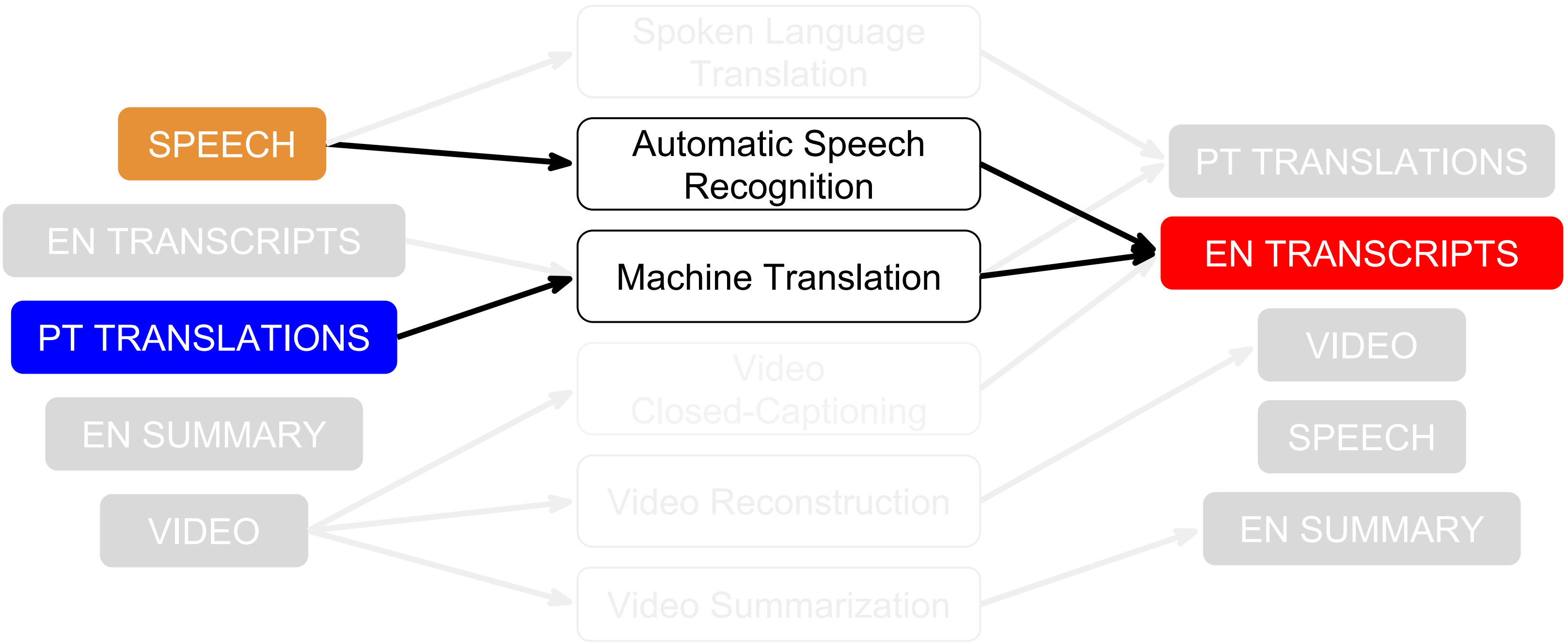
When: Shared Decoder

ASR + Video closed-captioning



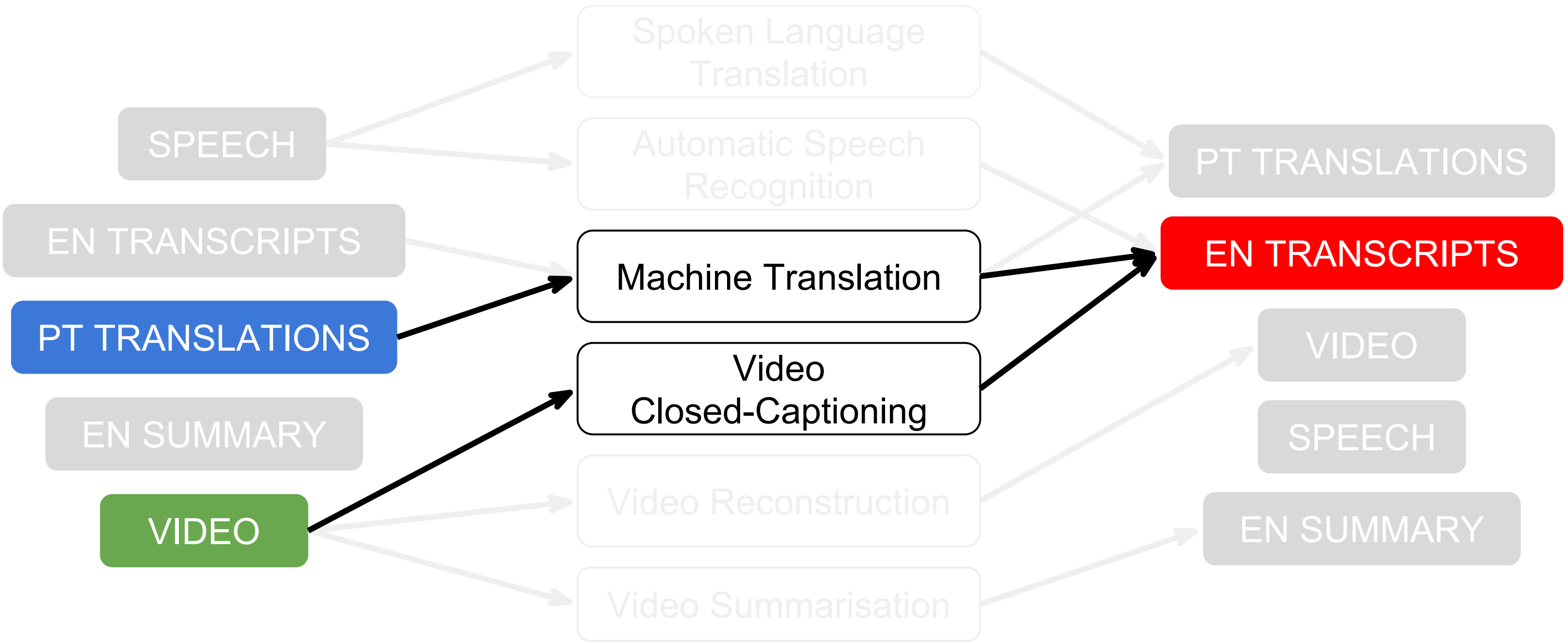
When: Shared Decoder

ASR + MT



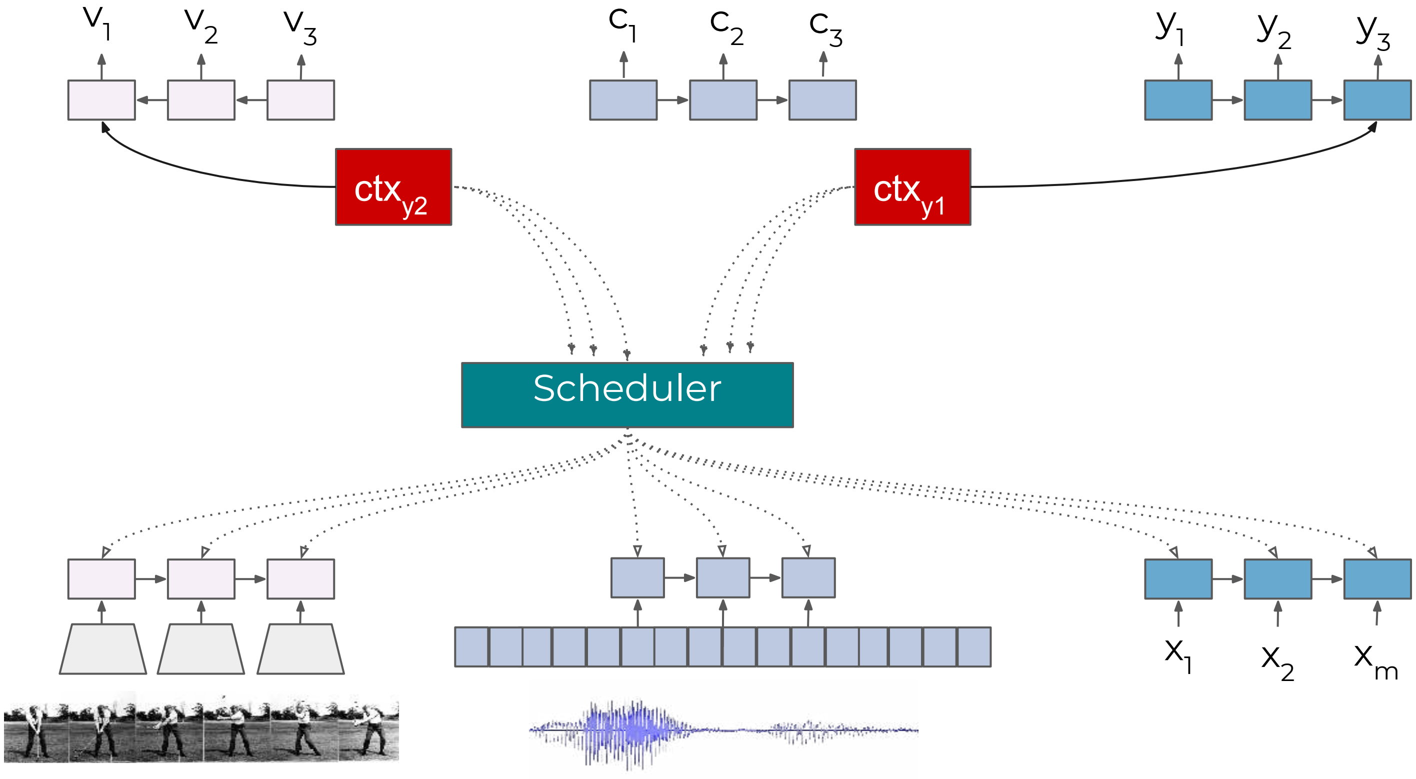
When: Shared Decoder

Video closed-captioning + MT

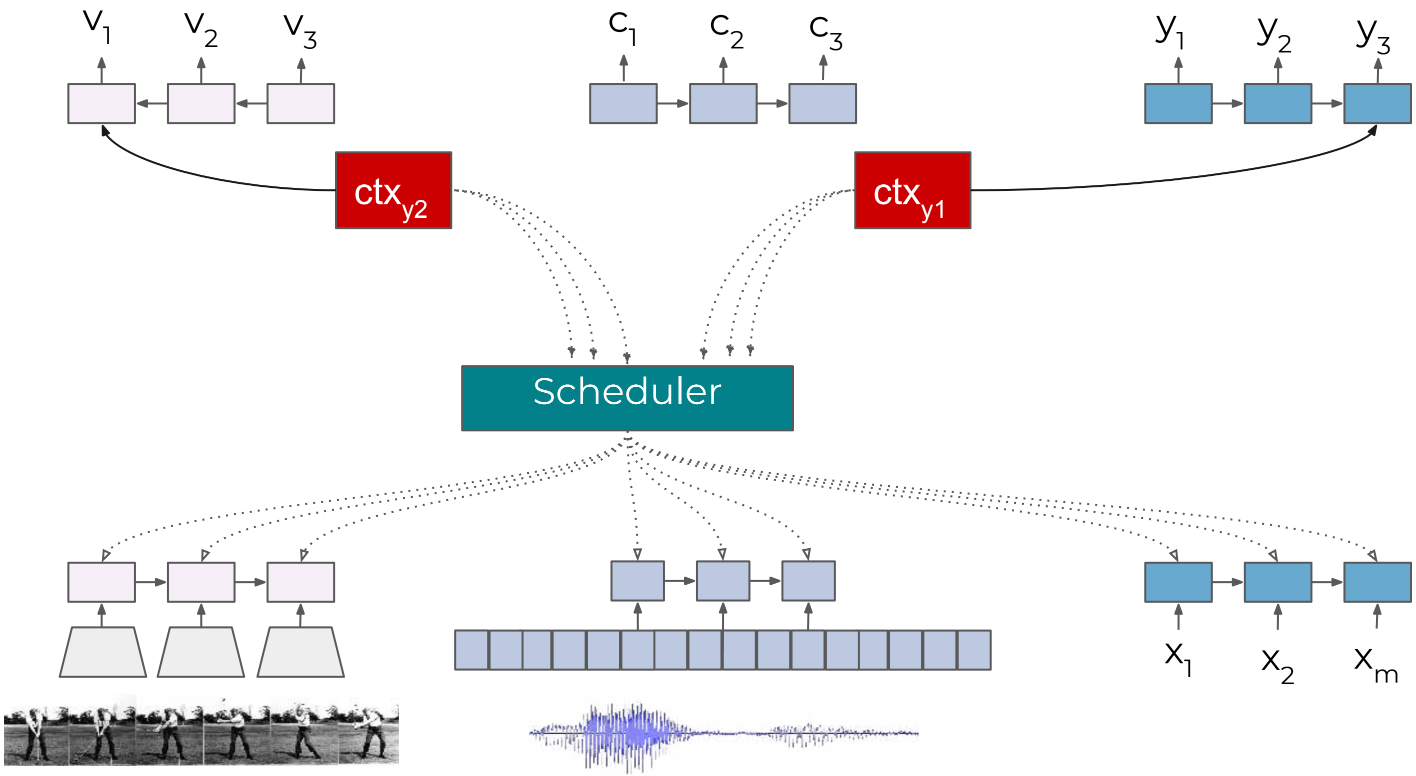


How: Multitask Learning

How: MTL by scheduling tasks

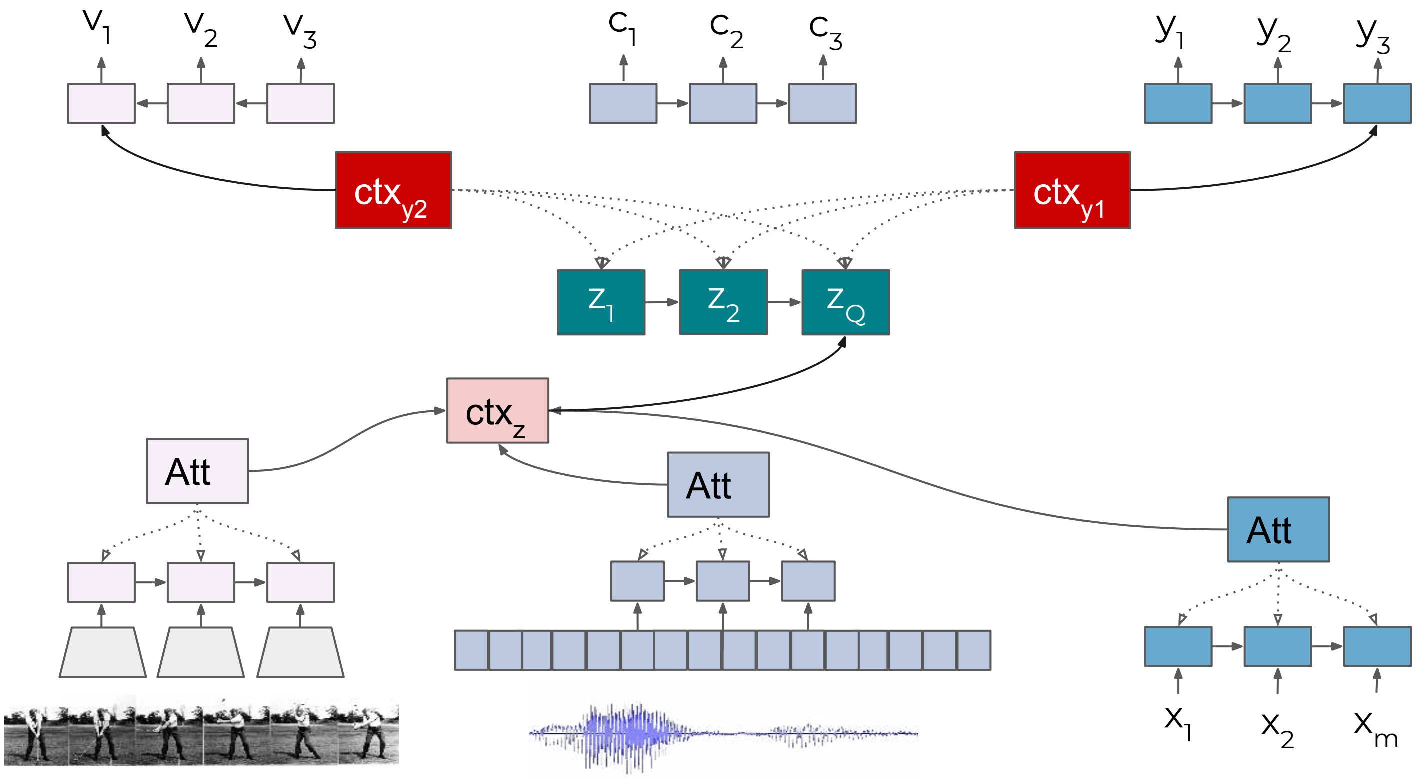


How: MTL by scheduling tasks



How: MTL with Shared Recurrent Space

- Learn a shared representation z given multiple tasks (Lu et al. 2018)



MTL with Mutual Projection Networks

- Assume *n* > 2 modalities of **aligned** data
- Assume we have an encoder for each modality

  $D =$ Speech, English, Portuguese, Video, Teasers

Sample a source-target task from the training schedule and an auxiliary source of data

For $(x, y, a) \sim D$:

$$\mathcal{L}(\theta) = \sum_j \text{-log p}(y_j | y_{<j}, x) + \alpha \text{d}(x, a) + \beta \text{d}(y, a)$$

Minimise the primary loss

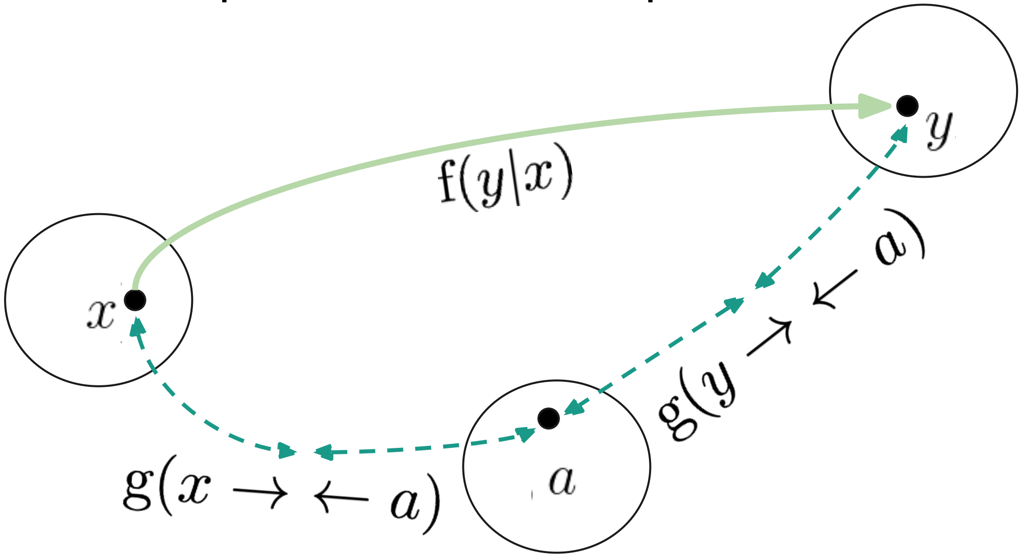
Project *auxiliary data* into the same space as the encoder and the decoder

Max-margin objective

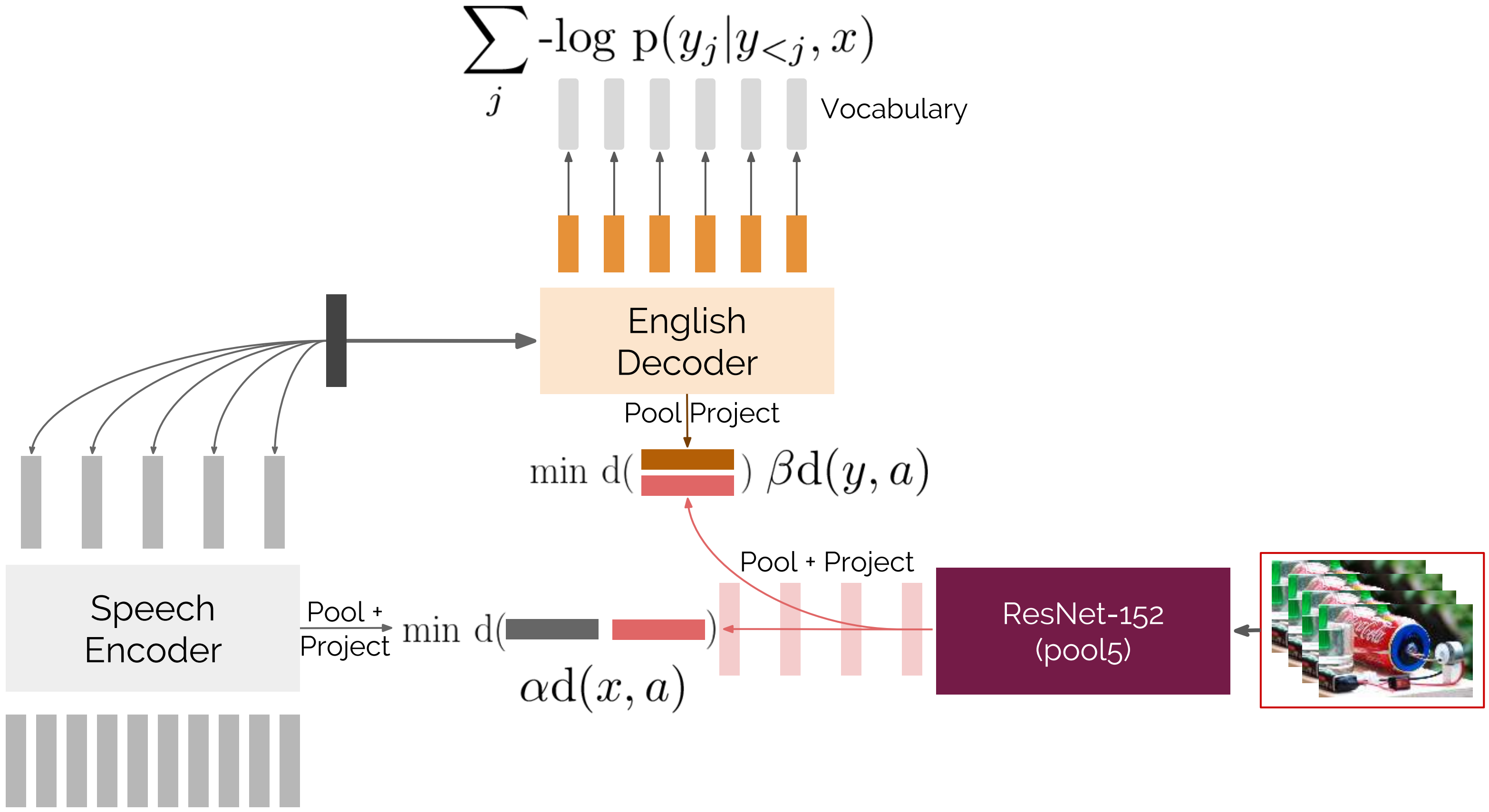
$$d(a,b) = \sum_{<\hat{a},b>} [\max(0, \alpha - cos(a,b) + cos(\hat{a},b))] + \sum_{<a,\hat{b}>} [\max(0, \alpha - cos(a,b) + cos(a,\hat{b}))]$$

Why Mutual Projection Networks?

- Explicitly learn a shared space between the different views of the data
- Regularise the main task encoder and decoder with projection losses
  - Learn multiple encoders for the price of one!



MPN Illustrative Model



Experiments

Experimental Methodology

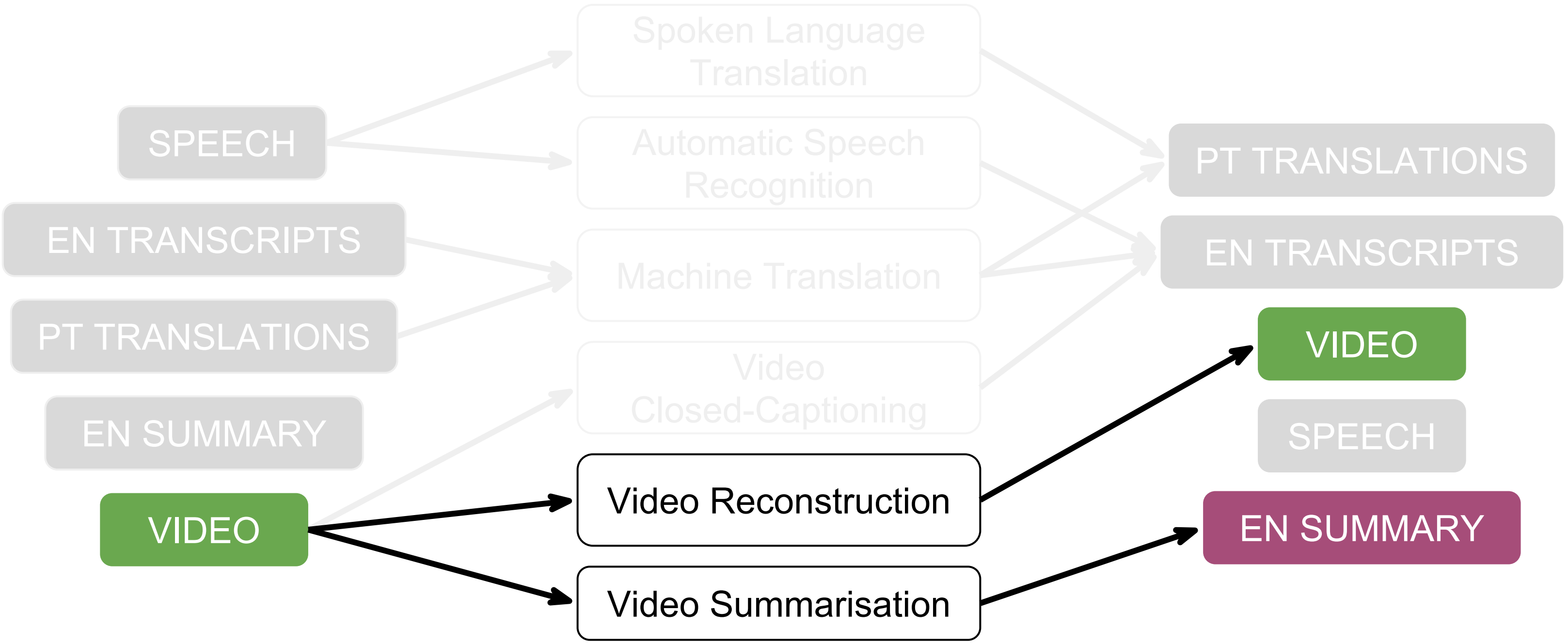
- Fixed hyperparameters from single-task baseline models
- Fixed data pre-processing pipeline

- Models:
  - Single-task baseline
  - Multi-task learning model (MTL)
  - MTL with Shared Recurrent Space
  - MTL with Mutual Projection Network

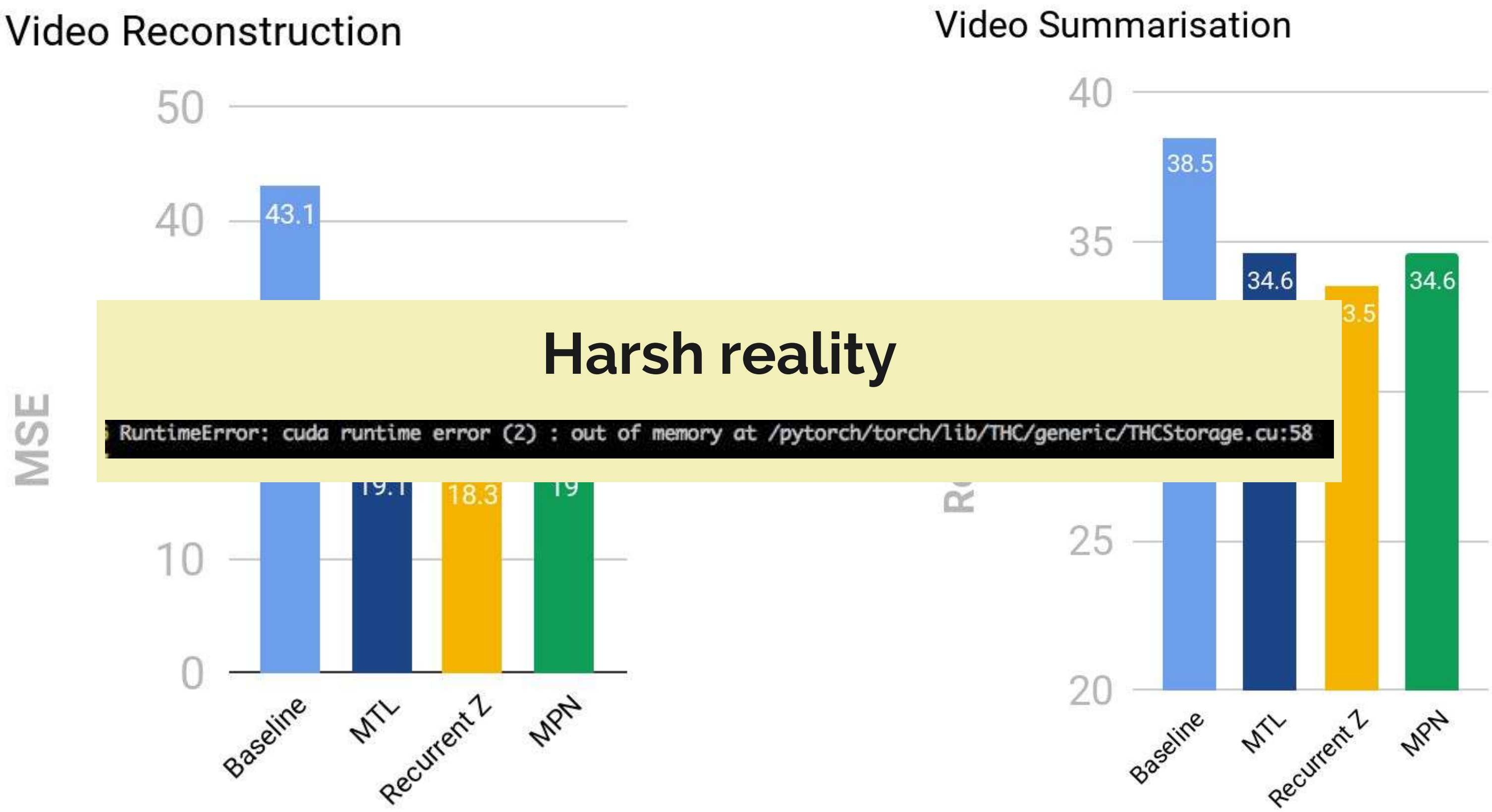
Hypothesis: the MTL models will outperform the single-task models because their representations need to be useful for more than one task.

When: Shared Encoder

Video Reconstruction + Teaser Generation



Results: Video Reconstruction + Teaser Generation

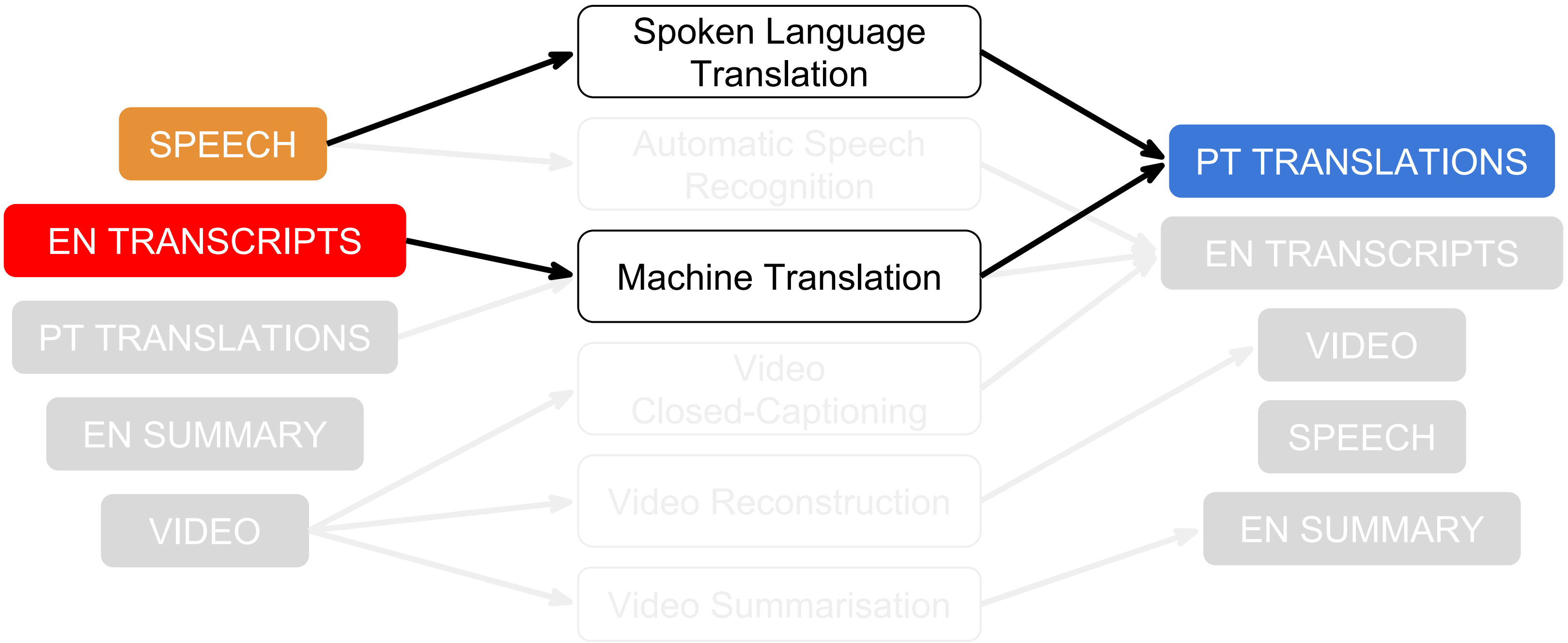


Harsh reality

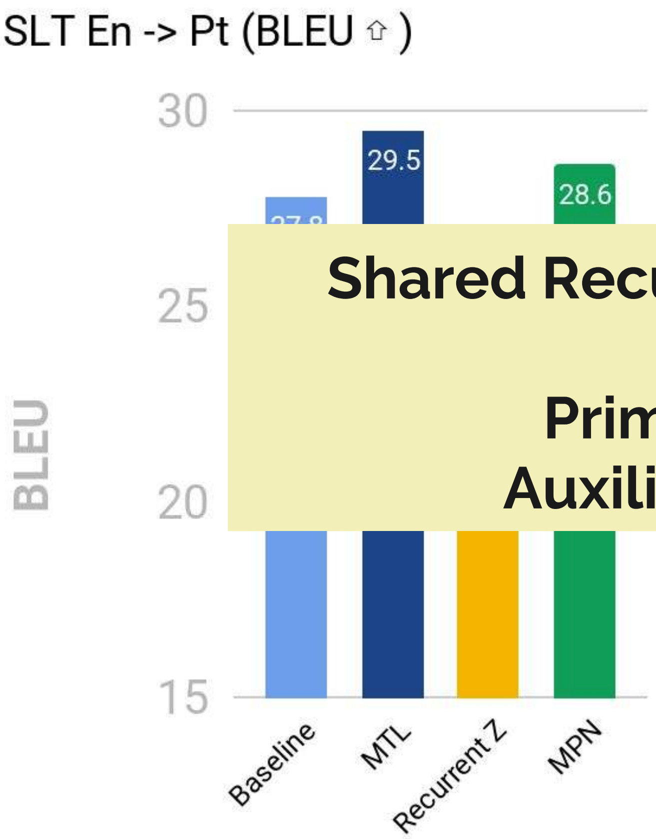
```
RuntimeError: cuda runtime error (2) : out of memory at /pytorch/torch/lib/THC/generic/THCStorage.cu:58
```

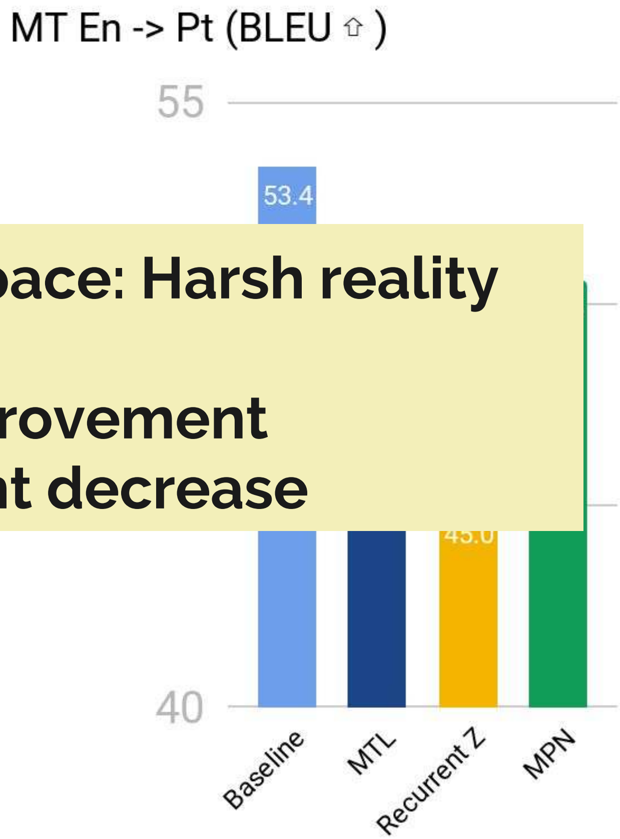
When: Shared Decoder

Spoken Language Translation + Machine Translation



Results: SLT + MT



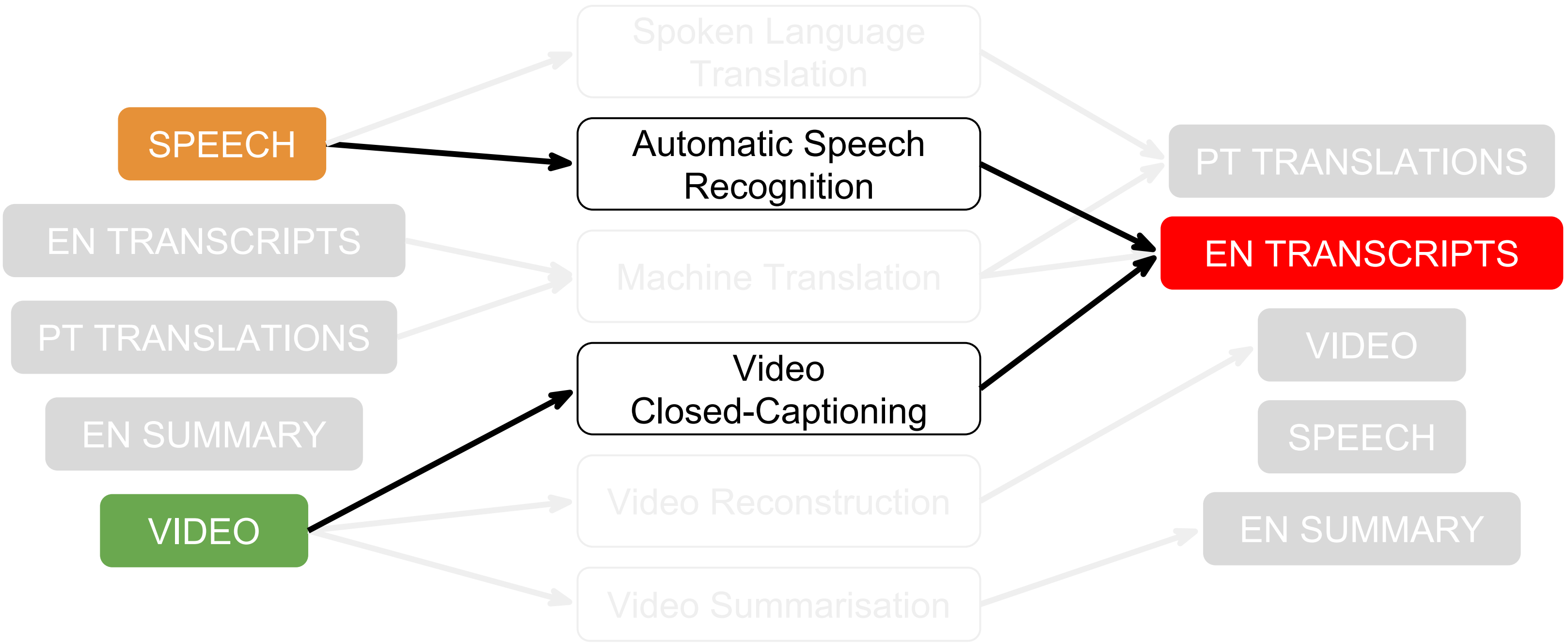


Shared Recurrent Space: Harsh reality

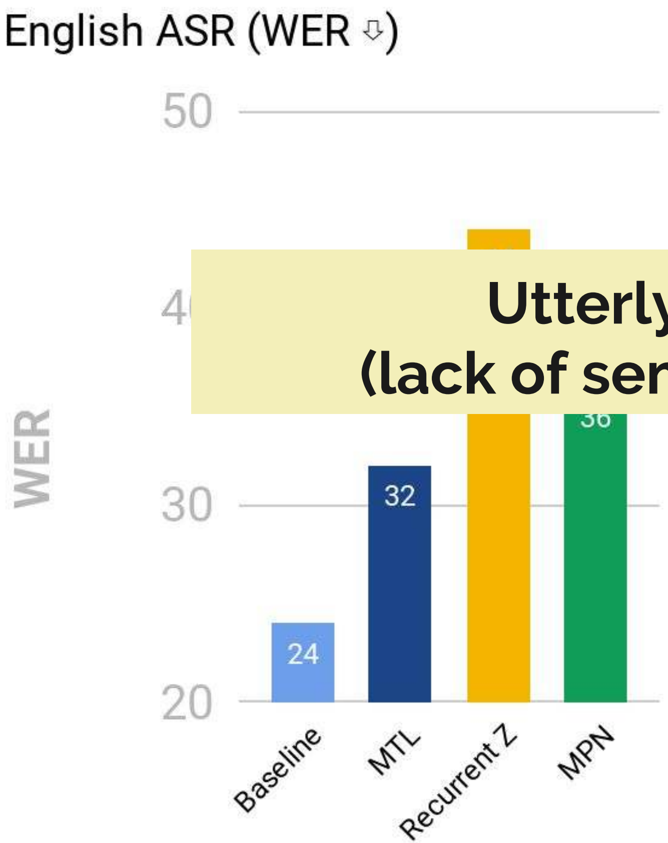
Primary: improvement
Auxiliary: slight decrease

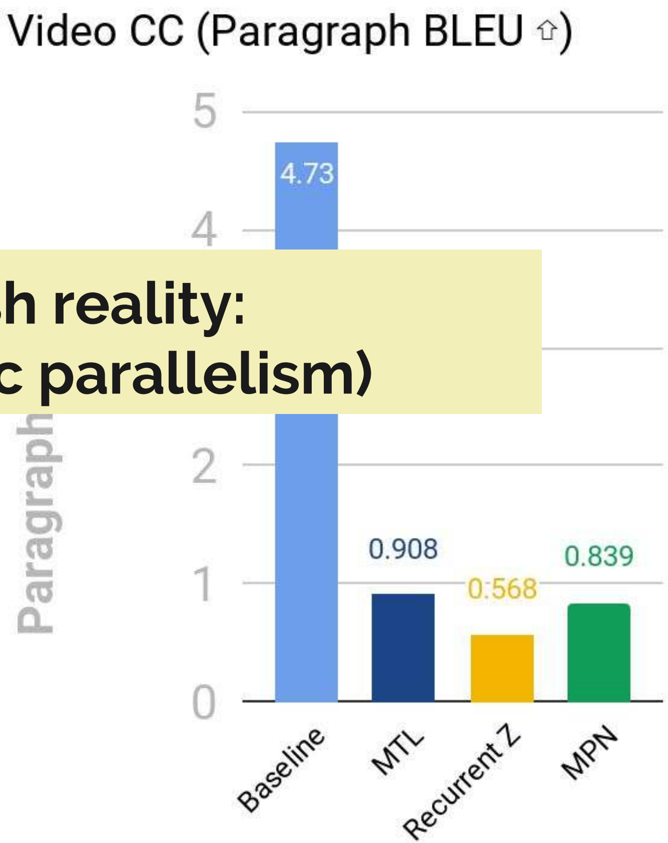
When: Shared Decoder

ASR + Video closed-captioning



Results: ASR + Video closed-captioning

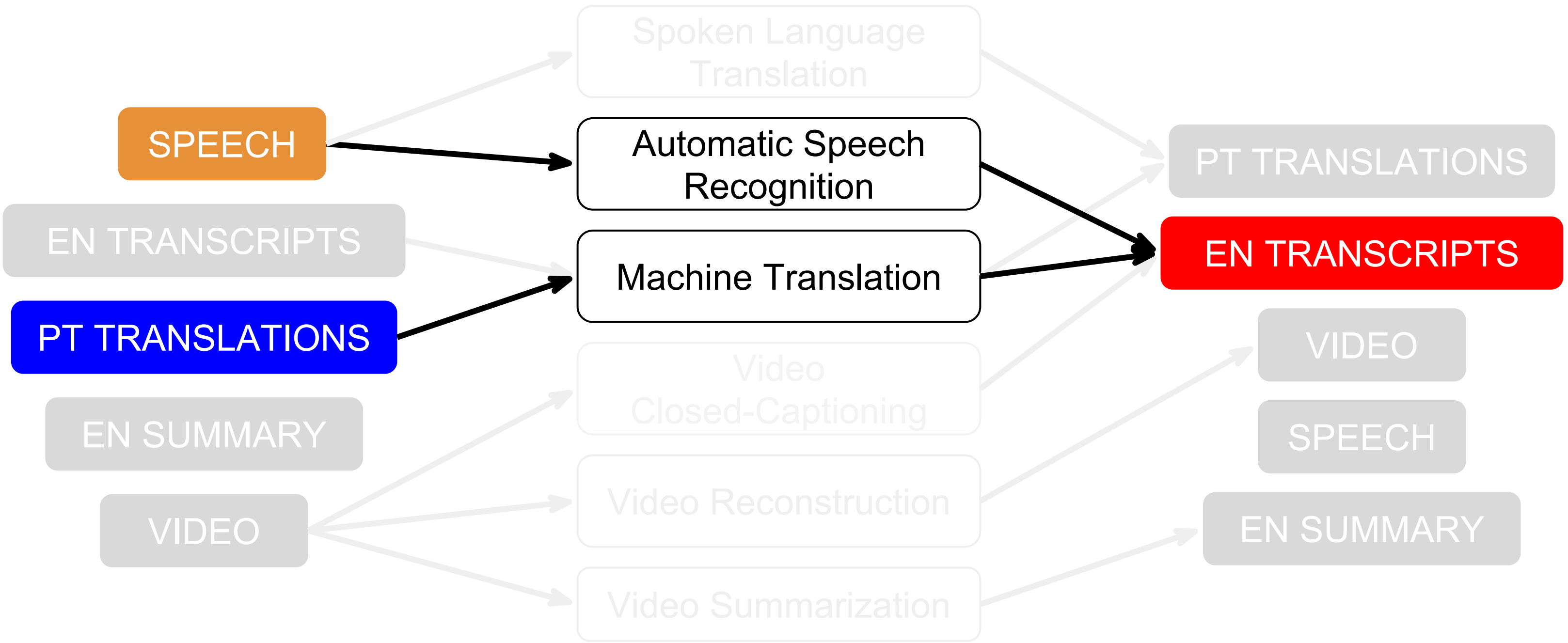




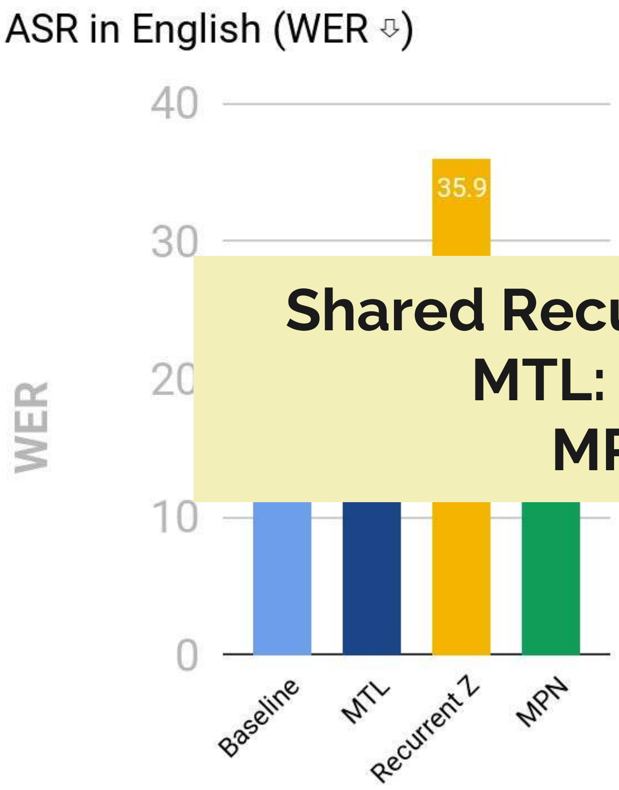
Utterly harsh reality:
(lack of semantic parallelism)

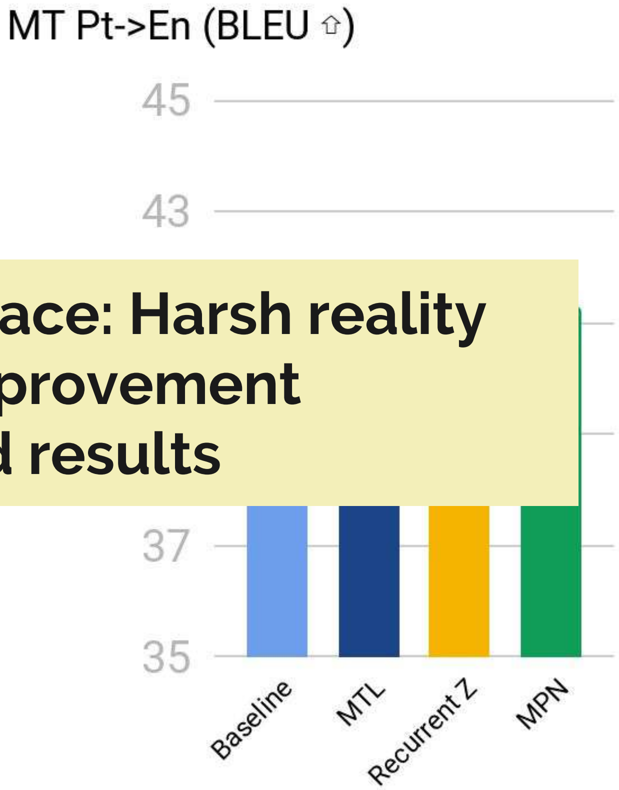
When: Shared Decoder

ASR + MT



Results: ASR + MT

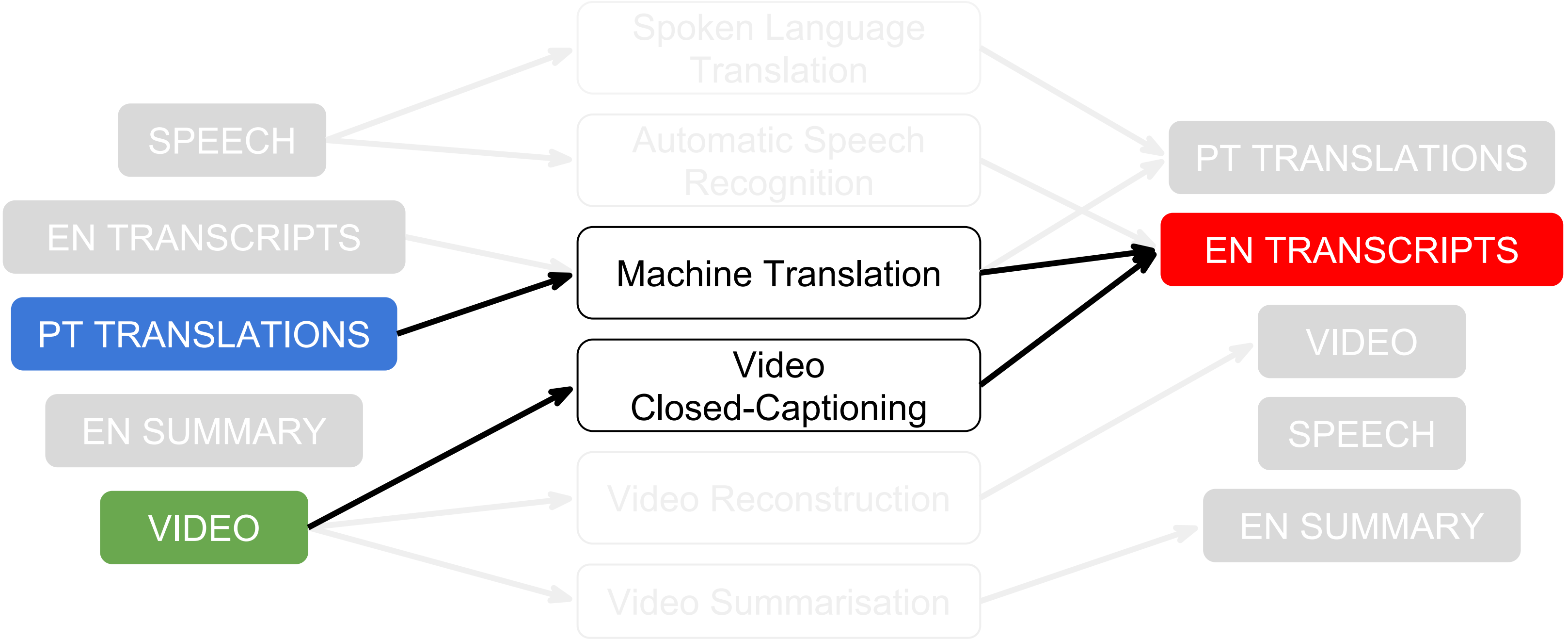




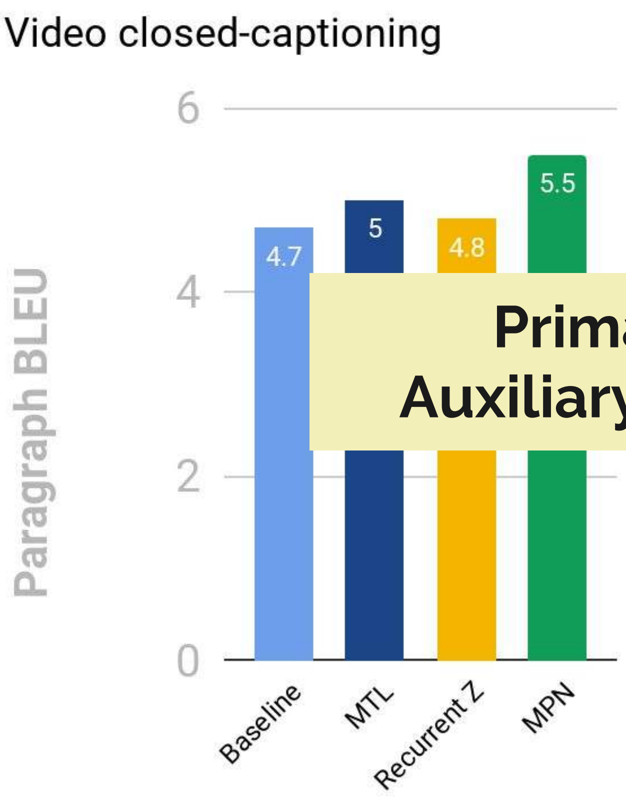
Shared Recurrent Space: Harsh reality
MTL: slight improvement
MPN: mixed results

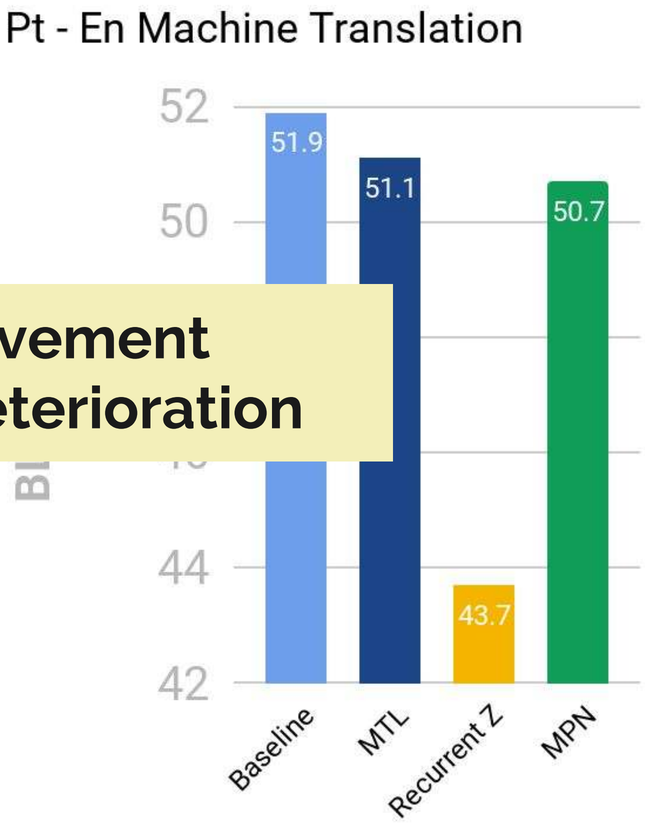
When: Shared Decoder

Video closed-captioning + MT



Results: Video closed-captioning + MT





Primary: improvement
Auxiliary: slight deterioration

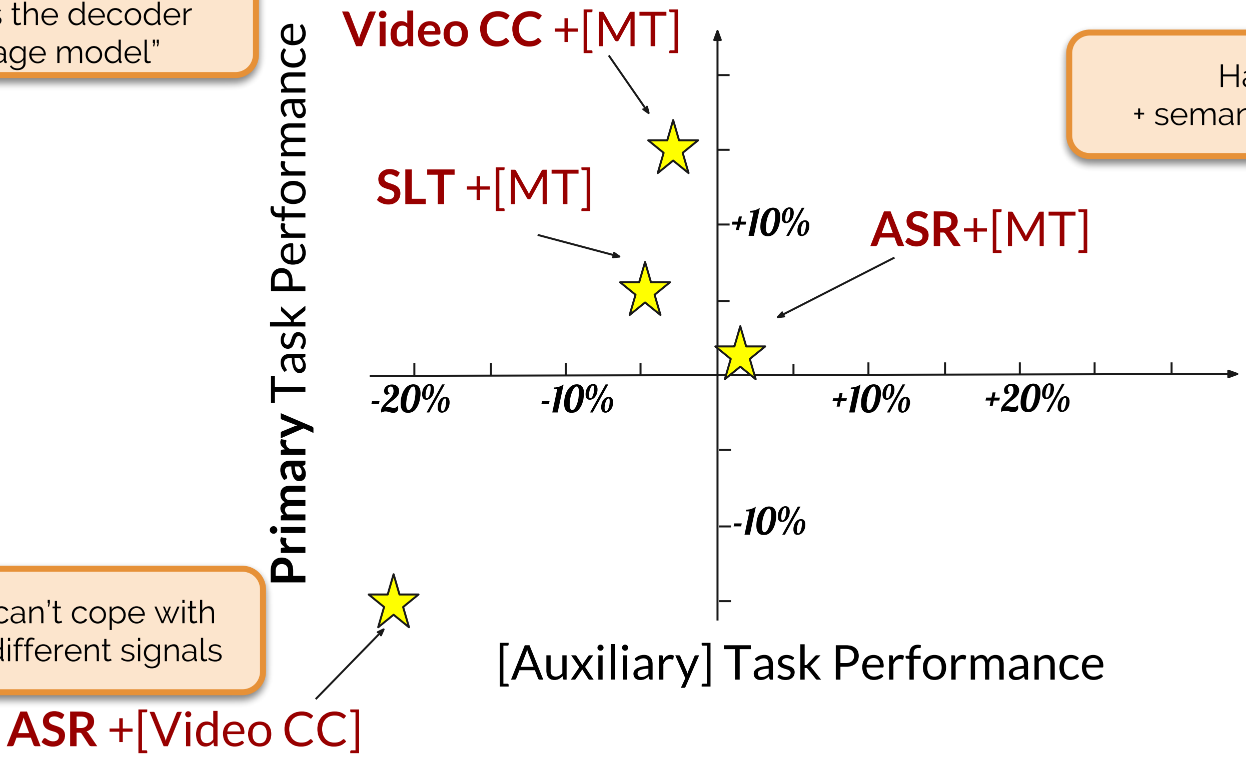
Summary

Conclusion: when is MTL useful?

Extremely hard task
MT helps the decoder
“language model”

Hard tasks
+ semantic parallelism

Decoder can’t cope with
two very different signals



Conclusion and Future Work

- Explored Multitask learning with different models
  - scheduling/shared space/mutual projection networks
  - Need more detailed analysis

- Can we cram multiple modalities into a sequence of vectors?
  - Can't be answered in a few weeks!
  - Need to study the behaviour of the Recurrent Shared Space
  - Plan: explore different architectures

- When does MPN regularisation help and why?
  - Few hints during this project, thorough investigation required
  - Plan: benchmark modality retrieval performance

Project Conclusions

Take home messages

- Multimodal ASR also works with S2S models
- Promising results for SLT & ASR
- Summarization works surprisingly well, need meaningful evaluation
- Region-specific MMT makes sense with the right evaluation
- CCA can obtain rich representations from diverse views and modalities
- MTL can be useful: potential gains $\propto$ semantic relatedness of the signals

We just need to keep trying!


GROUNDED

Thank you

Christian Fuegen

Publications

- Shruti Palaskar, Ramon Sanabria, and Florian Metze. End-to-end multi-modal speech recognition. In Proc. ICASSP, Calgary, Canada, 2018. IEEE.
- Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze. Visual features for context-aware speech recognition. In Proc. ICASSP, New Orleans, LA, 2017. IEEE.
- Yajie Miao and Florian Metze . Open-Domain Audio-Visual Speech Recognition: A Deep Learning Approach. In Proc. INTERSPEECH 2016. San Francisco, US, 2016. ISCA.
- Ozan Caglayan, Loïc Barrault, Fethi Bougares. Multimodal attention for neural machine translation. In arXiv 1609.03976.
- Caglayan, Ozan, et al. LIUM-CVC Submissions for WMT17 Multimodal Translation Task. In Proc. WMT, Copenhagen, Denmark, 2017.
- Jindřich Libovický, Jindřich Helcl,. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In Proc. ACL, Vancouver, Canada, 2017.
- Desmond Elliott, Stella Frank, Loic Barrault, Fethi Bougares, and Lucia Specia. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In Proc. WMT, Copenhagen, Denmark, 2017.

References

- Shoou-I Yu, Lu Jiang, and Alex Hauptmann. Instructional Videos for Unsupervised Harvesting and Learning of Action Examples. In Proc. ACM MM, Orlando, FL; U.S.A., Nov 2014. ACM.
- Alayrac, Jean-Baptiste, et al. "Unsupervised learning from narrated instruction videos." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
- Hara, K., Kataoka, H., & Satoh, Y. (2018, June). Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA.
- Hotelling, H., Relations between two sets of variants (1936)
- Wang, W., Arora, R., Livescu, K. & Bilmes, J. On deep multi-view representation learning: objectives and optimization (2016)
- Benton, A., Khayrallah, H., Gujral, B., Reisinger, D. A., Zhang, S., Arora, R., Deep generalized canonical correlation analysis (2017)
- Arora, S., Liang, Y., Ma, T., A Simple but Tough-to-Beat Baseline for Sentence Embeddings, ICLR 2017.
- Rich Caruana, Multitask Learning. 1998. Ph.D Thesis, Carnegie Mellon University.

Schedule

- 1:30 - 1:45: Intro
- 1:45 - 2:10: ASR/SLT
- 2:10 - 2:35: Summarization
- 2:35 - 3:00: Region MT
- 3:00 - 3:15: Break
- 3:15 - 3:40: Multiview
- 3:40 - 4:05: Multitask
- 4:05 - 4:10: Summary