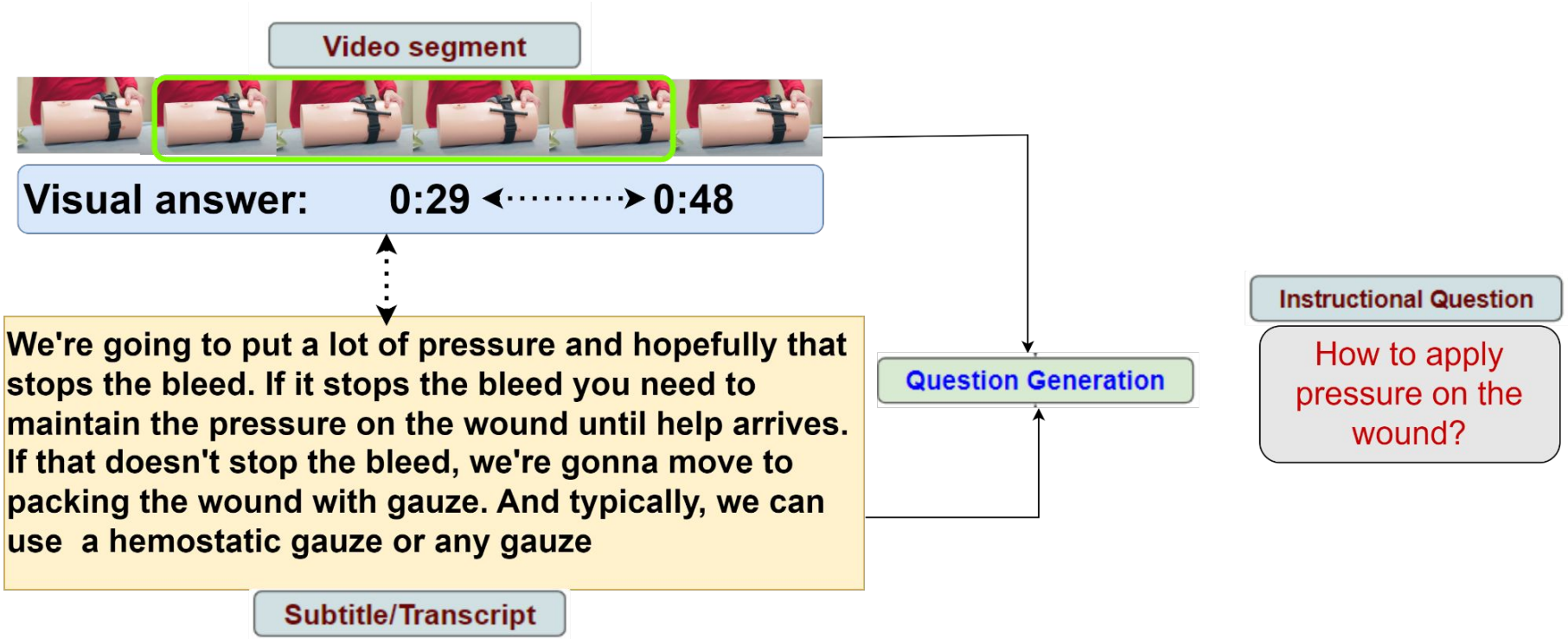
Attention-based Multimodal Deep Learning Models for Medical Instructional Question Generation

TRECVID 2023 MIQG Task

Shaswati Saha, Sanjay Purushotham
University of Maryland Baltimore County

Medical Instructional Question Generation (MIQG) Task



Related Works

Visual Question Answering (VQA)

- Z et. al: Prompting large language models with answer heuristics for knowledge-based visual question answering (CVPR 2023)
- Lin Z et. al: Medical visual question answering: A survey (AIME 2023)
- Ding et. al: MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering (CVPR 2022)
- Gao et. al: Transform-Retrieve-Generate: Natural Language-Centric Outside-Knowledge Visual Question Answering (CVPR 2022)
- Chappuis et. al: Prompt–RSVQA: Prompting visual context to a language model for Remote Sensing Visual Question Answering (CVPRW 2022)
- Garcia et. al : Knowit vqa: Answering knowledge-based questions about videos (AAAI 2020)
- Lei et. al: Tvqa+: Spatio-temporal grounding for video question answering (ACL 2020)
- Li et. al: Visual question generation as dual task of visual question answering (CVPR 2018)
- Lu et. al: Hierarchical question-image co-attention for visual question answering (NIPS 2016)
- S et. al: Vqa: Visual question answering (CVPR 2015)

Medical VQA

- Naseem et. al: Vision-Language Transformer for Interpretable Pathology Visual Question Answering (IEEE JBHI 2022)
- Gong et. al: A data-centric model with efficient training methodology for medical visual question answering (Imageclef 2021)
- Xiao et. al: Pretrained biobert for medical domain visual question answering (Imageclef 2021)
- Eslami et. al: BBN-Orchestra for long-tailed medical visual question answering (Imageclef 2021)
- Liao et. al: Knowledge Inference via a Skeleton-based Sentence Mapping Approach for Medical Domain Visual Question Answering (Imageclef 2020)

Related Works

Question Generation (QG)

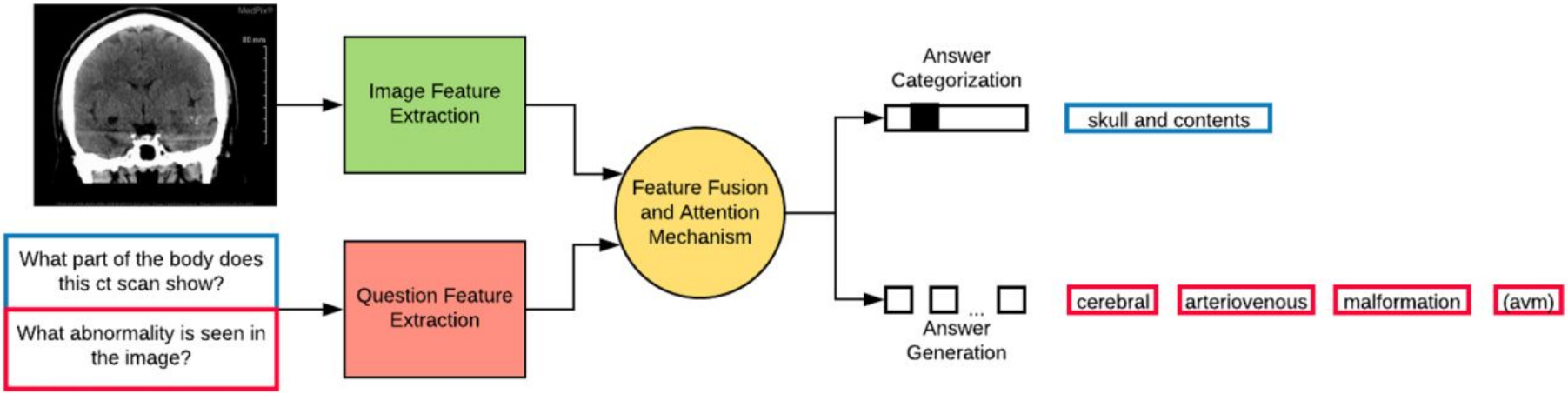
- N et. al: Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications (PRAI 2023)
- Vedd et. al: Guiding Visual Question Generation (NAACL 2022)
- Al-Sadi et. al: Pretrained VGG with Data Augmentation for Medical VQA and VQG (Imageclef 2021)
- Chebbi et. al: Visual Generation of Relevant Natural Language Questions from Radiology Images for Anomaly Detection (Imageclef 2021)
- Wang et. al: Video Question Generation via Semantic Rich Cross-modal Self-attention Networks Learning (ICASSP 2020)
- Sarrouti et. al: Visual Question Generation from Radiology Images (Imageclef 2020)

LLMs

- Guo et. al: From Images to Textual Prompts: Zero-shot Visual Question Answering with Frozen Large Language Models (CVPR 2023)
- Robinson et.al: Leveraging Large Language Models for Multiple Choice Question Answering (ICLR 2023)
- Yuan et. al: Selecting Better Samples from Pre-trained LLMs: A Case Study on Question Generation (ACL Findings 2023)
- Gautam et. al: A Lightweight Method to Generate Unanswerable Questions in English (EMNLP Findings 2023)
- Singhal et. al: Large language models encode clinical knowledge (Nature 2023)

Our related work on answer generation

MedFuseNet (Nature Scientific Reports' 21) : Answer Generation for Medical VQA - We introduced a attention-based multimodal model to generate answer using LSTM-based generative decoder



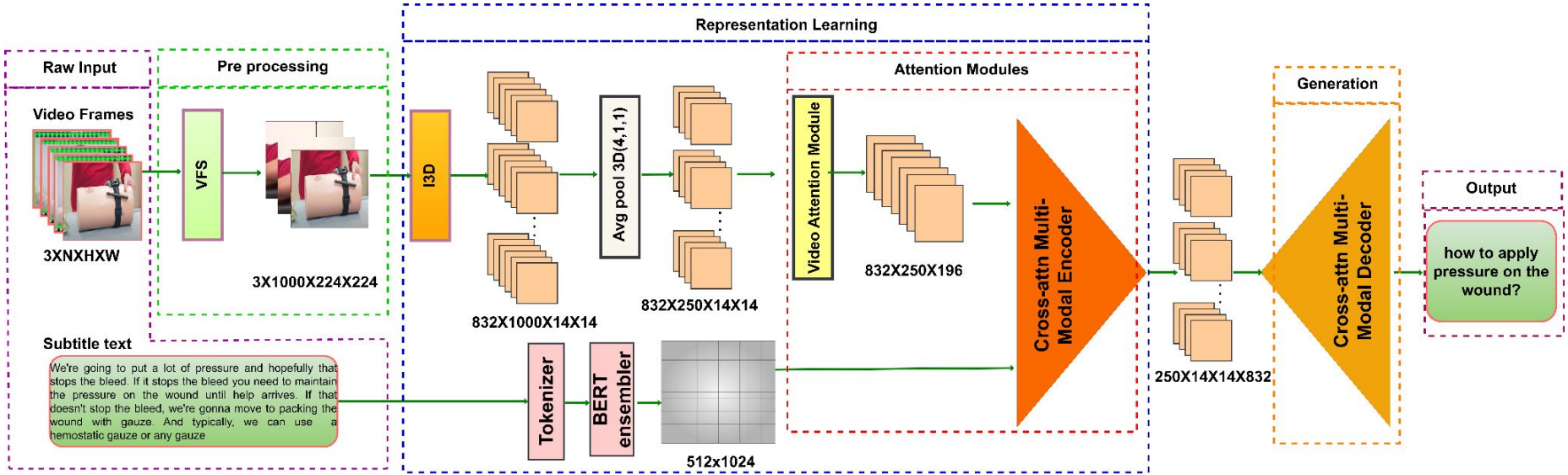
Paper: https://www.nature.com/articles/s41598-021-98390-1

Task Specific Challenges - Our perspective

- Existing VQA or MedVQA works are focused on representation learning for images not videos
  - Learning good video representation requires capturing the essential information and dynamics within the video
- Most existing works on question generation is on images, handful on video modality or multimodal data
- Question generated should be instructional and medically meaningful?
- Generated questions generated is coherent within the context of the given information while preventing the generation of redundant questions is challenging

Our Proposed (Initial) Solution: DEEP-CAM framework

- **DEEP-CAM:** Deep learning based cross attention multimodal framework
  - **Data pre-processing:** Pre-process video using VFS algorithm, subtitle using youtube API
  - **Representation learning:** Extract visual and textual feature using pre-trained models
  - **Cross-attention multimodal Encoder:** Combine visual and textual features using attention modules



DEEP-CAM: Data Preprocessing

Algorithm 1 Video Frame Sampling

Input: 2D video frames, $\boldsymbol{G} = \{\boldsymbol{g_n}\}_{n=1}^{N} \in \mathcal{R}^{3\times N\times H\times W}$
Parameter: number of desired frames, L
Output: 2D video frames, $\boldsymbol{F} = \{\boldsymbol{f_l}\}_{l=1}^{L} \in \mathcal{R}^{3\times L\times H\times W}$

1: **depletion factor** $d = \lfloor \frac{L}{N} \rfloor$
2: **remainder factor** $r = L \ (mod \ N)$
3: **if** $N = L$ **then**
4: **return** $\boldsymbol{F}$
5: **if** $N < L$ **then**
6: $f[1 : r - \lfloor \frac{r}{d} \rfloor] = g[1]$
7: $f[L - \lfloor \frac{r}{d} \rfloor : L] = g[N]$
8: $j = r - \lfloor \frac{r}{d} \rfloor + 1$
9: **for** $n \in \{0, 1, .., N\}$ **do**
10: $f[j : j + d] = g[n]$
11: $j := j + d$
12: **return** $\boldsymbol{F}$
13: **if** $N > L$ **then**
14: $j = r - \lfloor \frac{r}{d} \rfloor + 1$
15: **for** $l \in \{0, 1, .., L\}$ **do**
16: $f[l] = g[j]$
17: $j := j + d$
18: **return** $\boldsymbol{F}$

Oversampling

If not enough frames found, copy at calculated intervals

If more frames found than required, sample at calculated intervals

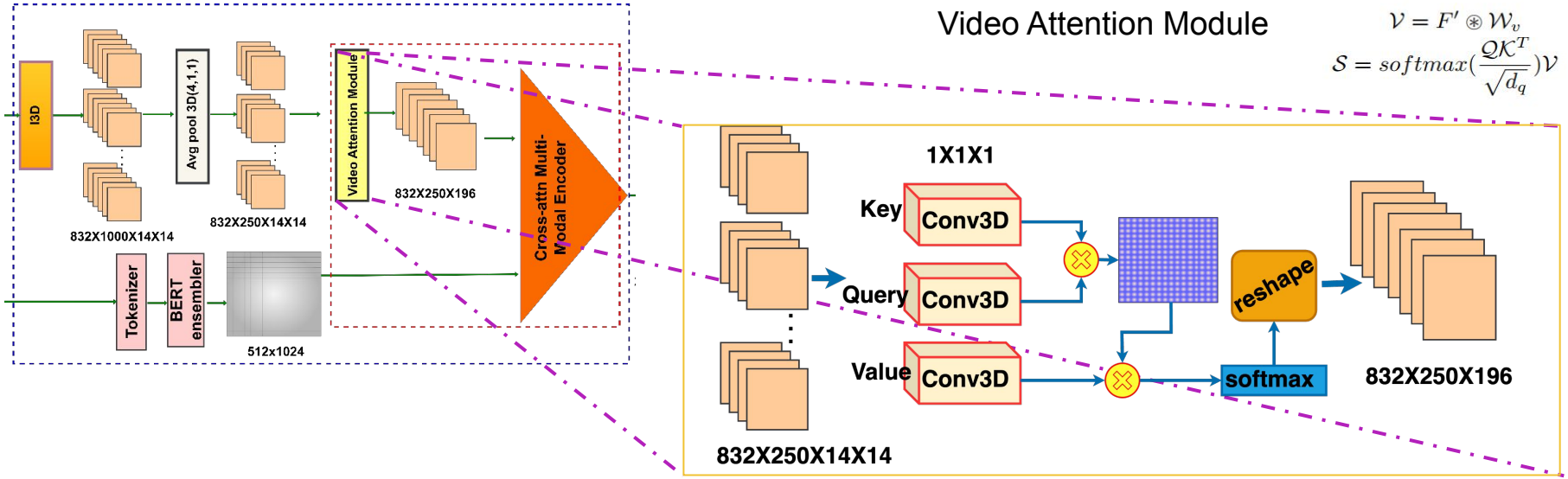
undersampling

- Extract video frames using OpenCV and sample desired #frames using Video Frame Sampling
- Extract (full/timed) subtitles using youtube-transcript-api
- Glove-embedding for questions only for training/validation data

DEEP-CAM: Representation learning

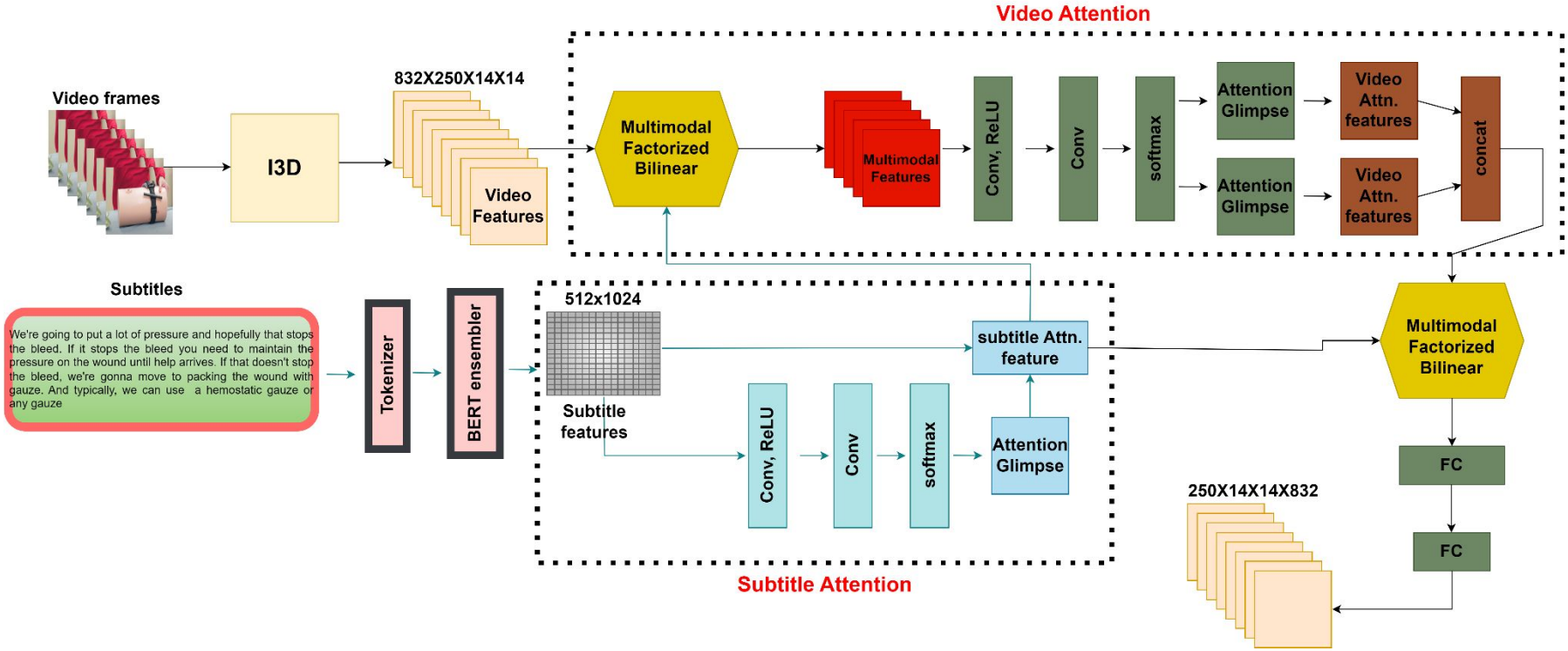
Visual feature extraction: Use pre-trained I3D-based spatio-temporal representation model [J Carreira, 2017]
Subtitle/text feature extraction: Pre-trained BERT model [Devlin, 2018]
Attention modules: Video attention module and Cross-attention multimodal encoder module

$$\mathcal{K} = F' \circledast \mathcal{W}_k$$
$$\mathcal{Q} = F' \circledast \mathcal{W}_q$$
$$\mathcal{V} = F' \circledast \mathcal{W}_v$$
$$\mathcal{S} = softmax(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d_q}})\mathcal{V}$$



DEEP-CAM: Cross-attention multi-modal encoder

- Inspired by our MedFuseNet [Sharma et. al 2021], fuse quantized features from two different modalities using Multi-modal Factorized Bilinear (MFB) Pooling [Zhou Yu 2018]



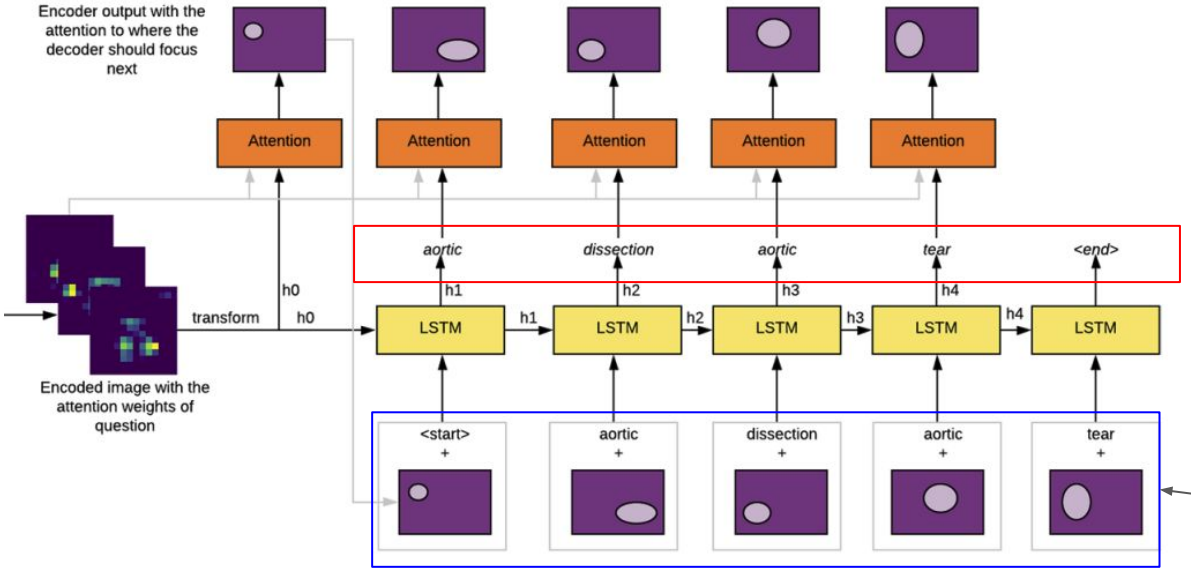
DEEP-CAM: Cross-attention multimodal decoder

- Question generation decoder module:
  a. Teacher forcing (training only)
  b. Attention mechanism
  c. Beam search during inference

Algorithm 2: Decoder Algorithm for Answer Generation

Input: Attended Image Features $\hat{v}_e$, Answer $a_1, \ldots, a_n$
Output: Generated Answer $\tilde{a} = [\tilde{a}_1, \ldots, \tilde{a}_n]$

Initialize the decoder LSTM states using image features ($\hat{v}_e$)
Initialize generated answer $\tilde{a}$ as an empty list
Initialize d_0 as image features ($\hat{v}_e$)
for *each step i in* $[a_1, \ldots, a_n]$ **do**
    Concatenate a_i and d_{i-1}, the output of Decoder Attention $\mathscr{E}_d$ for $(i-1)^{th}$ step
    Feed this concatenated vector to the i^{th} decoder step
    Add h_i, which is also $\tilde{a}_i$, to list $\tilde{a}$
    Feed $\hat{v}_e$ and h_i to decoder attention $\mathscr{E}_d$ to get d_i
end
return Generated Answer $\tilde{a}$



Experiments

- Videos: Training, validation, test: 800, 49, 50 respectively
- Subtitles extracted: Full video and video answer segment
  - Examples:

Timed

We're going to put a lot of pressure and hopefully that stops the bleed. If it stops the bleed you need to maintain the pressure on the wound until help arrives. If that doesn't stop the bleed, we're gonna move to packing the wound with gauze. And typically, we can use a hemostatic gauze or any gauze

Entire Video

I'm Lisa Hollister, the director of trauma and acute care surgery for Parkview Health. Today I'm going to show you how to stop the bleed. Stop the bleed is very simple. Three steps. The first one is pressure, then wound packing, then a tourniquet. So let's start with pressure. If you come upon a bleeding patient that has a wound the first step you're gonna do is put pressure on it. If it's a large wound, you're gonna put your entire palm of your hand and all of your weight on the wound. We're going to put a lot of pressure and hopefully that stops the bleed. If it stops the bleed you need to maintain the pressure on the wound until help arrives. If that doesn't stop the bleed, we're gonna move to packing the wound with gauze. And typically, we can use a hemostatic gauze or any gauze or you could use a shirt if you have nothing available. So we're gonna take the gauze and we're going to pack it inside the wound until you can't pack it anymore. And that's getting to the source of the bleeding. So we're going to just keep packing more and more and more and it could be a deep wound, so don't be afraid. So once we've gotten this completely packed, hopefully we can put some pressure on it and that will stop the bleed. So if you're all by yourself, and you need to apply a tourniquet, Just put the tourniquet on. We're going to tighten it as much as possible. It's a velcro, so super super tight, Then we're gonna take the handle we're gonna twist it until The bleeding source is stopped. And then we're gonna place it into the handle We're gonna take the velcro and close it. And we're going to write the time. For courses in our area go to parkview.com and search \"Stop the Bleed.\

- Pre-trained bert with max token length = 512 and embedding size = 1024

Submission Results

- Run 1: We employ subtitles that correspond to the start and finish time stamps of the video segments
- Run 2: We utilized the subtitles of the entire video
- Unsubmitted run: Self-evaluation on ChatGPT generated questions using subtitles only

RunID	BLEU	BLEU-4	ROUGE-2	ROUGE-L	BERTScore
Run 1	0	0	0.13167	0.3154	0.87683
Run 2	0	0	0.12262	0.26083	0.85332
Self-evaluation ChatGPT	0.04312	0.01795	0.15247	0.35128	0.88015

Experimental Results for Run 1

Example 1:

a. **Ground-truth:** how can i massage the area where the neck meets the skull to relieve neck pain and tension ?
b. **Prediction:** how to i use my neck to using neck using exercise muscle to reduce the pain ? headache ?
c. **Refinement with LLM**: How can I use neck exercises to alleviate muscle pain and headaches?

Example 2:

a. **Ground-truth:** how to improve blood flow to treat varicose veins by performing calf raises or heel raises ?
b. **Prediction:** how can perform knee flow in legs bowed legs and performing the ? ? muscle ? ?
c. **Refinement with LLM**: How can one execute a knee flow with bowed legs while engaging the leg muscles?

Our thoughts: Challenges and Limitations

1. We are not medical domain experts, so self-evaluation of generated questions was difficult
2. Subtitles from youtube's closed caption- may not be accurate and/or calibrated with visual answer
3. Our submission did not use latest state-of-the-art models (late entry: 3 weeks before submission) or data augmentation or LLMs
4. Only one instructional question for each visual answer evaluation!
    a. Example: How to wrap a larger bandage around the arm to stop bleeding for an object wound?
5. Start early!

Our thoughts: Challenges and Limitations

Dataset issues:

1. Visual segments varies from 3 seconds to ~300 seconds!
2. Raw videos not available to download
3. Subtitles not available for all videos

Computational issues: Expensive!
Submission issues: Submission error on codalab

This phase of the competition is closed. Here are the submissions you made:

#	SCORE	FILENAME	SUBMISSION DATE	SIZE (BYTES)	STATUS	✔	
1	---	predictions.zip	08/11/2023 04:37:16	4960	Failed		+
2	---	predictions.zip	08/11/2023 19:41:24	5082	Failed		+

Summary

- We proposed DEEP-CAM, a deep spatio-temporal, cross-modality, and cross-attention encoder-decoder model that takes a medical video segment and subtitle text to generate an instructional question
- We submitted two runs to the challenge with differing textual inputs: timed subtitle input performed better
- Ongoing work:
  - More advanced state-of-the-art models and multimodal models - Clipbert, BLIP for current framework
  - New question generation framework
  - Use LLMs for data augmentation, baseline, and post-processing
  - Domain expert validation: We have additional questions generated by a pre-med student on validation dataset

Acknowledgements

- TRECVID Organizers
- Part of the work supported by NSF CAREER grant #2238743

Thank you!

Q & A