

CALL-E

#Dialogue Summarization #Text-to-Image #Chat

NLP-08

MIML

Contents

Introduction

Datasets

Models

Serving

Results

Appendix

About Us



임동진

- 대화 요약 Feasibility Check
- 대화 요약 모듈화 및 metric 구현
- Backend & Product Serving
- Github 관리



정재운

- 채팅 데이터 전처리
- Frontend



조설아

- Text-to-Image 모듈화 및 metric 구현
- Image dataloader 구현
- Input format 변형 실험



허치영

- Dialogue summarization 구현
- Documentation
- Model Search



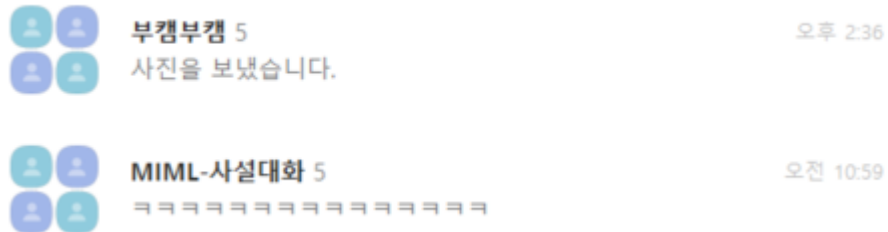
이보림

- Text-to-Image fine-tuning
- Image crawling 구현
- 이미지 데이터셋 구현
- Docker

Introduction

Problem


- 카카오톡 등의 메신저 사용시 의도와는 다른 채팅방에 메시지를 잘못 전송하는 경우가 흔히 발생
- 채팅방 간의 혼동이 주요 원인
- 마지막 대화 기록만으로는 채팅방 구분이 어려움



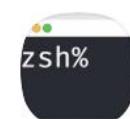
Solution


- 프로젝트 목표
 - 각 채팅방의 특징을 잘 표현하는 대표 이미지 생성
- 방법
 - 최신 대화 로그를 활용
 - 대화 로그 요약 문장으로 이미지 생성
- 기대효과
 - 채팅방 간의 혼동을 막아 실수를 방지
 - 알맞은 채팅방에서 대화 가능

 부캠부캠 5
사진을 보냈습니다. 오후 2:36

 MIML-사설대화 5
ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 오전 10:59



 부캠부캠 5
찍찍찍 이모티콘을 보냈습니다. 오전 10:27

 MIML-사설대화 5
ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 6월 3일

Why

"Use a picture. It's worth a thousand words. " –Arthur Brisbane

- 텍스트보다 빠른 정보 인식
 - 최근 대화 기록보다 직관적이고 빠르게 채팅방 구분 가능
 - [Imagery vs text which does the brain prefer?](#)
- 최근 대화를 이용한 이유
 - 최근 채팅이 해당 채팅방의 분위기를 가장 잘 표현한다고 생각

Strength

- 불편하지만 개발되지 않고 있는 기능
- 활발하게 연구되지 않은 Dialogue-to-Image를 활용한 서비스
- 생성한 이미지로 각 그룹의 고유 개성 표출 가능
- 대화방 별 이미지 선택에 소요되는 시간 절약

Datasets

Train Datasets

Dialogue Summarization



- Train set: 279,992
- Validation set: 35,004
- 대화 요약 모델 fine-tuning dataset

```
“body”: {  
  “dialogue”: [  
    {  
      “utteranceID”: “U1”,  
      “turnID”: “T1”,  
      “participantID”: “P1”,  
      “date”: “2020-07-23”,  
      “time”: “10:52:02”,  
      “utterance”: “내일이 #@이름# 엄마 생일임.”,  
    },  
    {  
      “utteranceID”: “U2”,  
      “turnID”: “T1”,  
      “participantID”: “P1”,  
      “date”: “2020-07-23”,  
      “time”: “10:52:05”,  
      “utterance”: “저녁에 외식할까 함.”  
    },  
    {  
      “utteranceID”: “U3”,  
      “turnID”: “t1”,  
      “participantID”: “P1”,  
      “date”: “2020-07-23”,  
      “time”: “10:52:06”,  
      “utterance”: “시간 ㅇㄱ?”  
    },  
    {  
      “utteranceID”: “U4”,  
      “turnID”: “T2”,  
      “participantID”: “P2”,  
      “date”: “2020-07-23”,  
      “time”: “10:53:08”,  
      “utterance”: “강의만 제때 끝나면 저녁에 시간돼요”,  
    },  
    {  
      “utteranceID”: “U5”,  
      “turnID”: “T3”,  
      “participantID”: “P3”,  
      “date”: “2020-07-23”,  
      “time”: “11:15:20”,  
      “utterance”: “저도 가능해요”,  
    }  
  ],  
  “summary”: “엄마 생일을 맞아 모두 모여 외식할 수 있는 저녁 식사 시간을 정하고 있다.”  
},  
  ...  
},  
}
```

Train Datasets

Dialogue Summarization

- Train set: 5,749
- Validation set: 1,500
- Metric 계산 모델(SentenceBERT) 학습

KorSTS

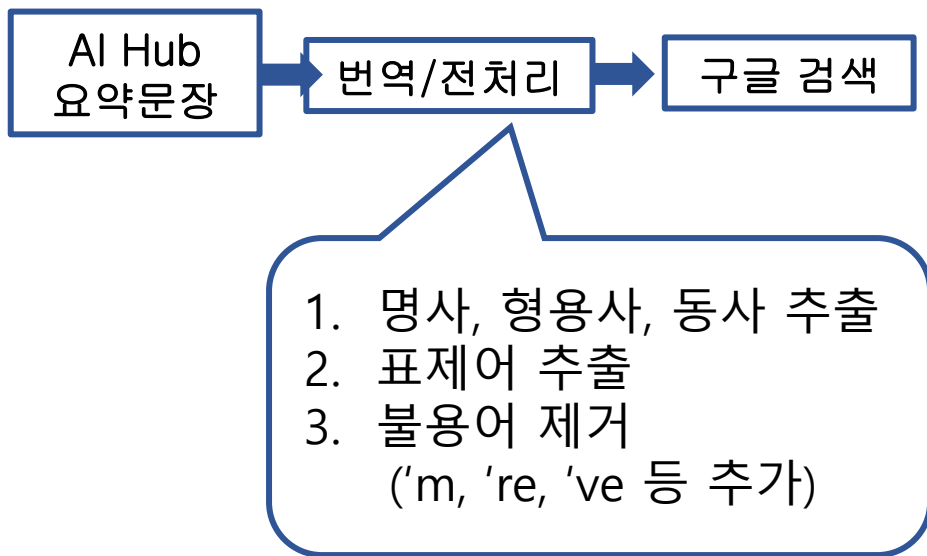
KorSTS	Total	Train	Dev.	Test
Source	–	STS-B	STS-B	STS-B
Translate by	–	Machine	Human	Human
# Example	8,628	5,749	1,500	1,379
Avg. # words	7.7	7.5	8.7	7.6

Example

Example	English Translation	Label
한 남자가 음식을 먹고 있다. 한 남자가 뭔가를 먹고 있다.	A man is eating food. A man is eating something.	4.2
한 비행기가 착륙하고 있다. 애니메이션화된 비행기 하나가 착륙하고 있다.	A plane is landing. A animated airplane is landing.	2.8
한 여성이 고기를 요리하고 있다. 한 남자가 말하고 있다.	A woman is cooking meat. A man is speaking.	0.0

Train Datasets

Text-to-Image



AI Hub 요약문장

나 혼자 산다(나혼산)에 나온 kwak do-won 배우가 제주도에서 1초를 안 쉬고 바쁘게 살고 있다며 티브이에서 본 여러 가지 활동을 전한다.



검색어

Actor Kwak Do-won appeared Live Nahonsan Mountain said live Jeju Island second rest conveying various activity saw TV.

Train Datasets

Text-to-Image



Crawled Dataset (Total 17,081)

	Total
Train (90%)	15,373
Validation(10%)	1,708

AI Hub 요약문장

나 혼자 산다(나혼산)에 나온 곽도원 배우가 제주도에서 1초를 안 쉬고 바쁘게 살고 있다며 티브이에서 본 여러 가지 활동을 전한다.



Text

Actor Kwak Do-won appeared Live Nahonsan Mountain said live Jeju Island second rest conveying various activity saw TV

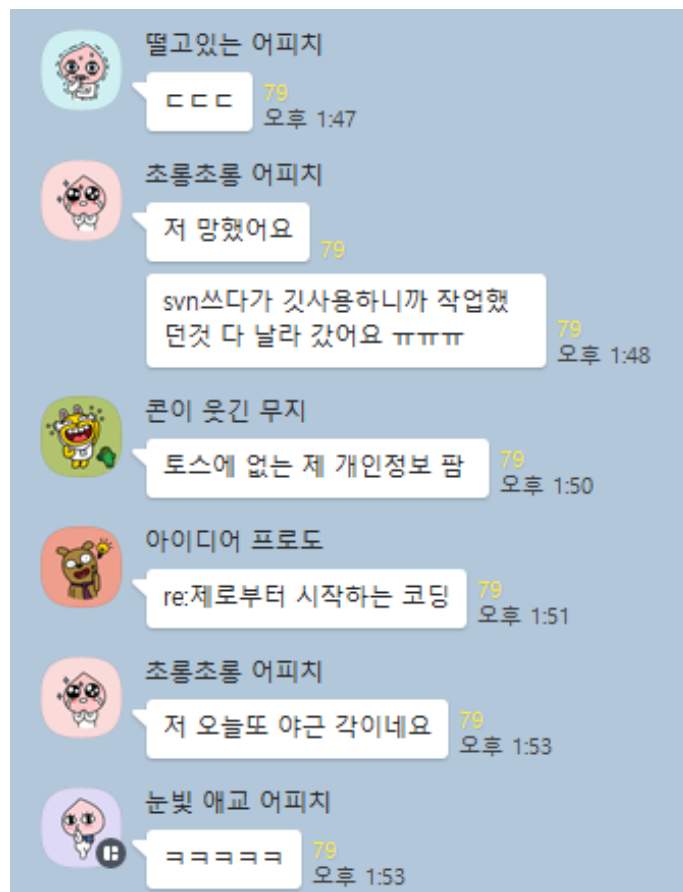
Image



Test Datasets




카카오톡 단체 채팅
크롤링 데이터 Test data



[브라운] [오후 5:29] 이번주 토
[브라운] [오후 5:29] 생일이던데
[브라운] [오후 5:29] 주말에 시간 빠야되지 않겠냐
[브라운] [오후 5:29] [redacted]
[초코] [오후 5:32] 난 가능한데
[샬리] [오후 5:32] 하...
[샬리] [오후 5:32] 그때 일정 봐야할거같은데
[문] [오후 5:41] 나 원주감 ㅎㅎ 나중에봐
[문] [오후 5:42] 담주에 다들 괜찮음??
[초코] [오후 5:42] 다음주 주말은 너무 좋지
[초코] [오후 5:43] 15일 이후부터 괜찮음
[샬리] [오후 6:22] 나
[샬리] [오후 6:22] 17일날 끝남
[샬리] [오후 6:22] 개꿀
[샬리] [오후 6:41] 굿
[브라운] [오후 7:36] ㅇㅋ
[초코] [오후 7:39] 그럼 18일로 알고있을게요
[문] [오후 8:38] 나 18일 약속있음
[샬리] [오후 8:44] 19일 ㄱ

Test Datasets

 Google AI Conceptual Caption validation dataset

			Tokens per Caption		
Split	Examples	Unique Tokens	Mean	StdDev	Median
Train	3,318,333	51,201	10.3	4.5	9.0
Valid	15,840	10,900	10.4	4.7	9.0
Test (Hidden)	12,559	9,645	10.2	4.6	9.0



where 's the best place to show off your nails ? right in front of the castle , of course !



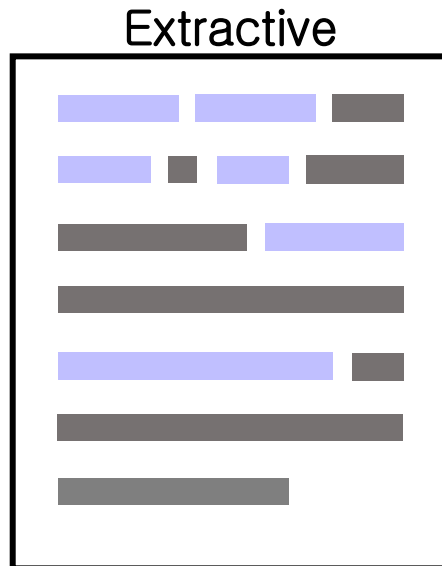
that combines elements of a simple vegetable and dish

Models

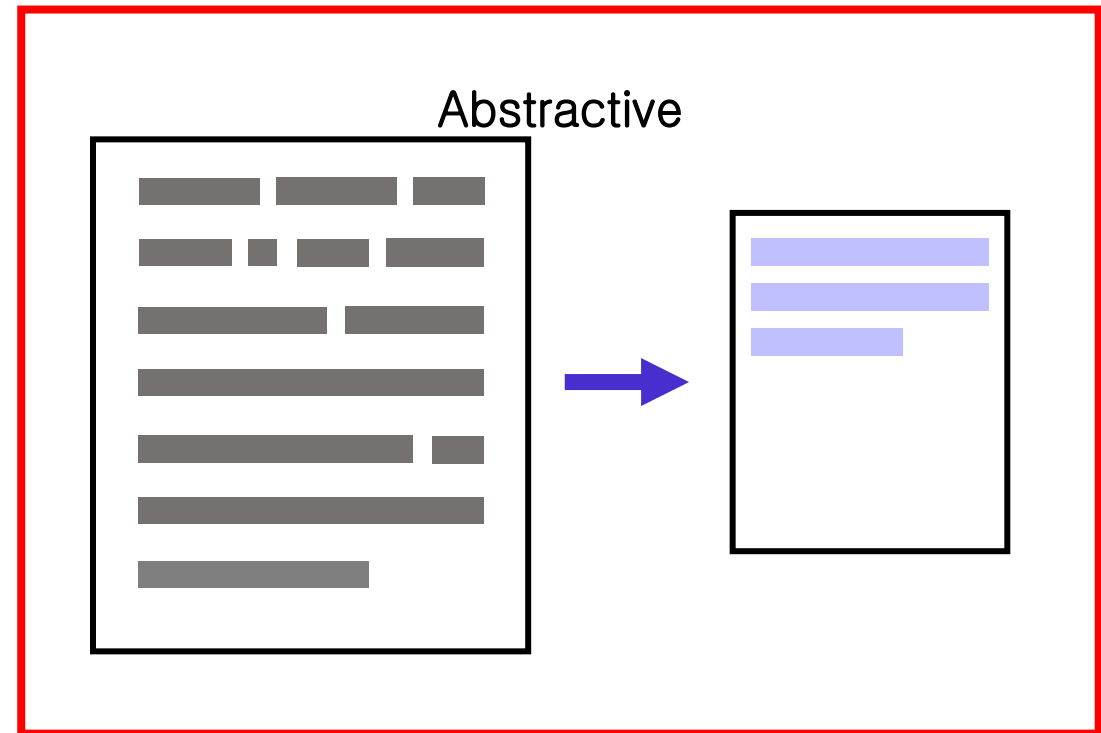
Model Search

Dialogue Summarization

- Extractive? Abstractive?

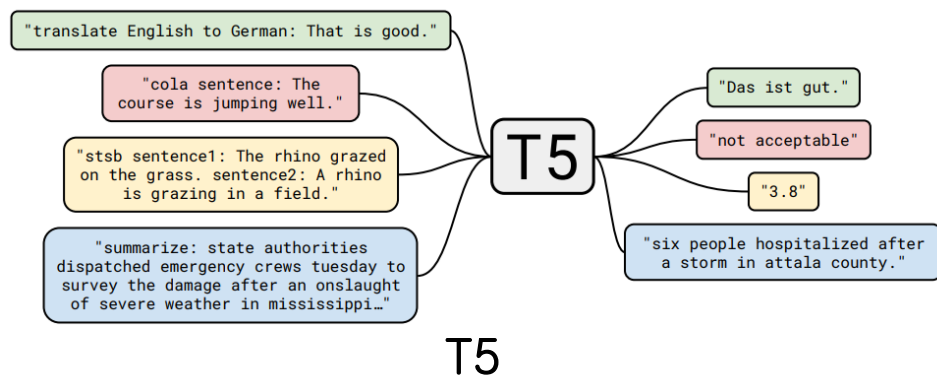


VS

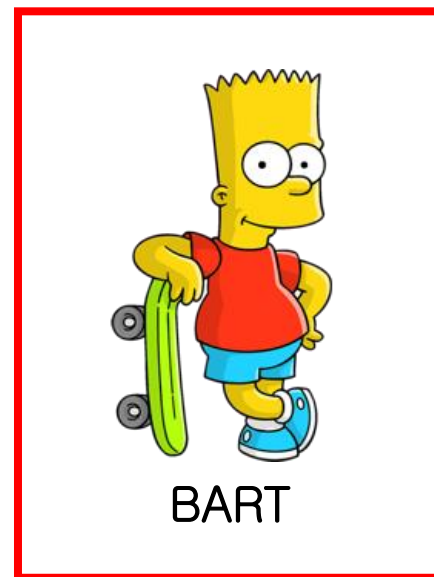


Model Search

Dialogue Summarization



VS

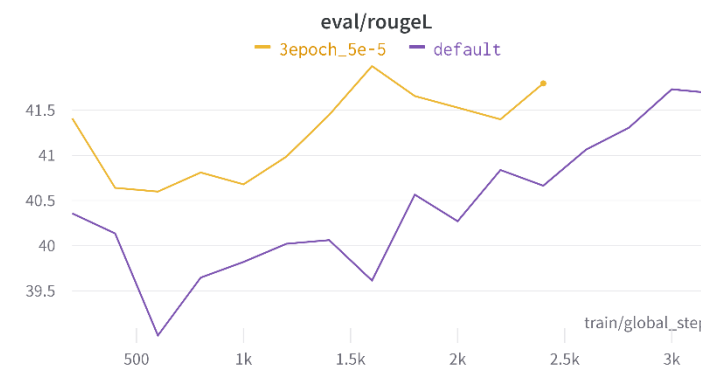
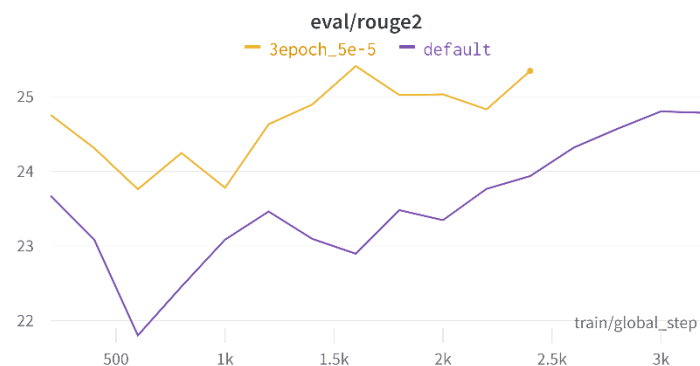
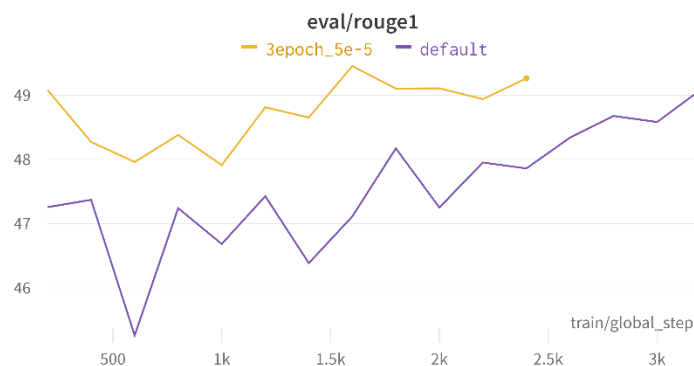


Feasibility Check

Dialogue Summarization

Model	ROUGE-1	ROUGE-2	ROUGE-L
BART-large-SAMSum	53.434	28.744	44.185

SAMSum dataset



Metric

Dialogue Summarization

- ROUGE Score

- 동의어를 고려하지 않음
- 한국어 접사로 인해 정확한 성능 평가 불가능

$$\text{ROUGE-N} = \frac{\text{count}_{\text{match}}(\text{gram}_n)}{\text{count}(\text{gram}_n)}$$

- RDASS (More details in Appendix.)

- 문서(v_d), 정답 요약 문장(v_r), 예측 요약 문장(v_p) 사이의 관계 고려
- Sentence BERT를 이용해 벡터 v_d, v_r, v_p 추출
- 각 vector간 cosine similarity의 평균을 지표로 사용
- Dialogue-Golden Summary간의 유사도 정확하게 평가 불가

$$s(p, r) = \cos(v_p, v_r) = \frac{v_p^T \cdot v_r}{\|v_p\| \|v_r\|}$$

$$s(p, d) = \cos(v_p, v_d) = \frac{v_p^T \cdot v_d}{\|v_p\| \|v_d\|}$$

$$RDASS = \frac{s(p, r) + s(p, d)}{2}$$

Metric

Dialogue Summarization

- Cosine Similarity (Golden Summary – Prediction)
 - RDASS에서 아이디어 착안
 - Golden Summary와 Model Prediction의 cosine similarity만을 이용
 - KorSTS 기준 Score 3.0 (cosine similarity 0.6) 이상의 개수, 전체 데이터 평균 유사도로 평가

$$s(p, r) = \cos(v_p, v_r) = \frac{v_p^T \cdot v_r}{\|v_p\| \|v_r\|}$$

v_r : Golden Summary

v_p : Generated Summary

KorSTS Example

Example	English Translation	Label
한 남자가 음식을 먹고 있다. 한 남자가 뭔가를 먹고 있다.	A man is eating food. A man is eating something.	4.2
한 비행기가 착륙하고 있다. 애니메이션화된 비행기 하나가 착륙하고 있다.	A plane is landing. A animated airplane is landing.	2.8
한 여성이 고기를 요리하고 있다. 한 남자가 말하고 있다.	A woman is cooking meat. A man is speaking.	0.0

Experiments

Dialogue Summarization

- Input Format
 - [PID: utterance \r\n;PID: utterance \r\n]

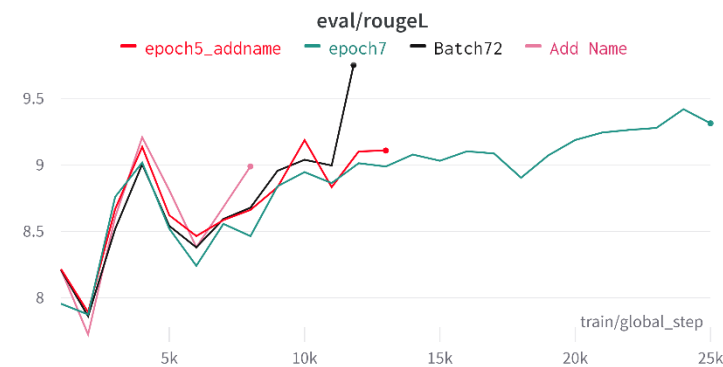
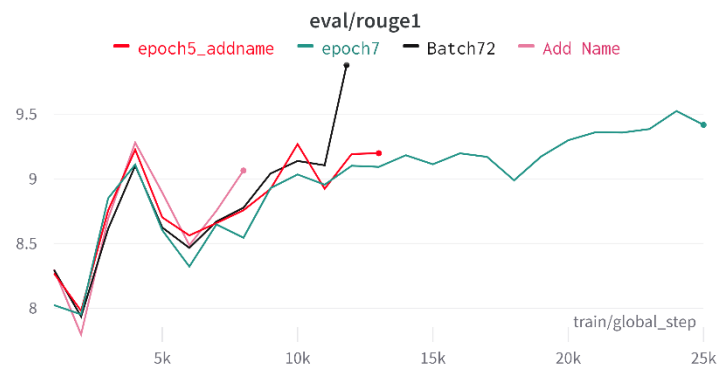
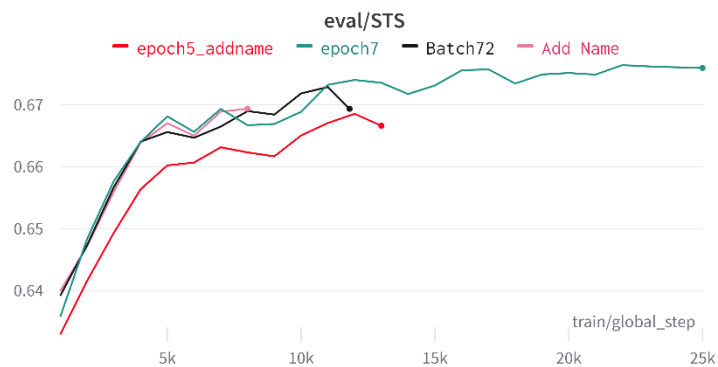
ParticipantID	Utterance
P01	너~~무 앞인데
P02	ㅋㅋㅋㅋ
P02	#@이름#가
P02	겉돈단 생각 없지않아 조금하고있어서 조금 예민할수도
P02	너랑나처럼 #@기타#편하고 막 잘놀고그런건아니자너
P02	그와중에 놀아달라할때 씹은 약간의 소외감느낄껄ㅋㅋ
P02	근데오긴올듯
P02	근데 헤어졌는데
P02	너무 무시해서ㅋㅋ맘상했것다..
P01	아아아아
P01	ㅇㅇ갑자기
P01	이해 팍 된다
P01	확 와닿았음
P01	ㅋㅋㅋㅋㅋㅋ
P01	겉돈단 느낌 지금도 들려나

Experiments

Dialogue Summarization

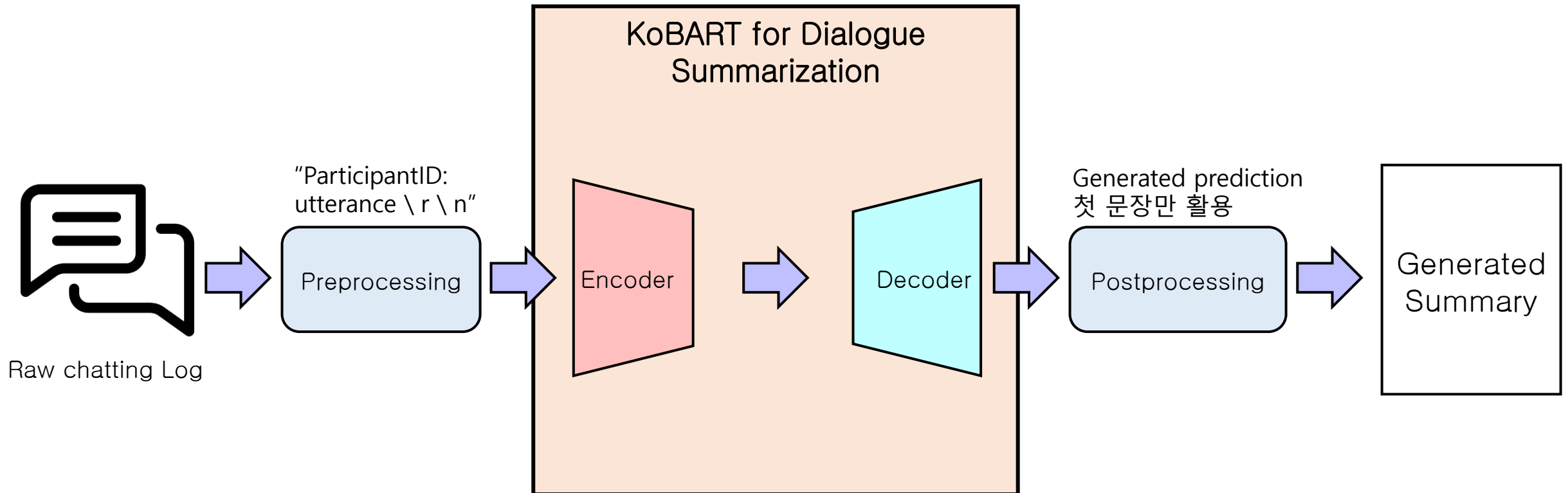
- Change Input Format (Add Name) *Controllable Neural Dialogue Summarization with Personal Named Entity Planning*
 - {ParticipantIDs} + Dialogue

Comprehensive Planning: {John, Mary, Tony, Bell}
Output: Mary sent John some gossip from her college reunion.
John missed the reunion. Tony and Bell split up. Bell met a new guy.
He came with her to to reunion.



Architecture

Dialogue Summarization



Experiments

Dialogue Summarization

- Hyperparameter Tuning of Decoding Methods

- Top_k: 확률 상위 50개만
- Top_p: 확률 0.95 이상만
- No_repeat_ngram: trigram은 최대 1회 반복
- Temperature: 모델의 randomness 약하게

```
outputs = model.generate(input_ids,  
                           num_beams=5,  
                           max_length=64,  
                           attention_mask=attention_mask,  
                           top_k=50,  
                           top_p=0.95,  
                           no_repeat_ngram_size=3,  
                           temperature=0.7  
                           )
```

generate() arguments	Avg. Cosine Similarity	# of Cosine Similarity >0.6
Baseline	0.6443	23,121
temperature(0.7)	0.6453	23,438
no_repeat_ngram_size(3)	0.6482	23,238
temperature(0.7) + no_repeat_ngram_size(3)	0.6487	23,477

Model Search

Text-to-Image

	LAFITE	Latent Diffusion Model (LDM)	minDALL-E
후보로 선택한 이유	<ol style="list-style-type: none"> 1. Model Size ↓ (75M) → 속도 ↑ 2. Fine-tuning 하기 좋음 → 다양한 pretrained 모델 제공 3. Language-Free training으로 이미지로만 학습 가능 	<ol style="list-style-type: none"> 1. 4억 개의 image-text 쌍을 학습한 모델 → 높은 성능 기대 2. GAN을 앞선 높은 image 합성 성능 	<ol style="list-style-type: none"> 1. Zero-shot 평가에도 높은 성능을 가진 DALL-E의 구조를 가져옴 2. V100에서 학습 가능
모델 구조	<p>Random Noise Z → Mapping Network → Intermediate W</p> <p>Real Image X → Translator → h' Semantic</p> <p>Style S (from Affine) and Condition C (from 2-layer FC) → Conditional Style u (from Affine)</p> <p>Different learned module per generator layer</p>	<p>Pixel Space: $x \rightarrow \mathcal{E} \rightarrow z$, $\tilde{x} \leftarrow \mathcal{D} \leftarrow z$</p> <p>Latent Space: Diffusion Process, Denoising U-Net ϵ_θ with Q, K, V blocks</p> <p>Conditioning: Semantic Map, Text, Representations, Images, τ_θ</p> <p>denoising step, crossattention, switch, skip connection, concat</p>	DALL-E의 축소

Model Search

Text-to-Image

- 모델 선정
 - 평균적으로 높은 CLIP Score

	minDALL-E	Lafite	Latent Diffusion(LDM)
CLIP score	0.3193	0.2722	-
FID score	322.235	467.279	399.945

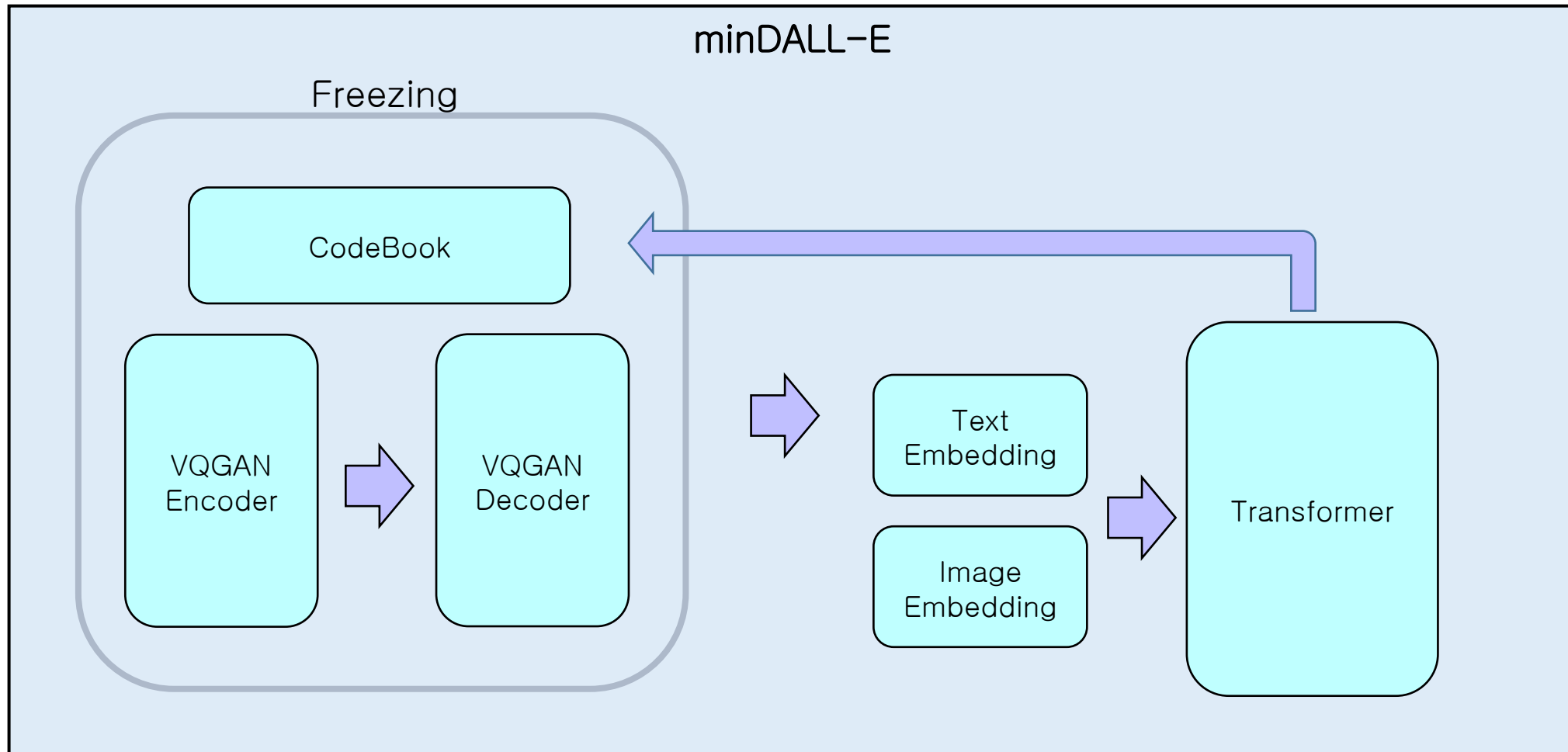
Metric

Text-to-Image

- **CLIP Score** (Main Metric)
 - 텍스트-이미지 사이의 상관 관계 평가
 - 높을 수록 좋은 성능
 - FID, Inception Score와 달리 요약 문장을 활용할 수 있음
- **FID**
 - 생성된 이미지와 정답 이미지의 유사도 평가
 - 낮을 수록 좋은 성능

Architecture

Text-to-Image



Input text transformation

Text-to-Image

- Sentence Transformation

- Papago api를 활용하여 Input data(한국어 문장)를 영어로 번역
- 번역된 영어 문장을 다양한 방식으로 변형하여 CLIP score 비교 실험
- 접두어, 문장 전처리, 문장 재구성의 조합으로 총 12가지 문장 생성 (* 내용은 아래와 같음)

	접두어	문장 전처리	문장 재구성
내용	① 접두어 없음	④ 원본 문장	⑦ 띄어쓰기로 구분
	② A painting of	⑤ 명사/동사/형용사 추출	⑧ 쉼표로 구분
	③ A painting =	⑥ 불용어 제거	⑨ +기호로 구분

Input text transformation

Text-to-Image

- 12가지 문장 변형 방법에 대한 모델 성능 비교

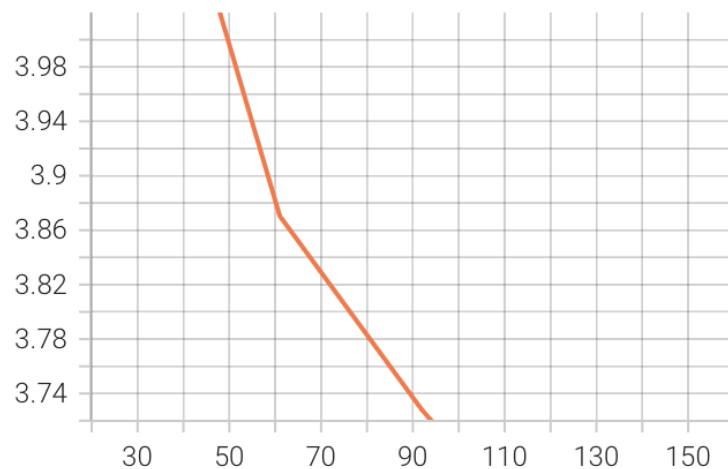
- CLIP Score > 0.2 + Human Evaluation
- BEST : 불용어 제거
- + 쉼표로 문장 성분 구분

문장 변형 방법 조합	문장	CLIP score
① + ④ + ⑦	I went home right away today, but I couldn't rest well.	0.2002
① + ⑤ + ⑦	went home right away today n't rest well	0.2017
① + ⑤ + ⑧	went, home, right, away, today, n't, rest, well	0.2126
① + ⑥ + ⑦	went home right away today , could n't rest well .	0.1899
① + ⑥ + ⑧	went, home, right, away, today, ,, could, n't, rest, well, .	0.2100
② + ④ + ⑦	A painting of I went home right away today, but I couldn't rest well.	0.2475
② + ⑤ + ⑦	A painting of went home right away today n't rest well	0.2521
② + ⑤ + ⑧	A painting of went, home, right, away, today, n't, rest, well	0.2719
② + ⑥ + ⑦	A painting of went home right away today , could n't rest well .	0.2570
② + ⑥ + ⑧	A painting of went, home, right, away, today, ,, could, n't, rest, well, .	0.2727
③ + ⑤ + ⑨	A painting = went + home + right + away + today + n't + rest + well	0.2536
③ + ⑥ + ⑨	A painting = went + home + right + away + today + , + could + n't + rest + well + .	0.2521

Experiments

Text-to-Image

- minDALL-E Fine-Tuning
 - Cross-entropy Loss



Tensor board CE loss graph

- Early stopping

```
# Setting EarlyStopping
early_stop_callback = EarlyStopping(
    monitor='val/loss',
    min_delta=0.00,
    patience=3,
    verbose=True,
    mode='min'
)
```


Architecture



KoBART

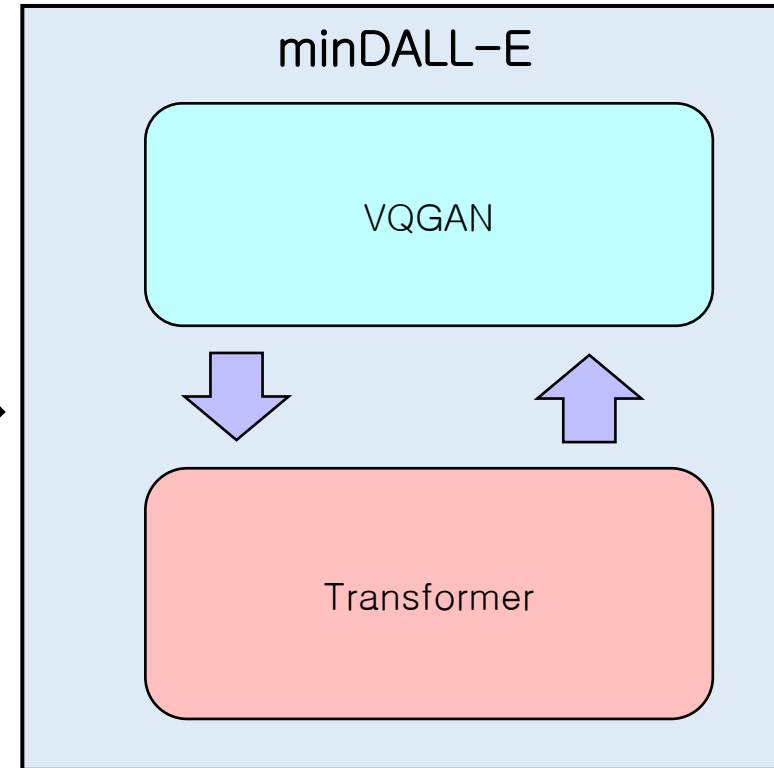
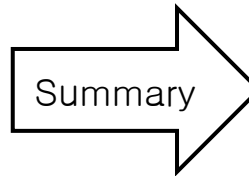
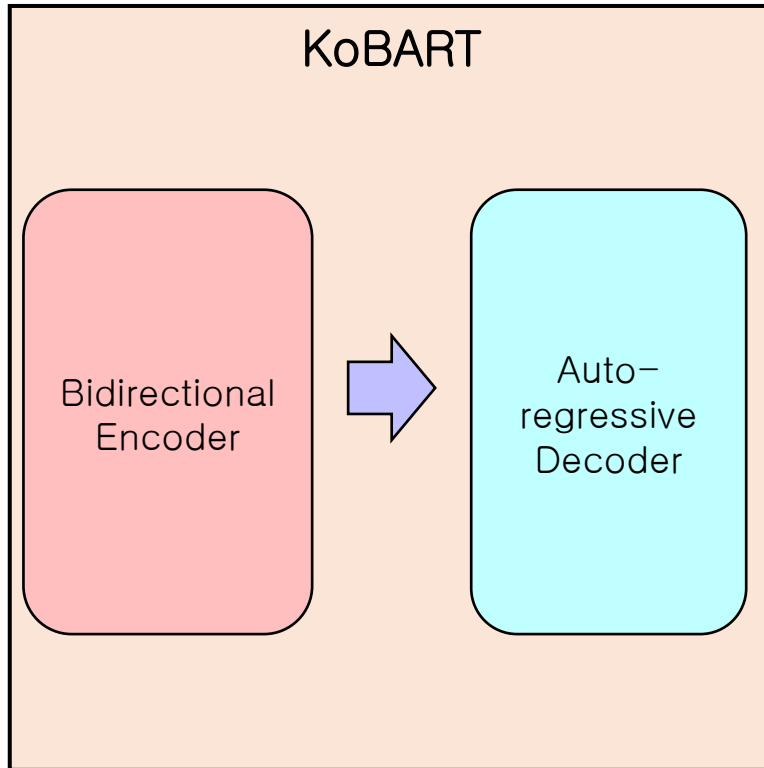


minDALL-E

Architecture



Chatting Log



Serving

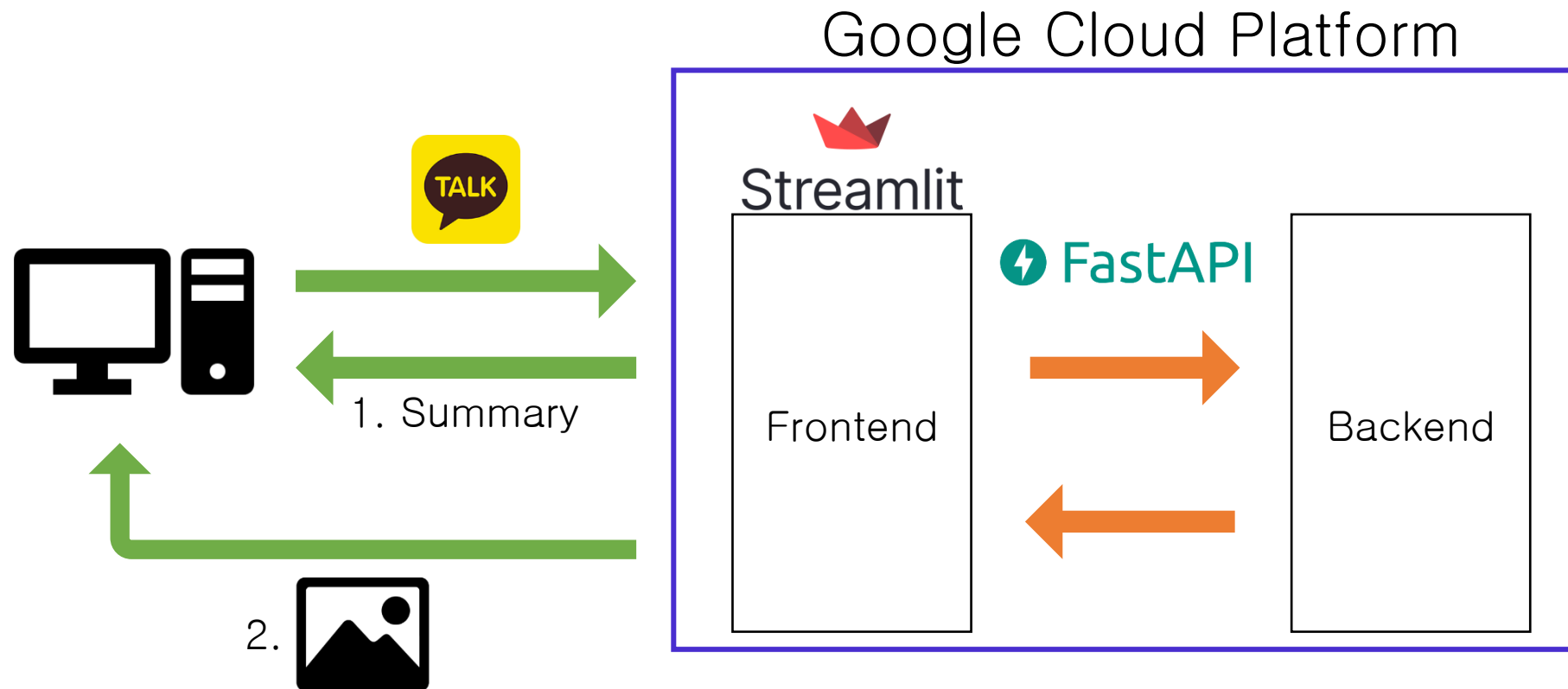
Cloud



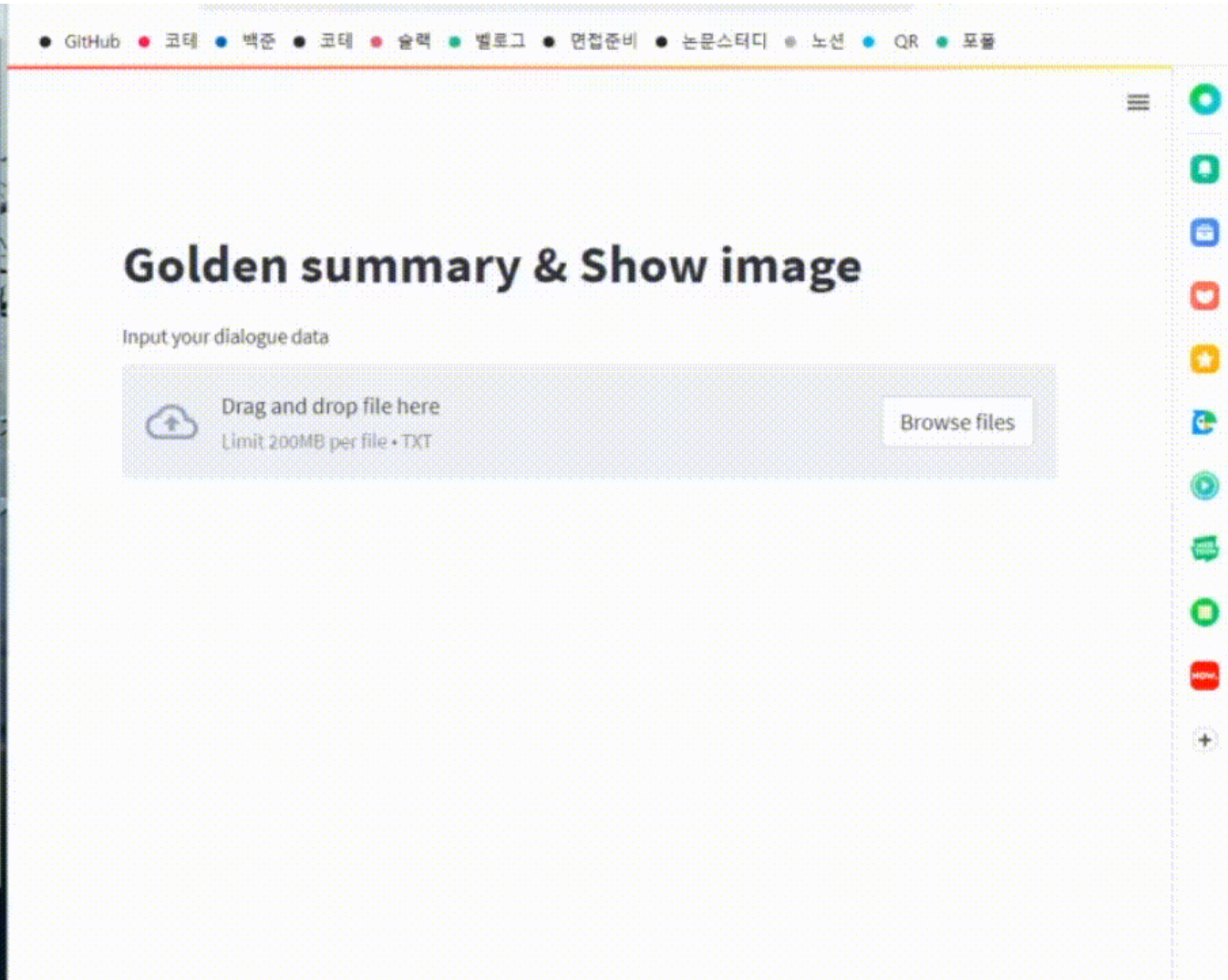
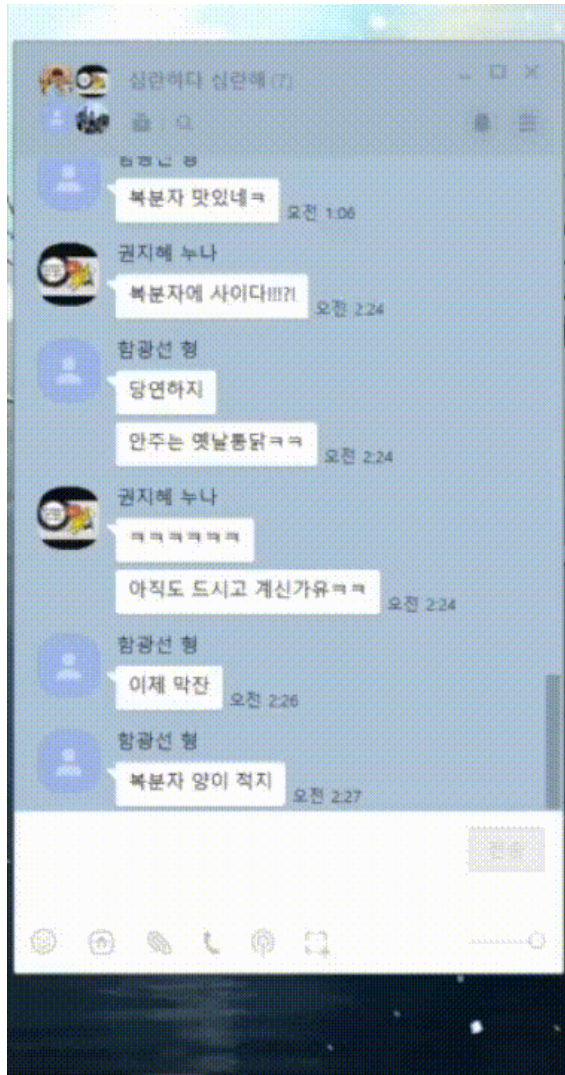
vs



Service Architecture



Result



Appendix

Future Works

한계점	해결 방안
한국어 → 영어 번역 단계의 필요	한국어 text-image 쌍으로 학습된 모델 필요
서비스 제공까지 긴 시간 소요	모델 경량화를 통한 추론 속도 개선
복잡한 summary로 인한 성능 하락	Summarization 모델 변경 및 개선
한국어 대화 요약에 맞는 평가 지표 부재	ROUGE, RDASS보다 대화 요약에 걸맞는 평가 지표의 연구 필요
모델의 재학습이 이루어지지 않음	유저 피드백을 통한 모델 학습 방법 고안 필요
개인정보 비식별화가 완벽히 이루어지지 않음	완전한 개인정보 비식별화 규칙 적용 필요
모델 크기에 따른 리소스 부족	모델 불러오는 방식의 변경

Lesson Learn

- AI 서비스 개발에서 문제정의와 그에 맞는 **양질의 데이터셋** 수집 및 구축의 중요성
- Subproblem의 해결보다 **product의 흐름**에 더 집중해야한다.
- 논문의 **목적과 활용** 가능성을 고려하여 아이디어를 얻어오는 것에 대한 중요성의 인지
- 주제 선정 시 **실현 가능성 확인**이 우선되어야 함
- 정량적 **평가 지표** 선택 및 **실험 기록**의 중요성
- Clean Code와 Modularization의 중요성
- 실행 속도 향상을 위한 **모델 경량화**
- 어려운 문제 해결 시 **세분화**의 필요

예상 Q&A

- minDALL-E에서는 왜 dVAE가 아니라 VQGAN을 사용하나요?
 - minDALL-E에서는 높은 퀄리티의 이미지 샘플링을 위해 DALL-E의 dVAE를 VQGAN으로 대체
 - dVAE: 메모리 사용량을 줄이기 위해 세부 특징보다 이미지의 큰 특징을 구분해주는 방향으로 학습 진행
 - VQGAN: 성능 자체에 집중. 학습한 이미지를 세분화하여 진품에 가까운 가짜 이미지를 생성하고, 그 가짜를 구별해내는 과정을 반복하며 성능을 향상
- 생성 문장의 가장 첫 문장만 선택한 이유가 있을까요?
 - 생성 결과에서 첫 문장 이후 ㅋㅋㅋ 등의 무의미한 단어나 보완 설명이 생성
 - Text-to-Image의 입력은 한 문장만 들어가는 것이 좋음
 - 가장 핵심 요약을 담고 있는 첫 문장만을 활용하기로 결정

Q&A

References

- Ham, Jiyeon, et al. "Kornli and korsts: New benchmark datasets for korean natural language understanding." *arXiv preprint arXiv:2004.03289* (2020).
- Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461* (2019).
- Liu, Zhengyuan, and Nancy F. Chen. "Controllable Neural Dialogue Summarization with Personal Named Entity Planning." *arXiv preprint arXiv:2109.13070* (2021).
- Lee, Dongyub, et al. "Reference and document aware semantic evaluation methods for Korean language summarization." *arXiv preprint arXiv:2005.03510* (2020).
- Ramesh, Aditya, et al. "Zero-shot text-to-image generation." *International Conference on Machine Learning*. PMLR, 2021.
- Zhou, Yufan, et al. "LAFITE: Towards Language-Free Training for Text-to-Image Generation." *arXiv preprint arXiv:2111.13792* (2021).
- Rombach, Robin, et al. "High-Resolution Image Synthesis with Latent Diffusion Models." *arXiv preprint arXiv:2112.10752* (2021).

References

- <https://www.learnevents.com/blog/2015/09/07/imagery-vs-text-which-does-the-brain-prefer/>
- <https://news.mit.edu/2014/in-the-blink-of-an-eye-0116>
- <https://huggingface.co/blog/how-to-generate>
- <https://github.com/kakaobrain/minDALL-E>
- <https://github.com/lucidrains/big-sleep>
- <https://aihub.or.kr/aidata/30714>
- <https://pixabay.com/ko/>