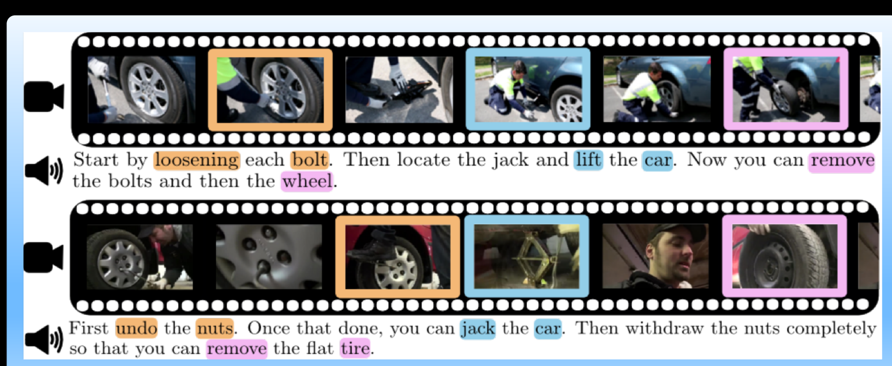
Grounded Sequence to Sequence Transduction
(Multi-Modal Speech Recognition)

Florian Metze
July 17, 2018

Language Technologies Institute
Carnegie Mellon University
School of Computer Science

Imagine How-To Videos



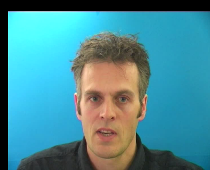
- Lots of potential for multi-modal processing & fusion
- For Speech-to-Text and beyond

Audio-Visual ASR vs Multi-modal ASR

- Traditional audio-visual ASR based on speakers' lip/ mouth movement
  - **Sub-phonetic synchronicity** required, fusion a problem
- Lip/ mouth information not always available in how-to videos
  - Humans are usually present, but often they "do things"
- Instead: fuse information at the **semantic level (words, …)**

e.g. AVASR "Grid" Corpus

"How-To" Video



Multi-Modal MT – Example

- **SRC**: Three children in football uniforms of two different teams are playing football on a football field, while another player and an adult stand in the background.
- **TXT**: Drei Kinder in Fußballtrikots zweier verschiedener Mannschaften spielen Fußball auf einem Fußballplatz während ein weiterer Spieler und eine Erwachsener im Hintergrund stehen.
- **IMG**: Drei Kinder in Footballtrikots zweier verschiedener Mannschaften spielen Football auf einem Footballplatz während ein weiterer Spieler und ein Erwachsener im Hintergrund stehen.

Courtesy of Lucia Specia

Two (+) Types of Features

- **Object Features**

- monitor, mouse, keyboard, ...
- 1000 classes [Deng et al., 2009]
- Could also do **Actions**, …

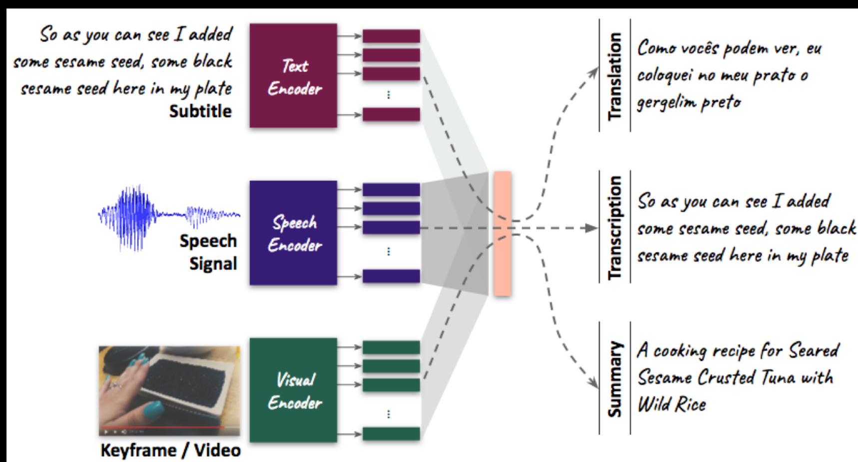
- **Place Features (Scenes)**

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

How-to Video Corpus [Miao et al., ‘14]

- “How-to” dataset of instructional videos
  - Harvested from the web **(2000h+ available)**
  - “Utterance” (from caption) is 8s…10s
  - On average 18 words
- ~55,000 videos
  - 300h+ have been translated into Portuguese
  - 4h dev & eval set; ~20k+ vocabulary size
- Extract one quasi-static visual “context” vector per utterance
  - Pick frame randomly (for now)
  - Object/ place detection, or action recognition

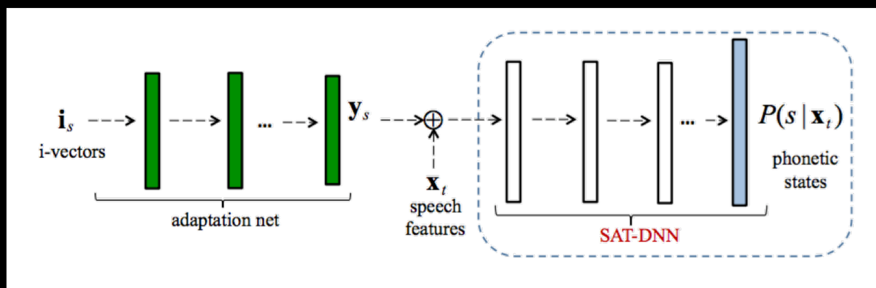
@ JSALT 2018:
one NN to rule them all!



The Goal

- Have a corpus of 2000h of how-to videos
  - Fully transcribed in English
  - Partially translated into Portuguese (and Turkish)
  - With short descriptions of videos
- Learn shared audio-visual (or text-visual) representations to help us understand video
  - Recognize, translate, and summarize videos
- Use sequence-to-sequence models (S2S) as unified architecture

Preliminary Experiments: ASR Adaptation



- 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

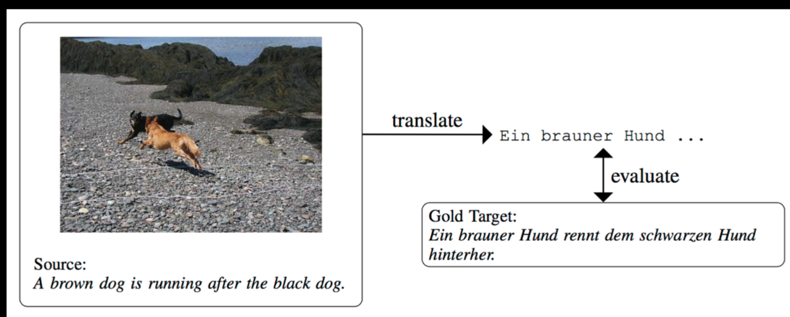
Model	Features	WER(%)	
DNN (Baseline)	-----	23.4	
Adaptive Training	161-dim visual features	22.3	↓4.7%
Adaptive Training	100-dim speaker i-vectors	22.0	↓6.0%
Adaptive Training	261-dim fused features	21.5	↓8.1%

[Gupta et al., 2017]

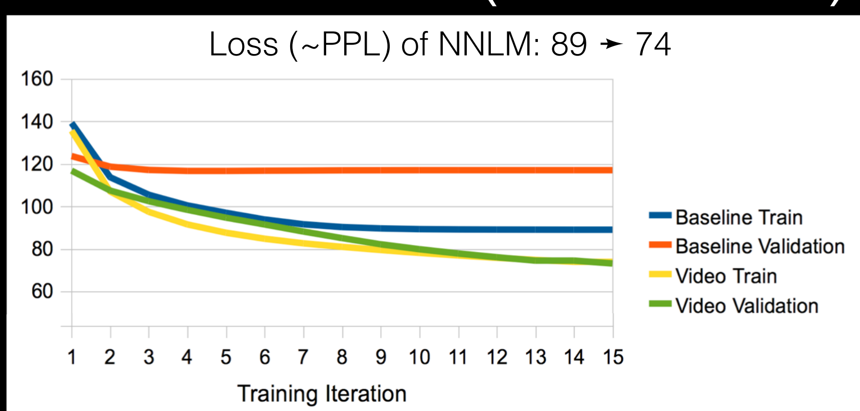
- AV adaptation does not beat i-Vector adaptation, but is in ballpark, somewhat complementary

Language Modeling

- 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]



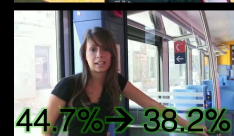
Bi-LSTM LM (5-fold CV)

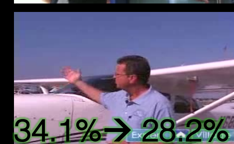


- 30-best lists from 23.4% WER DNN baseline
  - Re-score and re-rank with LSTM-LM
- ➢ 22.6% WER (15.6% Oracle WER)
  - Small but consistent improvements

Result Analysis – “indoor” vs “outdoor”

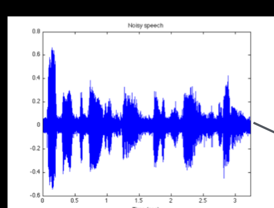
- Using object and place features only
  - AM+LM adapt.: 23.4% → 21.5% WER on 4h dev set (90h training)
- LM adaptation improves results across the board
  - 126/ 156 videos improve
- AM improves “noisy” videos
  - 55/ 156 videos improve (most are “outdoor”, according to their category)





Video as side-information in S2S ASR?

INPUT



$X_{t,n}$

Sequence-to-Sequence Model

$T_{i,n}$

OUTPUT

“Hey, now, I got those nice ripe cherries and I will decorate the cake with it”

TRANSCRIPT

$S_{j,n}$

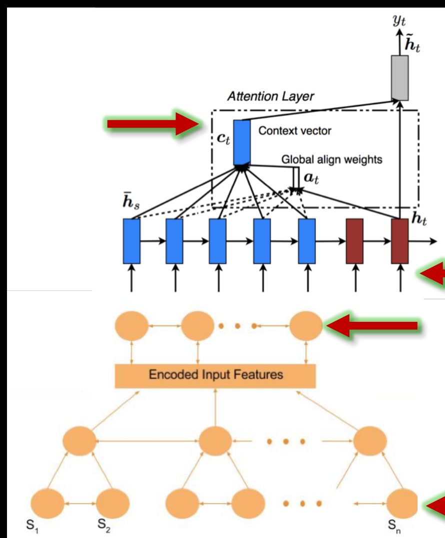
“Decorate the cake with cherries”

Summary
Translation
Retrieval
Q&

c_n

$n \in$ {all utterances}

Adaptive Seq-2-Seq with Attention

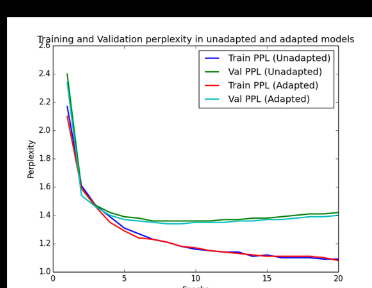


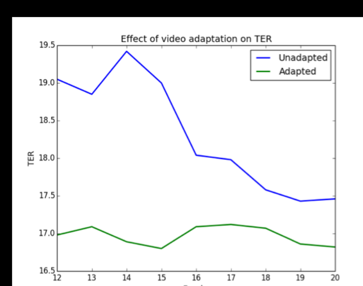
6+ ways of incorporating "visual context":

- Encoder feature shifts and appending features (AM)
  - Input layer, pyramid output
- At decoder (LM)
  - With attention mechanism
  - Co-Attention (2 sequences)
- At softmax layer (1G LM)

S2S Training Results (90h How-To)

- Appending 100d adaptation vector to 120d IMEL feature
- Best TER observed for later epochs, where perplexity increases
- Nice improvement in TER (17.5% ➛ 16.8%)
- Also works for CTC models, but somewhat inconsistent





Audio-Visual ASR Results

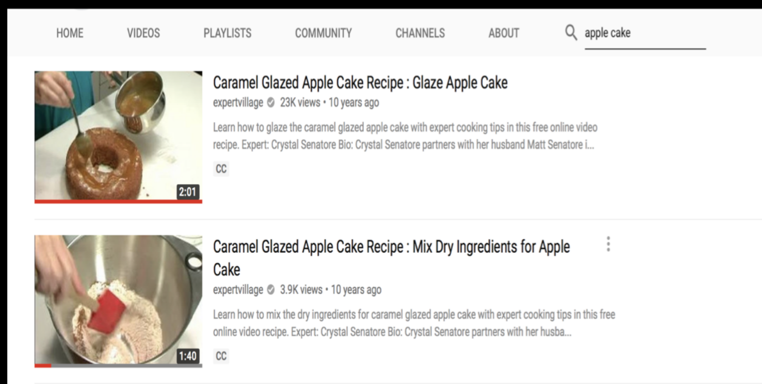
- It is possible to adapt (condition) a E2E ASR Model to static context, like a domain
  - CTC and S2S models both work
- The error rates improve, integration with an adapted language model gives further gains
- **More experimentation is needed, but models seem to learn semantic properties of the (correlated) video**
  - Multi-task (CTC+S2S) training?
  - Determine best units: chars, BPE, **words**, …
  - Shared representations have been learned?

Can you fly this thing?

Not yet.
[…]
Let’s go!

Multimedia Summarization

- Which how-to videos to watch, and why?



S2S Summarization

Results

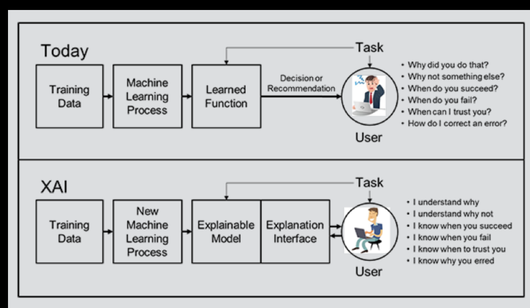
Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor, penalty=0	Rouge-L	Avg. words replaced
Baseline (original)	52.282	41.929	35.652	31.214	0.52	**0.506**	-
Without catch-phrases	33.811	22.731	16.699	12.862	0.36	**0.370**	6.70
Rule-based	22.152	10.059	5.527	3.345	0.21	**0.164**	-
Without catch-phrases	19.483	8.656	4.800	2.904	0.19	**0.155**	1.25

Ongoing Experiments

- Multi-Document Summarization
  - Take **triplets** of videos (anchor/ same/ different)
- Use a sequence-to-sequence model to generate **two** “descriptions” for **three** videos together
  - “similar” (portions of) videos or
  - “different” videos
- Experiment with different architectures ongoing
  - Triplet loss to encourage sharing and learning
  - Multi-modal features

Where To?

- Conversational Search: UIs without Screens
- Robotics – see what Humans see
- Explainable AI



Questions?

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

fmetze@cs.cmu.edu

Bibliography ASR

- **Fundamental Technologies in Modern Speech Recognition;** Sadaoki Furui, Li Deng, Mark Gales, Hermann Ney, Keiichi Tokuda. IEEE Signal Processing Magazine; Vol 29 (6), 2012. https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=6296521
- Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze. VISUAL FEATURES FOR CONTEXT-AWARE SPEECH RECOGNITION. In Proc. ICASSP, New Orleans, LA; U.S.A., March 2017. IEEE. https://arxiv.org/abs/1712.00489
- Shruti Palaskar, Ramon Sanabria, and Florian Metze. End-to-end multi-modal speech recognition. In Proc. ICASSP, Calgary, BC; Canada, April 2018. IEEE. https://arxiv.org/abs/1804.09713
- Yajie Miao, Hao Zhang, and Florian Metze. SPEAKER ADAPTIVE TRAINING OF DEEP NEURAL NETWORK ACOUSTIC MODELS USING I-VECTORS. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(11):1938-1949, November 2015. http://www.cs.cmu.edu/~fmetze/interACT//Publications_files/publications/bare_jrnl.pdf

Bibliography (Video) Summarization

- Florian Metze, Duo Ding, Ehsan Younessian, and Alexander Hauptmann. BEYOND AUDIO AND VIDEO RETRIEVAL: TOPIC ORIENTED MULTIMEDIA SUMMARIZATION. *International Journal of Multimedia Information Retrieval*, 2013. Springer. http://www.cs.cmu.edu/~fmetze/interACT//Publications_files/publications/10.1007_s13735-012-0028-y.pdf
- Over, Paul, Alan F. Smeaton, and Philip Kelly. "The TRECVID 2007 BBC rushes summarization evaluation pilot." In *Proceedings of the international workshop on TRECVID video summarization*, pp. 1-15. ACM, 2007. https://dl.acm.org/citation.cfm?id=1290032
- **Video Summarization with Long Short-term Memory;** Ke Zhang, Wei-Lun Chao, Fei Sha, Kristen Grauman. In Proc. ECCV 2016. https://arxiv.org/abs/1605.08110
- **A Deep Reinforced Model for Abstractive Summarization.** Romain Paulus, Caiming Xiong, Richard Socher. https://arxiv.org/abs/1705.04304
- Nenkova, Ani. "Summarization evaluation for text and speech: issues and approaches." In *Ninth International Conference on Spoken Language Processing*. 2006. https://www.isca-speech.org/archive/archive_papers/interspeech_2006/i06_2079.pdf