Multimodal Learning from Videos

Exploring Models and Task Complexities

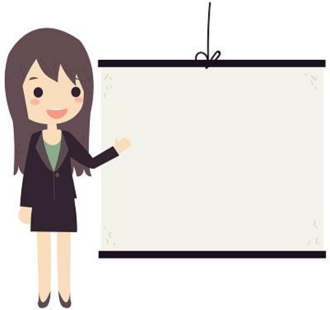
Shruti Palaskar

Thesis Proposal
April 28, 2021

Carnegie Mellon University
School of Computer Science

Language Technologies Institute

Human interaction is inherently multimodal



Videos have quickly become the largest form of data being generated & consumed

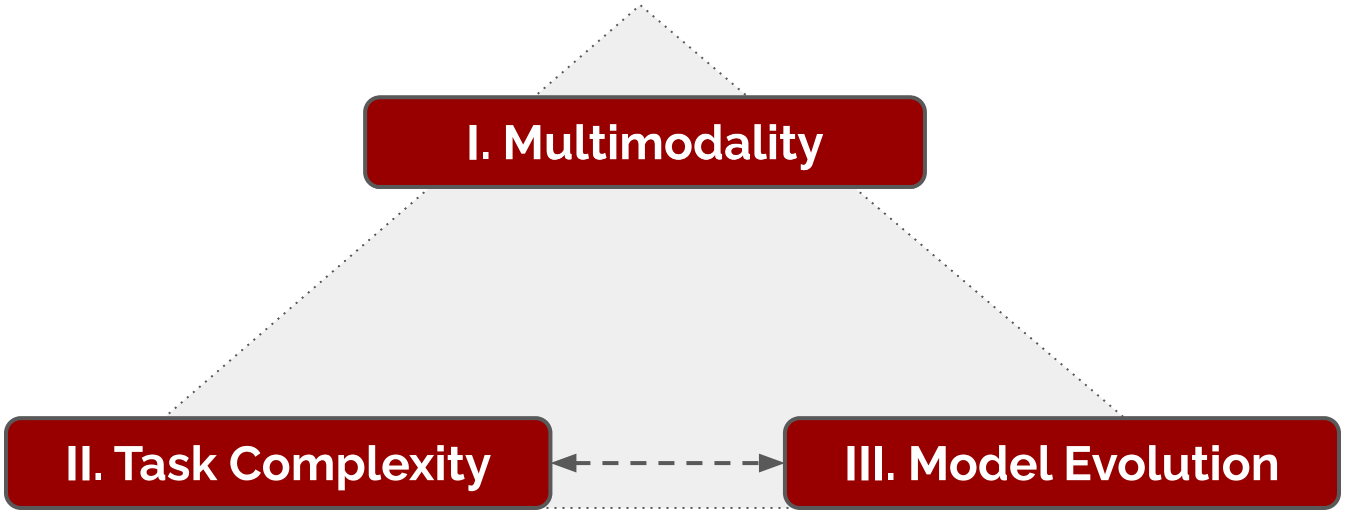
- 70% of YouTube viewers watch videos for "help with a problem" they are having in their hobby, work, or chores
- People engage equally if not more with Videos as with News, Music or Podcasts

NETFLIX YouTube coursera TikTok

Source: Understanding Value in Media: Perspectives from Consumers and Industry, World Economic Forum Report, 2020
Hubspot survey on Visual Content Marketing Statistics, 2021 https://blog.hubspot.com/marketing/visual-content-marketing-strategy

Thesis Statement

This thesis ranks four tasks of multimodal video understanding according to their complexity and shows how increasingly expressive models are important to perform well on each of these tasks.



Semantic Cues Across Modalities

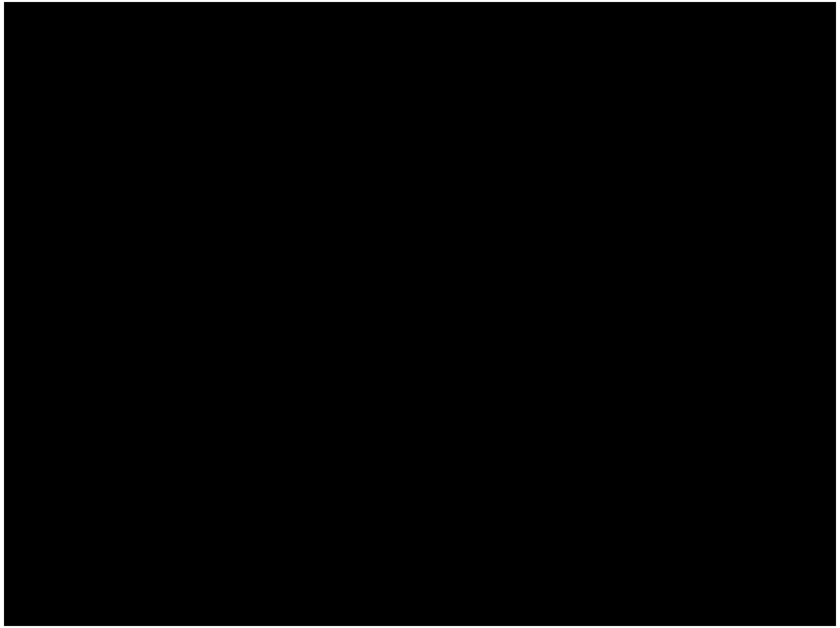
I. Multimodality

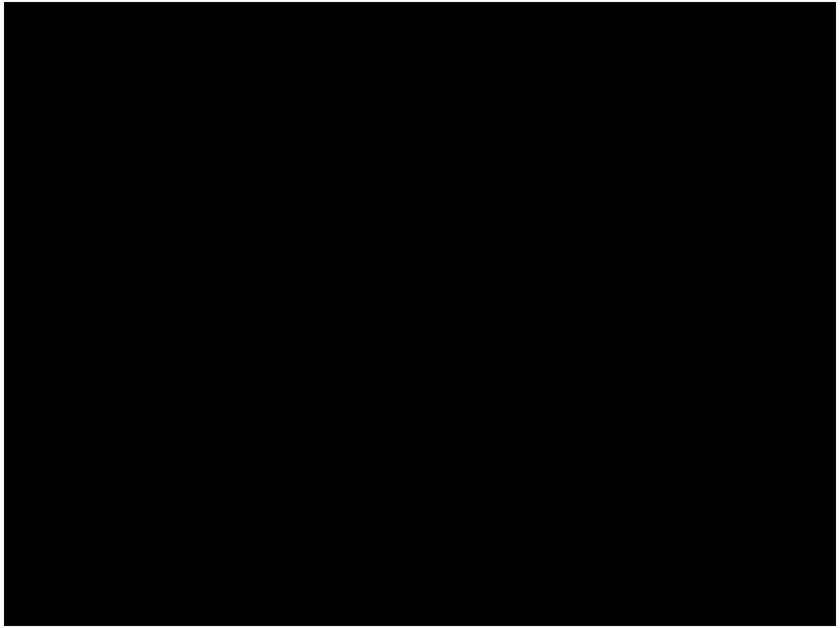
“Climate Change is the number one issue facing humanity.”

“Climate Change is the number one issue facing humanity.”

Semantic Cues Across Modalities

I. Multimodality





Understanding Videos is a Complex Problem

II. Task Complexity

Speech Recognition

Sound Event Detection

Video Tagging & Classification

Action Recognition

Dialog

Question Answering

Commonsense Reasoning

Pose Estimation

Scene Understanding

Summarization

Translation

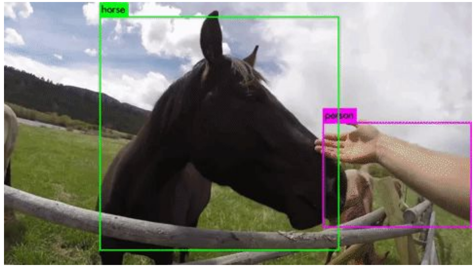
How to Repair a Polaris Pool Cleaner : Installing a Polaris 180 Pool Cleaner Head Float

- Visuals
- Audio & Speech
- Text Transcripts
- Title & Summary

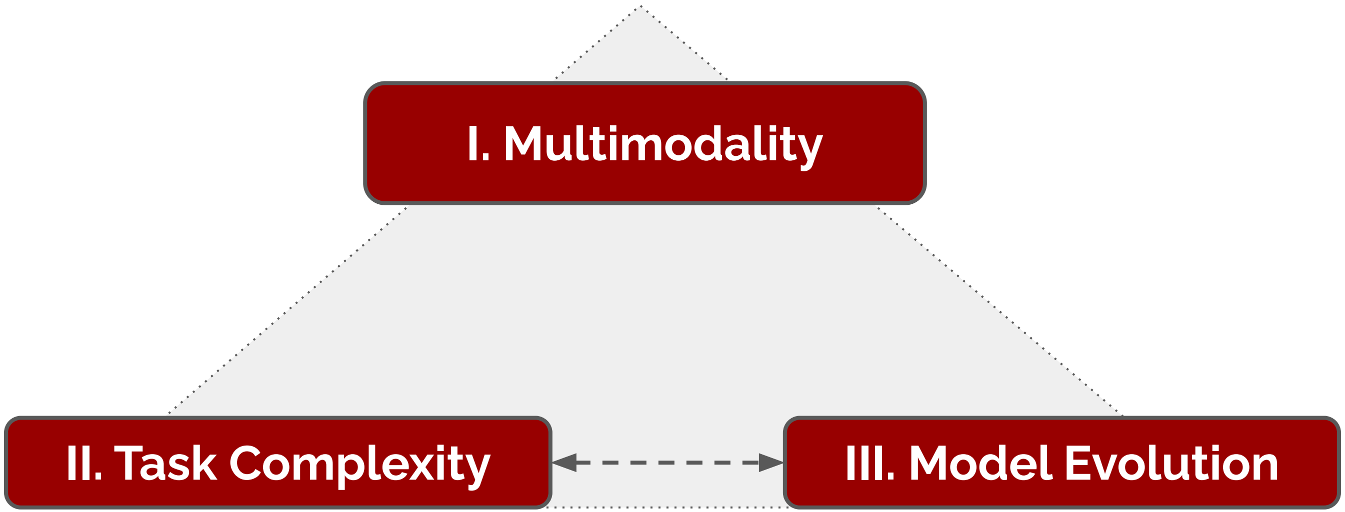
Understanding Videos is a Complex Problem

III. Model Evolution

… Because increasingly expressive models are important for satisfying task complexities



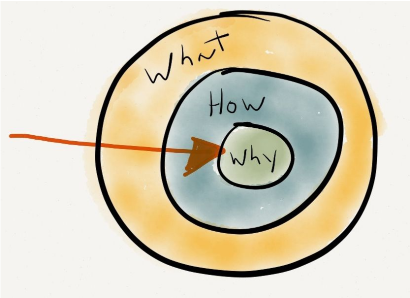
Thesis Motivation



Learning Tasks in this Thesis

Multimodal Video Understanding

- Multimodal Speech Recognition
- Multimodal Speech Translation
- Multimodal Summarization & QA
- Multimodal Rationalization



Adding Modalities Increases Task Complexity

Multimodal Speech Recognition

So let's get started.

MONOTONIC TASK

Multimodal Speech Translation

So let's get started.

Então vamos começar.

NON-MONOTONIC TASK

Multimodal Summarization & QA

So let's get started.
[Question] ...

Watch a seasoned professional ...

[Answer] ...

ABSTRACTION TASK

Multimodal Rationalization

So let's get started.
Watch a seasoned professional ...
[Question] …

[Answer] ...

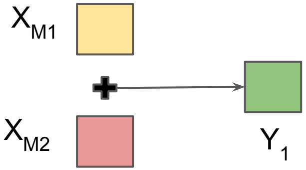
[Rationale] *Because ...*

EXPLANATORY TASK

Model Evolution Across Learning Tasks

Multimodal Speech Recognition

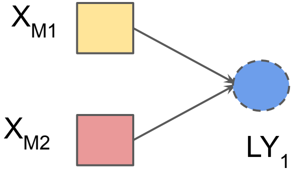
MONOTONIC TASK



Input Fusion

Multimodal Speech Translation

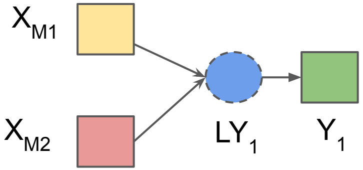
NON-MONOTONIC TASK



Latent Representation Fusion

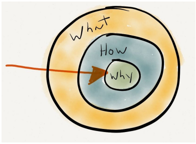
Multimodal Summarization & QA

ABSTRACTION TASK

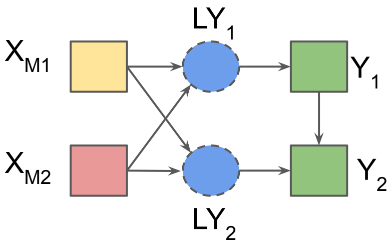


Hierarchical Latent Representation Fusion

Multimodal Rationalization



EXPLANATORY TASK



Hierarchical Interpretable Fusion

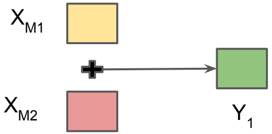
Outline

MONOTONIC TASK

I. Multimodal Speech Recognition

ICASSP '18, SLT '18

So let's get started.

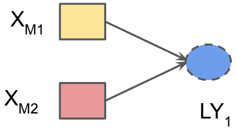


NON-MONOTONIC TASK

II. Multimodal Speech Translation

ICASSP '19, ICASSP '19

So let's get started.
Então vamos começar.

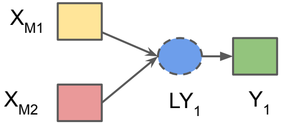


ABSTRACTION TASK

III. Multimodal Summarization & QA

ACL '19, DSTC AAAI '19, CS&L '20

So let's get started.
[Qn] ...
[Ans] ...

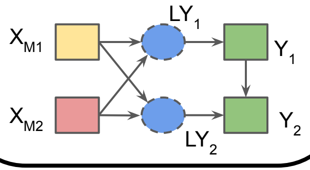


EXPLANATORY TASK

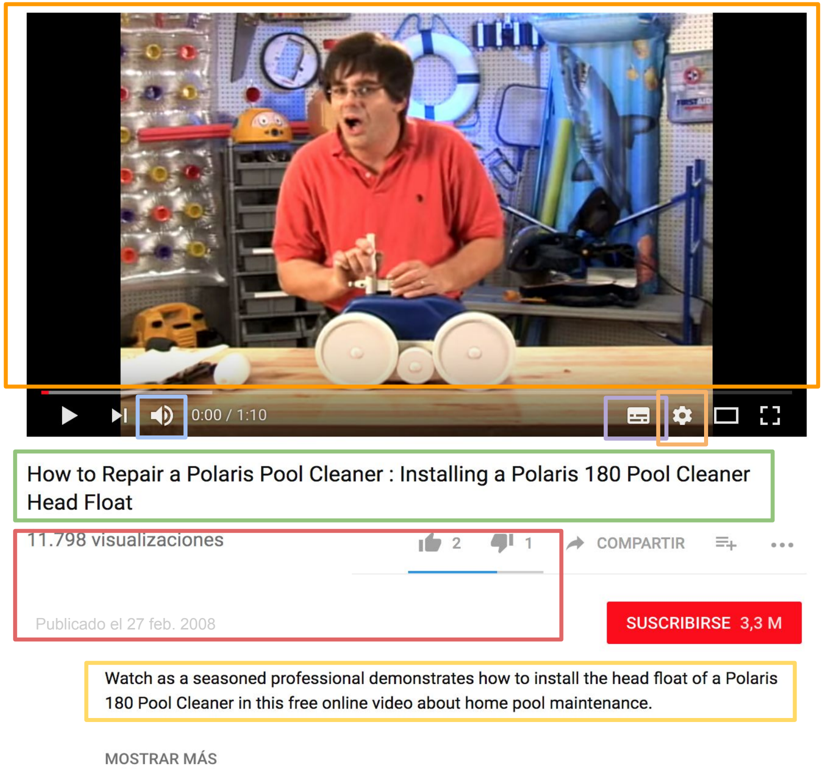
IV. Multimodal Rationalization

Proposed Work

So let's get started.
Watch a seasoned profess...
[Qn] ...
[Ans] ...
[R] *Because* ...



How2 Dataset



Visuals

Audio & Speech

English Transcripts

Portuguese Transcripts

Title

How to Repair a Polaris Pool Cleaner?

Metadata

(likes, dislikes, views, …)

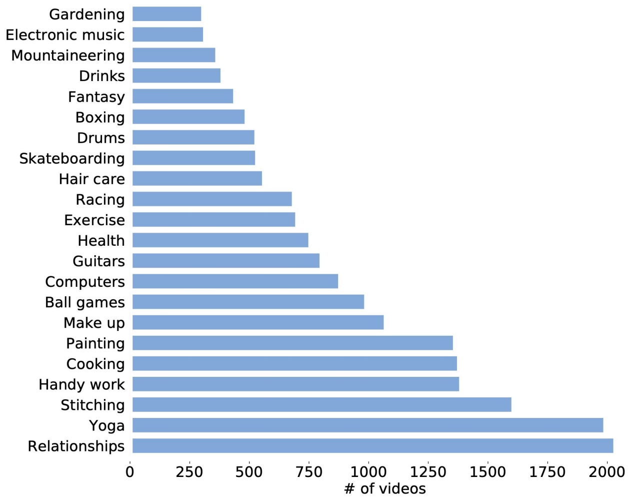
Summary

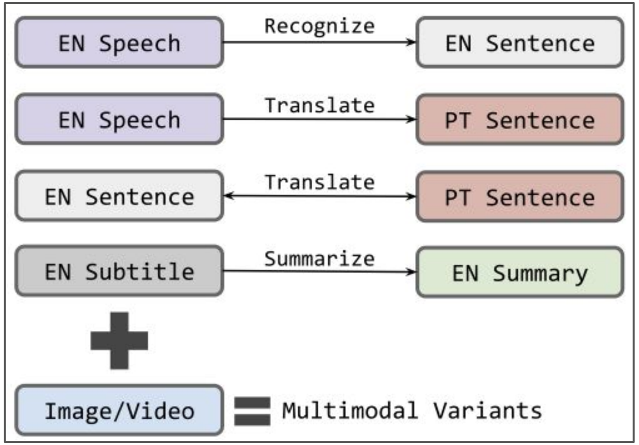
Watch as a seasoned professional demonstrates …

*"**How2: A Large-Scale Dataset for Multimodal Language Understanding**", Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliot, Loic Barrault, Lucia Specia, and Florian Metze, ViGIL Workshop @ NeurIPS 2018, Montreal, Canada*

How2 Dataset

- Multimodal Language Understanding
- Open-domain instructional videos corpora
- 5-way parallel modalities
- 80,000 videos; ~2000 hours
- Variety of topics



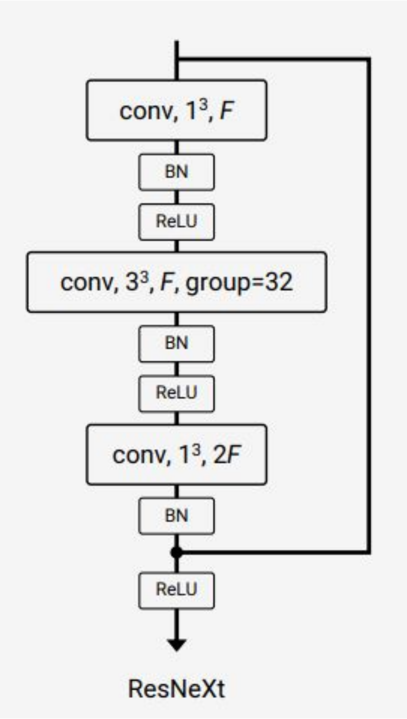


		Videos	Hours	Clips/Sentences
300h	train	13,168	298.2	184,949
	val	150	3.2	2,022
	test	175	3.7	2,305
	held	169	3.0	2,021
2000h	train	73,993	1,766.6	-
	val	2,965	71.3	-
	test	2,156	51.7	-

*"**How2: A Large-Scale Dataset for Multimodal Language Understanding**", Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliot, Loic Barrault, Lucia Specia, and Florian Metze, ViGIL Workshop @ NeurIPS 2018, Montreal, Canada*

How2 Dataset

- Object Features (Frame-level) ResNet-152 (He et al. 2016)
- Place Features (Frame-level) ResNet-50 (Zhou et al. 2017)
- Action Features (Video-level) ResNeXt 101 (Hara et al. 2018)



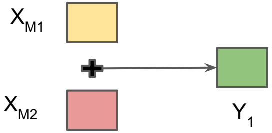
Outline

MONOTONIC TASK

I. Multimodal Speech Recognition

ICASSP ‘18, SLT ‘18

So let's get started.

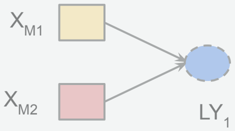


NON-MONOTONIC TASK

II. Multimodal Speech Translation

ICASSP ‘19, ICASSP ‘19

Então vamos começar.
So let's get started.

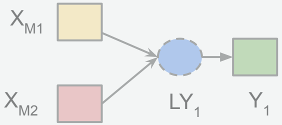


ABSTRACTION TASK

III. Multimodal Summarization & QA

ACL ‘19, DSTC AAAI ‘19, CS&L ‘20

So let's get started.
[Qn] ...
[Ans] ...

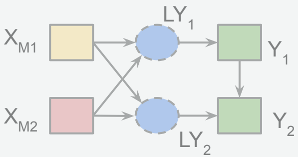


EXPLANATORY TASK

IV. Multimodal Rationalization

Proposed Work

So let's get started.
Watch a seasoned profess...
[Qn] ...
[Ans] ...
[R] *Because* ...



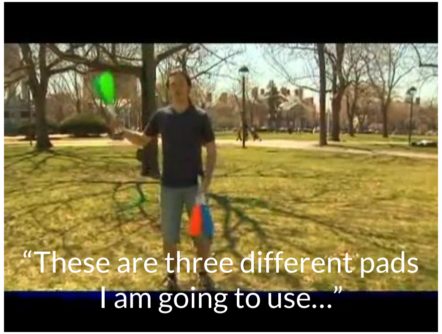
I. Multimodal Speech Recognition

Task Description

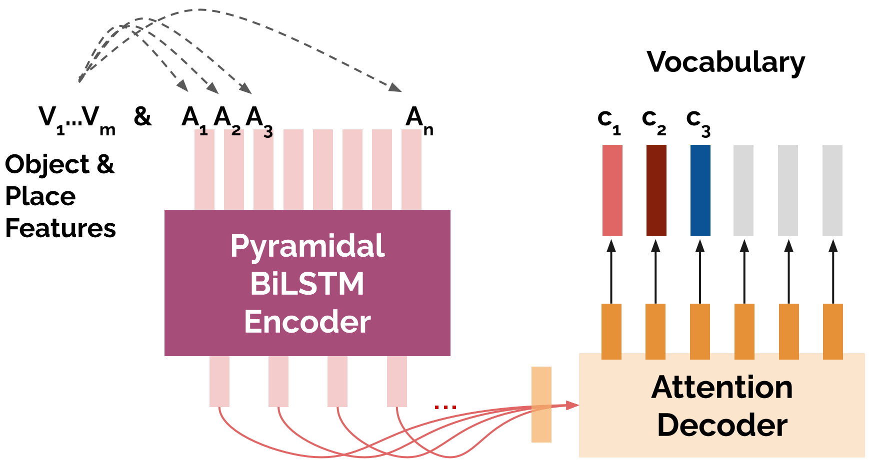
- How-To videos recorded in a wide variety of settings
  - Indoors vs. Outdoors
  - Close microphone vs. Distant microphone
  - Home recording setups or handheld devices

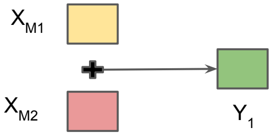
- Lot of acoustic noise compared to standard speech recognition corpora
  - WERs ~15-25% compared to ~3-10% of pure-ASR setup

- Can Visual information that is often highly correlated with the spoken narration help improve ASR?



Input Fusion Model

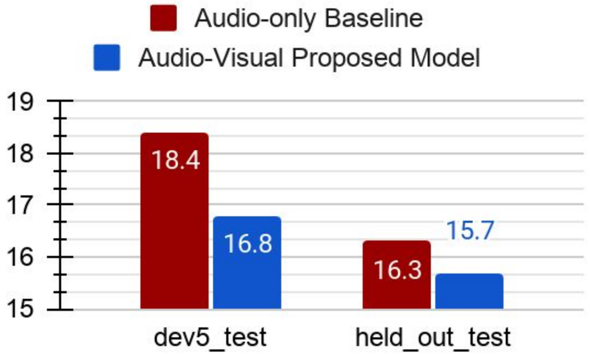




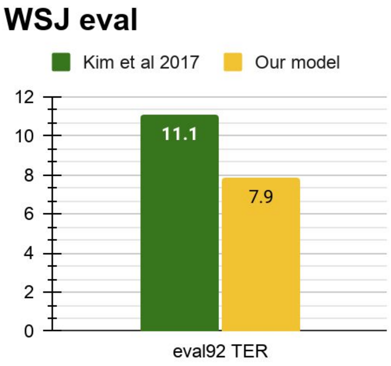
- Frame-level and utterance-level multimodal control for effective fusion
  - Object & Place features
- **Introducing end-to-end sequence-to-sequence model for audio-visual speech recognition**
  (2017-2018)

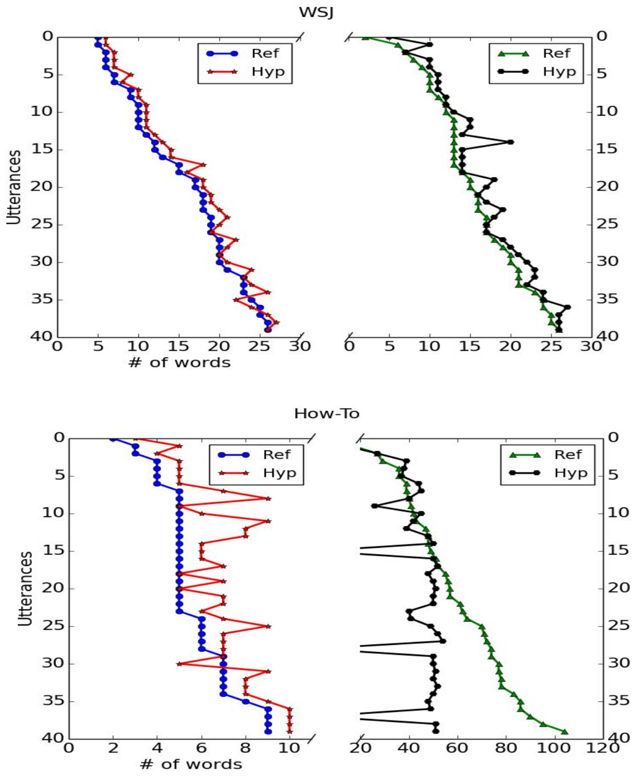
“***End-to-End Multimodal Speech Recognition***”, *Shruti Palaskar*, Ramon Sanabria*, and Florian Metze, ICASSP 2018, Calgary, Canada*
“***Multimodal Grounding for Sequence-to-Sequence Speech Recognition***”, *Caglayan, Sanabria, Palaskar, Barrault, and Metze, ICASSP 2019, Brighton, UK*

Results



💡 **8.7% relative TER improvement**





*"**End-to-End Multimodal Speech Recognition**", Shruti Palaskar*, Ramon Sanabria*, and Florian Metze, ICASSP 2018, Calgary, Canada*
*"**Multimodal Grounding for Sequence-to-Sequence Speech Recognition**", Ozan Caglayan, Ramon Sanabria, Shruti Palaskar, Loic Barrault, and Florian Metze, ICASSP 2019, Brighton, UK*

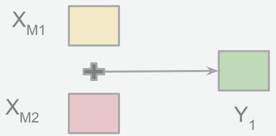
Outline

MONOTONIC TASK

I. Multimodal Speech Recognition

ICASSP ‘18, SLT ‘18

So let's get started.

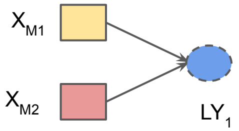


NON-MONOTONIC TASK

II. Multimodal Speech Translation

ICASSP ‘19, ICASSP ‘19

So let's get started.
Então vamos começar.

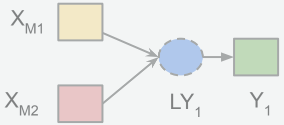


ABSTRACTION TASK

III. Multimodal Summarization & QA

ACL ‘19, DSTC AAAI ‘19, CS&L ‘20

So let's get started.
[Qn] ...
[Ans] ...

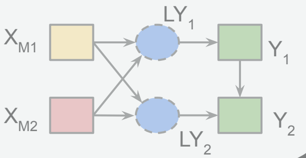


EXPLANATORY TASK

IV. Multimodal Rationalization

Proposed Work

So let's get started.
Watch a seasoned profess...
[Qn] ...
[Ans] ...
[R] *Because* ...



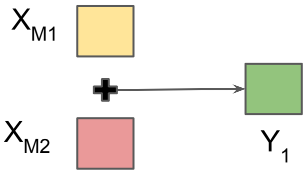
II. Multimodal Speech Translation

Task Description

- Direct Speech Translation
  - No intermediate speech-to-text step
  - English Speech to Portuguese Text
- Semi-supervised modeling that uses inherent cross-modal supervision
  - Fully supervised sequence-to-sequence based approaches can be applied to multimodal tasks
  - But, can the inherent cross-modal supervision available through speech, english text, and vision, facilitate direct speech translation?

Model Evolution

What's missing in the previous model?



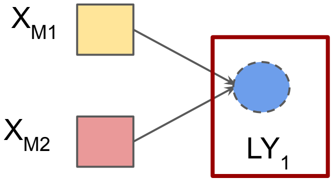
Input Fusion

Strict monotonic correspondence

MONOTONIC TASK

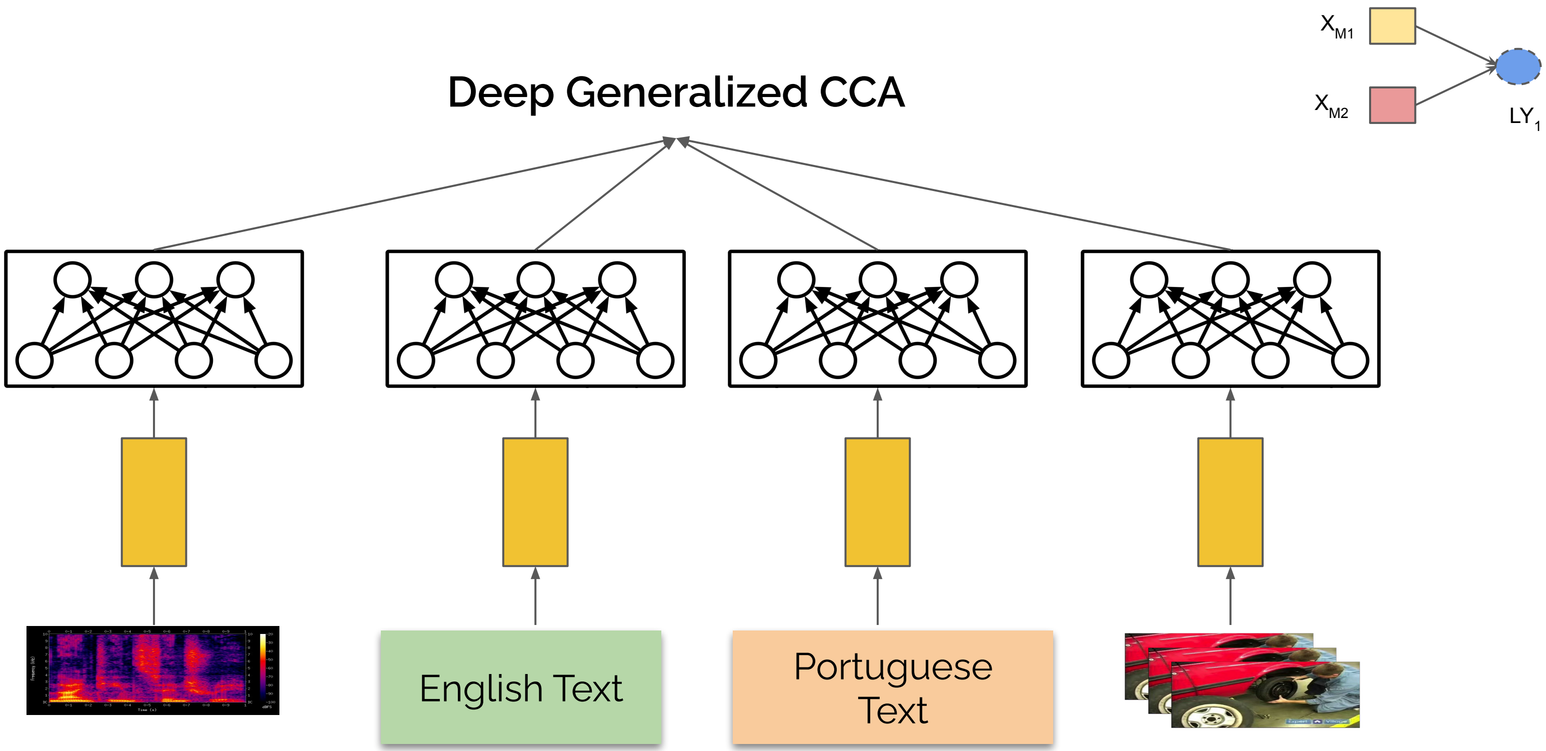
NON-MONOTONIC TASK

- Multimodal adaptation for re-ordered outputs
- Latent space adaptation as no monotonic constraint
- Latent space adaptation also opens the possibility of training with lesser supervision



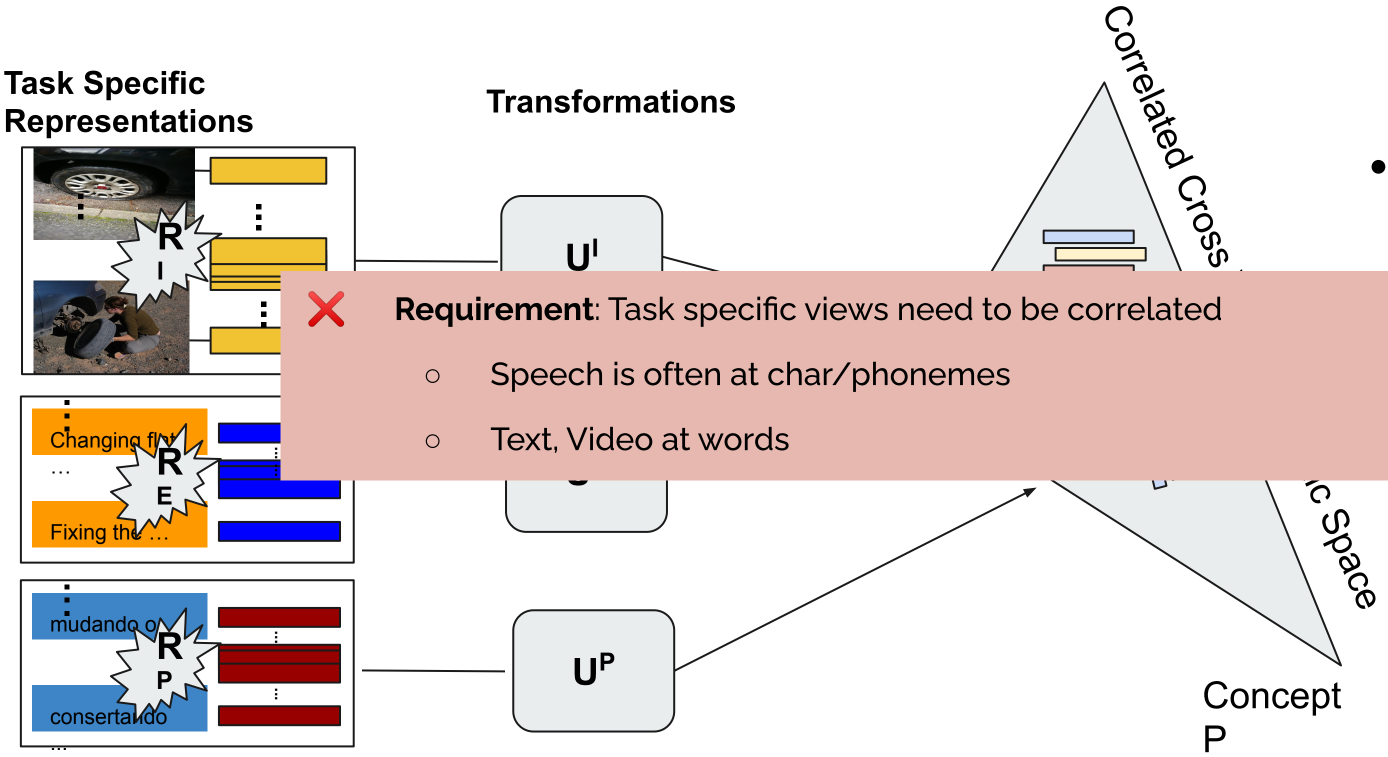
Latent Representation Fusion

Latent Representation Fusion Model



*"**Learning From Multiview Correlations in Open-Domain Videos**", Nils Holzenberger*, Shruti Palaskar*, Pranava Madhyastha, Florian Metze, and Raman Arora ICASSP 2019, Brighton, UK*

Deep Generalized Canonical Correlation Analysis



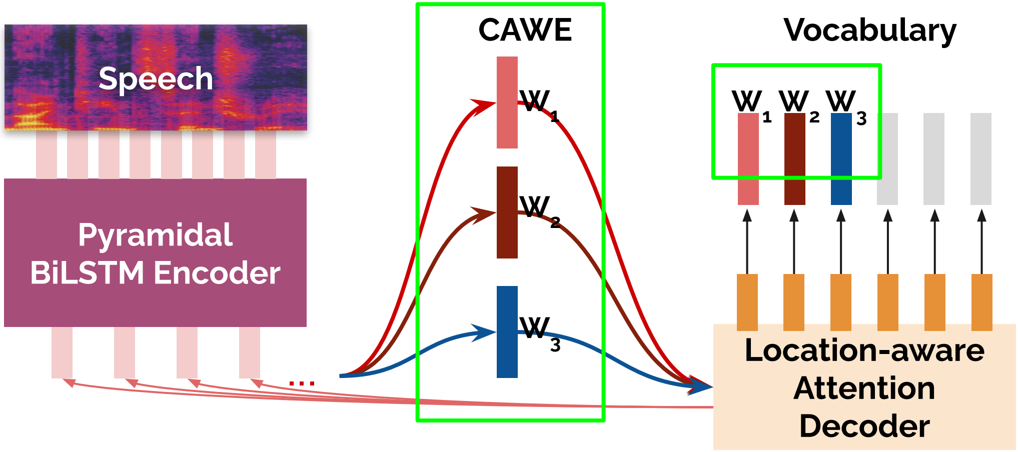
❌ **Requirement**: Task specific views need to be correlated

- ○ Speech is often at char/phonemes
- ○ Text, Video at words

- Obtain maximally correlated information that is consistent with h view
- os denoise and und information
- kimize mutually relevant info

"***Deep Generalized Canonical Correlation Analysis***", *Adrian Benton, Huda Khayrallah, Biman Gujral, Dee Ann Reisinger, Sheng Zhang, Raman Arora, 4th Workshop on RepL4NLP, ACL 2019, Florence, Italy*

Contextual Acoustic Word Embeddings



- Build Direct Acoustic-to-Word models
- Proposed approach learns CAWE as a by product of training acoustic-to-word ASRs
- Evaluated on 16 standard benchmarks

CAWE-W: Averaged with attention weights

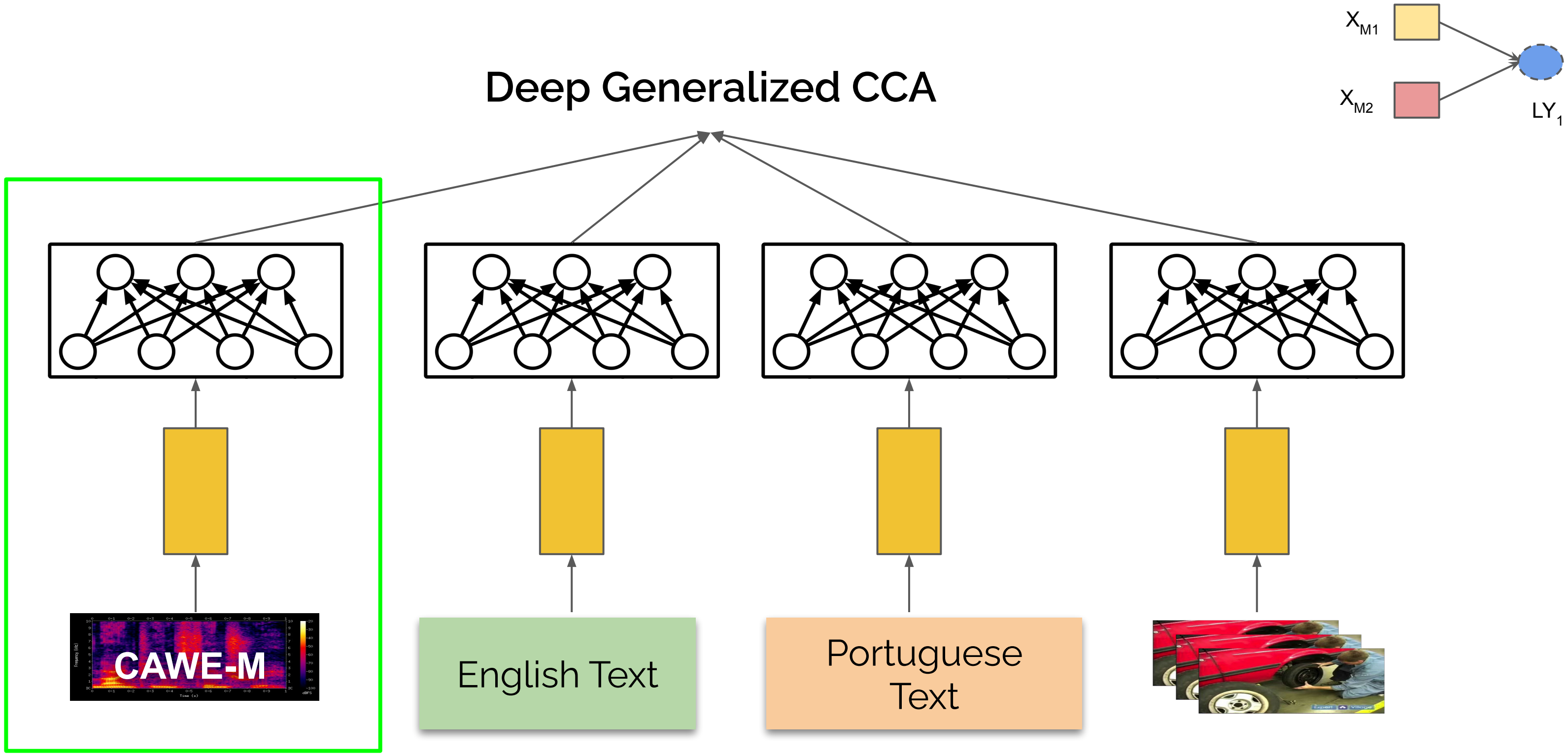
$$w_i = \frac{\sum_{k \in K} attention(a_k) \cdot encoder(a_k)}{n(K)}$$

CAWE-M: Arg max of attention weights

$$w_i = encoder(a_k) \;\; where \;\; k = \arg\max_{k \in K} attention(a_k)$$

"***Learned In Speech Recognition: Contextual Acoustic Word Embeddings***", *Shruti Palaskar*, Vikas Raunak*, and Florian Metze, ICASSP 2019, Brighton, UK*

Latent Representation Fusion Model



*"**Learning From Multiview Correlations in Open-Domain Videos**", Nils Holzenberger*, Shruti Palaskar*, Pranava Madhyastha, Florian Metze, and Raman Arora ICASSP 2019, Brighton, UK*

Results

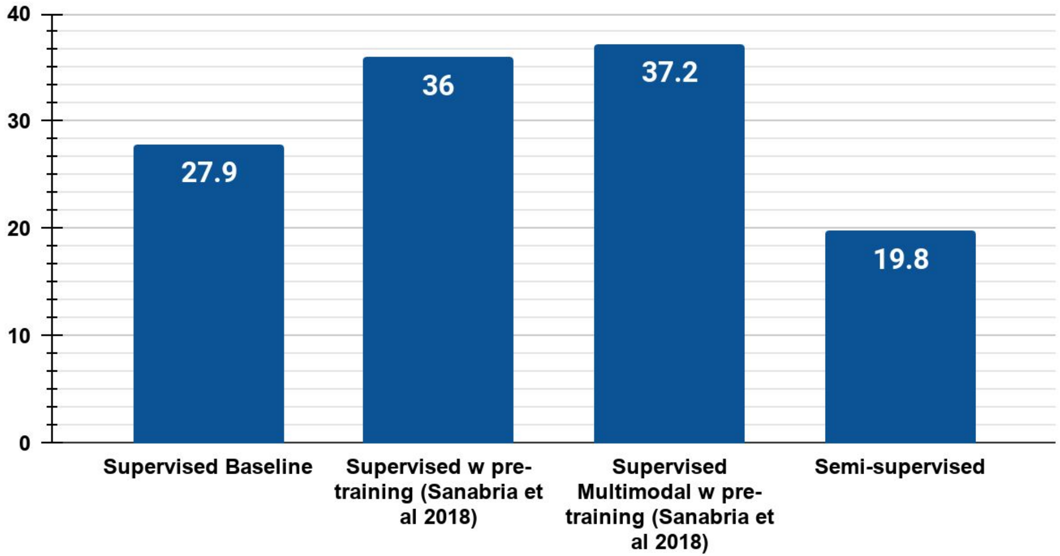
Retrieval-based evaluation

Portuguese reference sentences

Input speech segment

Hypothesis for Spoken Language Translation

BLEU scores comparison

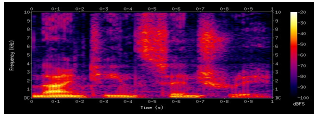
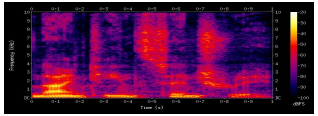


SLT improves with multimodal information

Semi-supervised Speech Translation model achieves up to 50-70% performance of a fully supervised models

*“**Learning From Multiview Correlations in Open-Domain Videos**”, Nils Holzenberger*, Shruti Palaskar*, Pranava Madhyastha, Florian Metze, and Raman Arora ICASSP 2019, Brighton, UK*

Results

Recall@10		English Text	Portuguese Text	
	-	85.4	70.7	1.0 (didn't work)
English Text	85.4	-	98.4	0.9
Portuguese Text	71.0	98.3	-	1.1
	1.1	1.1	0.9	-

Semi-supervised cross-modal learning can also be applied to speech recognition & machine translation

*"**Learning From Multiview Correlations in Open-Domain Videos**", Nils Holzenberger*, Shruti Palaskar*, Pranava Madhyastha, Florian Metze, and Raman Arora ICASSP 2019, Brighton, UK*

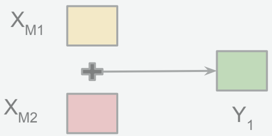
Outline

MONOTONIC TASK

I. Multimodal Speech Recognition

ICASSP '18, SLT '18

So let's get started.

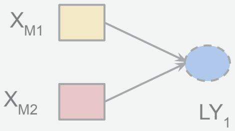


NON-MONOTONIC TASK

II. Multimodal Speech Translation

ICASSP '19, ICASSP '19

So let's get started.
Então vamos começar.

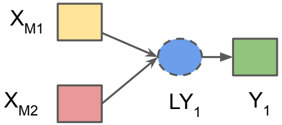


ABSTRACTION TASK

III. Multimodal Summarization & QA

ACL '19, DSTC AAAI '19, Elsevier CS&L '20

So let's get started.
[Qn] ...
[Ans] ...

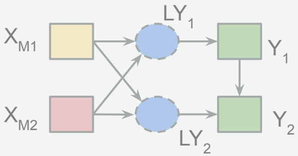


EXPLANATORY TASK

IV. Multimodal Rationalization

Proposed Work

So let's get started.
Watch a seasoned profess...
[Qn] ...
[Ans] ...
[R] *Because* ...



III. Multimodal Summarization & QA

Multimodal Summarization - Task Description

Spanish Omelet

1 minute 7 seconds of audio and video

Summary (26 words)

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

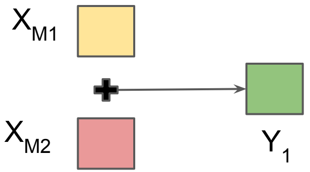
Transcript (215 words)

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 .

“***Multimodal Abstractive Summarization for How2 Videos***”, *Shruti Palaskar, Jindrich Libovicky, Spandana Gella, and Florian Metze, ACL 2019, Florence, Italy*

Model Evolution

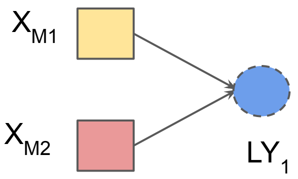
What's missing in the previous models?



Input Fusion

Utterance-level
Adaptation

<u>MONOTONIC TASK</u>



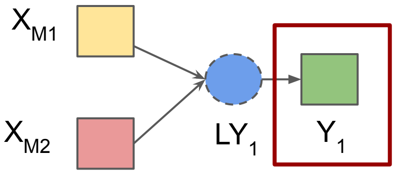
Latent Representation
Fusion

Utterance-level
Adaptation

<u>NON-MONOTONIC TASK</u>

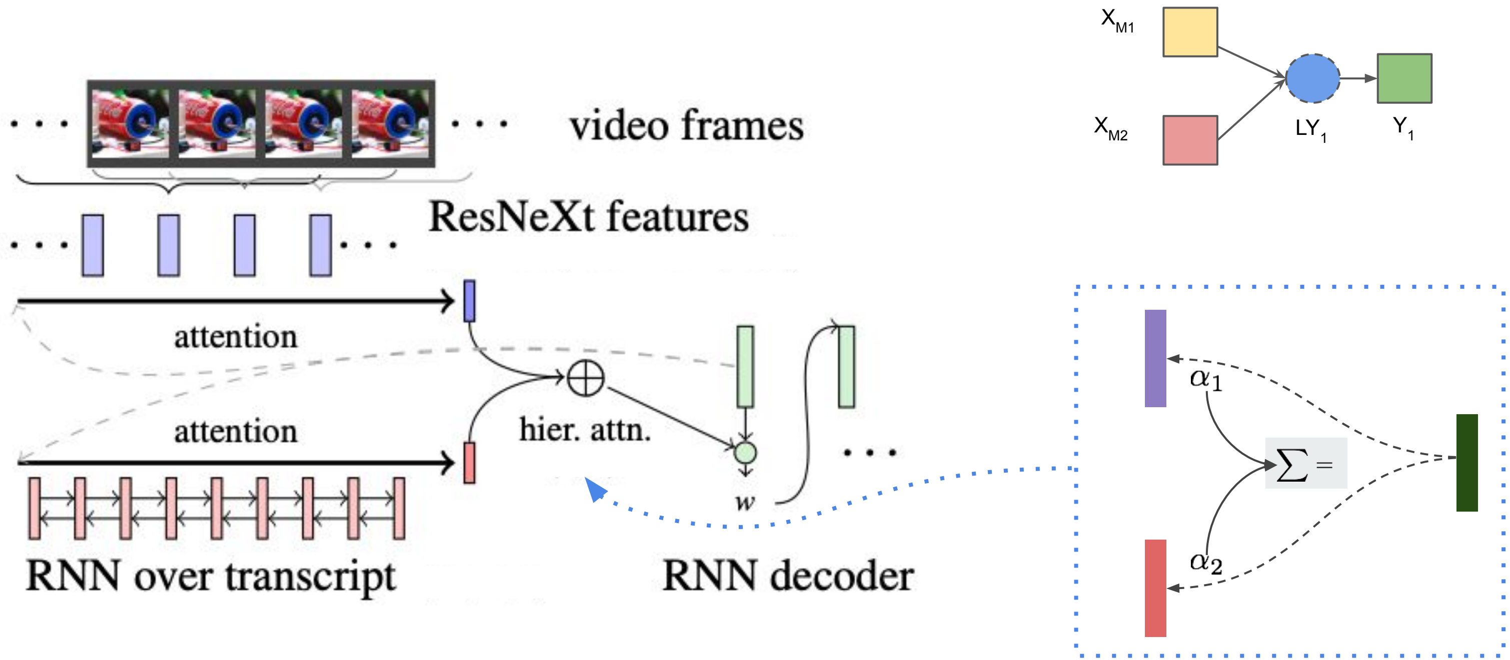
ABSTRACTION TASK

- Video-level Multimodal Adaptation
- Video-level Information Flow
- Information Selection, Compression & Restructuring



Hierarchical Latent
Representation
Fusion

Hierarchical Latent Representation Fusion Model



"***Attention strategies for multi-source sequence-to-sequence learning***", *Jindrich Libovicky and Jindrich Helcl, ACL 2017*
"***Multimodal Abstractive Summarization for How2 Videos***", *Shruti Palaskar, Jindrich Libovicky, Spandana Gella, and Florian Metze, ACL 2019, Florence, Italy*

Evaluation

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

- **Content F1 (Proposed Evaluation)**
  - Focus on content words
  - Zero weight to function words
  - Equal weight to Precision and Recall
  - Ignores fluency

Catchphrases in teasers

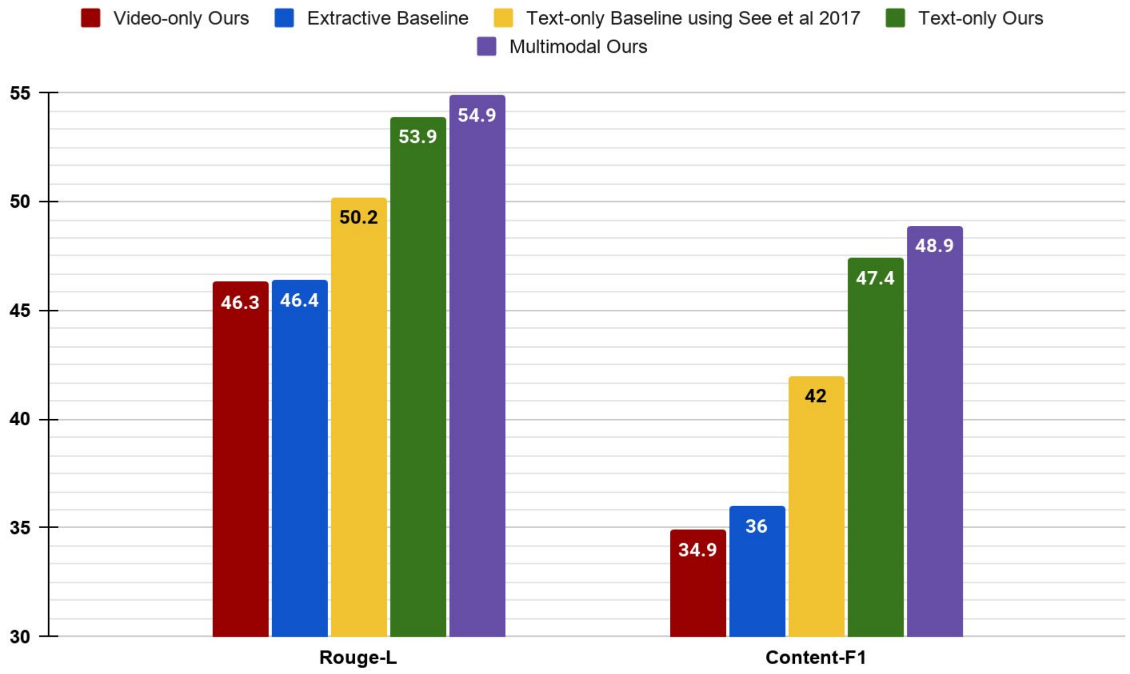
```
3799 in
3058 this
2922 free
2832 video
1948 learn
1460 how
1321 tips
 756 expert
```

>=500 times

~~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~~ ~~.~~

“***Multimodal Abstractive Summarization for How2 Videos***”, *Shruti Palaskar, Jindrich Libovicky, Spandana Gella, and Florian Metze, ACL 2019, Florence, Italy*

Results



💡 **3.2% relative improvement in Content F1 score**

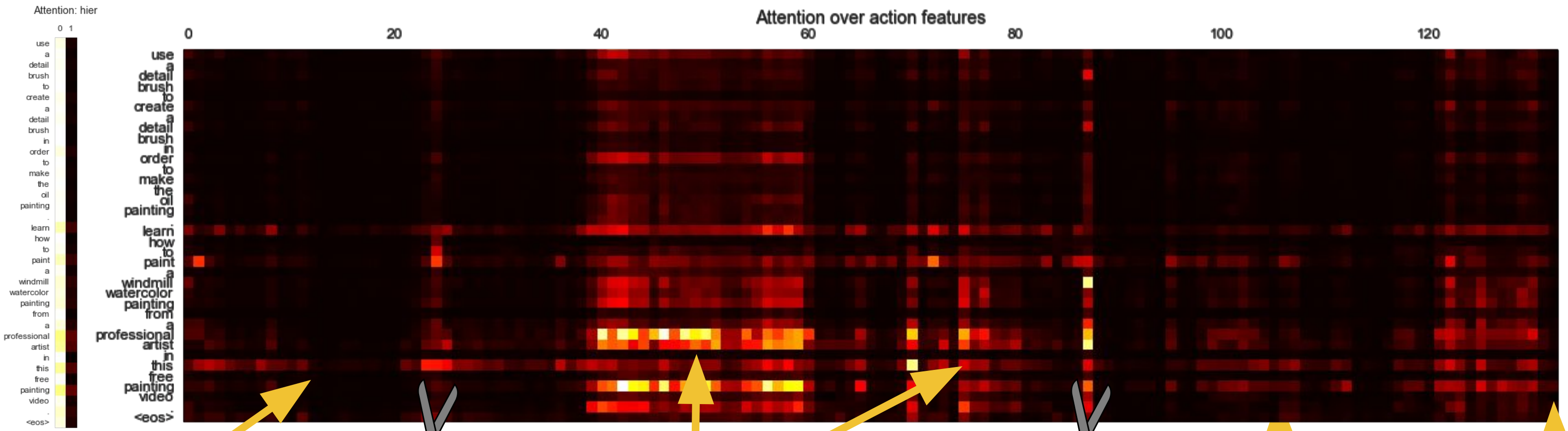
Human Evaluation on Informativeness, Relevance, Coherence, and Fluency

Model	INF	REL	COH	FLU
Text-only	3.86	**3.78**	3.78	3.92
Video-only	3.58	3.30	3.71	3.80
Text-and-Video	**3.89**	3.74	**3.85**	**3.94**

💡 **Multimodal summaries preferred by human evaluators**

*"**Multimodal Abstractive Summarization for How2 Videos**", Shruti Palaskar, Jindrich Libovicky, Spandana Gella, and Florian Metze, ACL 2019, Florence, Italy*

Results - Attention Analysis



cut

cut

Talking and preparing the brush

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

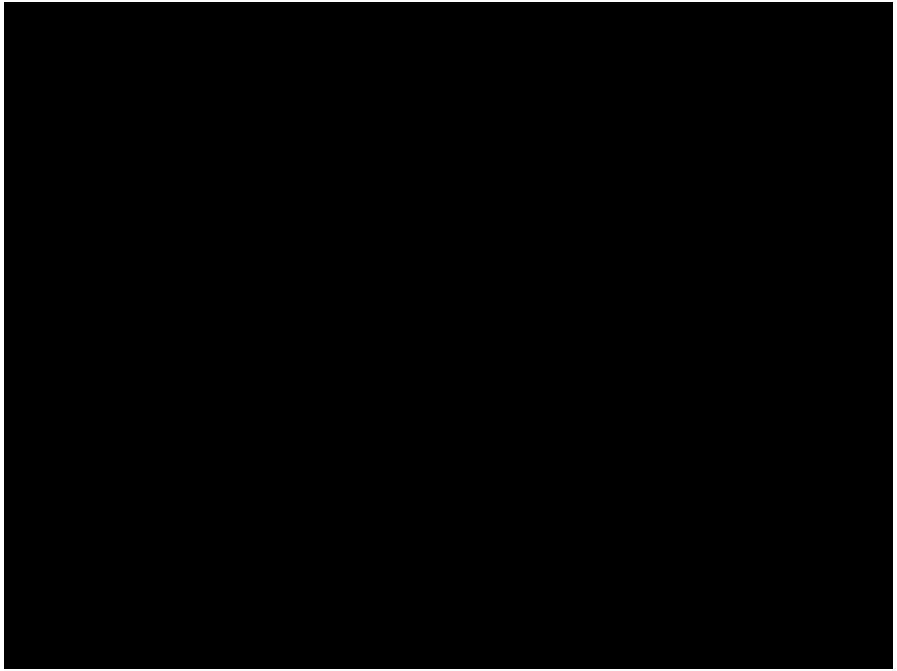
Close-up of brushstrokes **no hand**

Black frames at the end

```
Learn how to paint a windmill watercolor painting from a professional artist in this
free painting video.
```

Transfer Learning from Summarization to QA

Multimodal QA - Task Description



QUESTIONS

is there only one person ?
does she walk in with a towel around her neck ?
does she interact with the dog ?
does she drop the towel on the floor ?

ANSWERS

there is only one person and a dog .
she walks in from outside with the towel around her neck .
she does not interact with the dog
she dropped the towel on the floor at the end of the video .

SUMMARY

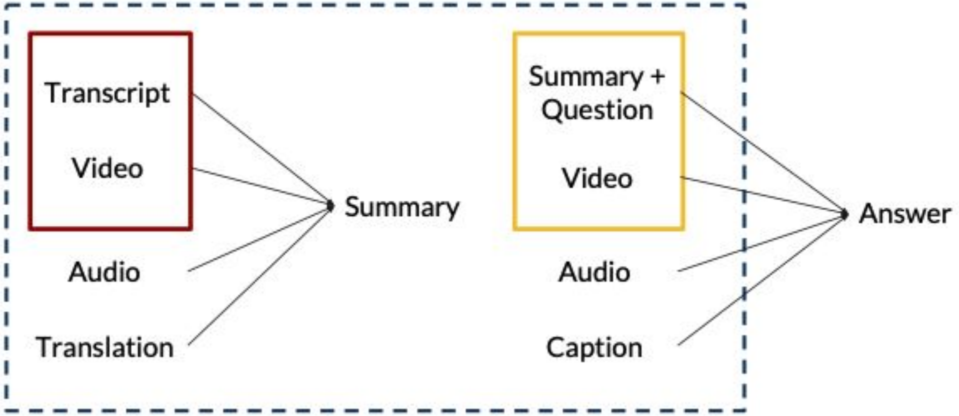
the girl walks into a room with a dog with a towel around her neck . she does some stretches and then drops the towel .

CAPTION

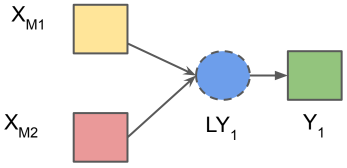
a person walked through a doorway into the living room with a towel draped around their neck , and closed the door . the person stretched and threw the towel on the floor .

*"**Audio-Visual Scene-aware Dialog**", Huda Alamri, Vincent Cartillier, Abhishek Das, ..., Devi Parikh, CVPR 2019*

Transfer Learning Setup



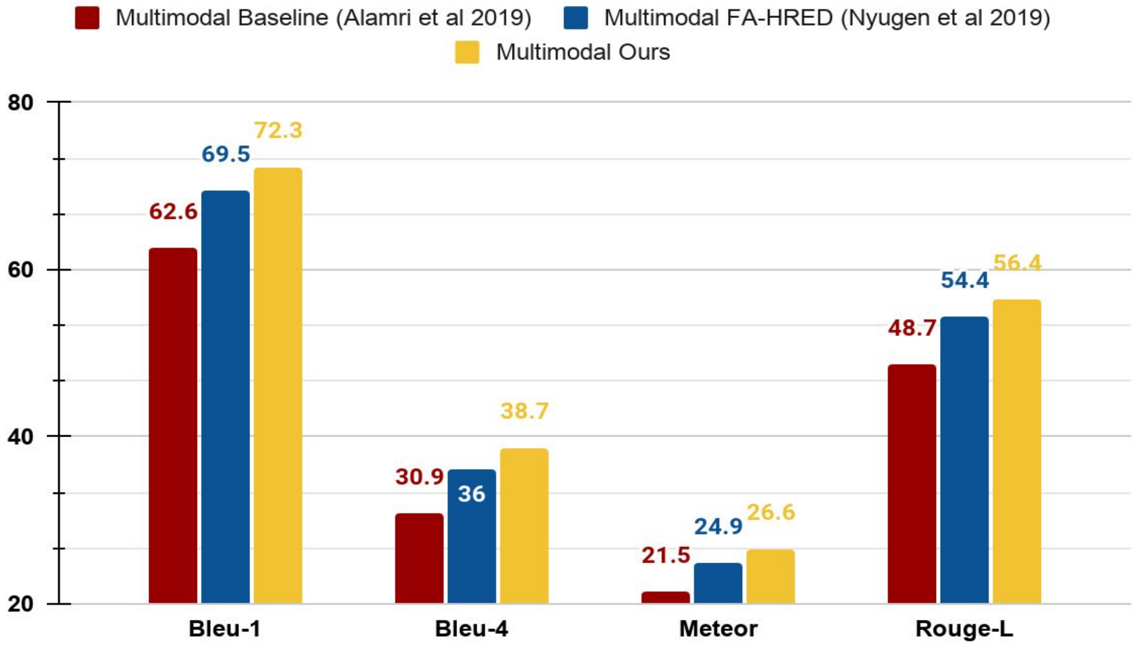
- Fine-tuning the trained Hierarchical Latent Representation Fusion model for QA
- Framing QA as a Summarization task led to optimal gains
- Abstraction Task
  - Compression
  - Rephrasing
  - **Information Selection**



	Charades		How2
Split	Sentences	Videos	Videos
train	76590	7659	73993
val	17870	1787	2965
test	7330	733	2156
held_out	6745	1710	169

*"**Transfer Learning for Multimodal Dialog**", Shruti Palaskar*, Ramon Sanabria*, and Florian Metze, DSTC AAAI '19 & Elsevier Computer Speech and Language '20*

Results



💡 **Significant absolute improvements across all metrics compared with a strong baseline provided by challenge organizers!**

💡 **Our approach was the winning system on both automatic and human evaluation of the inaugural Video QA challenge**

*"**Transfer Learning for Multimodal Dialog**", Shruti Palaskar*, Ramon Sanabria*, and Florian Metze, DSTC AAAI '19 & Elsevier Computer Speech and Language '20*

Example Outputs

Question: is he talking or reading out loud ? **Answer**: no , he is not talking at all .
Question: what 's in the mug ? **Answer**: i don 't know , i can 't see the inside .
Question: hello . did someone come to the door ? **Answer**: no and it is a window that he is standing in front of .
Question: are they talking in the video ? **Answer**: not really no i don 't hear anything

*"**Transfer Learning for Multimodal Dialog**", Shruti Palaskar*, Ramon Sanabria*, and Florian Metze, DSTC AAAI '19 & Elsevier Computer Speech and Language '20*

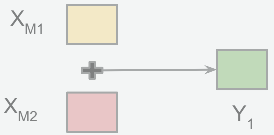
Outline

MONOTONIC TASK

I. Multimodal Speech Recognition

ICASSP '18, SLT '18

So let's get started.

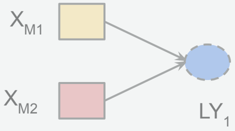


NON-MONOTONIC TASK

II. Multimodal Speech Translation

ICASSP '19, ICASSP '19

So let's get started.
Então vamos começar.

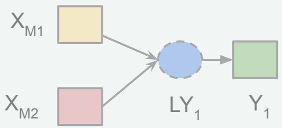


ABSTRACTION TASK

III. Multimodal Summarization & QA

ACL '19, DSTC AAAI '19, CS&L '20

So let's get started.
[Qn] ...
[Ans] ...

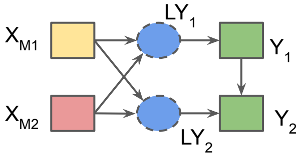


EXPLANATORY TASK

IV. Multimodal Rationalization

Proposed Work

So let's get started.
Watch a seasoned profess...
[Qn] ...
[Ans] ...
[R] *Because* ...

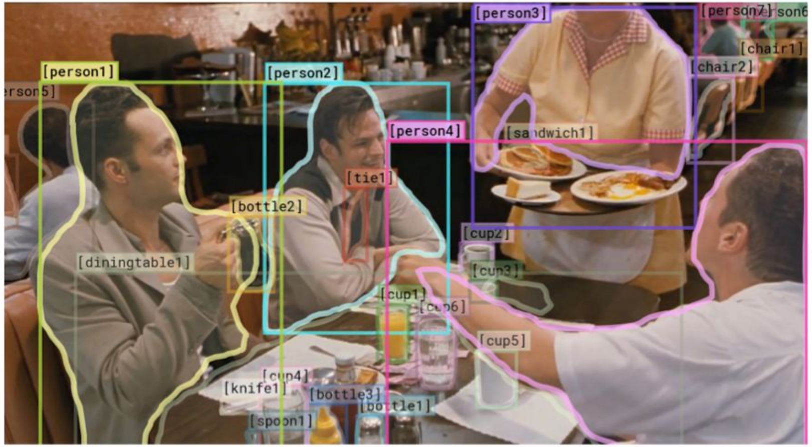


IV. Multimodal Rationalization

PROPOSED WORK

Task Description

Visual Commonsense Reasoning



Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

*I chose **a)** because...*

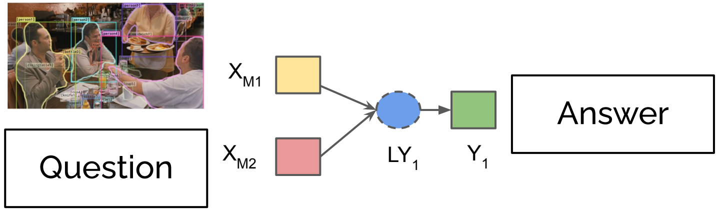
a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.

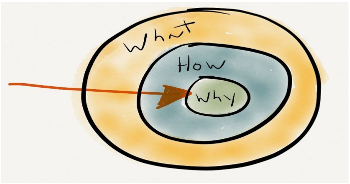
"***From Recognition to Cognition: Visual Commonsense Reasoning***", *Rowan Zellers, Yonatan Bisk, Ali Farhadi, Yejin Choi, CVPR 2019*

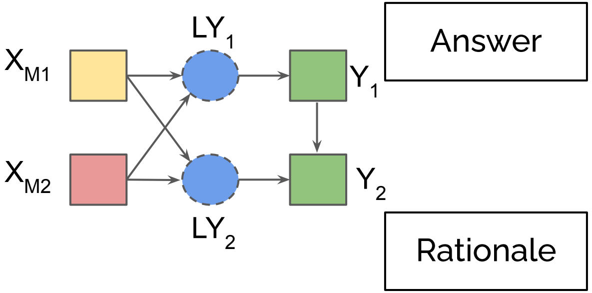
Proposed Work & Hypotheses

Beyond Video Question Answering through *Explanations*

Next type of task in the series so far; interpretable language understanding through explanations; increased complexity







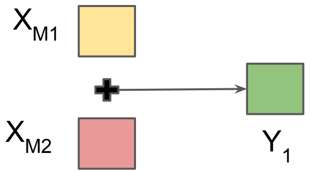
Hierarchical Interpretable Fusion

<u>Hypotheses</u>:

1. We can design open-ended rationalization as an extension of abstraction task for language generation
2. Multimodality helps ground such open-ended rationalization
3. Hierarchical Interpretable Fusion model will help joint Answer-Rationale generation

Model Evolution

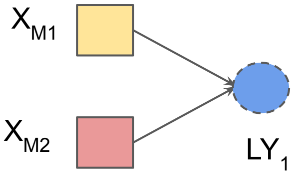
What's missing in the previous models?



Input Fusion

Utterance-level Adaptation

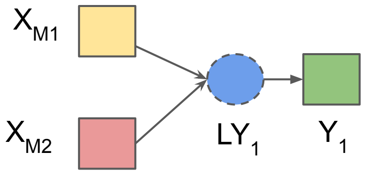
MONOTONIC TASK



Latent Representation Fusion

Utterance-level Adaptation

NON-MONOTONIC TASK



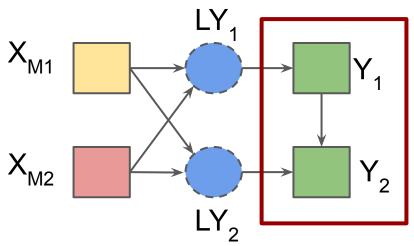
Hierarchical Latent Representation Fusion

Video-level Adaptation

ABSTRACTION TASK

EXPLANATION TASK

- Two observable outputs instead of one
- Dependent Information flow in output
- Generate Information not explicitly present in the inputs



Hierarchical Interpretable Fusion

Task Motivation

- Beyond QA to Explanations
- Inherently interpretable models by forcing the model to generate observable intermediate outputs "Y_1"
    - i.e. Rationale Generation (Y_2) -> Answers (Y_1)
- Proposed method of inherent interpretability can be expanded to many other multimodal generation tasks
    - e.g. Captioning (Y_2) -> Entities (Y_1)
    - e.g. Summary (Y_2) -> Noun Phrases (Y_1)
- Open-ended rationalization has a wide range of applications
    - decision support for ML systems
    - user-specific explainability

Summary

I. Multimodality

II. Task Complexity

III. Model Evolution

I. Multimodal Speech Recognition

II. Multimodal Speech Translation

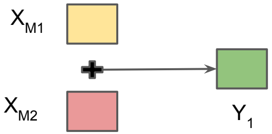
ICASSP ‘19, ICASSP ‘19

III. Multimodal Summarization & QA

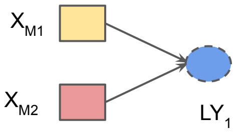
IV. Multimodal Rationalization

Proposed Work

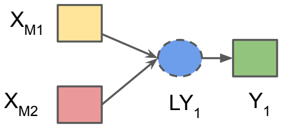
MONOTONIC TASK



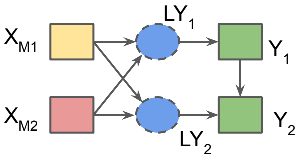
NON-MONOTONIC TASK



ABSTRACTION TASK



EXPLANATORY TASK



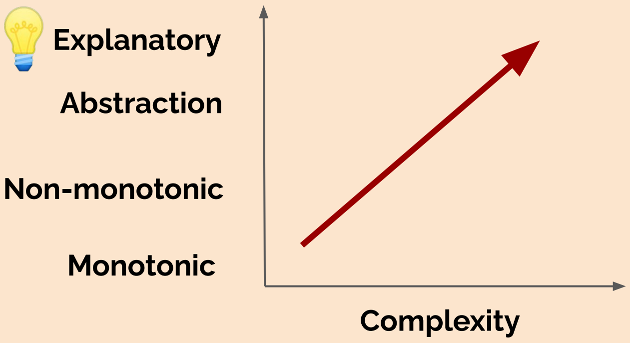
Conclusion

I. Multimodality

Multimodal modeling leads to improvements over unimodal & baseline models

It also facilitates cross-modal modeling requiring lesser supervision

II. Task Complexity



III. Model Evolution

We show how increasingly expressive models are important for satisfying task complexities

Timeline

Apr ‘21	Thesis Proposal
Now - May ‘21	Work on building the Hierarchical Interpretable Fusion model
May ‘21 - Aug ‘21	Summer internship at AI2 on Multimodal Rationalization
Sep ‘21 - Dec ‘21	Apply the Hierarchical Interpretable Fusion to Rationalization
Jan ‘22 - Feb ‘22	Thesis Writing
Mar ‘22 - Apr ‘22	Thesis Defense

Thank You

spalaska@cs.cmu.edu