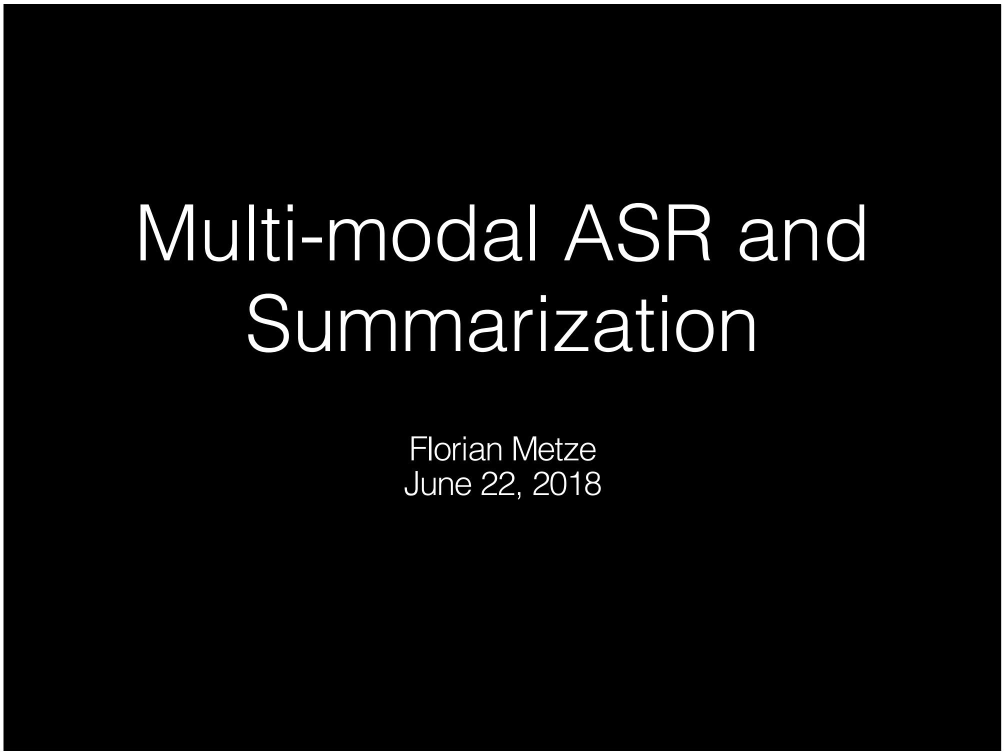
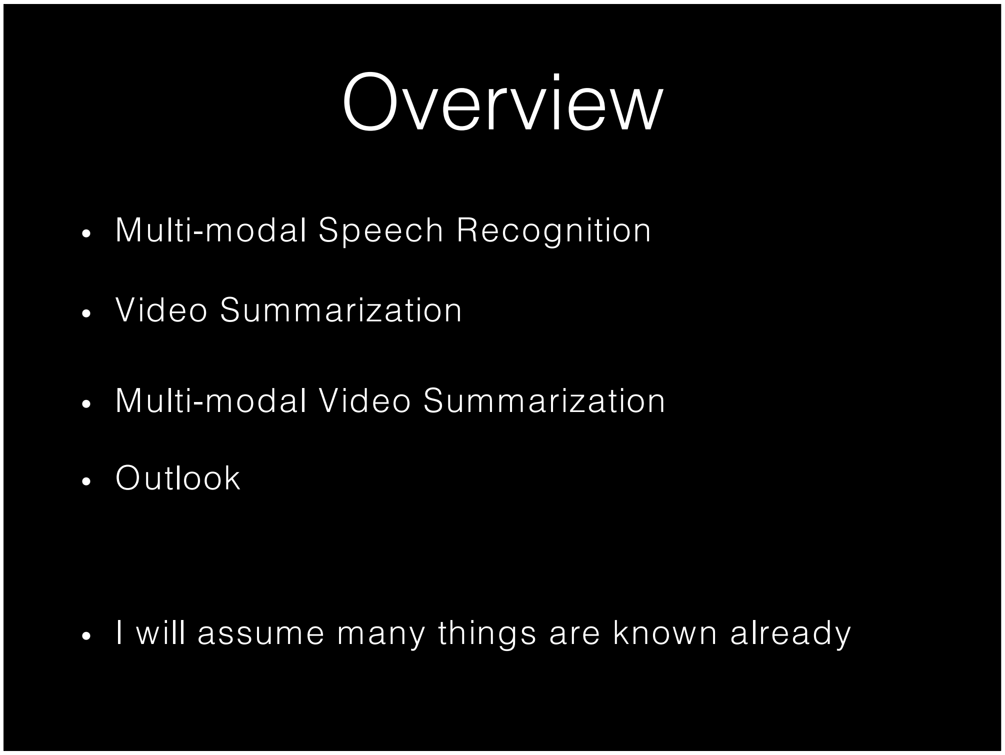


Multi-modal ASR and Summarization

Florian Metze
June 22, 2018



Overview

- Multi-modal Speech Recognition
- Video Summarization
- Multi-modal Video Summarization
- Outlook

- I will assume many things are known already

Overview

- **Multi-modal Speech Recognition**
- Video Summarization
- Multi-modal Video Summarization
- Outlook

- I will assume many things are known already

Automatic Speech Recognition

- Not going to repeat previous talks
  - But there are non-sequence models as well
- $W = \text{argmax}_{W'} P(W'|x) = \text{argmax}_{W'} p(x|W') P(W')$
- Clean separation into
  - Acoustic model $p(x|W)$
  - Language Model $P(W)$

Audio-visual ASR

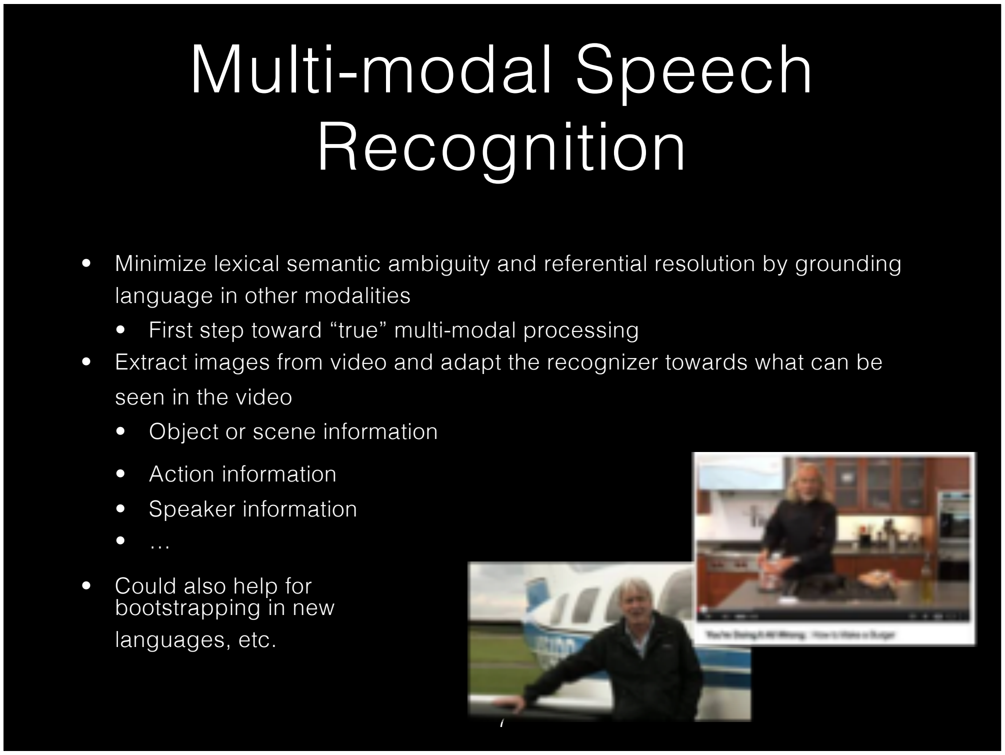
- It is nice if we need to adapt a single “S2S@ model only
- But it may be instructive to adapt the AM and the LM separately
  - In HMM framework: AM predicts state likelihoods (scaled posteriors) for every frame, e.g. p(s|W)
  - Multiply p for all states and frames during Viterbi search for best hypothesis
- It may also perform better

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”

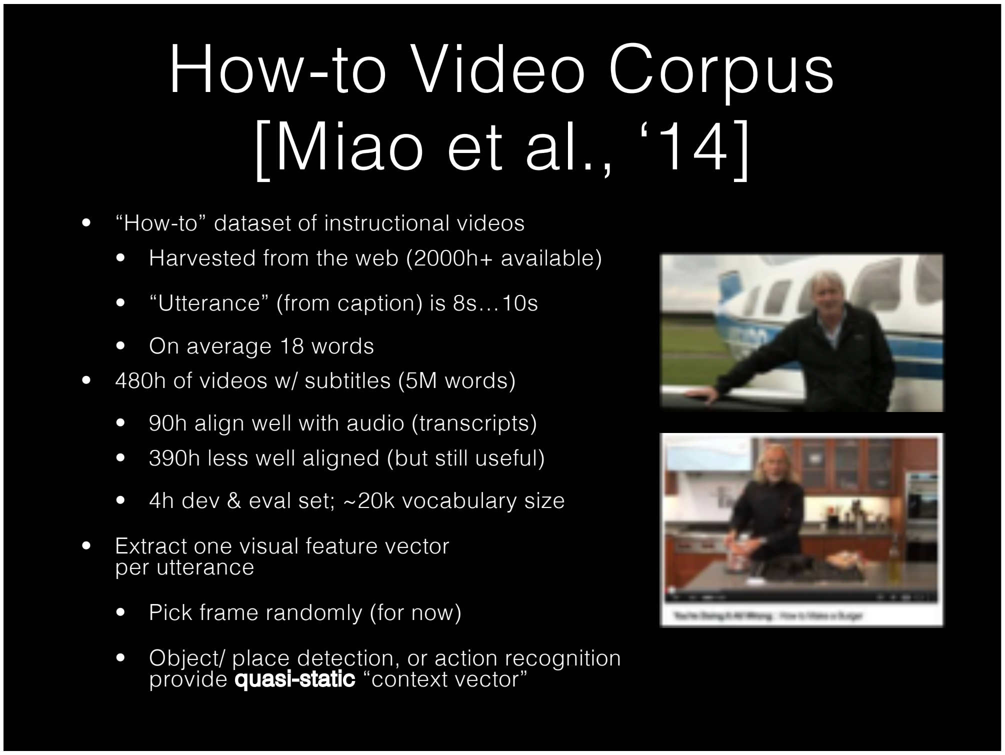
e.g. AVASR “Grid” Corpus

“Open-Domain” Video



Multi-modal Speech Recognition

- Minimize lexical semantic ambiguity and referential resolution by grounding language in other modalities
  - First step toward “true” multi-modal processing
- Extract images from video and adapt the recognizer towards what can be seen in the video
  - Object or scene information
  - Action information
  - Speaker information
  - …
- Could also help for bootstrapping in new languages, etc.

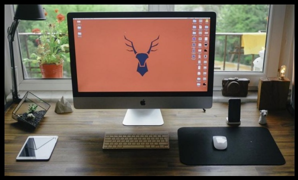


How-to Video Corpus [Miao et al., ‘14]

- “How-to” dataset of instructional videos
  - Harvested from the web (2000h+ available)
  - “Utterance” (from caption) is 8s…10s
  - On average 18 words
- 480h of videos w/ subtitles (5M words)
  - 90h align well with audio (transcripts)
  - 390h less well aligned (but still useful)
  - 4h dev & eval set; ~20k vocabulary size
- Extract one visual feature vector per utterance
  - Pick frame randomly (for now)
  - Object/ place detection, or action recognition provide **quasi-static** “context vector”

Two (Three) Types of Features

- Object Features



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

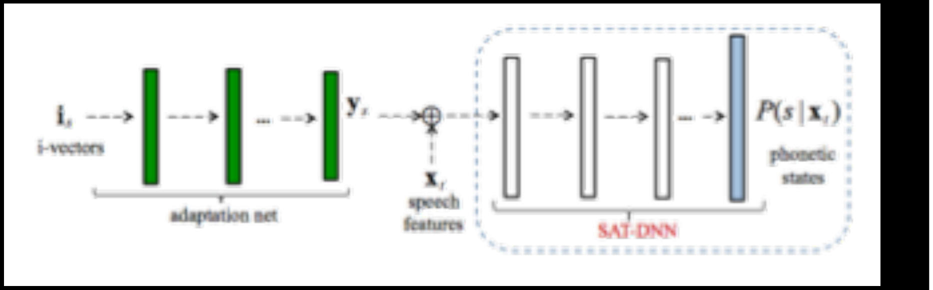
- Place Features (Scenes)

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

Adaptation in ASR

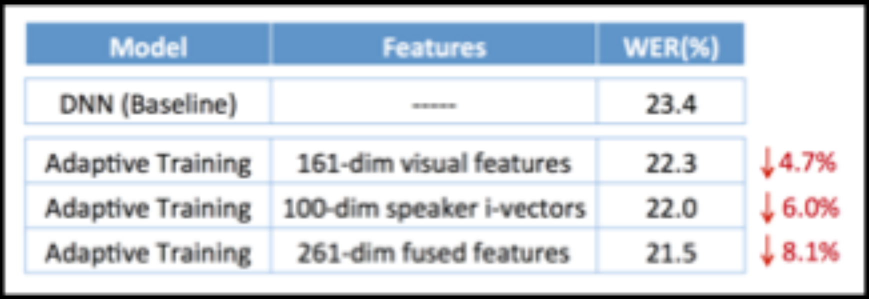
- One of the oldest and biggest topics in general
- Neural networks offer plethora of methods
- Will only discuss one idea (for Ams) that we have used in the past, using ResNet like idea
    - Features are time dependent x_t
    - Adaptation features are constant (over one utterance)
- *Our approach to multi-modal speech recognition could also be framed as adaptation using a "context vector" that is constant for one utterance!*

A General 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]

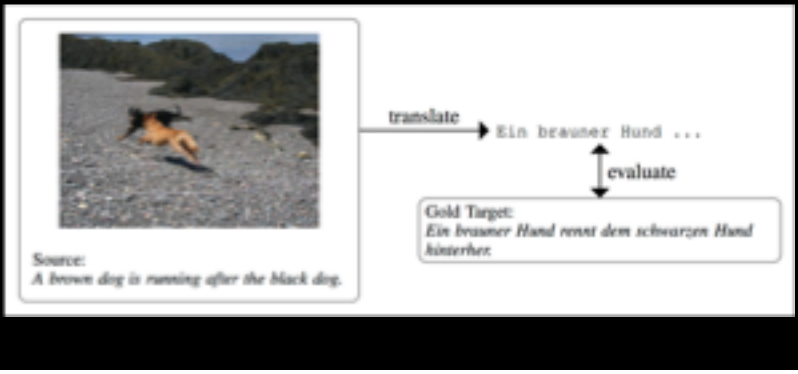
Comparison of Approaches



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 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]

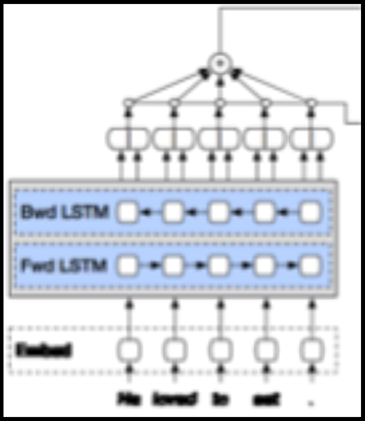


Gold Target:
Ein brauner Hund rennt dem schwarzen Hund hinterher.

Source:
A brown dog is running after the black dog.

LSTM Language Model

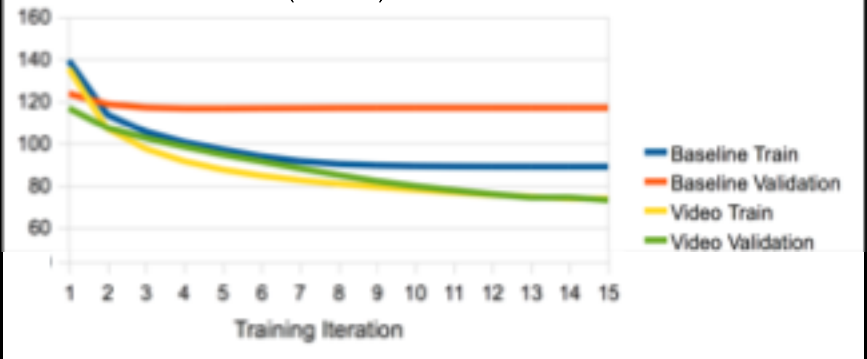
- Trained on 480h of transcriptions, optimized with 5-fold CV
- 2 BiLSTM layers, 1024 cells, Adagrad
- 1000d input vector consisting of
  - Learned 900d word embedding for vocabulary (~20k)
  - Context projected down to 100 dimensions
- 18 words sentence length on average (quite long!)



[redacted].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

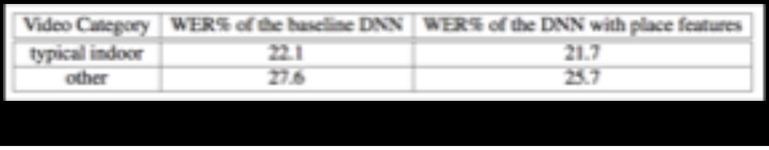
Analysis on 4h Test Set (156 Videos)

- 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% relative improvement on top of well-optimized 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

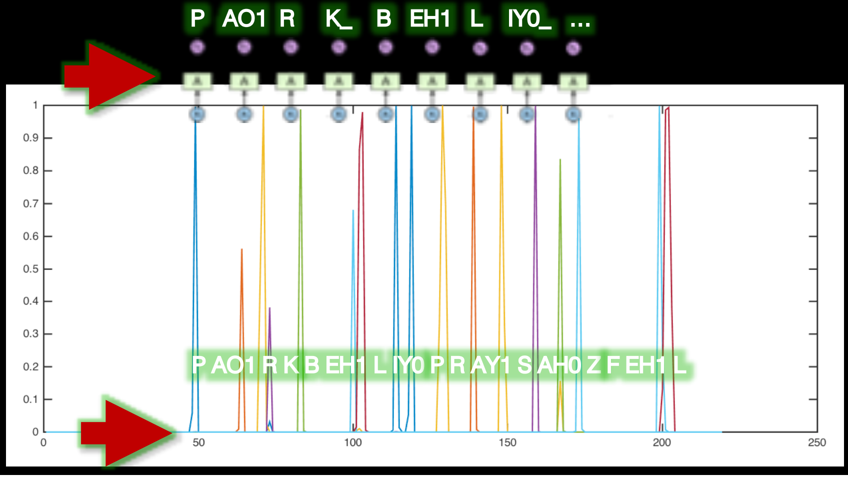
- AM improves “noisy” videos
  - 55/ 156 videos improve (most are “outdoor”, according to their category)



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

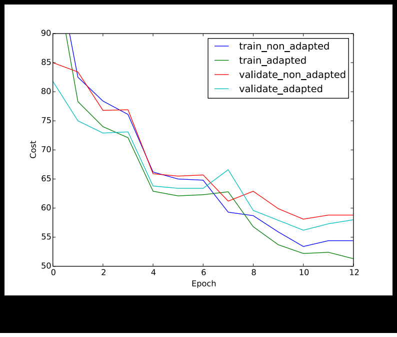
So – End-to-End Models?

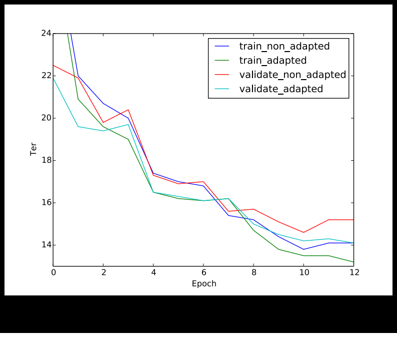
- Adapt a CTC AM with the “⊕” linear feature shift
- Adapt an RNN LM while decoding the CTC AM?



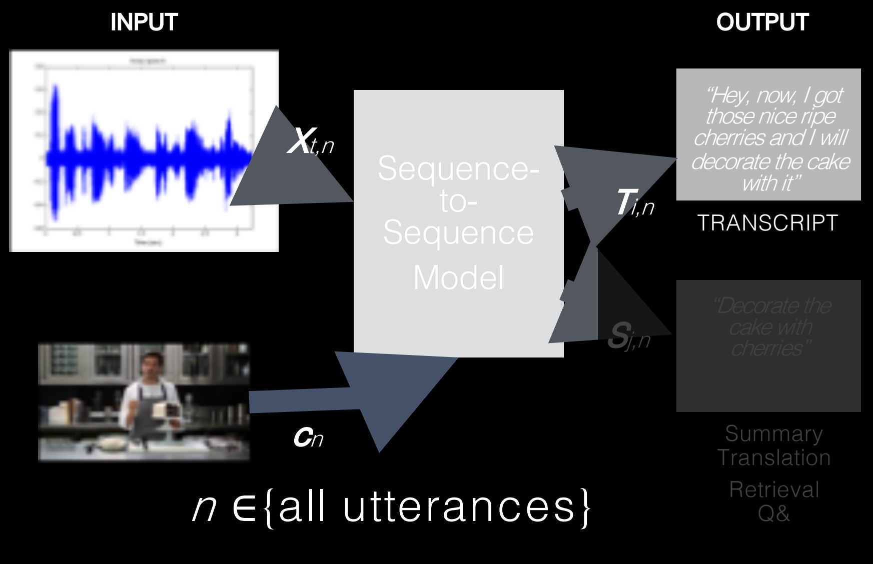
CTC Training Results (480h)

- Directly & jointly training BLSTM & “⊕”-MLP works best
- Improves performance from 15.2% ➛ 14.1% TER
- Training CTC on 90h did not work well (data not clean?)
- Hyper-parameter optimization & word decoding ongoing work



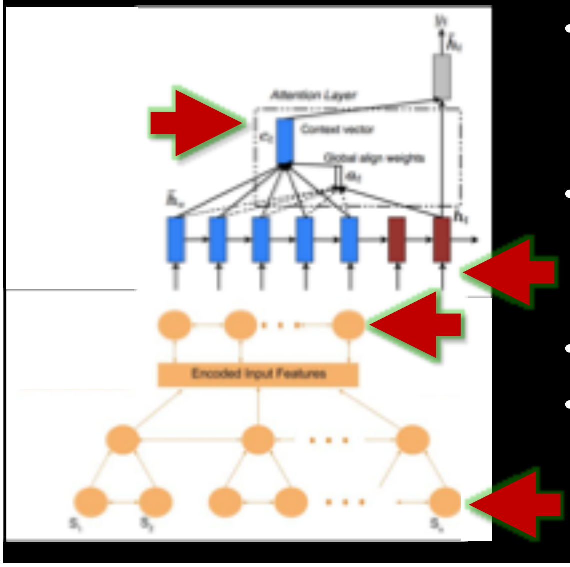


Video as side-information in S2S ASR?


INPUT
OUTPUT
$X_{t,n}$
Sequence-to-Sequence Model
$T_{i,n}$
$S_{j,n}$
c_n
“Hey, now, I got those nice ripe cherries and I will decorate the cake with it”
TRANSCRIPT
“Decorate the cake with cherries”
Summary
Translation
Retrieval
Q&

n ∈{all utterances}

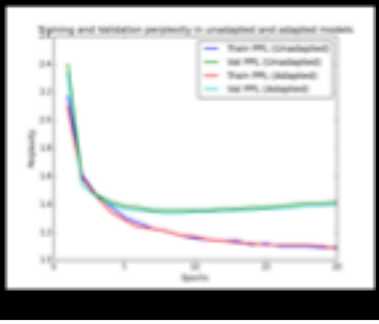
Adaptive Seq-2-Seq with Attention

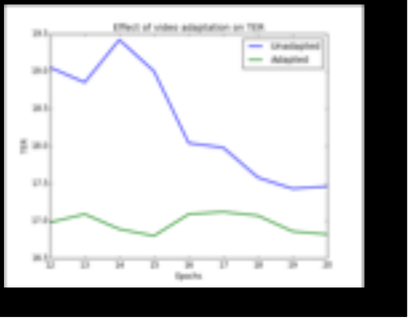


- 6+ ways of incorporating "visual context"
- Feature shifts & appending features
  - Input layer, pyramid output
- At decoder
- With attention mechanism

S2S Training Results (90h How-To)

- Appending 100d adaptation vector to 120d IMEL feature
- Best TER observed for later epochs, where perplexity increases
- Small improvement in (character) perplexity after adaptation
- Nice improvement in TER (17.5% ➛ 16.8%)





Audio-Visual ASR Results

- It is possible to adapt a E2E ASR Model to static context, like a domain
  - CTC and S2S models both work, exhibit different behavior
- The character error rate improves, integration with an adapted language model gives further gains
  - Dirty little secret of end-to-end ASR
- **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**, …

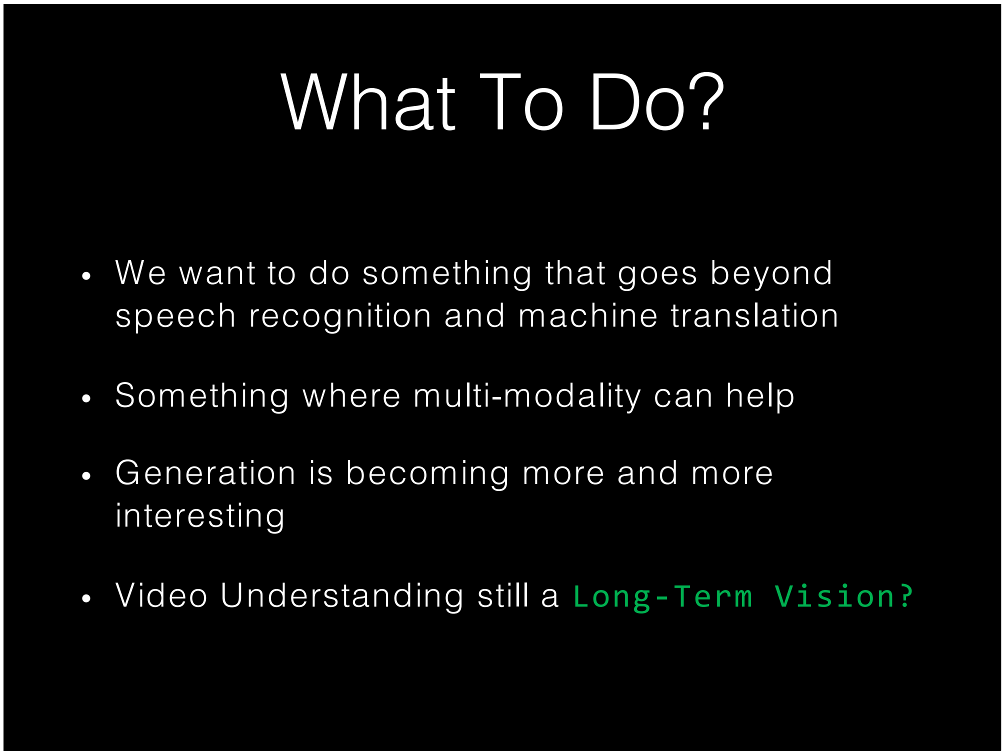
Overview

- Multi-modal Speech Recognition
- **Video Summarization**
- Multi-modal Video Summarization
- Outlook

- I will assume many things are known already

Can you fly this thing?

Not yet.
[…]
Let’s go!



What To Do?

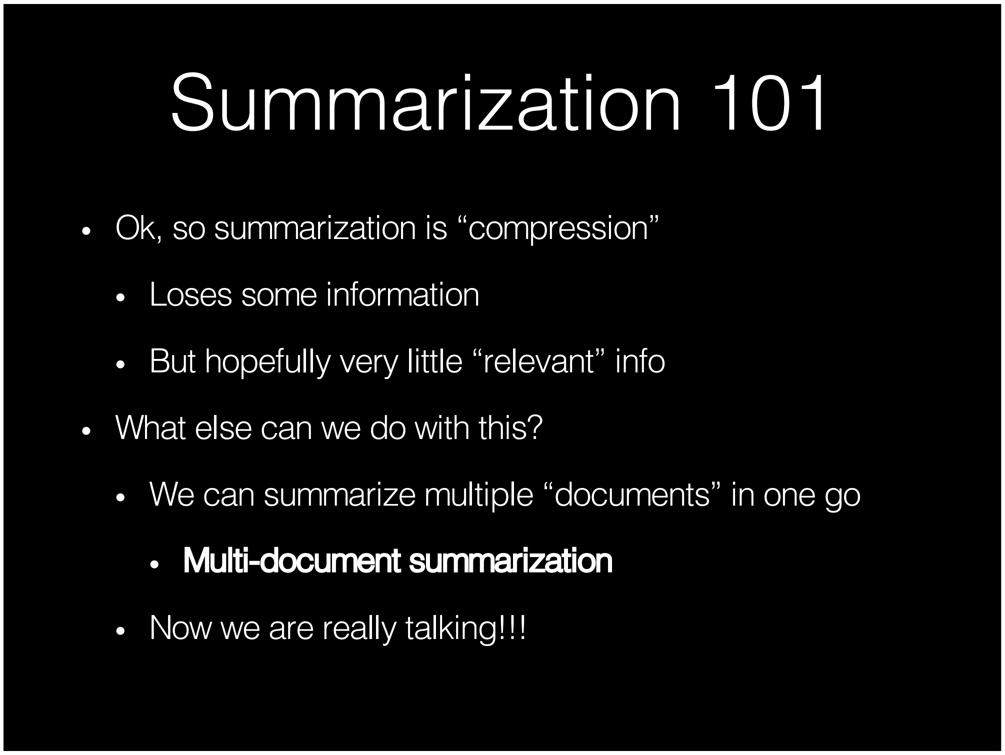
- We want to do something that goes beyond speech recognition and machine translation
- Something where multi-modality can help
- Generation is becoming more and more interesting
- Video Understanding still a Long-Term Vision?

Summarization 101

- Summarization is an interesting problem
  - Summarize text, speech, video (things that are sequences)
  - Images not so much (maybe called description)
- It can be extractive
  - Pick the ”most important ones” from the original elements
- It can be abstractive
  - Generate new elements (text for now)
  - Or even cross-modal (video-to-text)

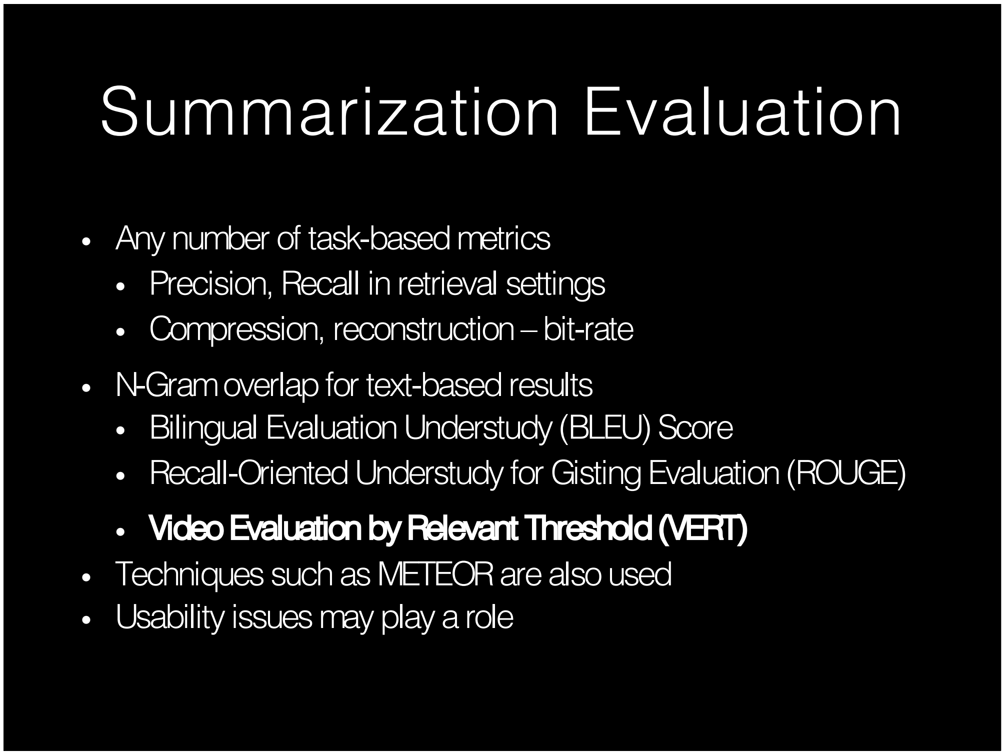
Summarization 101

- But why summarize in the first place?
  - Maybe to speed up human processing
  - Maybe to reduce storage requirements
  - Maybe to allow small screens, wearable UIs
- Evaluation is a big problem
  - Most meaningful evaluations require a task and human tests
  - Hard to optimize for such criteria, so use proxies



Summarization 101

- Ok, so summarization is “compression”
  - Loses some information
  - But hopefully very little “relevant” info
- What else can we do with this?
  - We can summarize multiple “documents” in one go
    - **Multi-document summarization**
  - Now we are really talking!!!

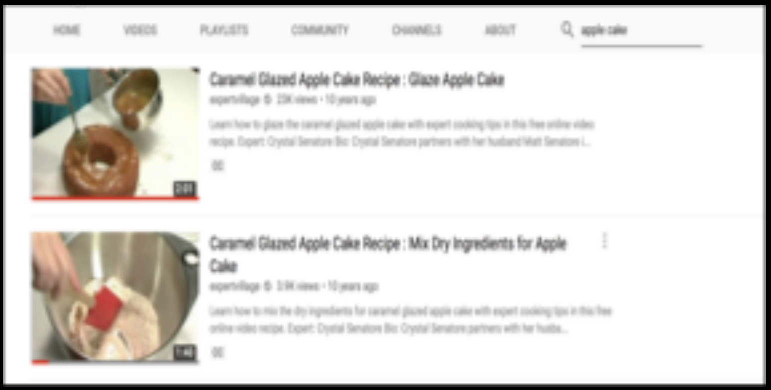


Summarization Evaluation

- Any number of task-based metrics
  - Precision, Recall in retrieval settings
  - Compression, reconstruction – bit-rate
- N-Gram overlap for text-based results
  - Bilingual Evaluation Understudy (BLEU) Score
  - Recall-Oriented Understudy for Gisting Evaluation (ROUGE)
  - **Video Evaluation by Relevant Threshold (VERT)**
- Techniques such as METEOR are also used
- Usability issues may play a role

Multimedia Example

- Which of these how-to videos should you watch, and why?



Caramel Glazed Apple Cake Recipe : Glaze Apple Cake

expertvillage · 23K views · 10 years ago

Learn how to glaze the caramel glazed apple cake with expert cooking tips in this free online video recipe. Expert: Crystal Senatore Bio: Crystal Senatore partners with her husband Matt Senatore i...

Caramel Glazed Apple Cake Recipe : Mix Dry Ingredients for Apple Cake

expertvillage · 3.9K views · 10 years ago

Learn how to mix the dry ingredients for caramel glazed apple cake with expert cooking tips in this free online video recipe. Expert: Crystal Senatore Bio: Crystal Senatore partners with her husba...

Multimedia Perspective (A. Hauptmann)

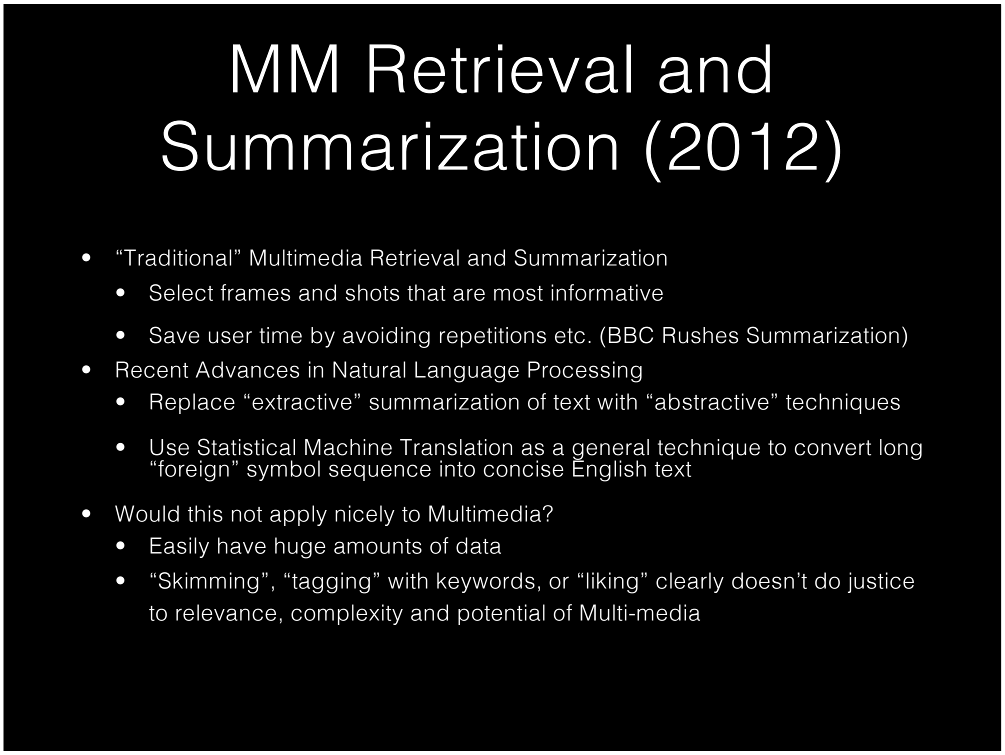
- “Video Summarization” has been researched
  - Skimming, thumbnail generation and other techniques exist to efficiently “browse” video
- It’s hard to improve on single-document summarization – unless in a very specific tasks
  - So, need to work on the multi-document case
  - Remove the browsing capability (no screen)

Our Approach

- Imagine we want to retrieve a number of videos from a database?
  - Like in a video information retrieval setup
- Then "structure" them in some way, e.g.
  - Explain (as text?) why these videos are good
  - Explain what these videos have in common
  - And how they are different?
- Would be useful in a multimedia community

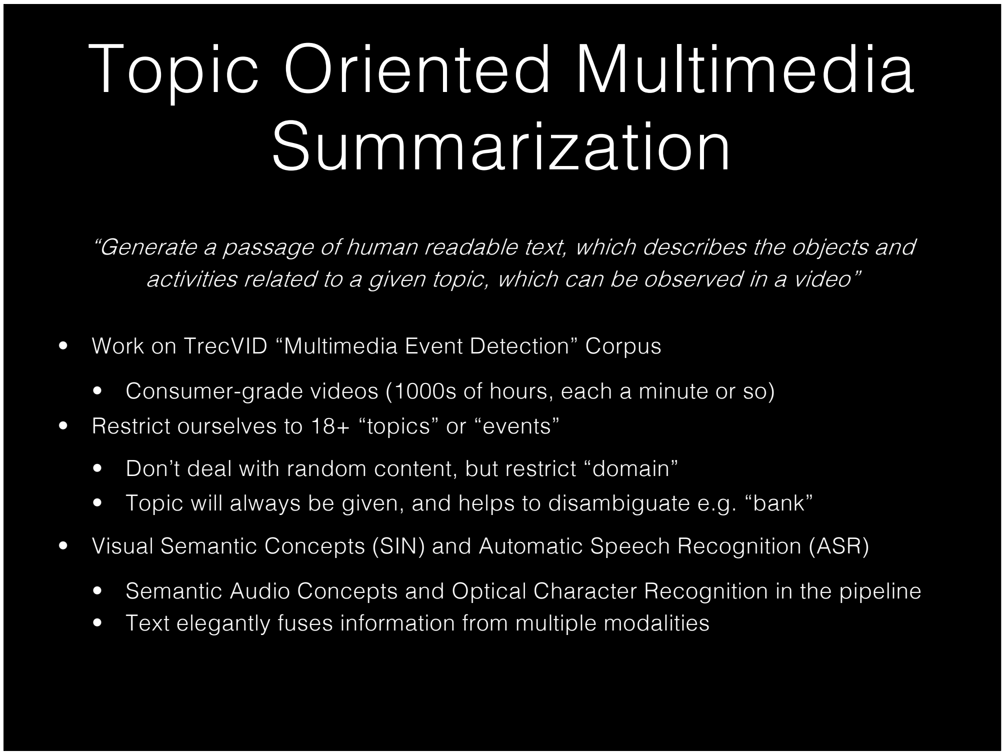
<BEGIN 2012>

- A Case Study: "**Multimedia Event Recounting**"
- This was earlier work done before Deep, Wide, and Recursive Generative Adversarial Networks became a thing
- Done during IARPA's "Aladdin" project, evaluated by NIST in the Trecvid "MER" task
  - F. Metze, D. Ding, E. Younessian, and A. Hauptmann. Beyond audio and video retrieval: **Topic oriented multimedia summarization**. International Journal of Multimedia Information Retrieval, 2013. Springer. http://dx.doi.org/10.1007/s13735-012-0028-y.
  - D. Ding, F. Metze, S. Rawat, P. F. Schulam, S. Burger, E. Younessian, L. Bao, M. G. Christel, and A. Hauptmann. **Beyond audio and video retrieval: Towards multimedia summarization**. In Proc. ICMR, Hong Kong; China, June 2012. ACM.



MM Retrieval and Summarization (2012)

- "Traditional" Multimedia Retrieval and Summarization
  - Select frames and shots that are most informative
  - Save user time by avoiding repetitions etc. (BBC Rushes Summarization)
- Recent Advances in Natural Language Processing
  - Replace "extractive" summarization of text with "abstractive" techniques
  - Use Statistical Machine Translation as a general technique to convert long "foreign" symbol sequence into concise English text
- Would this not apply nicely to Multimedia?
  - Easily have huge amounts of data
  - "Skimming", "tagging" with keywords, or "liking" clearly doesn't do justice to relevance, complexity and potential of Multi-media



Topic Oriented Multimedia Summarization

"Generate a passage of human readable text, which describes the objects and activities related to a given topic, which can be observed in a video"

- Work on TrecVID "Multimedia Event Detection" Corpus
  - Consumer-grade videos (1000s of hours, each a minute or so)
- Restrict ourselves to 18+ "topics" or "events"
  - Don't deal with random content, but restrict "domain"
  - Topic will always be given, and helps to disambiguate e.g. "bank"
- Visual Semantic Concepts (SIN) and Automatic Speech Recognition (ASR)
  - Semantic Audio Concepts and Optical Character Recognition in the pipeline
  - Text elegantly fuses information from multiple modalities

Topic Oriented Multimedia Summarization

How can we proceed in a principled way?

- Clearly, it would help if we could somehow (automatically) **generate example summaries**,
- **evaluate them with humans doing tasks**,
  to determine which ones are good,
- and **iterate**.

➢ Look at **efficiency, effectiveness, and satisfaction, etc.**

Example: Summaries for Different Videos on the Same Topic

The video shows the event of **Changing_a_vehicle_tire**. We probably heard the words "jack", "remove", "car", "open" and "people" in the video. We probably saw Vehicle, Ground_Vehicles, Hand, Car, Body_Parts, Adult, Outdoor and Man_Made_Thing in the video. We possibly saw **Construction_Vehicles** and Road in the video.

This video is about **Changing_a_vehicle_tire**. We heard the words "lug", "spare", "carjacke", "katherine", and "wheaty" in the video. We probably saw Vehicle, Ground_Vehicles, Hand, Body_Parts, Adult, Outdoor and Man_Made_Thing in the video. We possibly saw Car and Road in the video. But we also detected Text with a relatively high confidence, which is not usual for **Changing_a_vehicle_tire** events.

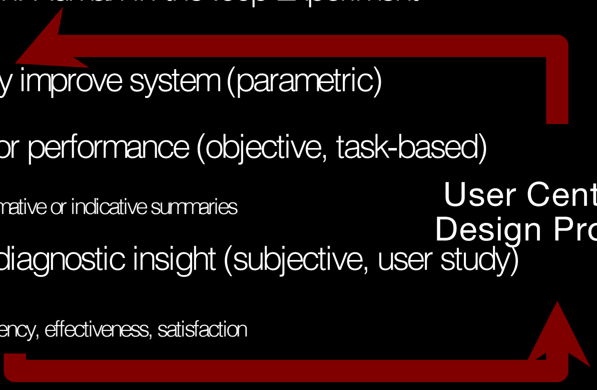
Example: Same Video, Different Summaries for Different Topics

This video is about **Changing_a_vehicle_tire**. We heard the words "lug", "spare", "carjack", "katherine", and "wheaty" in the video. W + probably saw Vehicle, Ground_Vehicles, Hand, Body_Parts, Ac Outdoor and Man_Made_Thing in the video. We possibly saw Car Road in the video. But we also detected Text with a relatively high confidence (above 0.8), which is not usual for Changing_a_vehicle_tire events.

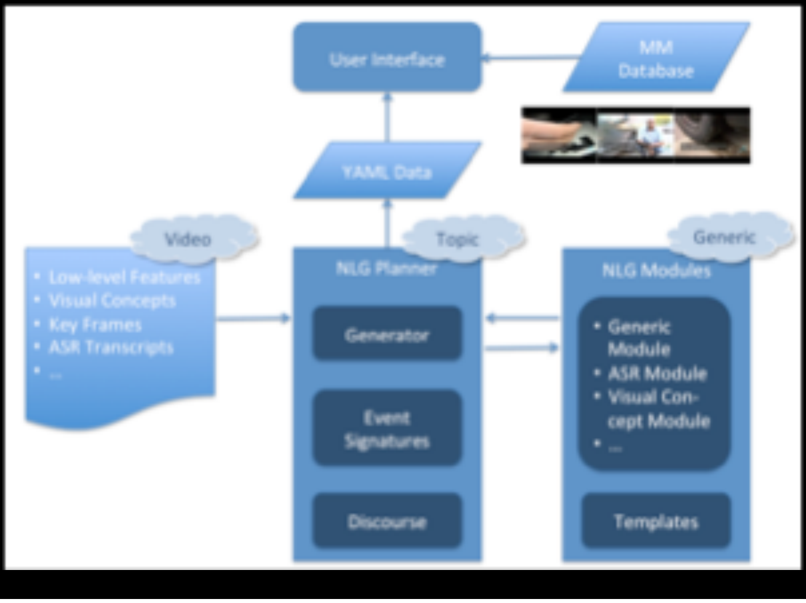
The video shows the event of **Making_a_sandwich**. We heard the words "wheaty", "spare", and "katherine" in the video. We probably saw Hand, Body_Parts and Adult in the video. But we also dete − Vehicle, Man_Made_Thing and Text with a relatively high confidence (above 0.8), which is not usual for Making_a_sandwich events. And we also detected Ground_Vehicles, Car, Road and Outdoor.

Topic Oriented Multimedia Summarization

- Approach: Human-in-the-loop Experiment
- Iteratively improve system (parametric)
  - Test for performance (objective, task-based)
    - Informative or indicative summaries
  - Gain diagnostic insight (subjective, user study)
    - Efficiency, effectiveness, satisfaction
- Done?



System Architecture



What to Mention? – The "Event Signature"

- Generate topic-specific "Event Signatures", to capture salient information
  - Rank and combine detected objects and actions, resolve ambiguities
  - E.g. "hand" is good for "changing tires" and "making sandwich", "vehicle" good only for "changing tires"
- Did something similar for ASR output words
- Problems of manually created signature (Ontologies, etc.)
  - Time consuming & subjective
  - Hard to quantify the relevance of concepts
- Automatic event-specific signature generation
  - Tried different things, before inserting in Human-in-the-loop experiment

Signature Generation by Bipartite Graph Propagation

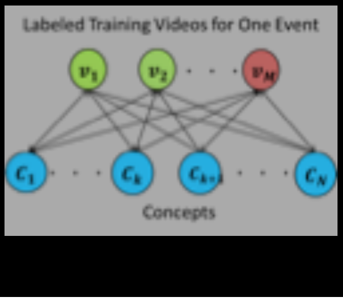
- Motivation
  - Explore the pair-wise relationships between training videos and concepts and generate meaningful event-specific signatures
  - Inspired by previous work in TrecVID Search Task (mapping query to concept)
- Bipartite Graph Construction

$\mathbf{G} = \{\mathbf{V}, \mathbf{C}, \mathbf{E}, \mathbf{W}\}$: $\mathbf{V}$ is the node set for the training video samples;

$\mathbf{C}$ is the node set for concepts;

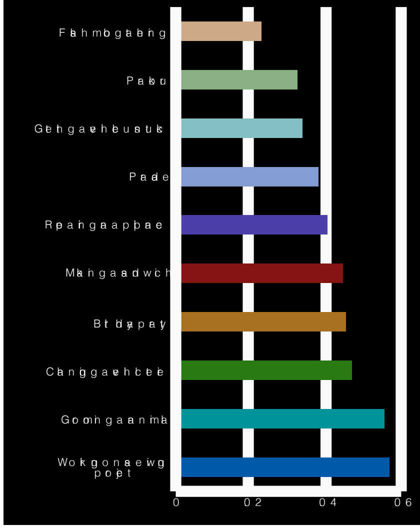
$\mathbf{E}$ is the edge set.
The edge is weighted by $\mathbf{W}_{ij}$.

$\mathbf{W}_{ij}$ indicates the concept $\mathbf{C}_j$'s prediction score on video $\mathbf{v}_i$.



Relevant Concepts learned by Bipartite Graph Propagation

Event Classifier Error



Top 8 Related SIN Concepts

Crowd, People_Marching, 3_Or_More_People, Demonstration_Or_Protest, Meeting, Cheering, Urban_Scenes, Walking

Urban_Scenes, Building, Windows, Outdoor, Streets, Road, Walking_Running, Cityscape

Car, Snow, Motorcycle, Outdoor, Landscape, Vehicle, Boat_Ship, Ground_Vehicles

Crowd, 3_Or_More_People, People_Marching, Demonstration_Or_Protest, Urban_Scenes, Meeting, Streets, Suburban

Room, Computers, Commercial_Advertisement, Kitchen, Synthetic_Images, Indoor, Network_Logo, Hand

Room, Kitchen, Food, Indoor, Hand, Attached_Body_Parts, Body_Parts, Man_Made_Thing

Dresses, Joy, Furniture, Sitting_Down, Room, Talking, Dining_Room, Eaters

Car, Text, Ground_Vehicles, Sports_Car, Vehicle, Construction_Vehicles, Minivan, Car_Racing

Baby, Sunglasses, Attached_Body_Parts, Dresses, Carnivore, Hand, Domesticated_Animal, Anger

Hand, Room, Animation_Cartoon, Commercial_Advertisement, Kitchen, Attached_Body_Parts, Baby, Synthetic_Images

Evaluation in Pilot User Study

Compare machine generated text passages with human generated ones

- 10 machine generated, 10 human generated
- Used 10 “non-expert” team members to generate text

Two tasks for computer-based user study

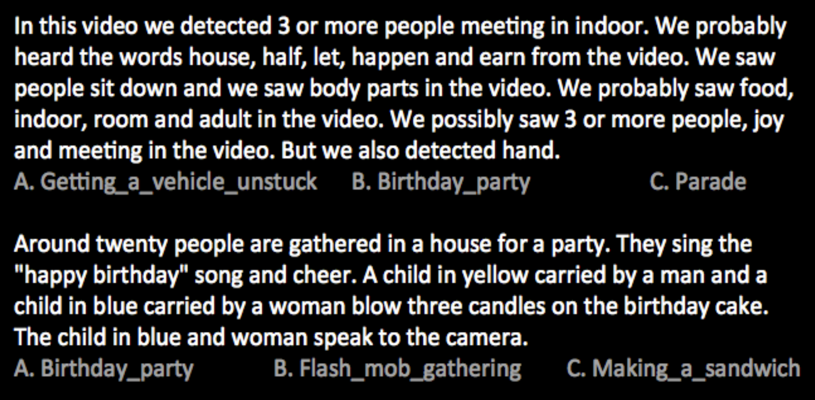
- Event Selection Task (summaries should be indicative)
- Video Selection Task (summaries should be informative)

Goals

- Evaluate performance
- Gather insight (diagnostics)

Event Selection Task

“How well does the text describe a topic”



In this video we detected 3 or more people meeting in indoor. We probably heard the words house, half, let, happen and earn from the video. We saw people sit down and we saw body parts in the video. We probably saw food, indoor, room and adult in the video. We possibly saw 3 or more people, joy and meeting in the video. But we also detected hand.

A. Getting_a_vehicle_unstuck B. Birthday_party C. Parade

Around twenty people are gathered in a house for a party. They sing the "happy birthday" song and cheer. A child in yellow carried by a man and a child in blue carried by a woman blow three candles on the birthday cake. The child in blue and woman speak to the camera.

A. Birthday_party B. Flash_mob_gathering C. Making_a_sandwich

Video Selection Task

“How well does the text allow to identify the specific video it describes”

This video is about Making_a_sandwich. We heard the word “cyru” in the video. We probably heard the words “bread”, “house”, “song” and “spread” in the video. We saw “Body_Parts”, “Indoor”, “Room” and “Man_Made_Thing” in the video. We probably saw “Food”, “Kitchen”, “Hand”, “Synthetic_Images” and “Furniture” in the video. But we also detected “Adult”.

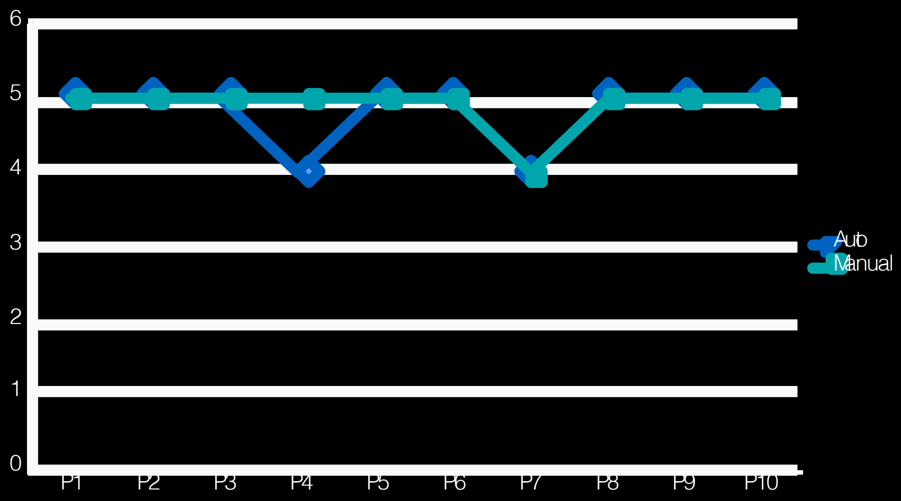
A. 1

B. 2

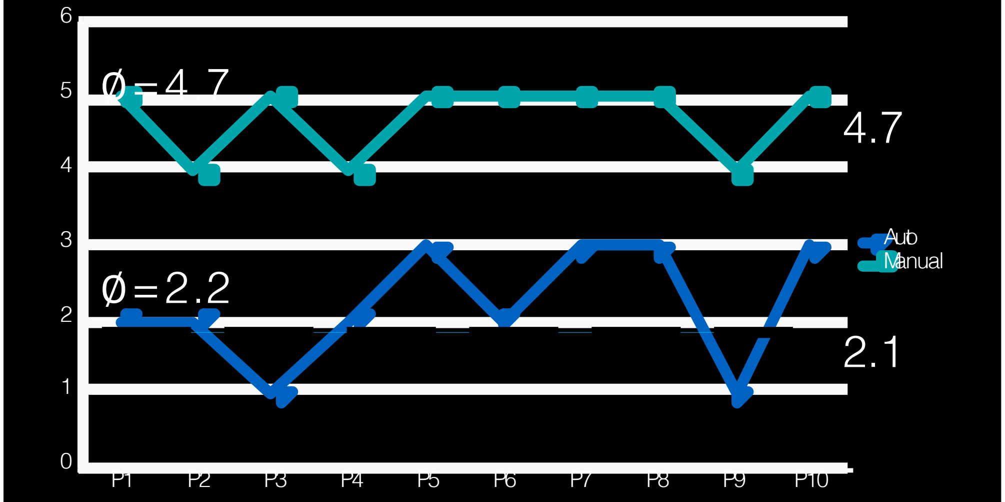
C. 3

http://rocks.is.cs.cmu.edu:8080/mer/index.jsp?id=297103.yml

Comparison between Machine and Human (Event Selection Tasks)



Comparison between Machine and Human (Video Selection Tasks)



Insights from Pilot User Study

Event Selection Task

- **Automatic language generation system can do almost as well as Humans**

Video Selection Task

- **System generated text clearly worse than Human generated text in helping users choose the right video**
- Expert and subject assessment of differences
- Human generated recounting passages are more detailed and specific
  - Humans use **temporal expressions**, **sequences** of observations
  - Humans use **identities** ("Volkswagen"), **object properties** (colors, sizes, etc.), and qualifiers (e.g. "*birthday* cake")
  - Not so many relations between objects ("next to", etc.)

</BEGIN 2012>

- Described principled process to learn from users how to summarize videos by “content”, including performance evaluation
- Automatically generated representation of “topic” and related, discriminating features

- This was a best paper candidate at ICMR 2012
- How else could this be relevant?

Summarization Idea

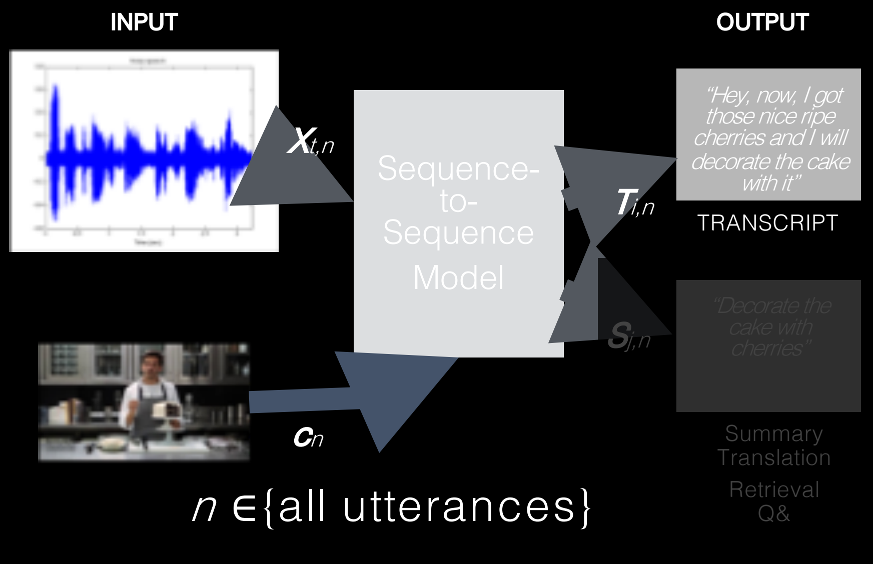
- Retrieval from a large video database (2000h how-to)
- Take a cluster of related videos
  - Explain what they have in common?
  - Explain how they differ? Or how one differs?
- Text output
  - Could be used in a conversational search assistant interface
- Explainable AI (XAI) idea:
  - Note that the classification decision and the explanation generation could be separate processes

Overview

- Multi-modal Speech Recognition
- Video Summarization
- **Multi-modal Video Summarization**
- Outlook

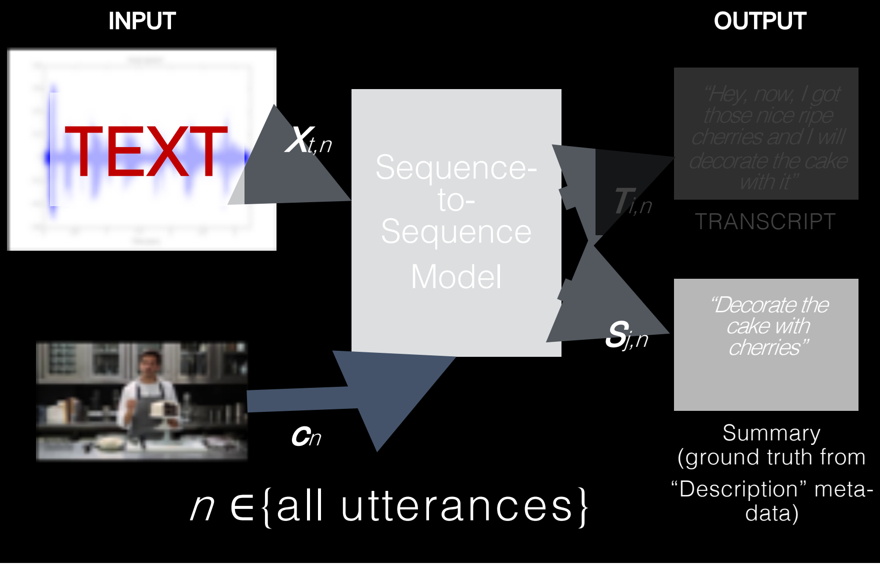
- I will assume many things are known already

Reminder


INPUT
OUTPUT
$X_{t,n}$
Sequence-to-Sequence Model
"Hey, now, I got those nice ripe cherries and I will decorate the cake with it"
$T_{i,n}$
TRANSCRIPT
"Decorate the cake with cherries"
$S_{j,n}$
c_n
Summary
Translation
Retrieval
Q&

n ∈{all utterances}

Multimodal Video Summarization



INPUT

OUTPUT

TEXT

$X_{t,n}$

Sequence-to-Sequence Model

$T_{i,n}$

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

TRANSCRIPT

$S_{j,n}$

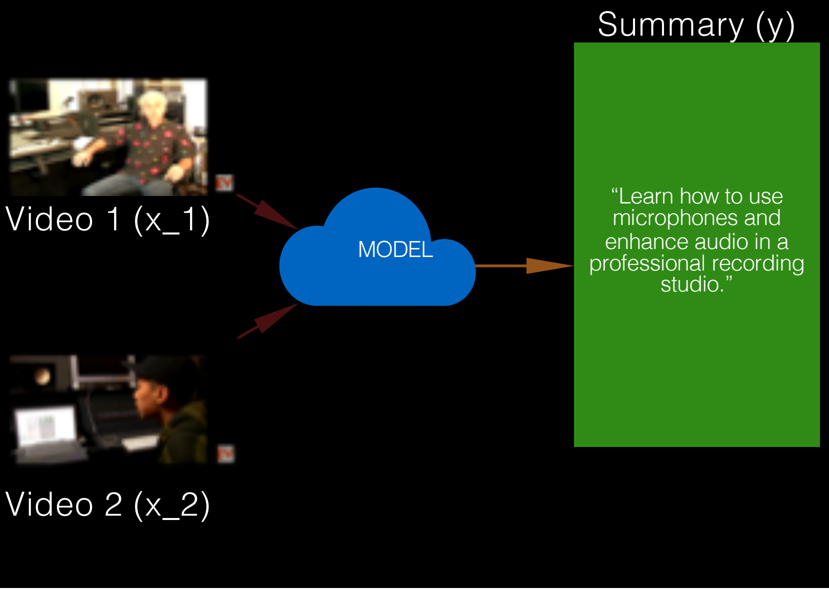
"Decorate the cake with cherries"

c_n

Summary (ground truth from "Description" meta-data)

$n \in$ {all utterances}

Multi-Document Case



Summary (y)

Video 1 (x_1)

MODEL

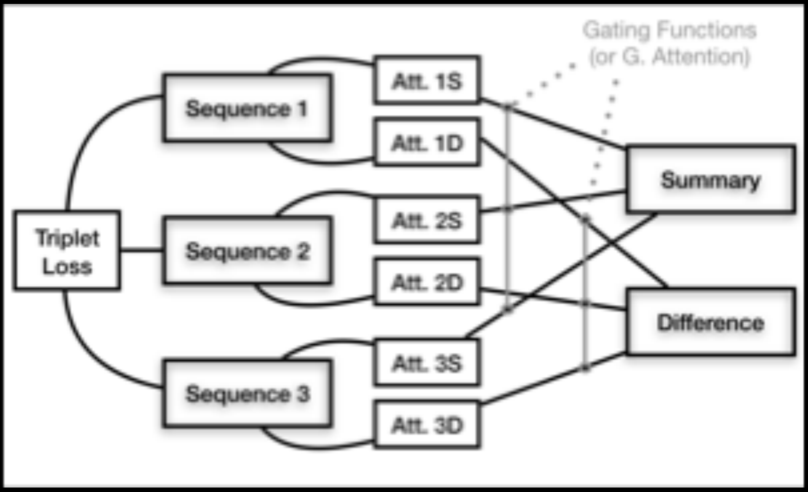
"Learn how to use microphones and enhance audio in a professional recording studio."

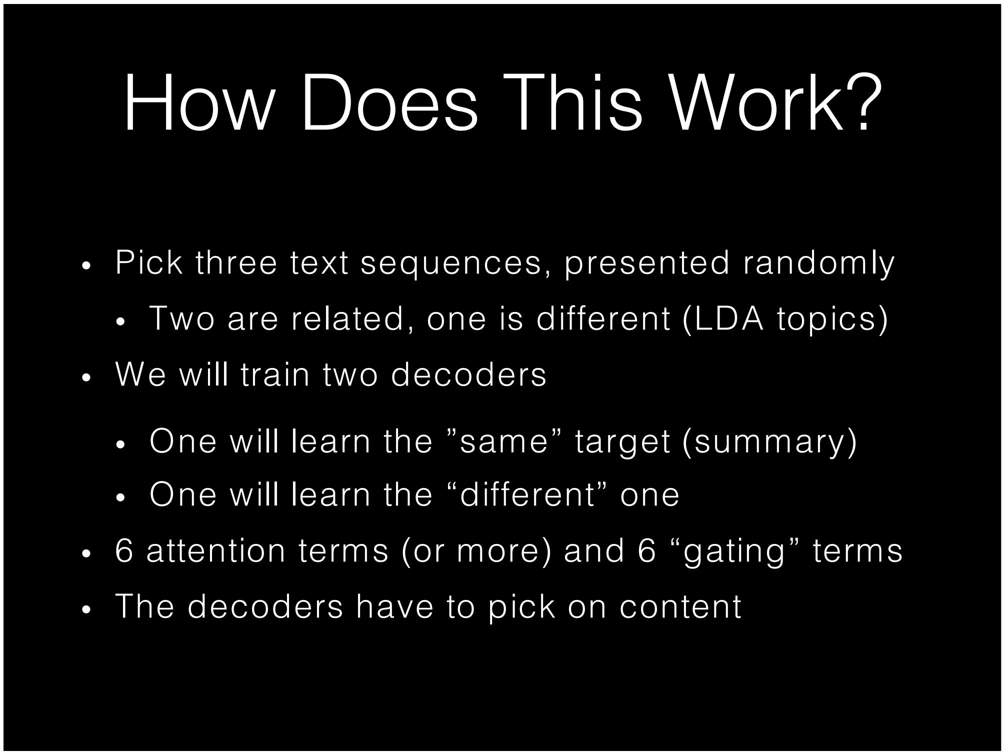
Video 2 (x_2)

First Experiment

- Take **triplets** of videos (anchor/ same/ different)
- Use a sequence-to-sequence model to generate **two** “descriptions” for
    - “similar” (portions of) videos or
    - “different” videos
- Initially, these may not be grammatically correct (depending on training data that is available)
- But they should show the idea and be informative

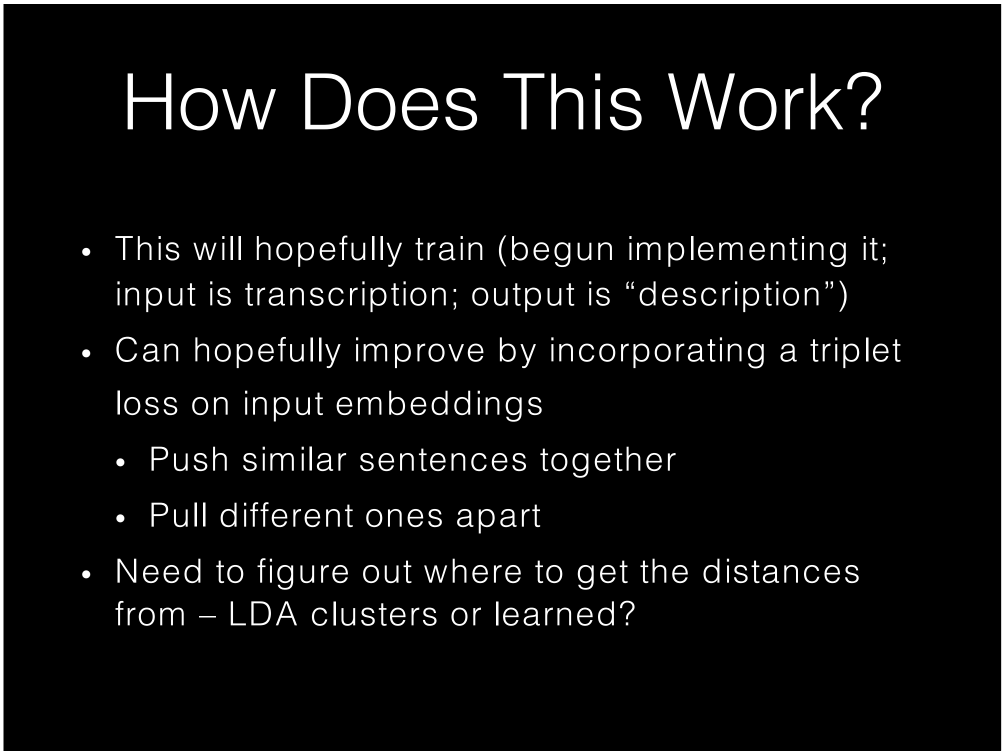
Architecture





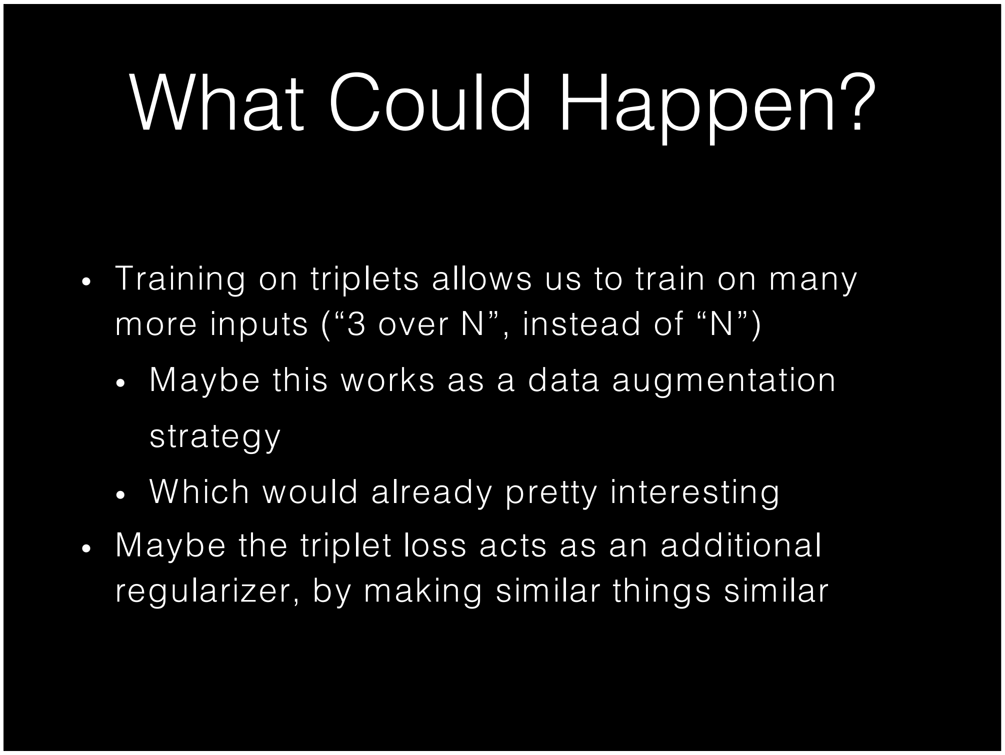
How Does This Work?

- Pick three text sequences, presented randomly
  - Two are related, one is different (LDA topics)
- We will train two decoders
  - One will learn the ”same” target (summary)
  - One will learn the “different” one
- 6 attention terms (or more) and 6 “gating” terms
- The decoders have to pick on content



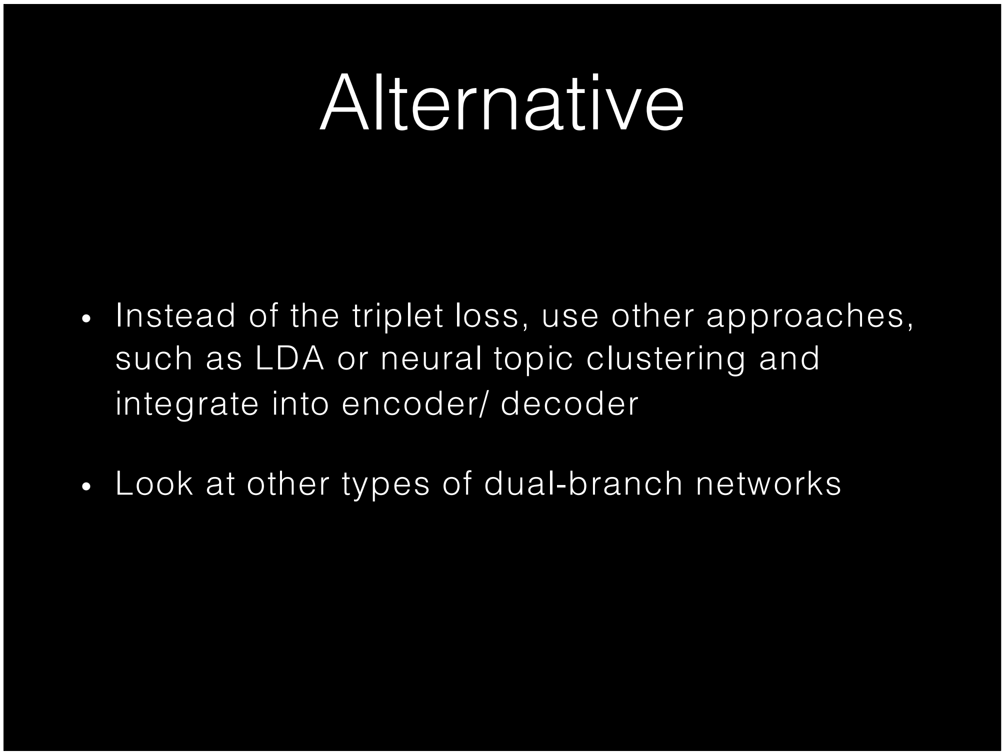
How Does This Work?

- This will hopefully train (begun implementing it; input is transcription; output is “description”)
- Can hopefully improve by incorporating a triplet loss on input embeddings
  - Push similar sentences together
  - Pull different ones apart
- Need to figure out where to get the distances from – LDA clusters or learned?



What Could Happen?

- Training on triplets allows us to train on many more inputs (“3 over N”, instead of “N”)
  - Maybe this works as a data augmentation strategy
  - Which would already pretty interesting
- Maybe the triplet loss acts as an additional regularizer, by making similar things similar

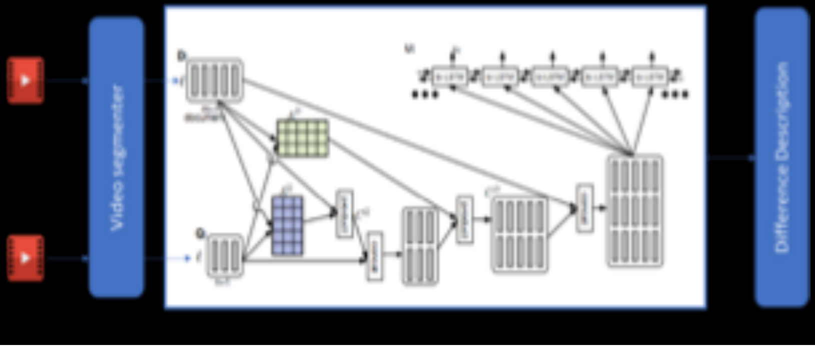


Alternative

- Instead of the triplet loss, use other approaches, such as LDA or neural topic clustering and integrate into encoder/ decoder
- Look at other types of dual-branch networks

Alternative 2

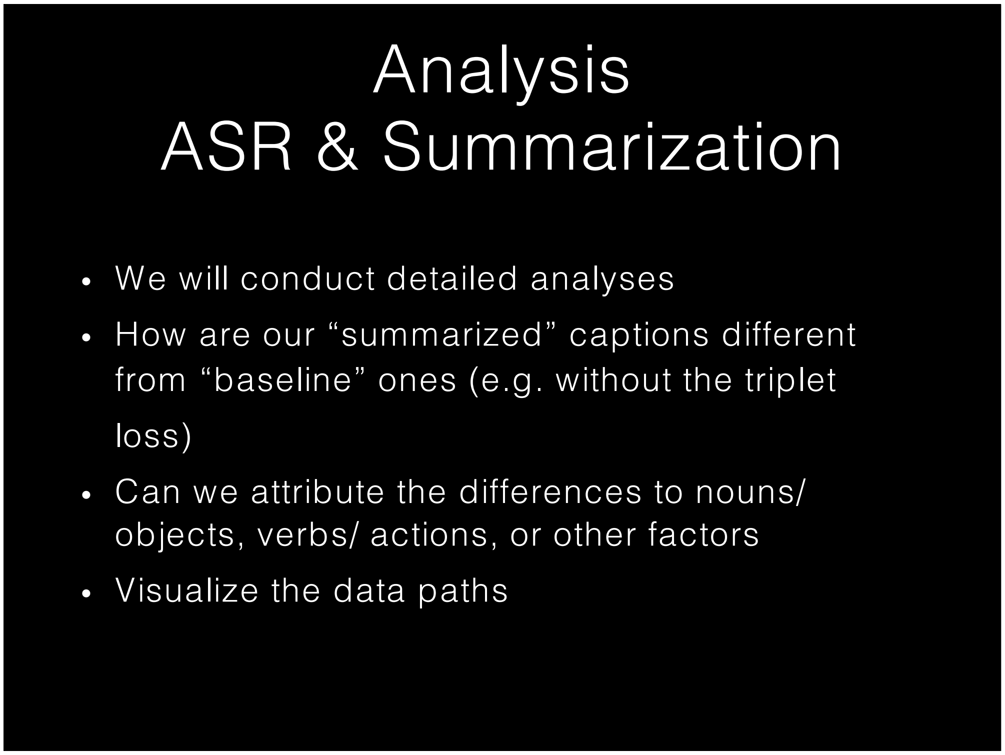
- Instead of the triplet loss, use other approaches, such as LDA or neural topic clustering and integrate into encoder/ decoder
- Look at other types of dual-branch networks
- Dynamic Coattention Networks For Question Answering, Caiming Xiong, et. al., ICLR 2017



Evaluation

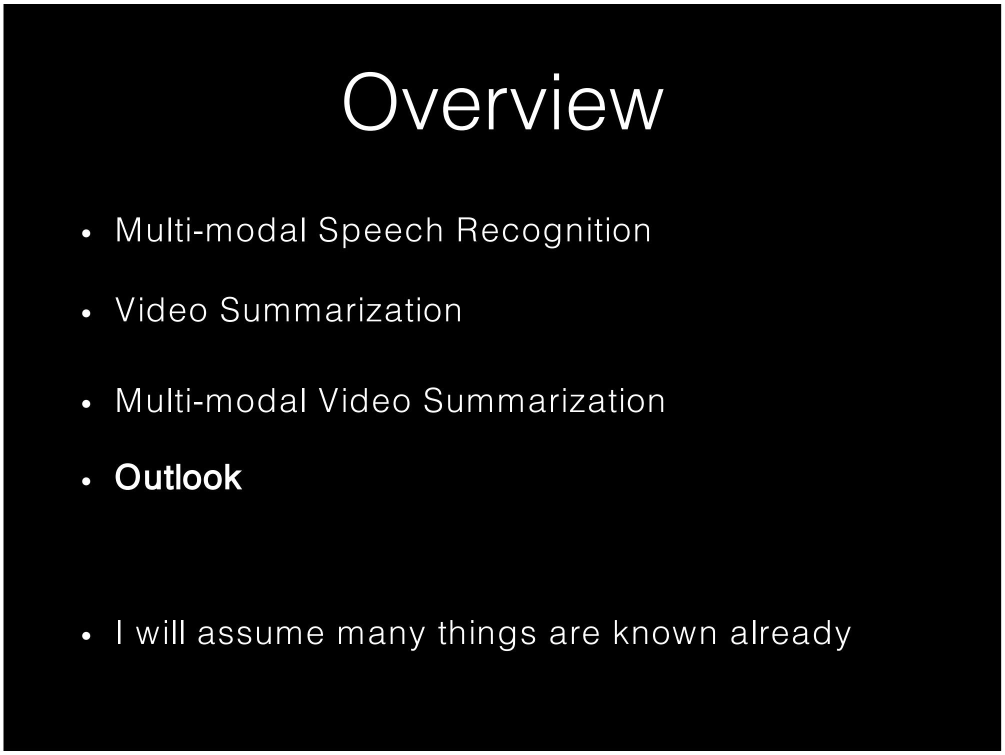
- Video Retrieval is usually evaluated in terms of P@N (precision at n)
    - Recall may be meaningless
- Here – precision may also not help, because we do not have an end-to-end task
- Summarization can be evaluated in terms of BLEU, ROUGE, and VERT (for videos)
    - Against a human reference
    - For now, we will simply use BLEU (and maybe METEOR, CIDEr)
    - This makes our approach similar to a captioning task, and amenable to automatic evaluation (scores are really low anyway)
- Formality, fluency and meaning preservation. [1]

[1] Rao, Sudha, and Joel Tetreault. "Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer."



Analysis
ASR & Summarization

- We will conduct detailed analyses
- How are our “summarized” captions different from “baseline” ones (e.g. without the triplet loss)
- Can we attribute the differences to nouns/ objects, verbs/ actions, or other factors
- Visualize the data paths

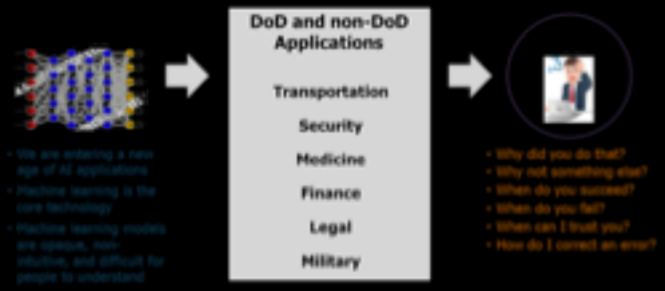


Overview

- Multi-modal Speech Recognition
- Video Summarization
- Multi-modal Video Summarization
- **Outlook**

- I will assume many things are known already

Explainable AI

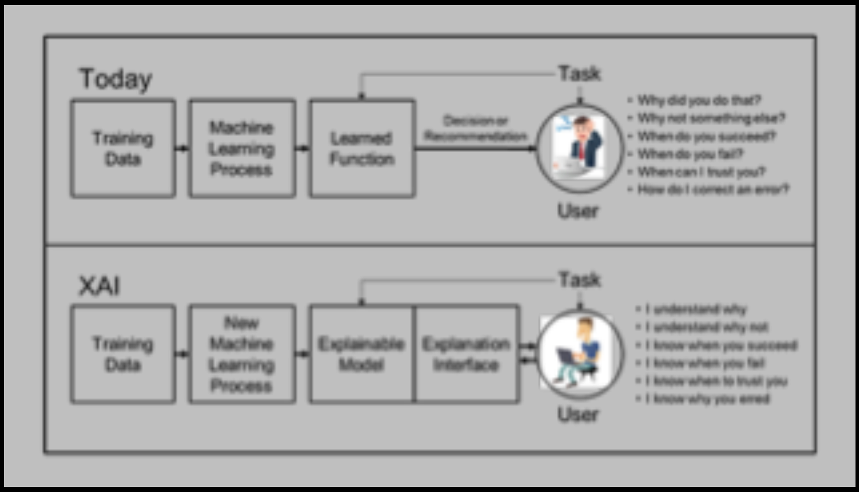


- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

DARPA “XAI”

Explainable AI



- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- I understand why
- I understand why not
- I know when you succeed
- I know when you fail
- I know when to trust you
- I know why you erred

Conversational Search

- Think of “Ambient Intelligence”
- Graphical user interfaces will go away
- We will do a lot less browsing than today
- Your “Search Assistant” will be a friend that helps you find the stuff you want

Outlook

- Summarization is the least obvious task
- We are talking to other MM and IR folks
- So far, we have only considered ideas that do not require the collection of more data (because Crowdflower …)
- Will have Capstone projects to continue this

Questions?

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