

MAST: Multimodal Abstractive Summarization with Trimodal Hierarchical Attention

NLP Beyond Text, 2020

Aman Khullar*¹, Udit Arora*²

¹ Gram Vaani

² New York University

* = equal contribution

Why Multimodal?

- Humans process information from multiple modalities
- Interplay between modalities allows us to understand language in context
 - Visual cues like objects, actions
 - Audio cues like emphasis, change in tone
- Can machines benefit from the same capability?

What is Multimodal Summarization?



Audio



Video



Text

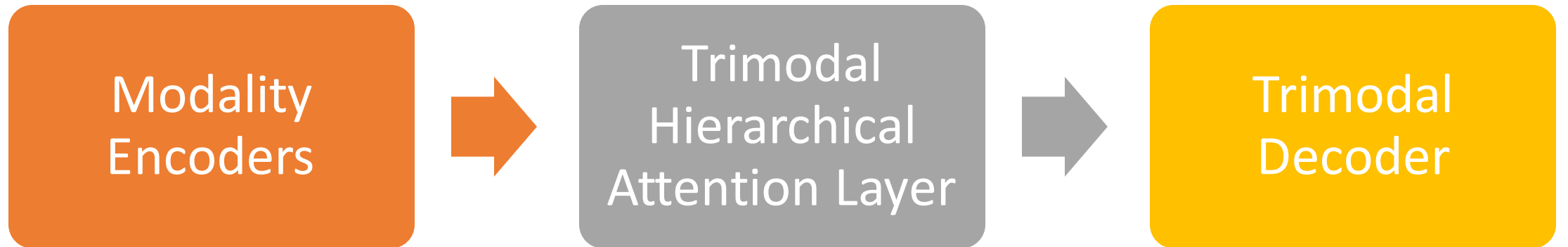
Motivation

- Baselines only considered *Video* and *Text* modalities
- Can we use *Audio* modality to improve text summarization?
- Intuitively, *Audio* contains useful additional information about the spoken words
 - level of emphasis
 - tone

How2 dataset (300h version)

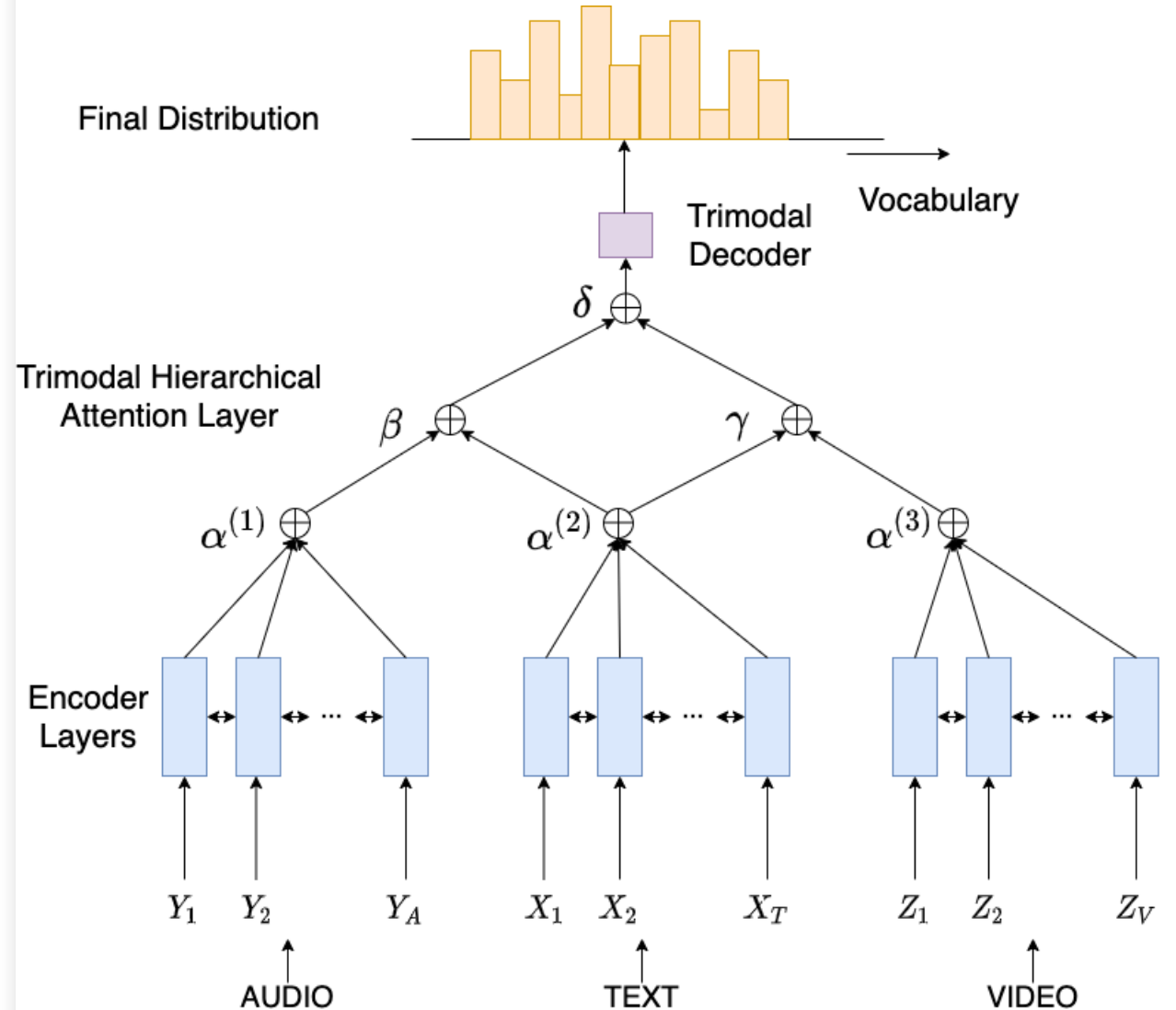
- Audio
 - Kaldi filter-bank features
- Text
 - Transcripts
- Video
 - Features from 3D CNN trained for action-recognition

Methodology

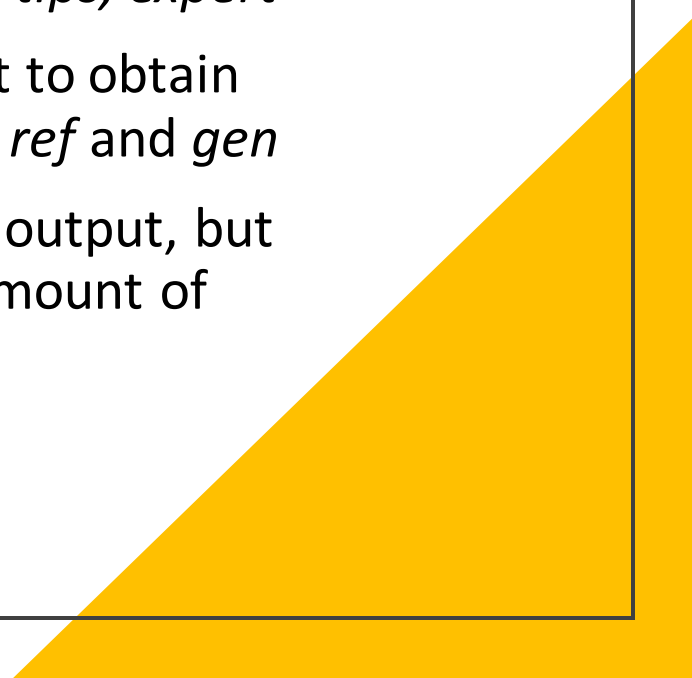


Our Models

- Trimodal H2: two-level attention hierarchy
 - Add audio modality to video-text hierarchical attention baseline
- MAST: three-level attention hierarchy
 - Additional level of attention hierarchy to choose between pairs of modalities: *Audio-Text* and *Video-Text*
 - Pays more attention to text



Metrics: Content F1

- **F1 score** of the **content** words in the summaries based on a monolingual alignment
 1. Remove function words and catchphrases like - *how, tips, expert*
 2. Use the METEOR toolkit to obtain the alignment between *ref* and *gen*
 - Ignores the fluency of the output, but gives an estimate of the amount of useful content words
- 

Results

- **MAST** outperforms all baselines in terms of *ROUGE* and *Content F1* scores
- Adding audio modality gives us a better *Content F1* score
- **MAST** obtains better performance than **TrimodalH2** in terms of *ROUGE*, while obtaining a close *Content F1* score

Model Name	ROUGE			Content F1
	1	2	L	
Text Only	46.01	25.16	39.98	33.45
BertSumAbs	29.68	11.74	22.58	31.53
Video Only	39.23	19.82	34.17	27.06
Audio Only	29.16	12.36	28.86	26.65
Audio-Text	34.56	15.22	31.63	28.36
Video-Text	48.40	27.97	42.23	32.89
TrimodalH2	47.85	28.46	42.17	35.65
MAST-Binned	46.22	25.94	40.34	33.56
MAST	48.85	29.51	43.23	35.40

Challenges in using the Audio Modality



Which audio features to use?

MFSC vs MFCC



Computational Efficiency

MAST-Binned



Audio-only and Audio-Text Models

Repetitive summaries

Usefulness of Audio modality

- Higher Content F1 score - Audio modality gives information of more useful content

Original text: let's talk now about how to bait a tip up hook with a maggot. typically, you're going to be using this for pan fish. not a real well known or common technique but on a given day it could be the difference between not catching fish and catching fish. all you do, you take your maggot, you can use meal worms, as well, which are much bigger, which are probably more well suited for this because this is a rather large hook. you would just, again, put that hook right through the maggot. with a big hook like this, i would probably put ten of these on it, just line the whole thing. this is going to be more of a technique for pan fish, such as, perch and sunfish, some of your smaller fish but if you had maggots, like this, or a meal worm, or two, on a hook like this, this would be a fantastic setup for trout, as well.

Text only: ice fishing is used for ice fishing. learn about ice fishing bait with tips from an experienced fisherman artist in this free fishing video.

Video-Text: learn about the ice fishing bait in this ice fishing lesson from an experienced fisherman.

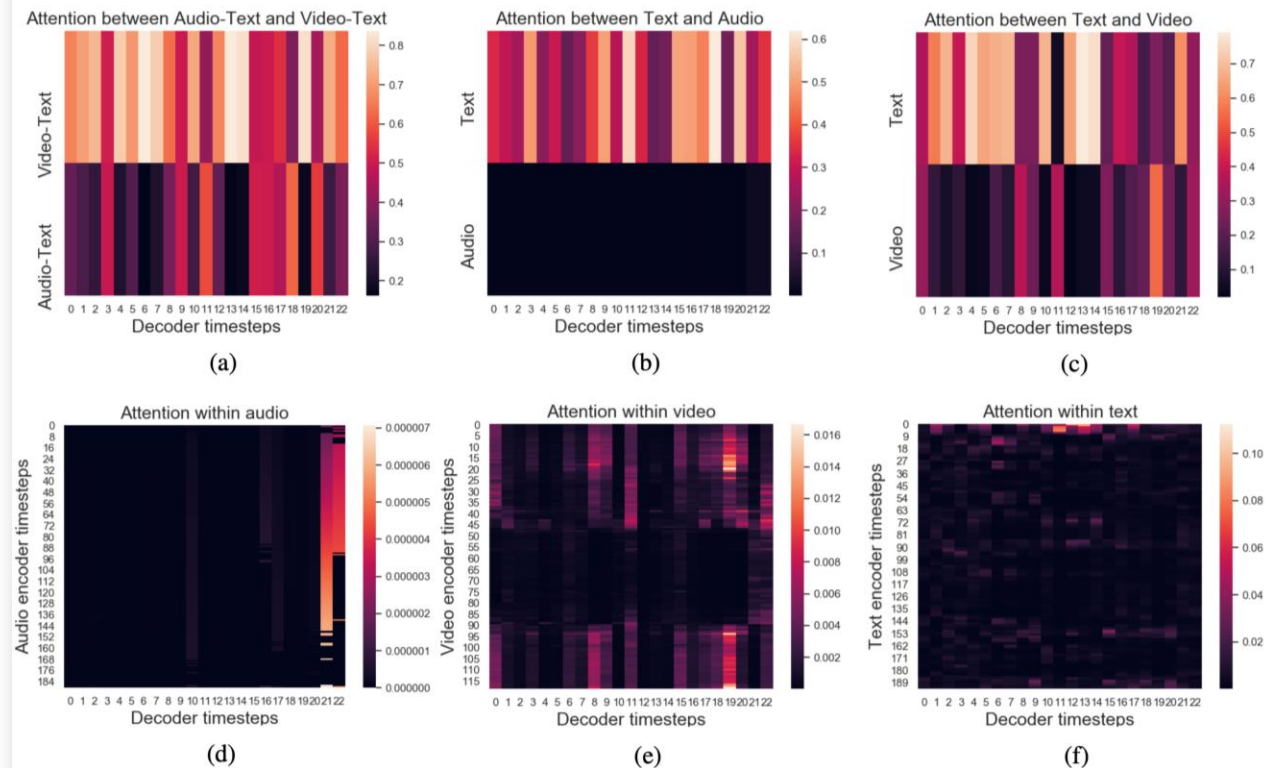
MAST: maggots are good for catching perch. learn more about ice fishing bait in this ice fishing lesson from an experienced fisherman.

Comparison of outputs by using different modality configurations:

- Frequently occurring words (highlighted in red) are easier for a simpler model to predict but don't contribute much in terms of useful content.
- The summary generated by our MAST model contains more content words

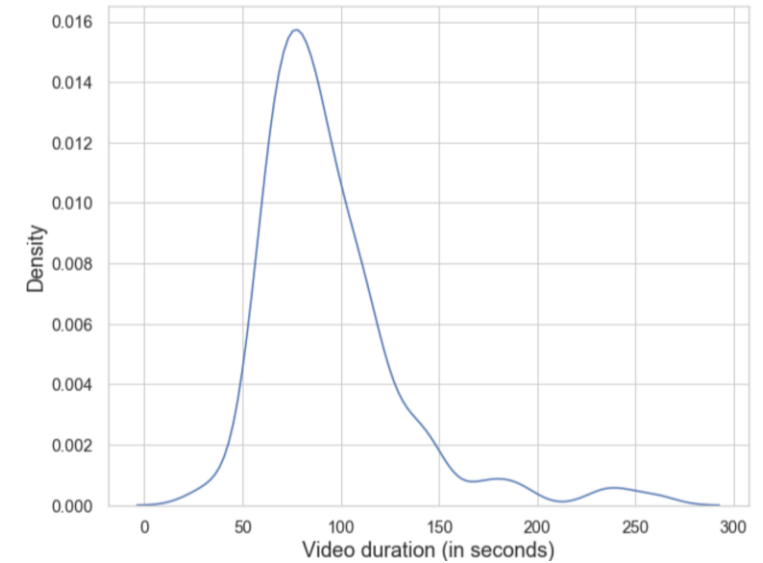
Attention distribution across modalities

- Visualization of attention weights in the Trimodal Hierarchical Attention Layer
 - (a), (b) and (c) show the attention distribution among different modality combinations
 - (d), (e) and (f) show the attention distribution within each modality encoder
- MAST pays higher attention to text to get high ROUGE score

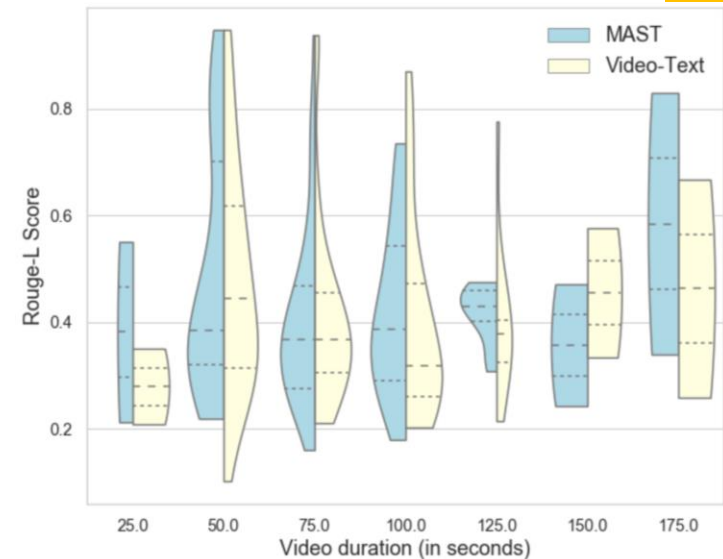


Performance across video durations

- How does ROUGE-L score vary with video duration?
- Binned videos in groups of 25 seconds by duration
- MAST outperforms the baseline in 5 out of 7 seven groups



Distribution of video durations in the test set



Distribution of Rouge-L scores of test set summaries

Conclusion

- Proposed **MAST**: a new model for multimodal text summarization that utilizes all three modalities
 - Uses a three-level trimodal hierarchical attention architecture to pay more attention to text
 - Generates more useful content words in summaries than baselines
- Explored the challenges and advantages of using audio modality
- Analyzed performance across videos of different durations