project-multimodal

salmonn

학생 원종찬, 최재원

Prof. Jae-Hong Lee, Jun-Hyung Park

Contents

리뷰

- 250922
  - SAKURA 및 JASCO 논문 발표
- 250929
  - SALMONN에 대해 후처리를 이유로 재현상 어려움을 제시
- 251013
  - SAKURA 및 SALMONN의 재현 결과 공유
- 251027
  - SALMONN 재현 종료
- 251103
  - multi-hop reasoning 성능 향상 방향 제시
- 251110
  - 장기 오디오에 대한 벤치마크 구상 (SAKURA 기반)
  - 수치 제시 및 모델 출력의 포맷을 강제하는 실험

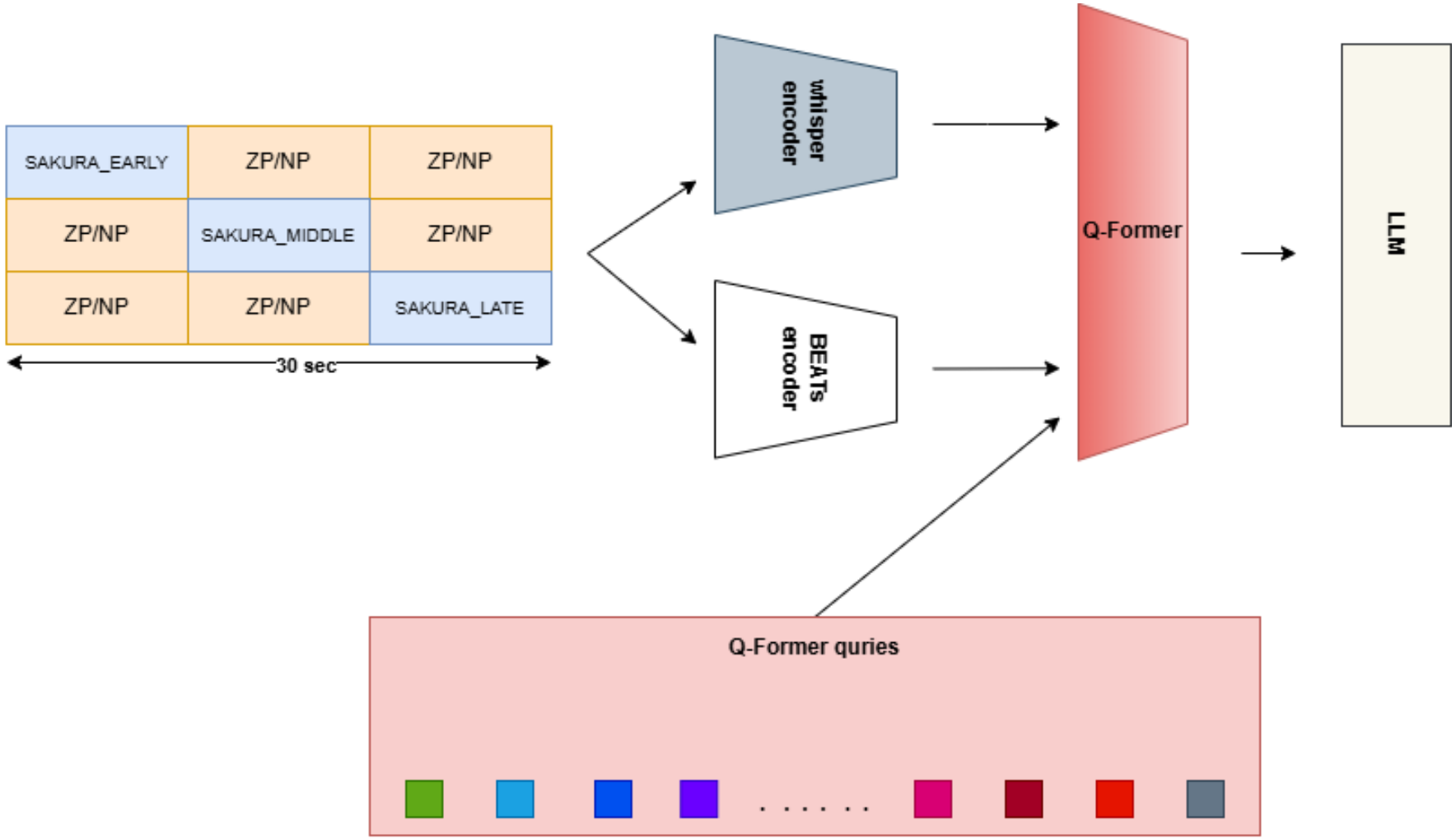
사고의 과정 : 원인 검증과 원인 개선 방안

- **목표** : **LALM의 multihop 추론 능력 향상** (SAKURA 결론 : LALMs, 음성, 오디오 추론 multihop 미흡.)

- multihop 미흡의 **원인**을 찾아야 한다

가설 : LALM의 음성과 오디오에 대한 multihop 추론 성능의 미흡은 오디오 길이에 따른 정보 손실 때문이다. **즉, LALM은 입력 오디오의 전 구간에 대해 동일한 집중도를 가지고 있지 <u>않다</u>.**

원인가설 검정 실험 결과


SAKURA_EARLY
ZP/NP
ZP/NP
ZP/NP
SAKURA_MIDDLE
ZP/NP
ZP/NP
ZP/NP
SAKURA_LATE
30 sec
whisper encoder
BEATs encoder
Q-Former
LLM
Q-Former quries

원인가설 검정 실험 결과

Category	Task Type	Info Location	ZP (%)	NZ (%)	SRC (%)
animal	single	EARLY	67.60	62.20	68.60
		MIDDLE	60.40	53.40	N/A
		LATE	56.20	50.40	N/A
	multi	EARLY	37.60	33.80	38.00
		MIDDLE	33.00	29.80	N/A
		LATE	32.60	28.40	N/A
gender	single	EARLY	57.00	55.40	56.60
		MIDDLE	59.40	58.20	N/A
		LATE	52.60	50.20	N/A
	multi	EARLY	47.80	45.60	48.00
		MIDDLE	49.60	47.40	N/A
		LATE	52.20	50.40	N/A
emotion	single	EARLY	22.00	20.80	21.00
		MIDDLE	20.80	19.40	N/A
		LATE	18.80	17.20	N/A
	multi	EARLY	29.00	29.60	27.80
		MIDDLE	27.40	29.20	N/A
		LATE	28.00	29.40	N/A
language	single	EARLY	22.00	22.60	21.00
		MIDDLE	24.80	26.60	N/A
		LATE	26.60	27.20	N/A
	multi	EARLY	28.80	29.80	29.80
		MIDDLE	29.40	31.20	N/A
		LATE	29.20	31.80	N/A

원인가설 검정 실험 결과

- animal category 관찰
- 관찰 1 : 오디오 길이를 늘린 결과 acc 하락이 관찰됨
- 관찰 2 : 두 task 모두 early가 가장 높은 성능, middle에서 late로 갈수록 성능 하락
- 관찰 3 : 제로 패딩과 노이즈 패딩에서의 이 경향이 같음

Category	Task Type	Info Location	ZP (%)	NZ (%)	SRC (%)
animal	single	EARLY	67.60	62.20	68.60
		MIDDLE	60.40	53.40	N/A
		LATE	56.20	50.40	N/A
	multi	EARLY	37.60	33.80	38.00
		MIDDLE	33.00	29.80	N/A
		LATE	32.60	28.40	N/A

원인가설 검정 실험 결과

- gender category 관찰
- 관찰 1 : 오디오 길이를 늘린 결과 acc 하락과 동시에 상승이 관찰됨
- 관찰 2 : single에서는 middle에 정보가 있을시 가장 높은 성능, multi에서는 late에서.
- 관찰 3 : 제로 패딩과 노이즈 패딩에서의 이 경향이 같음

gender	single	EARLY	57.00	55.40	56.60
		MIDDLE	59.40	58.20	N/A
		LATE	52.60	50.20	N/A
	multi	EARLY	47.80	45.60	48.00
		MIDDLE	49.60	47.40	N/A
		LATE	52.20	50.40	N/A

원인가설 검정 실험 결과

- emotion category 관찰
- 관찰 1 : 오디오 길이를 늘린 결과 acc 하락과 상승이 동시에 보여짐
- 관찰 2 : single에서는 early 에 정보가 있을시 가장 높은 성능, late로 갈수록 하락. multi도 마찬가지.
- 관찰 3 : multi의 노이즈 페딩에 대해 정보 위치별 차이가 1%p 아래임 (유일)

emotion	single	EARLY	22.00	20.80	21.00
		MIDDLE	20.80	19.40	N/A
		LATE	18.80	17.20	N/A
	multi	EARLY	29.00	29.60	27.80
		MIDDLE	27.40	29.20	N/A
		LATE	28.00	29.40	N/A

원인가설 검정 실험 결과

- Language category 관찰
- 관찰 1 : 오디오 길이를 늘린 결과 acc 하락과 상승이 동시에 보여짐
- 관찰 2 : single에서는 late로 갈수록 acc 좋음. multi 비슷한 경향이나 최대 acc 위치는 달랐음.

language	single	EARLY	22.00	22.60	21.00
		MIDDLE	24.80	26.60	N/A
		LATE	26.60	27.20	N/A
	multi	EARLY	28.80	29.80	29.80
		MIDDLE	29.40	31.20	N/A
		LATE	29.20	31.80	N/A

결론

- SAKURA data 종류와 task별 성능이 가장 높은 정보 위치가 다르다. (해당 데이터에 한해)

  >> 오디오의 전 구간에 대해 집중도가 다르다는 정황이 제시된다.

- 정보가 없는 묵음이나 잡음으로 패딩된 오디오를 앞이나 뒤에 삽입했다는 것만으로도 SAKURA에 대한 성능 상승이 관찰되었다.

- 1. 패딩의 위치에 따른 성능 변화와 2. 패딩을 통한 오디오 연장 여부에 따른 성능 변화의 원인 소명이 필요

- 그 원인이 소명된다 하더라도 SAKURA의 multihop 추론의 미흡 원인과 연결될 가능성이 얼마나 될지 잘 예상이 되질 않음

결론

- SAKURA data 종류와 task별 성능이 가장 높은 정보 위치가 다르다. (해당 데이터에 한해)

  >> 오디오의 전 구간에 대해 집중도가 다르다는 정황이 제시된다.

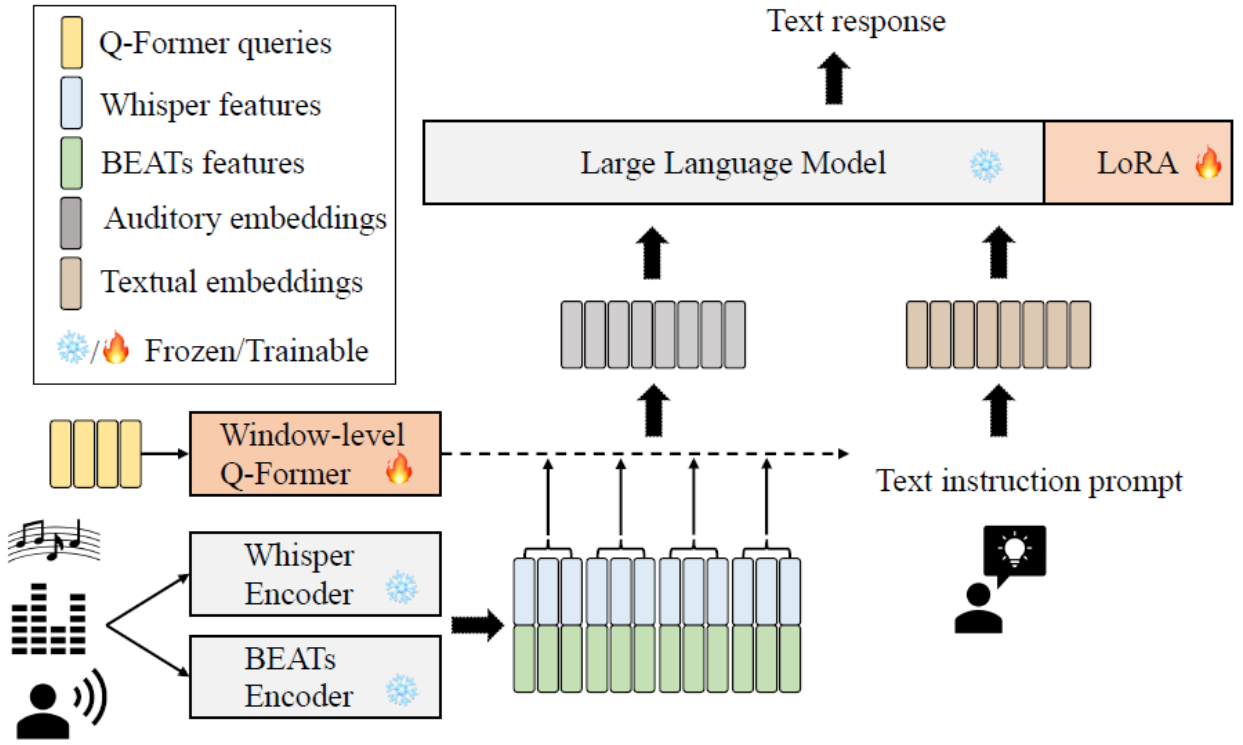
- 정보가 없는 묵음이나 잡음으로 패딩된 오디오를 앞이나 뒤에 삽입했다는 것만으로도 SAKURA에 대한 성능 상승이 관찰되었다.

- **1. 패딩의 위치에 따른 성능 변화와 2. 패딩을 통한 오디오 연장 여부에 따른 성능 변화의 원인** 소명이 필요

- 그 원인이 소명된다 하더라도 SAKURA의 multihop 추론의 미흡 원인과 연결될 가능성이 얼마나 될지 잘 예상이 되질 않음

결론 : 원인 분석 (가설)

- SALMONN의 핵심 : 1. LLM이 오디오를 이해한다 **(Q-Former)**
  2. activation tuning (LoRA)



결론 : 원인 분석 (추가 가설)

- 패딩의 위치에 따른 성능 변화는 Q-Former와 관련?

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Junnan Li Dongxu Li Silvio Savarese Steven Hoi
Salesforce Research
https://github.com/salesforce/LAVIS/tree/main/projects/blip2

3.1. Model Architecture

We propose Q-Former as the trainable module to bridge the gap between a frozen image encoder and a frozen LLM. It extracts a fixed number of output features from the image encoder, independent of input image resolution. As shown in Figure 2, Q-Former consists of two transformer submodules that share the same self-attention layers: (1) an image transformer that interacts with the frozen image encoder

결론 : 원인 분석 (추가 가설)

- 패딩의 위치에 따른 성능 변화는 Q-Former와 관련?

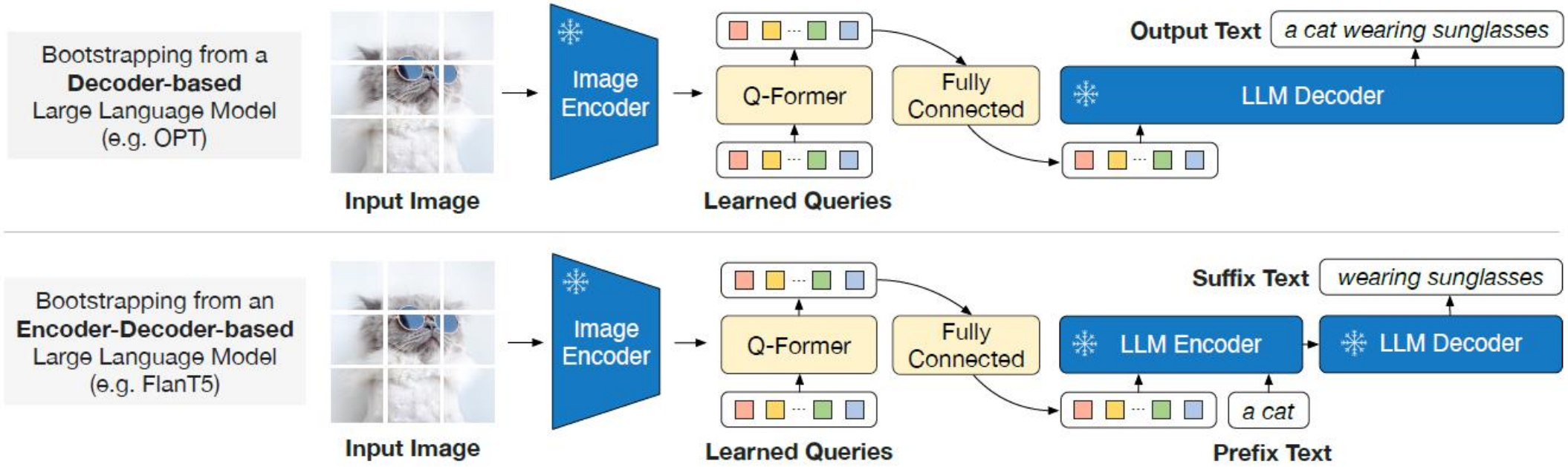


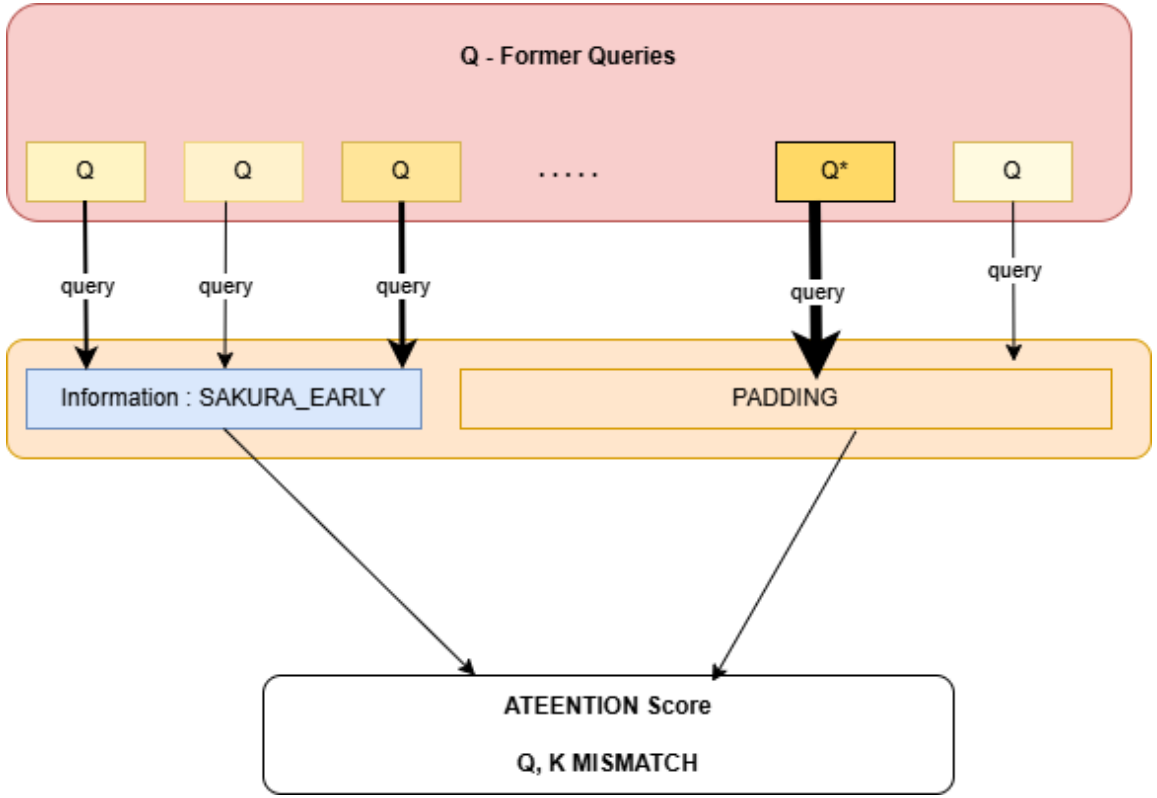
Figure 3. BLIP-2's second-stage vision-to-language generative pre-training, which bootstraps from frozen large language models (LLMs). (**Top**) Bootstrapping a decoder-based LLM (e.g. OPT). (**Bottom**) Bootstrapping an encoder-decoder-based LLM (e.g. FlanT5). The fully-connected layer adapts from the output dimension of the Q-Former to the input dimension of the chosen LLM.

결론 : 원인 분석 (추가 가설)

- 패딩의 위치에 따른 성능 변화는 Q-Former와 관련?

  >> Q-Former의 학습에 대해서 더 자세히 살펴봐야 (quirey들의 가중치?)

  >> Q-Former의 가중치들이 activation tuning의 영향으로 가중치가 특정 위치에 유리?



실험 방향성

- 이상의 내용을 SAKURA의 multihop 미흡 원인으로 연결시키기에는 어려움
- 다른 방향으로 multihop 성능 제고가 필요하다고 생각

배치 비일관성에 대한 최종 결론

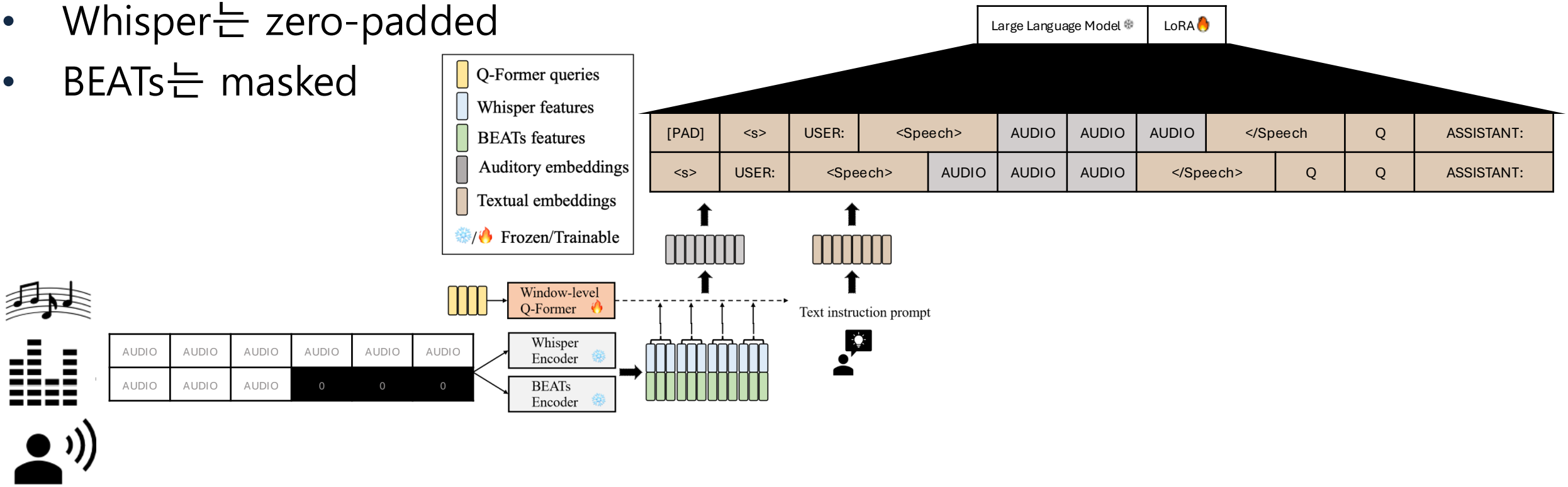
- 1. Llama 모델에서의 배치에 대한 이해 부족
  - 학습시 padding_side=right
  - 추론시 padding_side=left
    - SALMONN의 코드에서는 padding_side=right로 되어 있음
    - 배치 추론에 대한 코드가 없던 까닭

<s>	$input_{00}$	$input_{01}$	$input_{02}$	[PAD]	[PAD]		[PAD]	[PAD]	<s>	$input_{00}$	$input_{01}$	$input_{02}$
<s>	$input_{10}$	$input_{11}$	$input_{12}$	$input_{13}$	$input_{14}$	->	<s>	$input_{10}$	$input_{11}$	$input_{12}$	$input_{13}$	$input_{14}$
SALMONN 코드; right-padding							수정 코드; left-padding					

[입력] [패딩] [BOS]

배치 비일관성에 대한 최종 결론

- SALMONN에서 fused auditory token은 항상 88개
- Whisper는 zero-padded
- BEATs는 masked



배치 비일관성에 대한 최종 결론

- 2-1. precision에 따른 비일관성
  - 13B 모델의 8bit quantization이 이를 더 강화함

    - 7B, fp16, autocast float16
      - consistency: 98.35%
        - 'total': 4000
        - 'is_equal_inference': 3934
        - 'is_not_equal_inference': 66

    - 7B, fp16, autocast float32
      - consistency: 99.65% (+1.3%p)
        - 'total': 4000
        - 'is_equal_inference': 3986
        - 'is_not_equal_inference': 14

type	exponent (지수부)	mantissa (가수부)	sign
float32	8	23	1
float16	5	10	1
bfloat16	8	7	1

배치 비일관성에 대한 최종 결론

- 2-2. 통제 불가능한 것들에 따른 비일관성 (2-1 분석에 따른)
  - 동일한 임베딩 시퀀스를 1개 넣었을 때(a)와 2개 넣었을 때(b)
  - a와 b는 거의 유사하나 다르기도 함

["foo bar baz"] -> model forward & decode -> [a]
size: (**1**, 3, 4096)
["foo bar baz", "foo bar baz"] -> model forward & decode -> [**b, b**]
size: (**2**, 3, 4096)

배치 비일관성에 대한 최종 결론

- 2-2. 통제 불가능한 것들에 따른 비일관성 (예시)
  - batch_size=1
    - It is difficult to accurately classify the speaker's underlying emotions based solely on the provided audio clip. However, it is possible that the speaker may be expressing a mix of emotions such as frustration, anger, and sadness. It is important to consider the context of the conversation and the speaker's tone of voice in order to **fully understand** their emotions.
  - batch_size=2
    - It is difficult to accurately classify the speaker's underlying emotions based solely on the provided audio clip. However, it is possible that the speaker may be expressing a mix of emotions such as frustration, anger, and sadness. It is important to consider the context of the conversation and the speaker's tone of voice in order to **accurately interpret** their emotions.

  - batch_size=1
    - c) Italian
  - batch_size=2
    - **(**c) Italian

배치 비일관성에 대한 최종 결론

- 완전히 동일하지는 않지만, 일반적으로 용납 가능한 정도의 차이를 보인다.
  - BASE는 LLM-as-Judge로 ChatGPT를 사용함
  - OURS는 LLM-as-Judge로 GPT-OSS-20B를 사용함

exp \ category (%)	animal-multi		animal-single		emotion-multi		emotion-single		gender-multi		gender-single		language-multi		language-single	
PAPER (BASE)	34.6		68.6		28.2		19.8		48.6		59.8		29.6		21.8	
7B, B=1, FP16	34.6		68.2		28.4		20		48.8		60		29.8		20.6	
7B, B=1, FP32	35	+0.4%p	68.2	0	28.4	0	20	0	48.4	-0.4%p	60	0	29.4	-0.4%p	20.8	+0.2%p
7B, B=2, FP16	34.6		68.6		28.2		20		48.6		60		29.8		20.8	
7B, B=2, FP32	35.2	+0.6%p	68.2	-0.4%p	28.4	+0.2%p	20	0	48.6	0	60	0	29.4	-0.4%p	20.8	0

배치 비일관성에 대한 최종 결론

- 배치 추론으로 실험을 수행
  - 환경: 7B, autocast float16

- 최종 1회만 배치 크기 1로 평가

시도한 방법 1

- 입력 토큰 아이디들에 대한 로짓 레벨에서의 페널티
  - EncoderRepetitionPenaltyLogitsProcessor
    - SAKURA 데이터셋은 객관식이므로

시도한 방법 2

- 응답을 길게 하도록 EOS토큰을 로짓 레벨에서의 페널티
  - ExponentialDecayLengthPenalty

시도한 방법 3

- 특정 시퀀스에 대한 로짓 레벨에서의 페널티
  - SequenceBiasLogitsProcessor
    - [I cannot, -20.00]
    - [I'm sorry, -20.00]
    - [It is difficult, -20.00]

시도한 방법들의 문제

- 로짓 레벨에서의 페널티는 모델의 편향과 환각을 강화함
  - 약한 페널티는 출력 결과를 크게 변경하지 않음
  - 강한 페널티는 출력 결과를 크게 변경하여 모델을 붕괴시킴

When do SALMONN fail?

- exp1
  - BEATs, Whisper, fusioned representation as zeros compared to original one.

- exp2
  - expanded audio with two types of padding, two types of duration and three types of positions
    - 30초, 15초
    - 노이즈 패딩, 제로 패딩
    - 전, 중, 후

[exp1] When do SALMONN fail? – SKR-BEATS_ZP-B5

- Whisper 인코더만(BEATs 인코더 출력은 제로 패딩으로 설정)

exp \ category (%)	animal-multi	animal-single	emotion-multi	emotion-single	gender-multi	gender-single	language-multi	language-single
PAPER (BASE)	34.6	68.6	28.2	19.8	48.6	59.8	29.6	21.8
7B, B=5, BASE	35	68.4	28.2	20.2	48.4	60.2	29.6	21.2
7B, B=5, B_ZP	28.6 6.4	40.2 28.2	28.8 0.6	20.4 0.2	48.4 0	61.4 1.2	28.2 1.4	23 1.8

[exp1] When do SALMONN fail? – SKR-WHISPER_ZP-B5

- BEATs 인코더만(Whisper 인코더 출력은 제로 패딩으로 설정)

exp \ category (%)	animal-multi	animal-single	emotion-multi	emotion-single	gender-multi	gender-single	language-multi	language-single
PAPER (BASE)	34.6	68.6	28.2	19.8	48.6	59.8	29.6	21.8
7B, B=5, BASE	35	68.4	28.2	20.2	48.4	60.2	29.6	21.2
7B, B=5, W_ZP	20.6 14.4	21.8 46.6	21.6 6.6	5.4 14.8	52.4 4	45.4 14.8	18.6 11	22 0.8

[exp1] When do SALMONN fail? – SKR-AUDIO_ZP-B5

- 텍스트 프롬프트만 (QFormer 출력을 제로 패딩으로 설정)

exp \ category (%)	animal-multi	animal-single	emotion-multi	emotion-single	gender-multi	gender-single	language-multi	language-single
PAPER (BASE)	34.6	68.6	28.2	19.8	48.6	59.8	29.6	21.8
7B, B=5, BASE	35	68.4	28.2	20.2	48.4	60.2	29.6	21.2
7B, B=5, A_ZP	10 25	12.2 56.2	19.4 8.8	12 8.2	27.4 21	42.8 7.4	12.4 17.2	17.4 3.8

- 음성 없이 다지선다의 질문만을 줬지만 문제를 맞힘
  - LLM이 가지고 있는 정답에 대한 통계의 문제 (keyword: prior)
    - LLM은 이제까지 잘 맞힌 게 아니라, 음성 정보 + 통계로 잘 찍은 것일 수도 있음
    - 즉, 이 정확도가 낮아야 정확한 평가가 가능할 것임
    - 또한, 벤치마크는 텍스트+음성의 정확도와 텍스트 정확도의 차이로 나타나야 할 것임

SKR-SHRTANS-B5

- 보기를 제시하지 않음
  - Determine the language of the provided speech. ~~(a) German (b) Spanish (c) Italian (d) Korean~~

exp \ category (%)	animal-multi	animal-single	emotion-multi	emotion-single	gender-multi	gender-single	language-multi	language-single
PAPER (BASE)	34.6	68.6	28.2	19.8	48.6	59.8	29.6	21.8
7B, B=5, BASE	35	68.4	28.2	20.2	48.4	60.2	29.6	21.2
7B, B=5, SHRT	20.2 14.8	69 0.6	1.2 27	12.8 7.4	0.8 47.6	68.4 8.6	7.6 22	3.4 17.8

- 한계
  - Considering the voice heard in the audio, which celebrity **from the options below** is most likely associated with the same gender as the speaker?
    - 보기에서 고르라는 문제는 무조건 틀림

SKR-TOTAL-B5

- 다지선다 문제는 LLM의 편향(보기)에 일부 음성 정보를 통해 해결함
  - LLM (prior) > Whisper > BEATs
    - 음성을 기반으로 답을 내는 게 아닐 정도로 LLM의 지배력이 과도히 큼
      - LM의 편향을 줄이는 방향이 필요해 보임

exp \ category (%)	animal-multi		animal-single		emotion-multi		emotion-single		gender-multi		gender-single		language-multi		language-single	
PAPER (BASE)	34.6		68.6		28.2		19.8		48.6		59.8		29.6		21.8	
7B, B=5, BASE	35		68.4		28.2		20.2		48.4		60.2		29.6		21.2	
7B, B=5, B_ZP	28.6	6.4	40.2	28.2	28.8	0.6	20.4	0.2	48.4	0	61.4	1.2	28.2	1.4	23	1.8
7B, B=5, W_ZP	20.6	14.4	21.8	46.6	21.6	6.6	5.4	14.8	52.4	4	45.4	14.8	18.6	11	22	0.8
7B, B=5, A_ZP	10	25	12.2	56.2	19.4	8.8	12	8.2	27.4	21	42.8	7.4	12.4	17.2	17.4	3.8
7B, B=5, SHRT	20.2	14.8	69	0.6	1.2	27	12.8	7.4	0.8	47.6	68.4	8.6	7.6	22	3.4	17.8

[exp2] When do SALMONN fail? – SKR-LD-ZP-B5

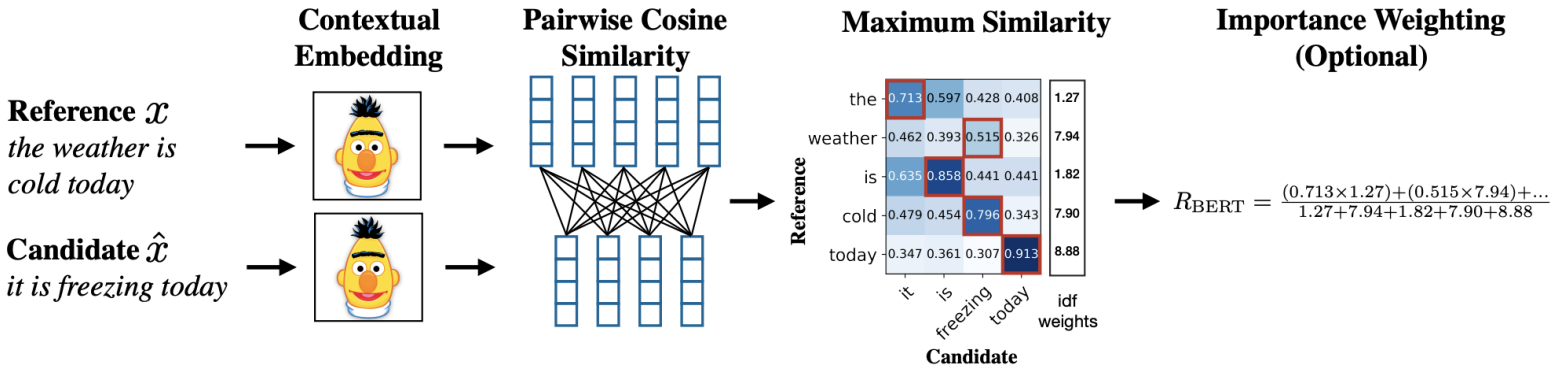
exp \ category (%)	animal-multi	animal-single	emotion-multi	emotion-single	gender-multi	gender-single	language-multi	language-single
PAPER (BASE)	34.6	68.6	28.2	19.8	48.6	59.8	29.6	21.8
7B, B=5, BASE	35	68.4	28.2	20.2	48.4	60.2	29.6	21.2
Z_EARLY, 15s	36.6	70	29.2	22	48.6	58.2	28	21.2
Z_EARLY, 30s	37.6	67.6	29	22	47.8	57	28.8	22
Z_MID, 15s	38.2	66.4	27.4	21.8	49.4	63.4	27.8	24.4
Z_MID, 30s	33	60.4	27.4	20.8	49.6	59.4	29.4	24.8
Z_LATE, 15s	35	60.4	28.2	20.4	49.6	63.6	29.8	24.8
Z_LATE, 30s	32.6	56.2	28	18.8	52.2	52.6	29.2	26.6

[exp2] When do SALMONN fail? – SKR-LD-NP-B5

exp \ category (%)	animal-multi	animal-single	emotion-multi	emotion-single	gender-multi	gender-single	language-multi	language-single
PAPER (BASE)	34.6	68.6	28.2	19.8	48.6	59.8	29.6	21.8
7B, B=5, BASE	35	68.4	28.2	20.2	48.4	60.2	29.6	21.2
N_EARLY, 15s	33.2	65	28.2	18	48.6	63	29	21.8
N_EARLY, 30s	33.8	62.2	29.6	19.8	52.4	55.2	29.8	22.6
N_MID, 15s	33.4	59.4	27	16.6	50.4	60.2	28.4	24.2
N_MID, 30s	30.8	53.4	29.2	19.4	52.8	55.2	30.2	26.6
N_LATE, 15s	30	57.2	27.6	18	50.6	63.2	28.2	24.6
N_LATE, 30s	29.6	50.4	29.4	19.6	52.6	56	29.2	27.2

Logit Calibration Idea

- BERTScore에 영감을 받아, 현재까지의 생성된 토큰 시퀀스와 유사한지에 대한 점수를 추가
  - log-sum prob + beamscores **+ bertscore-typed scores**



Logit Calibration Idea

- 목표
  - 각 Beam 안에서 의미적으로 유사하지 않게 다음 토큰이 결정되게 하여
  - 조금의 차이가 스텝이 쌓이면서 각 Beam이 서로 다른 의미를 가지도록 함

- [hidden_size x vocab_size]의 사용
  - 임베딩 룩업
  - 출력층 Linear

- 가정
  - |V|로의 Linear 직전의 마지막 hidden_size 크기의 벡터([:, -1, :])는 모든 의미를 가지고 있음
    - (1, hidden_size)

- 수정
  - (1, hidden_size)와 top-k'(k'=2k)인 (1, hidden_size)에 대한 cosine similarity를 penalty term으로
    - score -= α * ReLU(cos_sim)

project-multimodal
salmonn

학생 원종찬, 최재원
Language & AI융합전공

2025. 11. 17. 3 PM.
한국외대 교수회관 401호