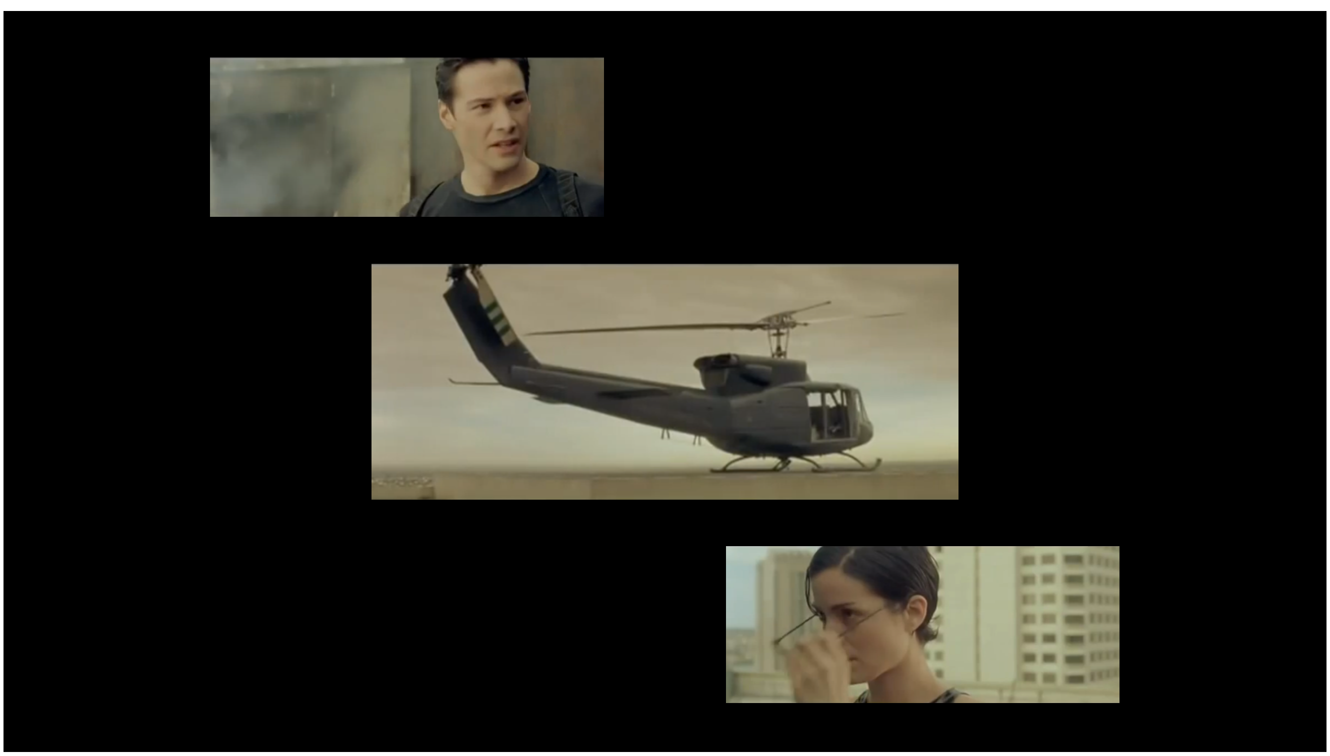
Open-Domain Audio-Visual Speech Recognition and Video Summarization

Presenter: **Florian Metze**

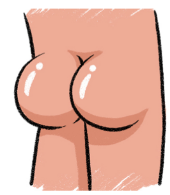
Hyderabad, September 7, 2018

1



Motivation

Understanding language is hard



3

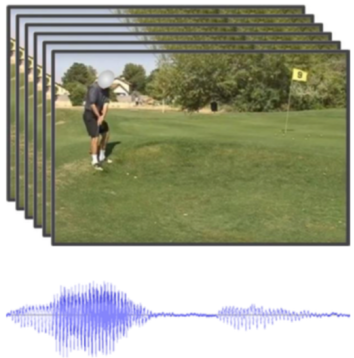
Motivation

Human information processing is inherently multimodal, and language is best understood in a situated context.
Machines should be able to jointly process multimodal data, and not just text, images, or speech in isolation.

- **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

4

Motivation - MT



I'm very close to the green but I didn't get it on the green so now I'm in this grass bunker.

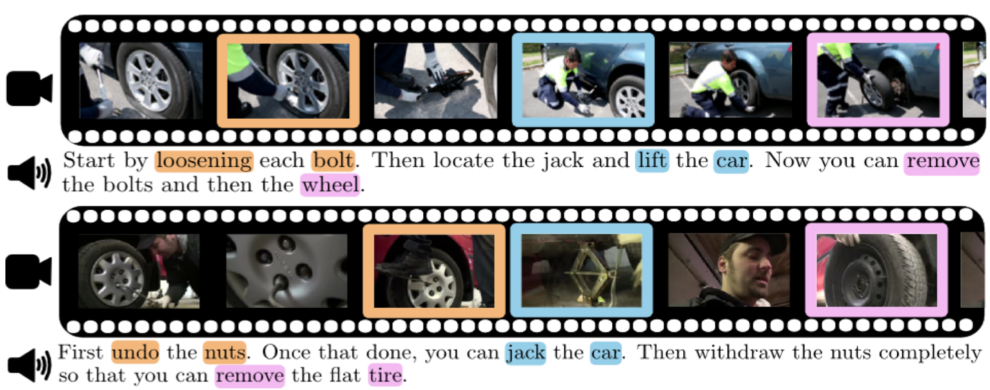
Eu estou muito perto do green, mas eu no pus a bola no green, ento agora estou neste bunker de grama.

- "green" is the correct term for the area, also in Portuguese
- You need "world knowledge" or "context information" in order to correctly interpret or translate this sentence
- In August 2018, both Google and Microsoft translate "green" incorrectly as "verde"
- Similar problem with summarization – any type of language understanding

5

Motivation - ASR

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



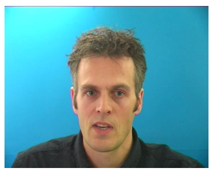
From [Alayrac et al., 2016]

6

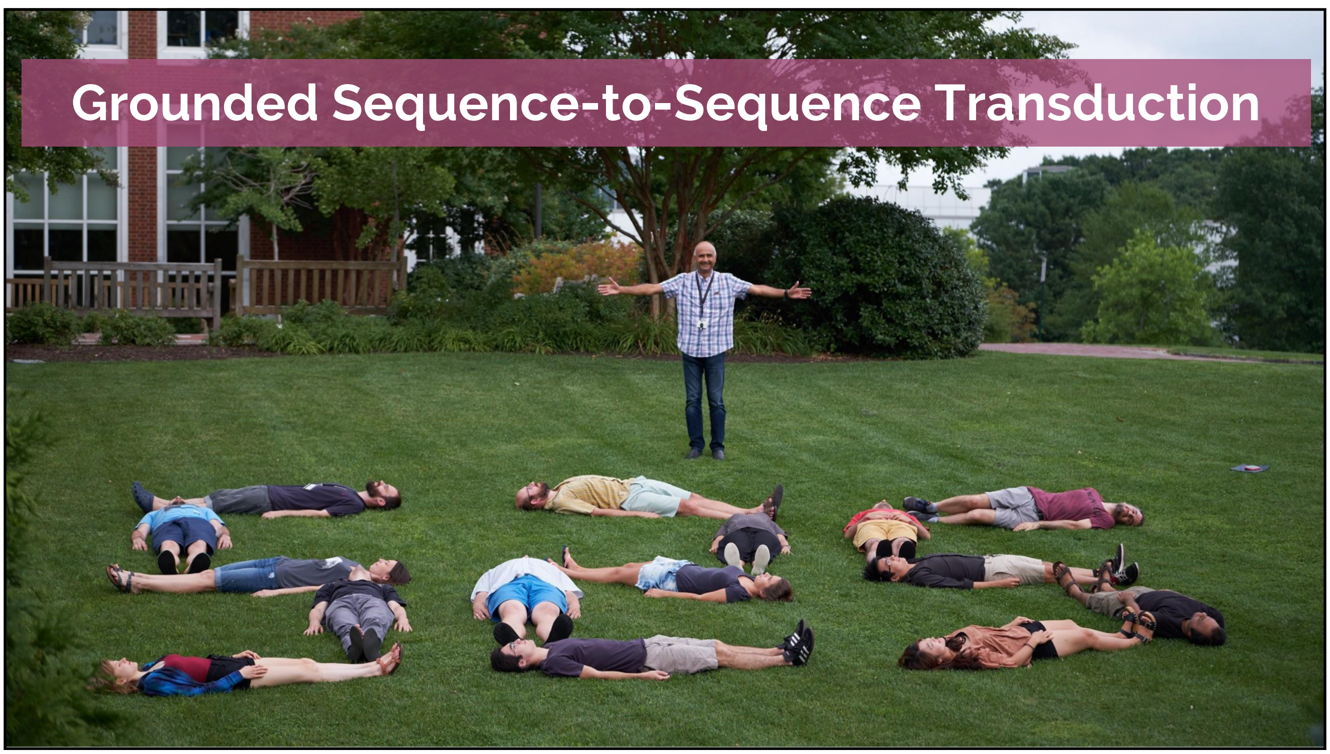
Audio-Visual ASR vs Multi-modal ASR

- Traditional audio-visual ASR based on speakers' lip/ mouth movement
  - Synchronicity between the audio and video frames required, fusion a problem
  - End-to-end lip-reading somewhat popular recently
- Lip/mouth information not always available in open-domain videos
  - Humans are usually present, but often they "do things"

e.g. AVASR "Grid" Corpus

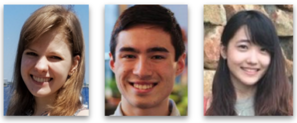


"Open-Domain" Video



Team

Undergraduate Students

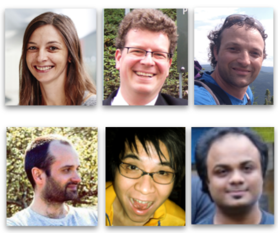


Alissa Ostapenko - WPI
Karl Mulligan - Rutgers
Sun Jae (Jasmine) Lee - UPenn

Graduate Students

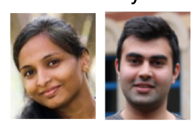
Jindrich Libovicky - Charles
Ramon Sanabria - CMU
Shruti Palaskar - CMU
Nils Holzenberger - JHU
Amanda Duarte - UPC
Ozan Caglayan - Le Mans

Senior Researchers



Lucia Specia - Sheffield
Florian Metze - CMU
Loïc Barrault - Le Mans
Des Elliott - Edinburgh / Copenhagen
Josiah Wang - Sheffield
Pranava Madhyastha - Sheffield

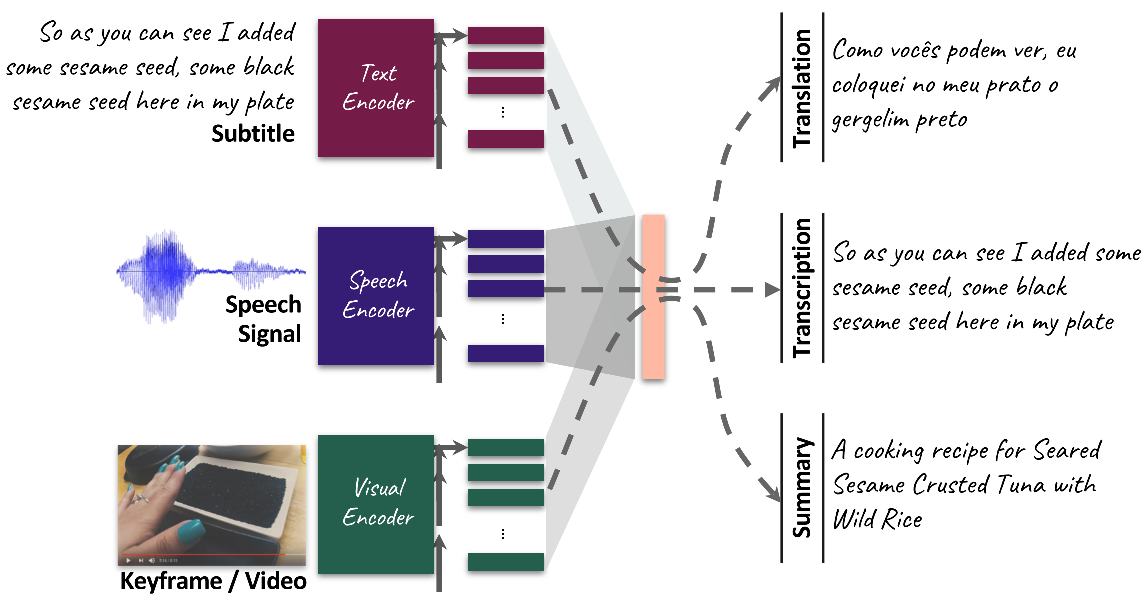
Remotely



Spandana Gella - Edinburgh
Chiraag Lala - Sheffield

9

The Big Picture

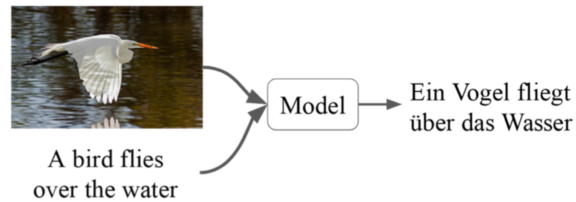


11

Before JSALT…

Multimodality useful for MT, but Multi-30k data not really “hard”

#	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)

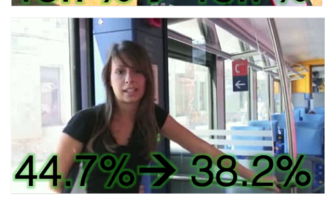
12

Before 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

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



13

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 summaries 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

➢ https://www.clsp.jhu.edu/workshops/18-workshop/

14

Highlights

~13K lines of code added

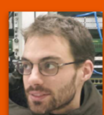
nmtpytorch

- 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

https://github.com/srvk/nmtpy-jsalt

16

Dataset and Features

Florian, Ramon

17

Dataset

- 2,000 hours how-to video corpus looked promising
  - Harder than previous MT data
  - ASR baselines available, some "quality" metrics defined (480h "good")
  - Harvested from on-line sources
  - Youtube Standard License applies (same as AudioSet, Youtube 8M)
- Dataset & code will be made available
  - Just submitted dataset description paper
- For now, contact me: fmetze@cs.cmu.edu

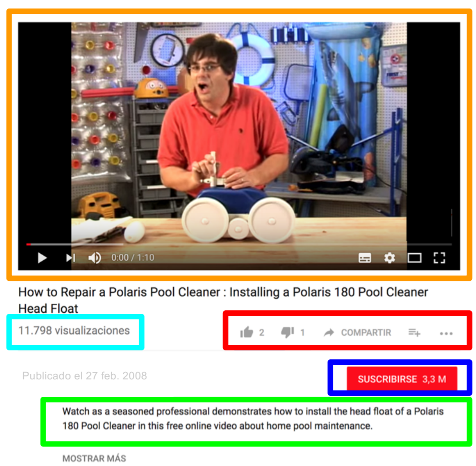
18

Dataset – Example (https://youtu.be/BJebuYFoRis)

19

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



20

Dataset - Translation

- For MT experiments, we need translation into non-English languages
  - Started with Portuguese, also have some Turkish data
- Crowd-sourced translations using CrowdFlower (figure eight)
  - Settled on post-editing of Google MT outputs
  - Gets improvement of ~1 BLEUE point
- At beginning of workshop, had **300h** available; continuing to 480h
  - Screenshot of annotation interface

21

Dataset – Topics

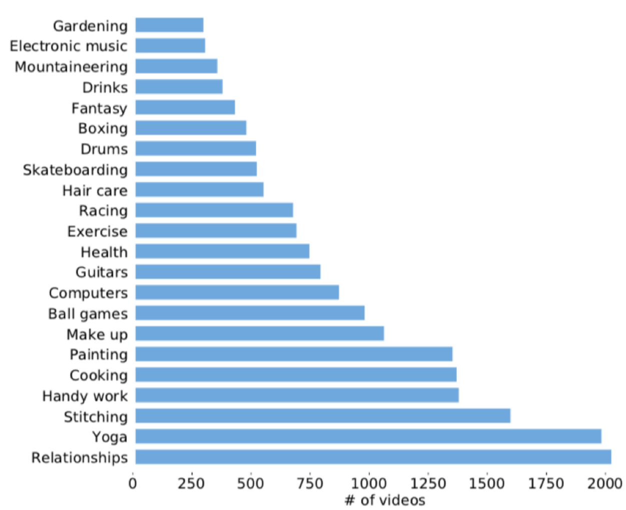


Figure 4: Distribution of topics in the *300h* subset as given by an LDA model trained on complete video subtitles.

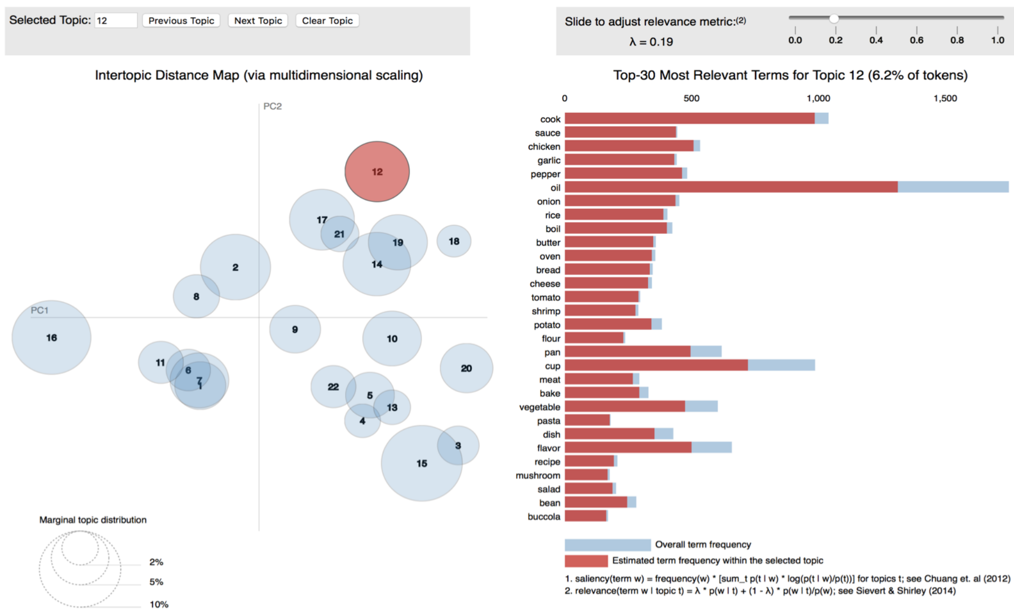
- Most ASR numbers will be on 300h subset, could use 480h. Summarization uses 2000h.

Split	Sentences	Videos	Hours
train	185011	13168	298.5
val	2022	150	3.2
test	2361	175	3.8
held_out	2021	169	3.0

Table 1: Sentence-level statistics for *300h* subset.

22

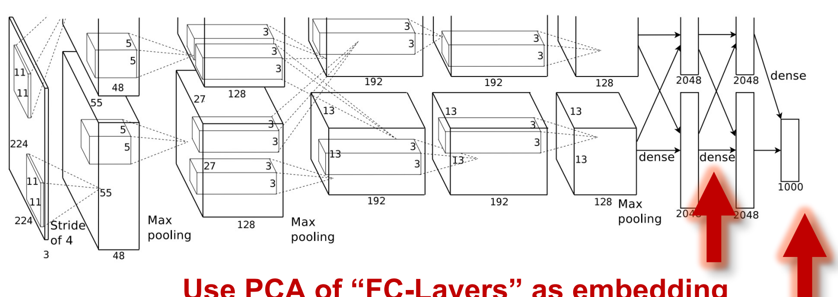
Topics in How-To Videos (LDA on Transcripts)



23

"Semantic Indexing" CNN Features

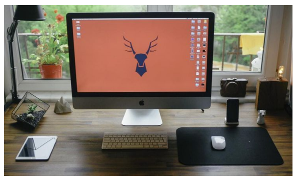
- ImageNet/ AlexNet [Krizhevsky et al., 2012]



Use PCA of "FC-Layers" as embedding

Three Types of Features

- Object Features



- monitor, mouse, keyboard, ...
- 1000 classes [Deng et al., 2009]

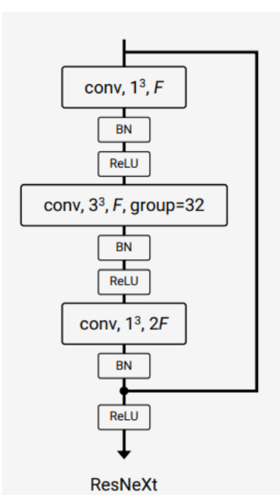
- Place Features (Scenes)

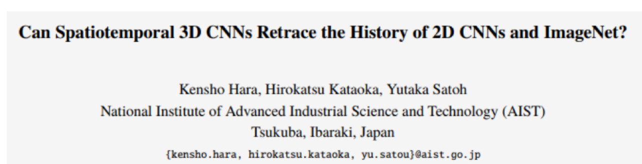
- train (office, baseball field, airport apron, ...)
- 205 classes [Zhou at al., 2014]

- Default approach: randomly extract one static frame per time-aligned “utterance”

25

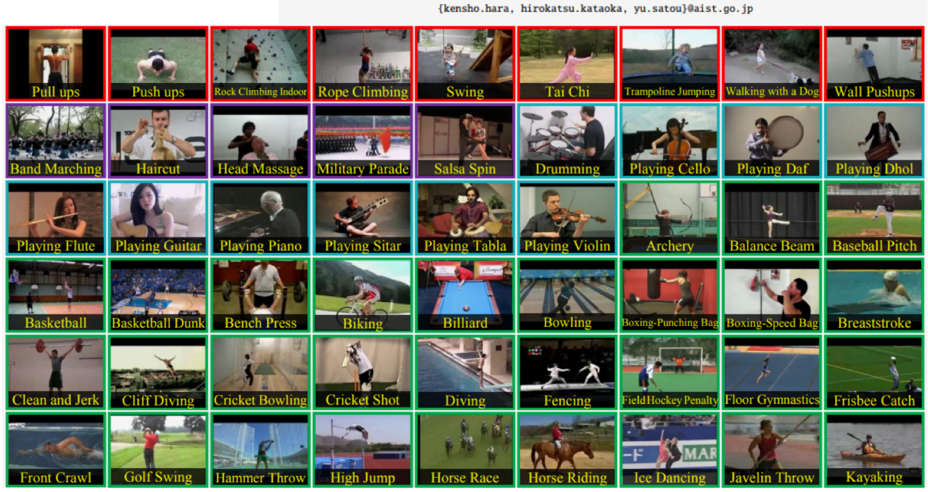
Action-level Video Features [Hara et al., 2018]





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



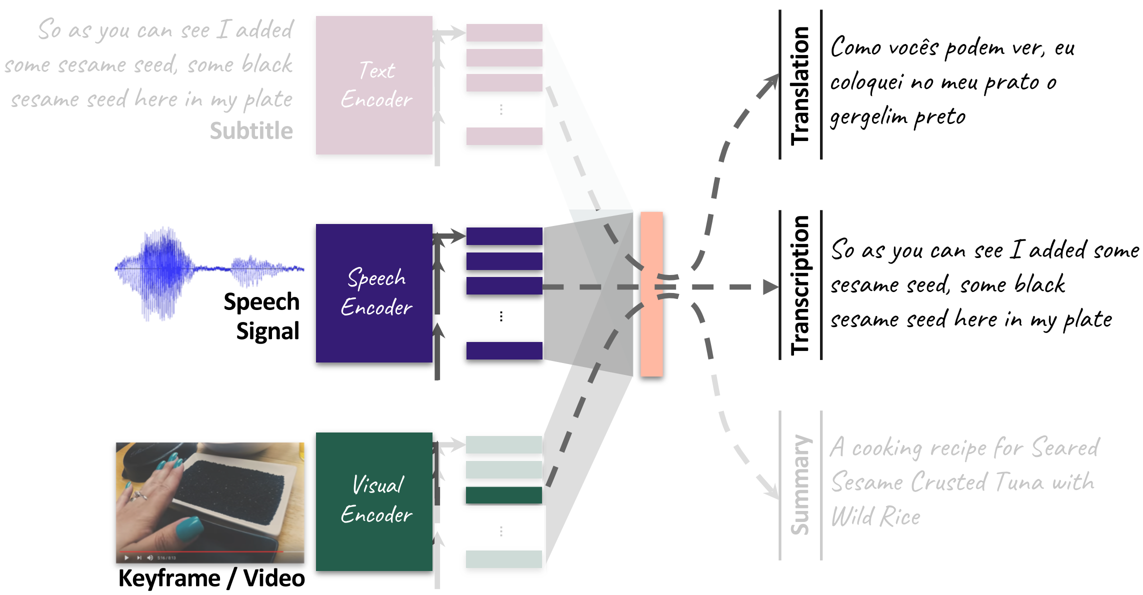
26

Automatic Speech Recognition Spoken Language Translation

Florian, Jindrich, Ozan, Ramon, Shruti

28

The big picture



29

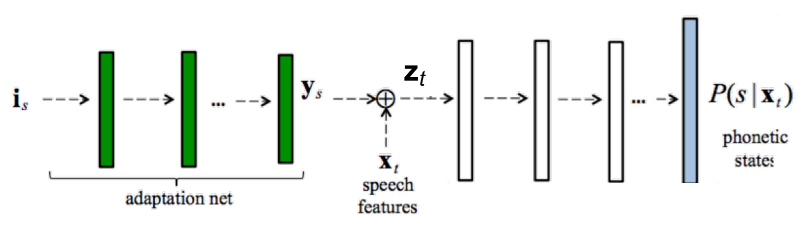
Related & Previous Results

- 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**
- Used **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)

30

Adapting a DNN Acoustic Model

- A General Adaptation Framework



- All is standard error back-propagation
- Independent of the structure & features, context
  - SAT technique can be naturally applied to CNNs, RNNs
  - Also tried: speaker microphone distance, speaker features (age, gender, race; 61-dimensional) [Miao et al., 2016]

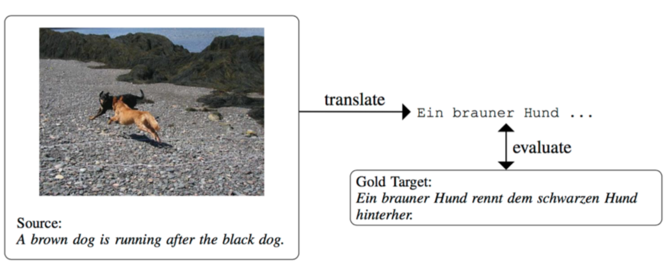
Comparison of Approaches

- Compare with 100d speaker i-Vectors
- Combine place/ object features, add speaker features to get 161-dim visual feature (with PCA)

Model	Features	WER(%)	
DNN (Baseline)	-----	23.4	
Adaptive Training	161-dim visual features	22.3	↓4.7%
Adaptive Training	100-dim speaker i-vectors	22.0	↓6.0%
Adaptive Training	261-dim fused features	21.5	↓8.1%

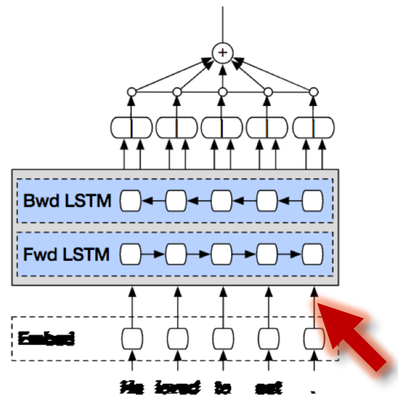
Language Model Adaptation

- Context aware language models easy with RNNs
  - [Zweig et al., 2012; ...]
  - Append context vector to word embeddings
- NMT of image captions [Specia et al., 2016]



LSTM Language Model

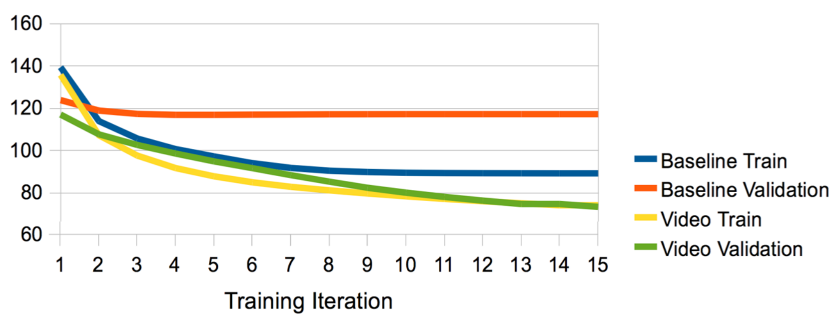
- Context-aware [Zweig et al., 2012; ...]
- Trained on 480h of transcriptions, optimized with 5-fold CV
- 2 BiLSTM layers, 1024 cells, Adagrad
- 1000d input vector consisting of
  - Learned 900d word embedding for vocabulary (~20k)
  - Context projected down to 100 dimensions
- 18 words sentence length on average (quite long!)



https://smerity.com/articles/2016/google_nmt_arch.html

Bi-LSTM LM (5-fold CV)

Loss (~PPL) of NNLM: 89 ➝ 74



- 30-best lists from 23.4% WER DNN baseline
  - Re-score and re-rank with LSTM-LM

➢22.6% WER (15.6% Oracle WER)
  - Small but consistent improvements

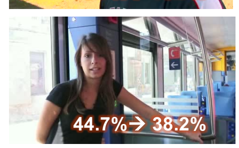
Analysis on 4h Test Set (156 Videos)

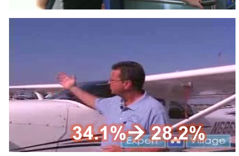
- Baseline: 23.4% WER with DNN
- AM Adaptation: 22.3% (object & place features)
- LM Adaptation: 22.6% (object & place features)
- AM+LM: ~21.5% WER with rescoring

- Almost 10% rel. improvement over reasonable HMM-DNN baseline

Result Analysis – “indoor” vs “outdoor”

- Using object and place features only
- LM adaptation improves results across the board
    - 126/ 156 videos improve
- AM improves “noisy” videos
    - 55/ 156 videos improve (most are “outdoor”, according to their category)



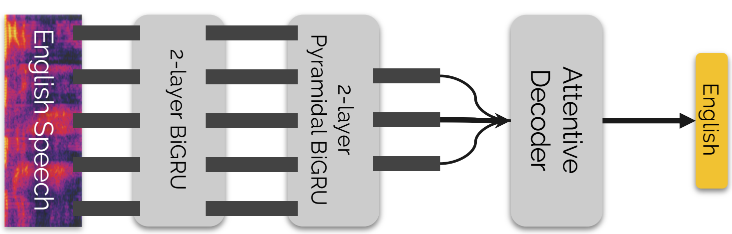


Video Category	WER% of the baseline DNN	WER% of the DNN with place features
typical indoor	22.1	21.7
other	27.6	25.7

Multimodal S2S ASR

38

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.

40

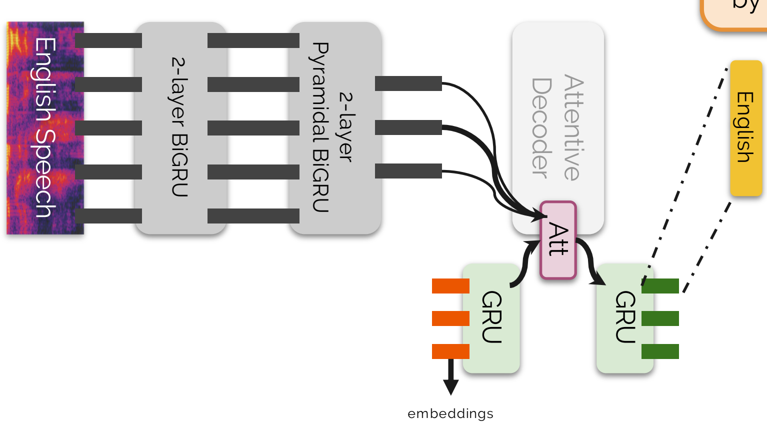
Multimodal ASR: Motivations

- **Decoder-side Integration**: improve the LM by providing visual context?
  - Action-level **global** visual features
- **Attention Integration**: can we benefit from multimodal attention?
  - Let the model learn when to pay attention to multiple modalities
  - Action-level **temporal** visual features
- **Encoder-side Integration**: like feature shift
- In early experiments, “action” features seemed to outperform others

41

Integration of Features

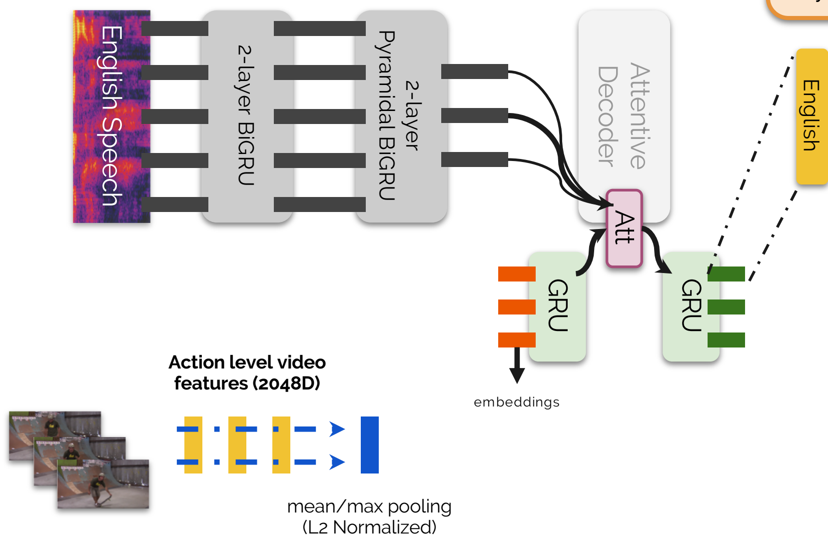
Motivation: Can we improve decoder LM by visual grounding?



42

Integration of Features

Motivation: Can we improve decoder LM by visual grounding?

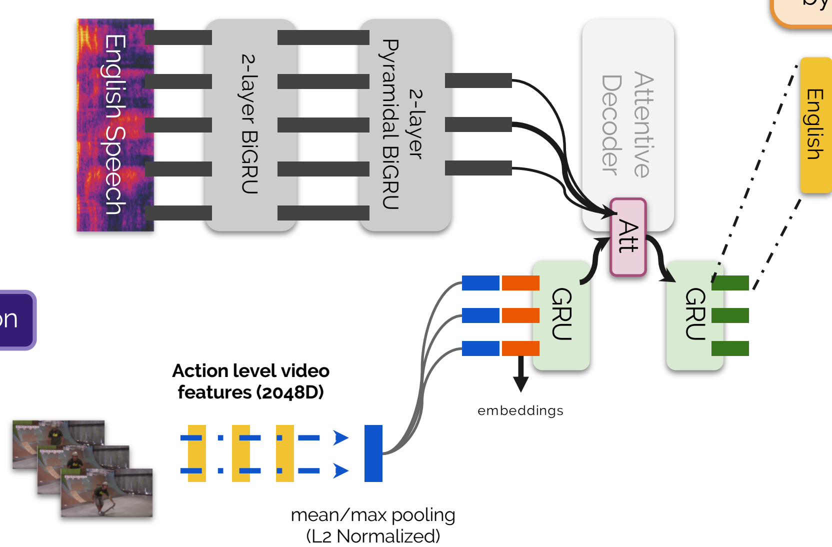


43

Integration of Features

Motivation: Can we improve decoder LM by visual grounding?

Concatenation

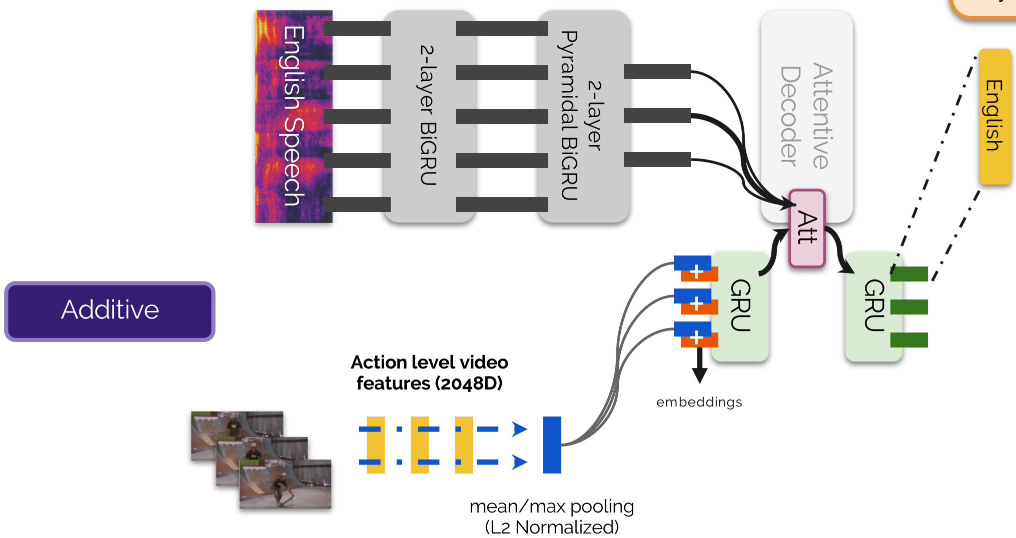


[Caglayan et al. 2017]

44

Integration of Features

Motivation: Can we improve decoder LM by visual grounding?

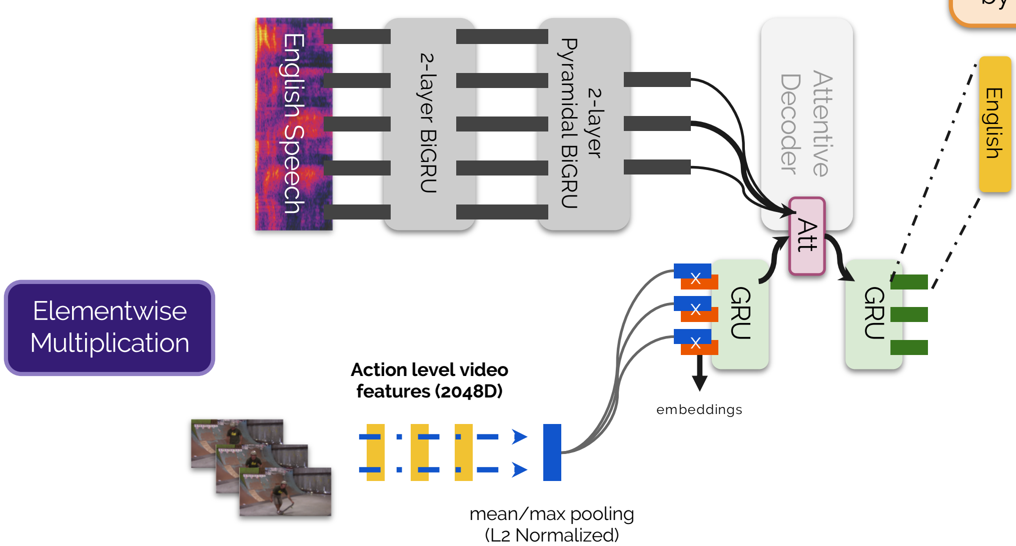


[Caglayan et al. 2017]

45

Integration of Features

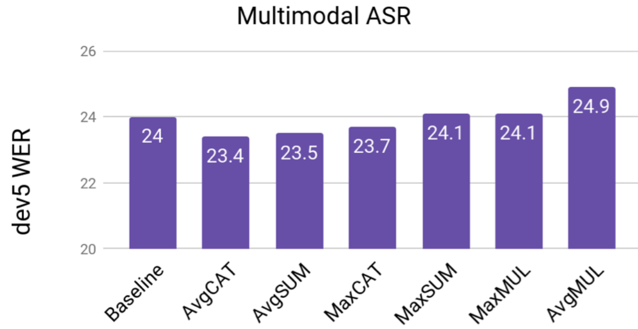
Motivation: Can we improve decoder LM by visual grounding?



[Caglayan et al. 2017]

46

Decoder-side Interaction

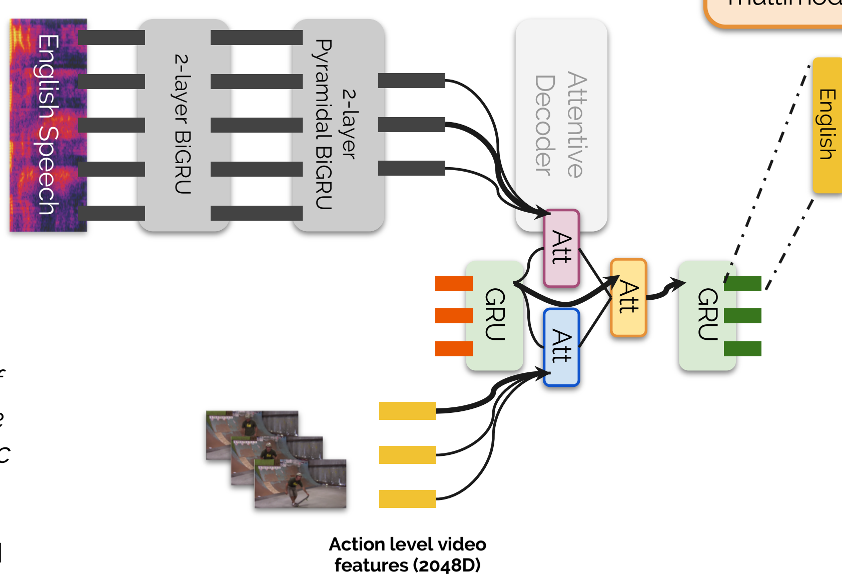


- Previous work
  - LM benefits from visual adaptation in terms of PPL [Gupta et al., 2018]
  - Visual features improve acoustic modeling in HMM [Miao & Metze, 2016]
- Hard to conclude for S2S models
  - Need to experiment with bigger models and different features

47

Hierarchical Attention

Motivation: Can we benefit from selective multimodal attention?



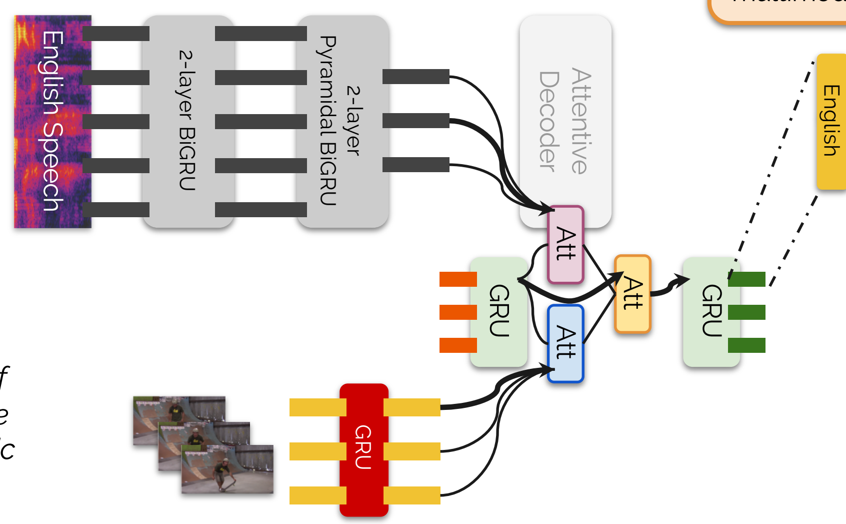
Another layer of attention to fuse modality-specific contexts.

[Libovický et al. 2017]

48

Hierarchical Attention + ActionGRU

Motivation: Can we benefit from selective multimodal attention?



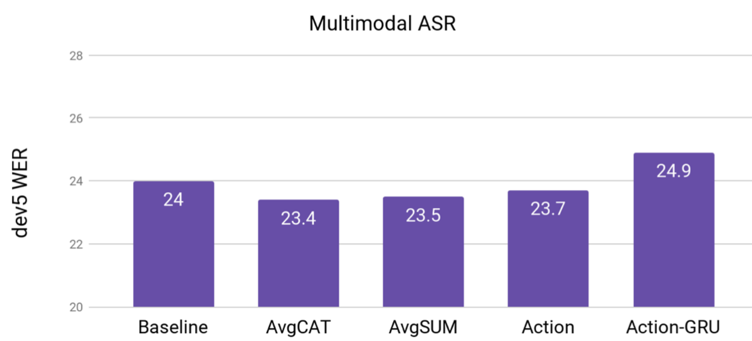
Action level video features (2048D)

Another layer of attention to fuse modality-specific contexts.

[Libovický et al. 2017]

49

Hierarchical Attention

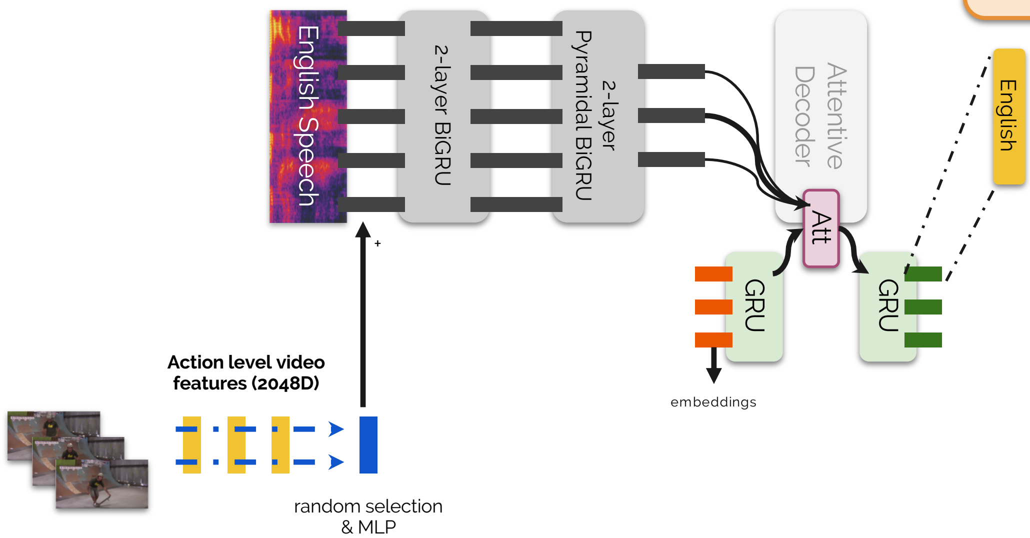


- `AvgCAT/AvgSUM/Action` are comparable: needs further exploration
- Encoding temporal action features with an RNN hurts WER
  - Reason → the model shifts attention

50

Integration of Features

Motivation: Can we adapt the features?



51

Encoder-Side Integration

- Integrate linear feature shift approach before main encoder
- Random selection of frame rather than pooling
- Action features (rather than object, scene)

	Params	val WER ⇓	test WER ⇓
S2S ASR	13.7M	19.1	20.0
S2S MMASR	13.8M	18.0	18.7

Table 3: Comparison of monomodal and multimodal ASR.

52

Multimodal ASR and SLT Conclusions

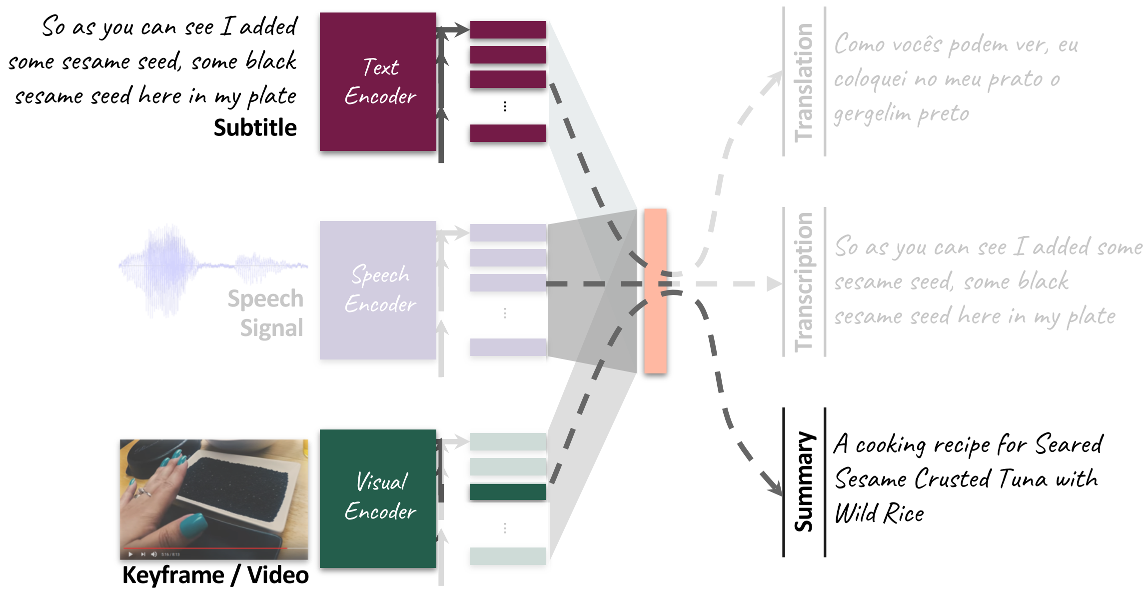
- Multimodal ASR with S2S Models
  - Seeing nice improvements over baseline(s)
  - Decoder side improvements consistent with previous work
  - Further exploration: Temporal smoothing of visual features, …
  - Further analysis (shared representations?) required

- 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

Florian, Jasmine, Jindrich, Shruti, Spandana

The big picture



66

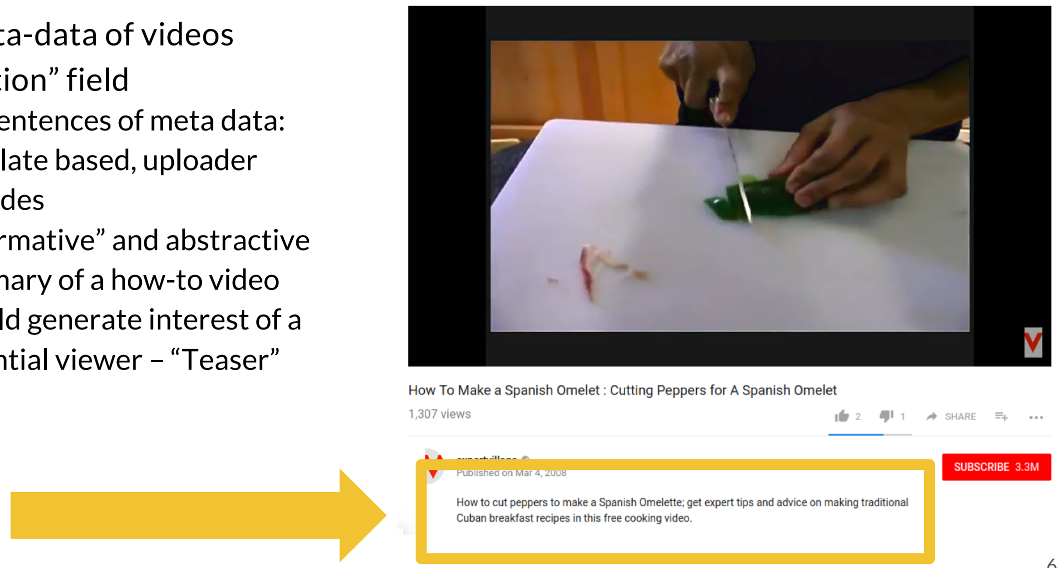
Summarization

- Present (subset of) information in shorter form
  - Maybe across modalities
- Can be abstractive or extractive
  - Generate "new" phrasing or content
- Evaluation is hard
  - Task dependent
  - Or use ROUGE/ BLEU like metrics to measure precision/ recall

67

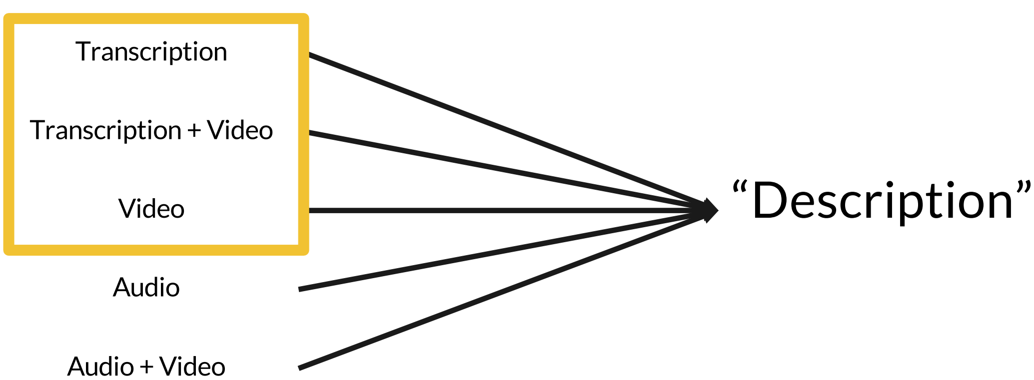
Summarization or Description Generation

- Have meta-data of videos
- “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 – “Teaser”



68

General Experimental Setup



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

69

Spanish Omelet

~1.5 minutes of audio and video

Description (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 description

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 descriptions

```
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)
    - Removed catchphrases
  - 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

74

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

75

Evaluation Metrics

Catchphrases in descriptions

```
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

76

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

77

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

78

Do we need the complete transcript?

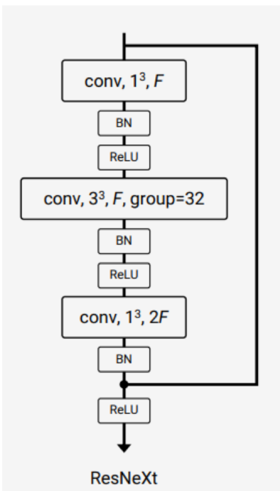
	Rouge-L	Content F1
No input = Language model	27.5	8.3
Extracted sentence (itself 18.8 F1 points)	46.6	36.0
First 200 tokens	40.3	27.5
Complete transcript (up to 650 tokens)	53.9	47.4

80

Action Recognition Features

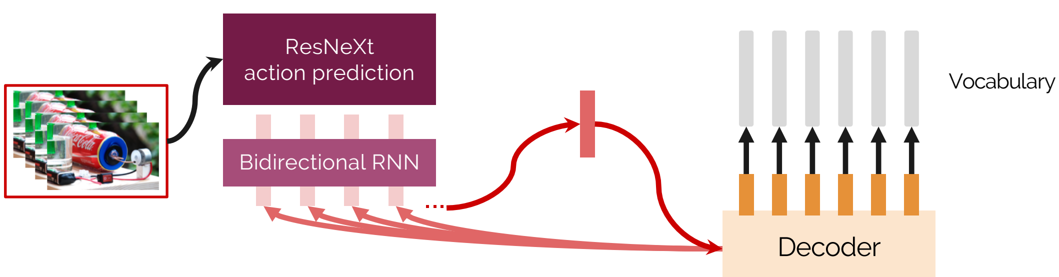
Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?

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



81

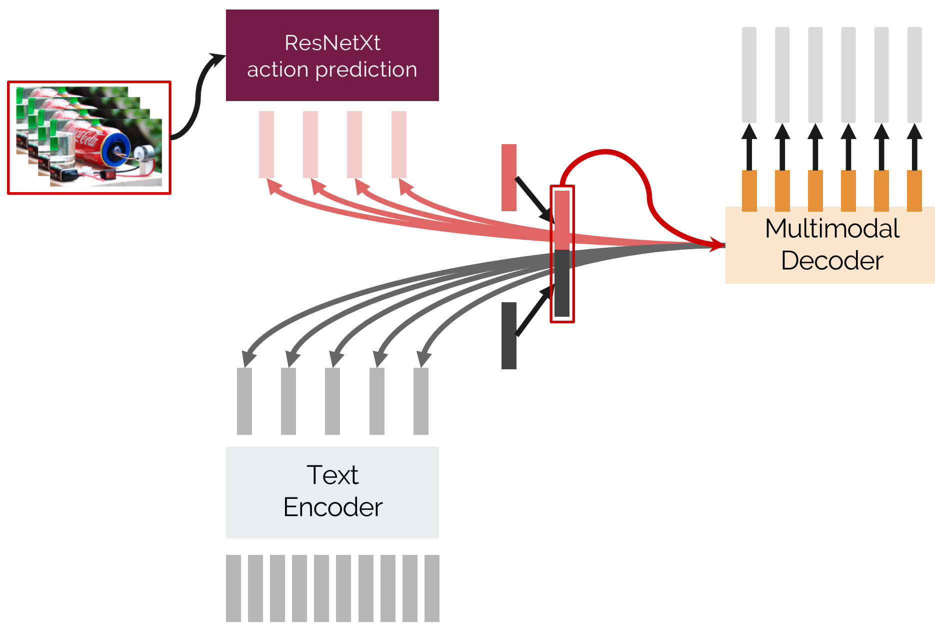
Video Features as Input



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

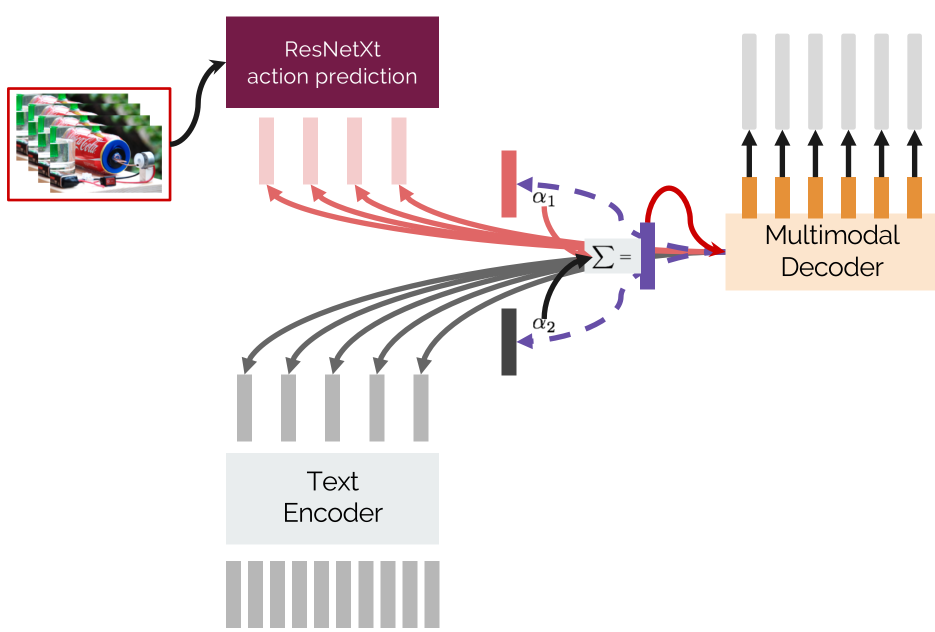
82

Context-vector Concatenation



83

Hierarchical Multi-modal Attention



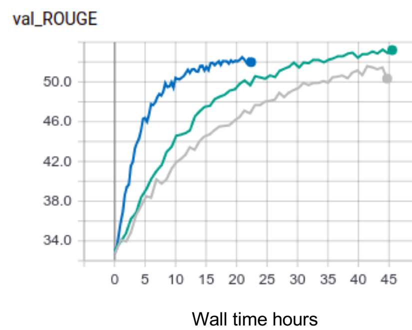
84

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



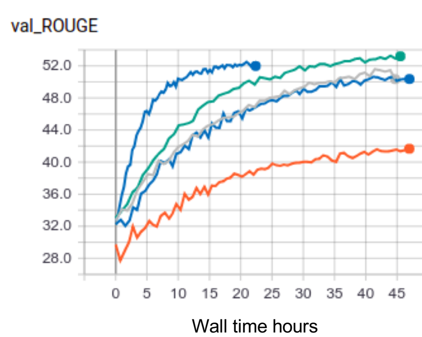
85

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



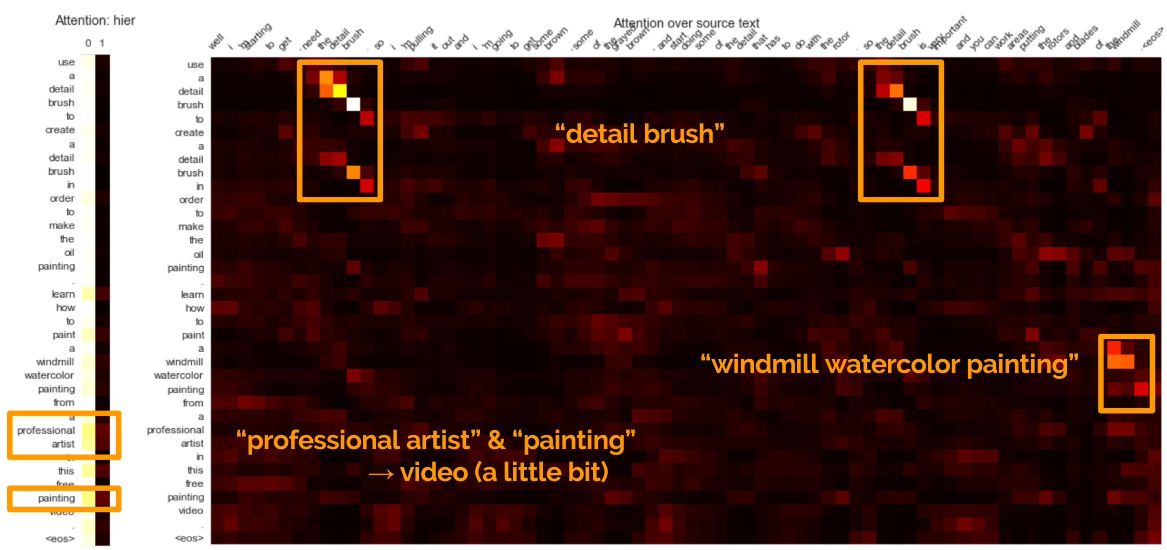
86

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

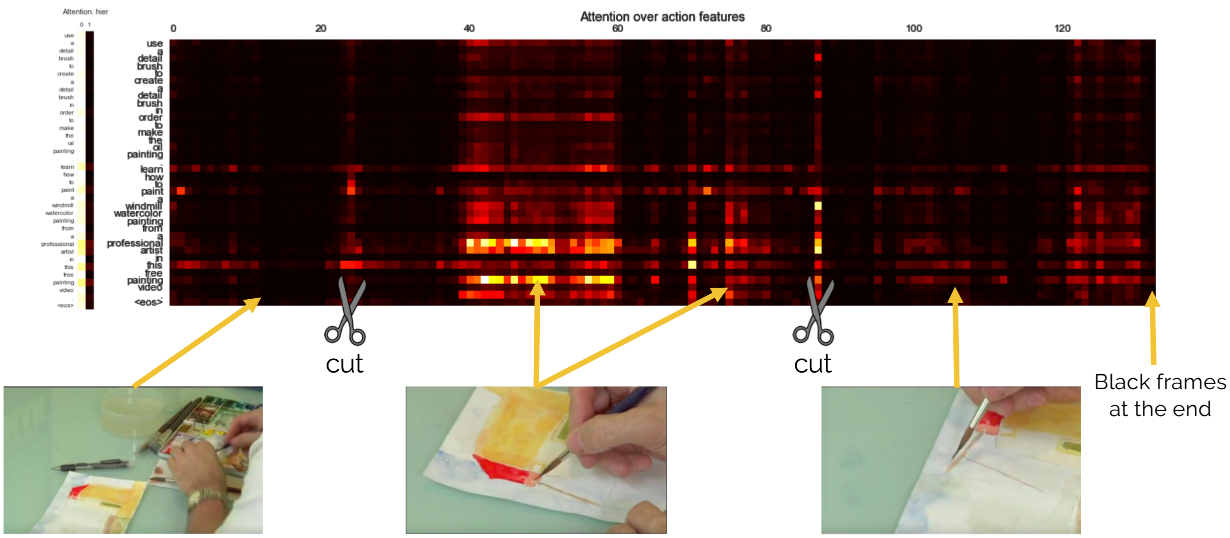
87

Attention over the Transcriptions



88

Attention over the Video Features



Talking and preparing the brush

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

Close-up of brushstrokes **no hand**

89

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

0:36 / 0:51

90

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 every-thing you need to know to make your own ceviche in this free video .	Content F1 25

1:38 / 1:50

91

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

92

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
  - 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

93

Summarization Conclusion

- It works! Kind of.
- *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

94

Wrap-Up

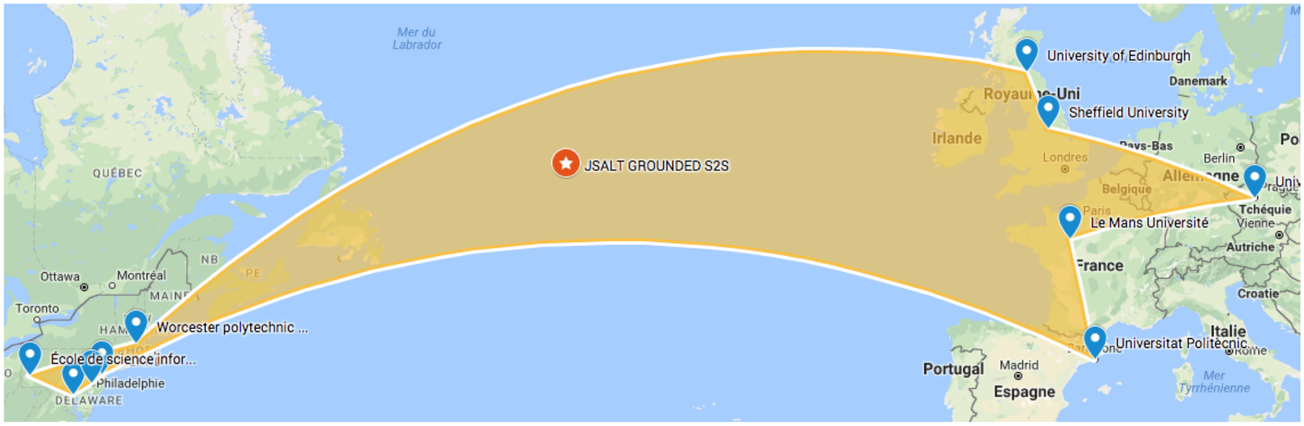
Take-Home Messages

- Interesting insights and comparisons across tasks
- Promising results for SLT & ASR, improved performance
- 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

More to come (data and code)

https://github.com/srvk/nmtpy-jsalt

Thank you



Christian Fuegen

97

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.

98

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.

Backup