Open-Domain Audio-Visual Speech Recognition and Video Summarization

Presenter: Florian Metze

Hyderabad, September 7, 2018

Can you fly this thing?

Not yet.

[…]

Let's go!
Motivation

Understanding language is hard

Motivation

- **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

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.
Motivation - MT

- “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

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]
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

Grounded Sequence-to-Sequence Transduction
Team

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

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

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

Remotely
- Spandana Gella - Edinburgh
- Chiraag Lala - Sheffield

The Big Picture

So as you can see I added some sesame seed, some black sesame seed here in my plate.

Translation
Como vocês podem ver, eu coloquei no meu prato o gergelim preto

Transcription
So as you can see I added some sesame seed, some black sesame seed here in my plate

Summary
A cooking recipe for Seared Sesame Crusted Tuna with Wild Rice
Before JSALT...

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

<table>
<thead>
<tr>
<th>#</th>
<th>Raw</th>
<th>z</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>77.8</td>
<td>0.665</td>
<td>LIUMCVC_MNMT.C</td>
</tr>
<tr>
<td>2</td>
<td>74.1</td>
<td>0.552</td>
<td>UvA-TiCC.IMAGINATION.U</td>
</tr>
<tr>
<td>3</td>
<td>70.3</td>
<td>0.437</td>
<td>NICT_NMTrerank.C</td>
</tr>
<tr>
<td></td>
<td>68.1</td>
<td>0.325</td>
<td>CUNI.NeuralMonkey.TextualMT.U</td>
</tr>
<tr>
<td></td>
<td>68.1</td>
<td>0.311</td>
<td>DCU-ADAPT.MultiMT.C</td>
</tr>
<tr>
<td></td>
<td>65.1</td>
<td>0.196</td>
<td>LIUMCVC_NMT.C</td>
</tr>
<tr>
<td></td>
<td>60.6</td>
<td>0.136</td>
<td>CUNI.NeuralMonkey.MultimodalMT.U</td>
</tr>
<tr>
<td></td>
<td>59.7</td>
<td>0.08</td>
<td>UvA-TiCC.IMAGINATION_C</td>
</tr>
<tr>
<td></td>
<td>55.9</td>
<td>-0.049</td>
<td>CUNI.NeuralMonkey.MultimodalMT.U</td>
</tr>
<tr>
<td></td>
<td>54.4</td>
<td>-0.091</td>
<td>OREGONSTATE_2NeuralTranslation.C</td>
</tr>
<tr>
<td></td>
<td>54.2</td>
<td>-0.108</td>
<td>CUNI.NeuralMonkey.TextualMT.C</td>
</tr>
<tr>
<td></td>
<td>53.3</td>
<td>-0.144</td>
<td>OREGONSTATE_1NeuralTranslation.C</td>
</tr>
<tr>
<td></td>
<td>49.4</td>
<td>-0.266</td>
<td>SHEF_ShefClassProj.C</td>
</tr>
<tr>
<td></td>
<td>46.6</td>
<td>-0.37</td>
<td>SHEF_ShefClassInitDec.C</td>
</tr>
</tbody>
</table>

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

(Elliott et al., 2017; Palaskar et al., 2018)
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/

Highlights

- **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-torch

- ~13K lines of code added
Dataset and Features

Florian, Ramon

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

- 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 - 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

Dataset – Topics

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

<table>
<thead>
<tr>
<th>Split</th>
<th>Sentences</th>
<th>Videos</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>185011</td>
<td>13168</td>
<td>298.5</td>
</tr>
<tr>
<td>val</td>
<td>2022</td>
<td>150</td>
<td>3.2</td>
</tr>
<tr>
<td>test</td>
<td>2361</td>
<td>175</td>
<td>3.8</td>
</tr>
<tr>
<td>held.out</td>
<td>2021</td>
<td>169</td>
<td>3.0</td>
</tr>
</tbody>
</table>

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

Figure 4: Distribution of topics in the 300h subset as given by an LDA model trained on complete video subtitles.
**Topics in How-To Videos (LDA on Transcripts)**

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

**“Semantic Indexing” CNN Features**

- **ImageNet/AlexNet**
  
  1. Use PCA of “FC-Layers” as embedding.
Three Types of Features

- **Object Features**
  - monitor, mouse, keyboard, ...
  - 1000 classes [Deng et al., 2009]
  - Default approach: randomly extract one static frame per time-aligned “utterance”

- **Place Features (Scenes)**
  - train (office, baseball field, airport apron, ...)
  - 205 classes [Zhou et al., 2014]

Action-level Video Features [Hara et al., 2018]
Automatic Speech Recognition
Spoken Language Translation

Florian, Jindrich, Ozan, Ramon, Shruti

The big picture

So as you can see I added some sesame seed, some black sesame seed here in my plate

Subtitle

Como vocês podem ver, eu coloquei no meu prato o gergelim preto

Translation

So as you can see I added some sesame seed, some black sesame seed here in my plate

Transcription

A cooking recipe for Seared Sesame Crusted Tuna with Wild Rice

Summary
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)

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)

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>WER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN (Baseline)</td>
<td>-----</td>
<td>23.4</td>
</tr>
<tr>
<td>Adaptive Training</td>
<td>161-dim visual features</td>
<td>22.3</td>
</tr>
<tr>
<td>Adaptive Training</td>
<td>100-dim speaker i-vectors</td>
<td>22.0</td>
</tr>
<tr>
<td>Adaptive Training</td>
<td>261-dim fused features</td>
<td>21.5</td>
</tr>
</tbody>
</table>

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://smarthy.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)

<table>
<thead>
<tr>
<th>Video Category</th>
<th>WER% of the baseline DNN</th>
<th>WER% of the DNN with place features</th>
</tr>
</thead>
<tbody>
<tr>
<td>typical indoor</td>
<td>22.1</td>
<td>21.7</td>
</tr>
<tr>
<td>other</td>
<td>27.6</td>
<td>25.7</td>
</tr>
</tbody>
</table>
Multimodal S2S ASR

S2S ASR Baseline

<table>
<thead>
<tr>
<th></th>
<th># of Params</th>
<th>Tokens</th>
<th>cv05 WER</th>
<th>dev5 WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR</td>
<td>4.3M</td>
<td>SentPiece-5K</td>
<td>23.0</td>
<td>24.0</td>
</tr>
<tr>
<td>ASR w/ 6-layer BiLSTM encoder</td>
<td>13.7M</td>
<td>SentPiece-5K</td>
<td>19.6</td>
<td>21.1</td>
</tr>
<tr>
<td>ESPNet 6-layer BiLSTM encoder</td>
<td>-</td>
<td>Char</td>
<td>19.6</td>
<td>19.8</td>
</tr>
</tbody>
</table>

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

We use a small ASR for faster experimental turnaround time.
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

Integration of Features

**Motivation**: Can we improve decoder LM by visual grounding?
Motivation: Can we improve decoder LM by visual grounding?

Integration of Features

Action level video features (2048D)

mean/max pooling (L2 Normalized)

[Concatenation]

[Caglayan et al. 2017]
Motivation: Can we improve decoder LM by visual grounding?

Additive

Elementwise Multiplication

Action level video features (2048D)

mean/max pooling (L2 Normalized)

[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

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?

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
**Integration of Features**

Motivation: Can we adapt the features?

**Encoder-Side Integration**

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

<table>
<thead>
<tr>
<th></th>
<th>Params</th>
<th>val WER</th>
<th>test WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2S ASR</td>
<td>13.7M</td>
<td>19.1</td>
<td>20.0</td>
</tr>
<tr>
<td>S2S MMASR</td>
<td>13.8M</td>
<td>18.0</td>
<td>18.7</td>
</tr>
</tbody>
</table>

Table 3: Comparison of monomodal and multimodal ASR.
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

So as you can see I added some sesame seed, some black sesame seed here in my plate.

Subtitle

[Diagram of text encoder, speech encoder, and visual encoder with arrows and connections to subtitle and translation]

Translation

Como vocês podem ver, eu coloquei no meu prato o gergelim preto.

So as you can see I added some sesame seed, some black sesame seed here in my plate.

Summary

A cooking recipe for Seared Sesame Crusted Tuna with Wild Rice.

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
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”

General Experimental Setup

- Transcription
- Transcription + Video
- Video
- Audio
- Audio + Video

“Description”

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

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 these 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 very very small. so we are going to dice the up also very very small. so we have small pieces of onions and peppers ready to go.
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

- 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)

>=500 times
**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**

- **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** = 16.4
- **Content F1** = 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.

<table>
<thead>
<tr>
<th></th>
<th>Rouge-L</th>
<th>Content F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>No input = Language model</td>
<td>27.5</td>
<td>8.3</td>
</tr>
<tr>
<td>Extracted sentence (itself 18.8 F1 points)</td>
<td>46.6</td>
<td>36.0</td>
</tr>
<tr>
<td>First 200 tokens</td>
<td>40.3</td>
<td>27.5</td>
</tr>
<tr>
<td>Complete transcript (up to 650 tokens)</td>
<td>53.9</td>
<td>47.4</td>
</tr>
</tbody>
</table>

Do we need the complete transcript?
**Action Recognition Features**

**Video Features as Input**

<table>
<thead>
<tr>
<th></th>
<th>Rouge-L</th>
<th>Content F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text-only input</td>
<td>53.9</td>
<td>47.4</td>
</tr>
<tr>
<td>Features only</td>
<td>38.5</td>
<td>24.8</td>
</tr>
<tr>
<td>Features + RNN</td>
<td>46.3</td>
<td>34.9</td>
</tr>
</tbody>
</table>
Context-vector Concatenation

Hierarchical Multi-modal Attention
Results of Attention Combination

- Modest improvements when we combine text and video

<table>
<thead>
<tr>
<th></th>
<th>Rouge-L</th>
<th>Content F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text-only input</td>
<td>53.9</td>
<td>47.4</td>
</tr>
<tr>
<td>Context vector concat</td>
<td>51.0</td>
<td>44.4</td>
</tr>
<tr>
<td>Hierarchical attention</td>
<td>54.9</td>
<td>48.9</td>
</tr>
</tbody>
</table>

Results of Attention Combination

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

<table>
<thead>
<tr>
<th></th>
<th>Rouge-L</th>
<th>Content F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text-only input</td>
<td>53.9</td>
<td>47.4</td>
</tr>
<tr>
<td>Context vector concat</td>
<td>51.0</td>
<td>44.4</td>
</tr>
<tr>
<td>+ RNN over actions</td>
<td>42.2</td>
<td>30.3</td>
</tr>
<tr>
<td>Hierarchical attention</td>
<td>54.9</td>
<td>48.9</td>
</tr>
<tr>
<td>+ RNN over actions</td>
<td>53.4</td>
<td>46.8</td>
</tr>
</tbody>
</table>
### Overview of the Result

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

### Attention over the Transcriptions

![Attention over the Transcriptions](image)
Attention over the Video Features

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.
  - 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.

- **Actions RNN**
  - learn the basics of hatha yoga with expert tips on headache relief in this free home improvement video.

- **Actions**
  - "your knees as much as"
**Example**

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.

Partial dentures will help to prevent dentures. Learn about partial dentures from a dentist in this free oral hygiene video.

Don't leave a home drug test. Learn about vacuum cleaners with expert tips from a dentist in this free oral hygiene video.

In order to make a 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.

---

**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.
  - Don't wrinkle your suit right after ironing it! Learn how to hang a jacket while ironing a man'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.
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

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
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

Publications

References

- Hotelling, H., Relations between two sets of variants (1936)