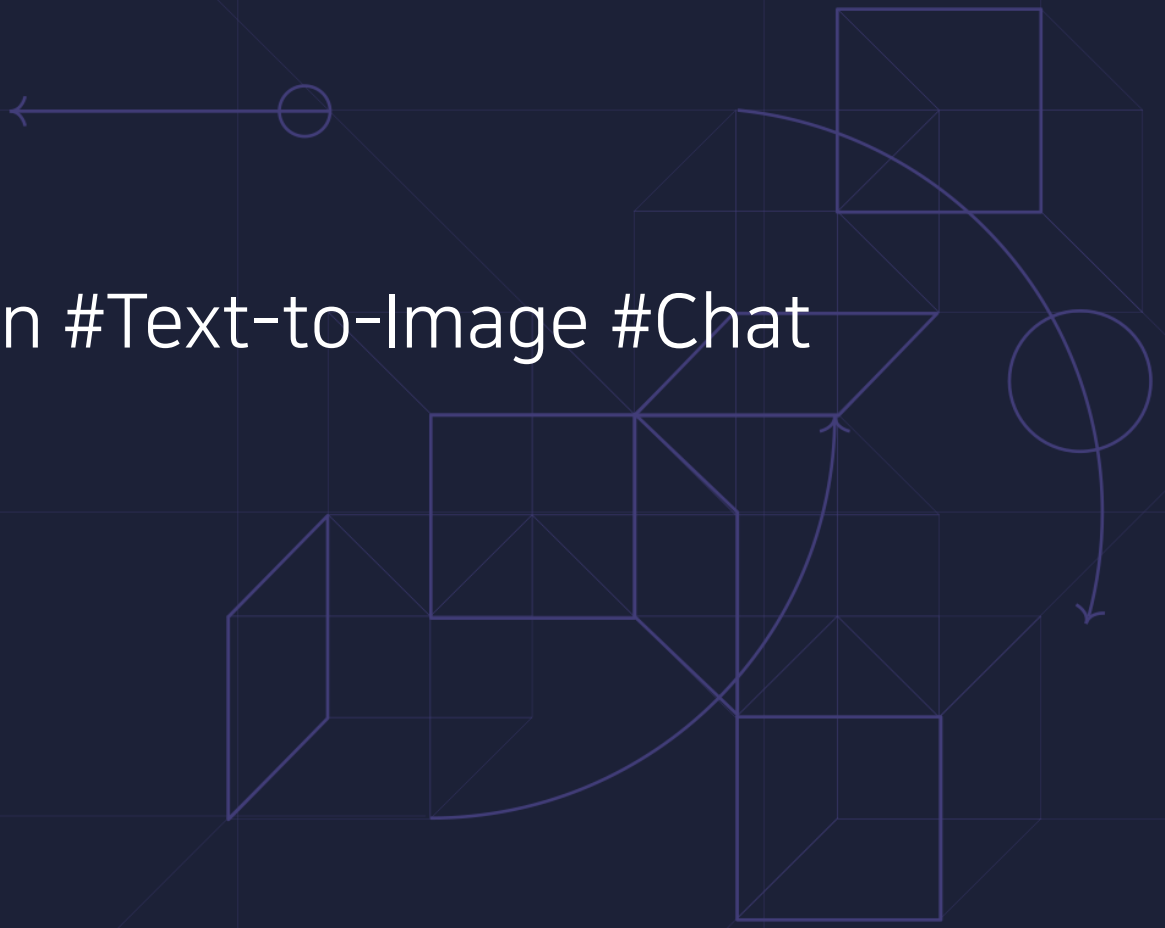
CALL-E

#Dialogue Summarization #Text-to-Image #Chat

NLP-08
MIML



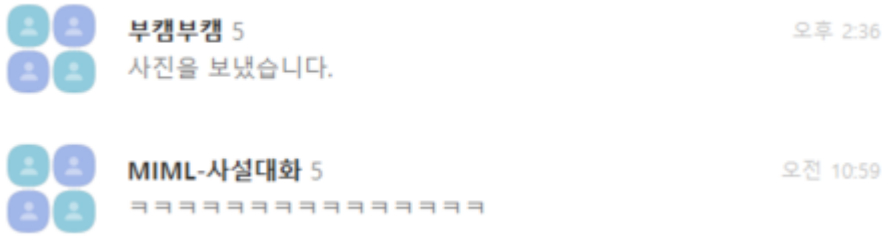
boostcamp ai tech

Contents

Introduction

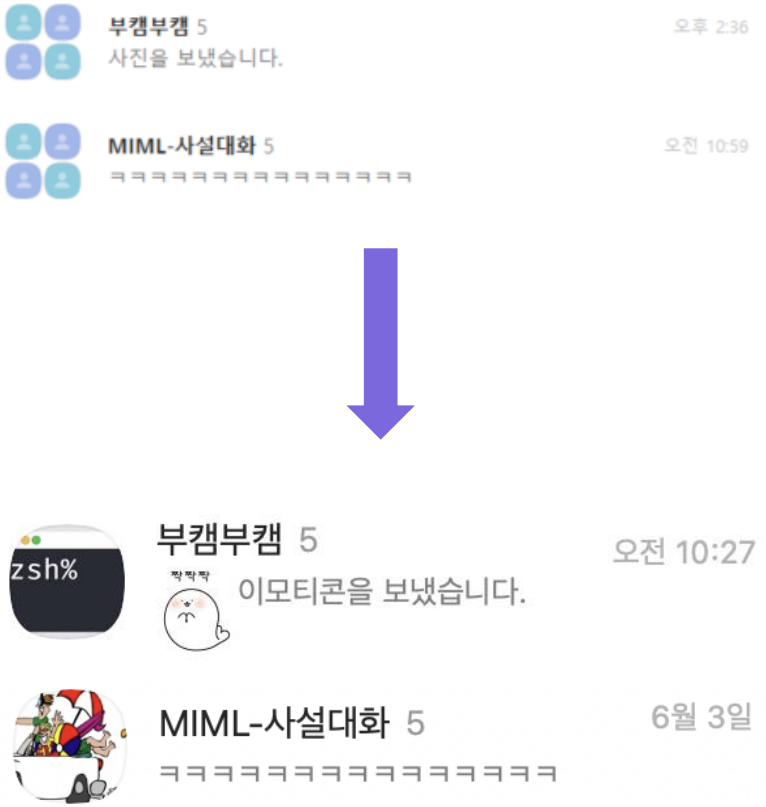
Problem

- 카카오톡 등의 메신저 사용시 의도와는 다른 채팅방에 메시지를 잘못 전송하는 경우가 흔치 않게 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



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

kakaobrain

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

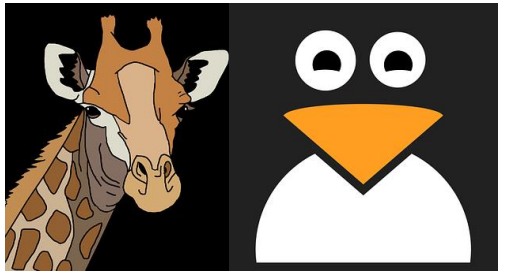
AI Hub

Train: 279,992
Validation: 35,004
data
- Header
  - dialogueInfo
    - dialogueID: str
    - numberOfParticipants: int
    - numberOfUtterances: int
    - numberOfTurns: int
    - type: str
    - topic: str
  - participantsInfo: list
    - participantsID: str
    - gender: str
    - age: str
    - residentialProvince: str
- body
  - dialogue: list
    - utteranceID: str
    - turnID: str
    - participantID: str
    - date: str
    - time: str
    - utteance : str
  - summary: str

Train Datasets

Text-to-Image

pixabay



< Text: An illustration of ### >

Crawled Dataset (Total 56,251)

	Total	Illustration	Scenery	Vector Image
Train	45,000	15,595	17,566	11,839
Validation	11,251	3,899	4,392	2,960

Test Datasets

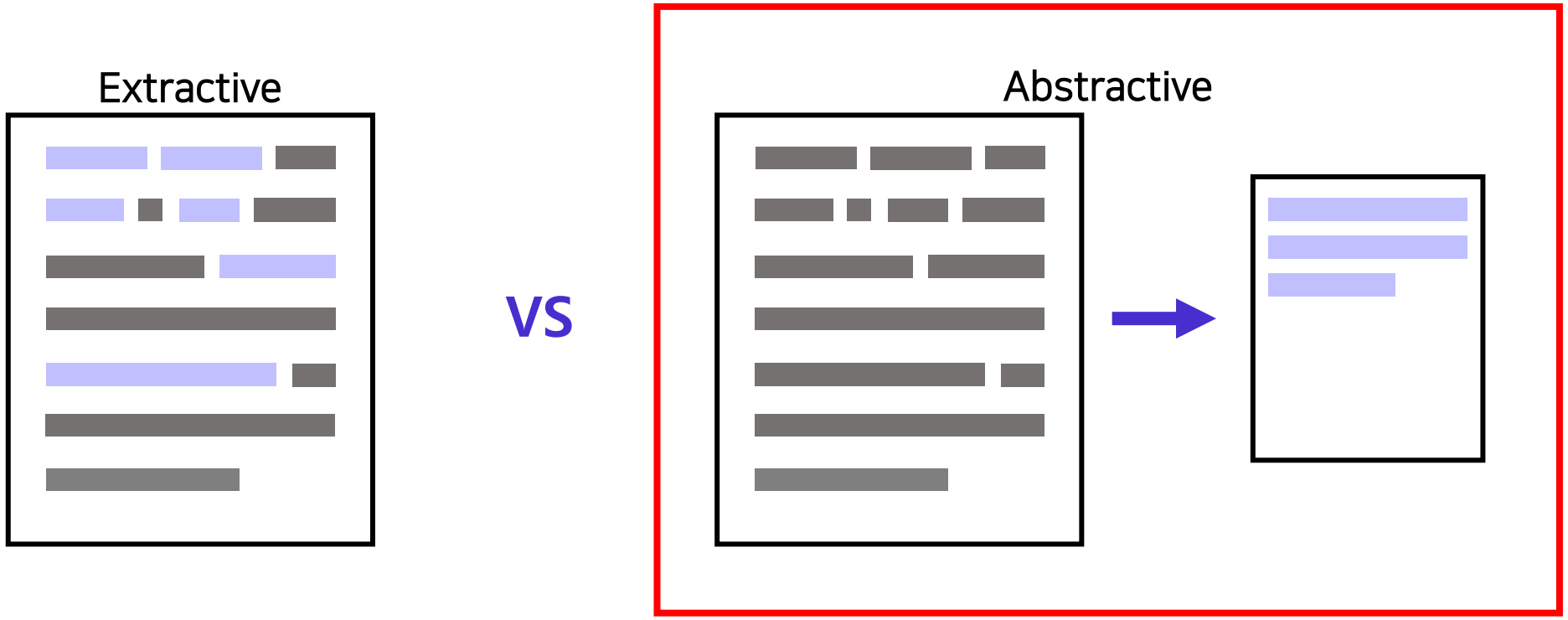
카카오톡 단체 채팅 크롤링 데이터 Test data

Models

Model Search

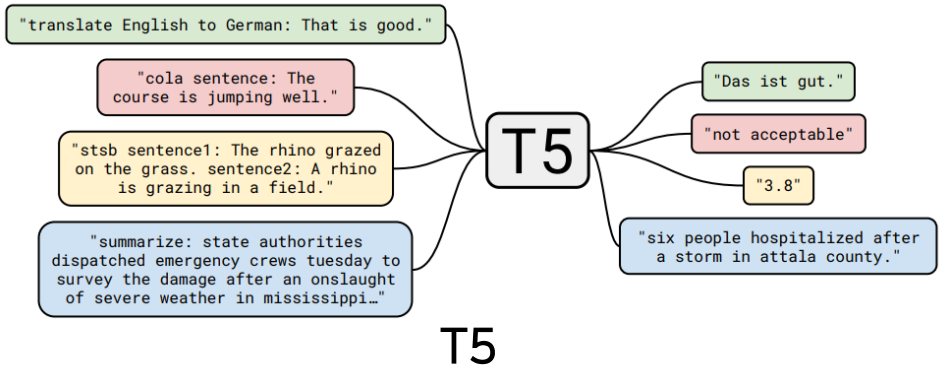
Dialogue Summarization

- Extractive? Abstractive?



Model Search

Dialogue Summarization



T5

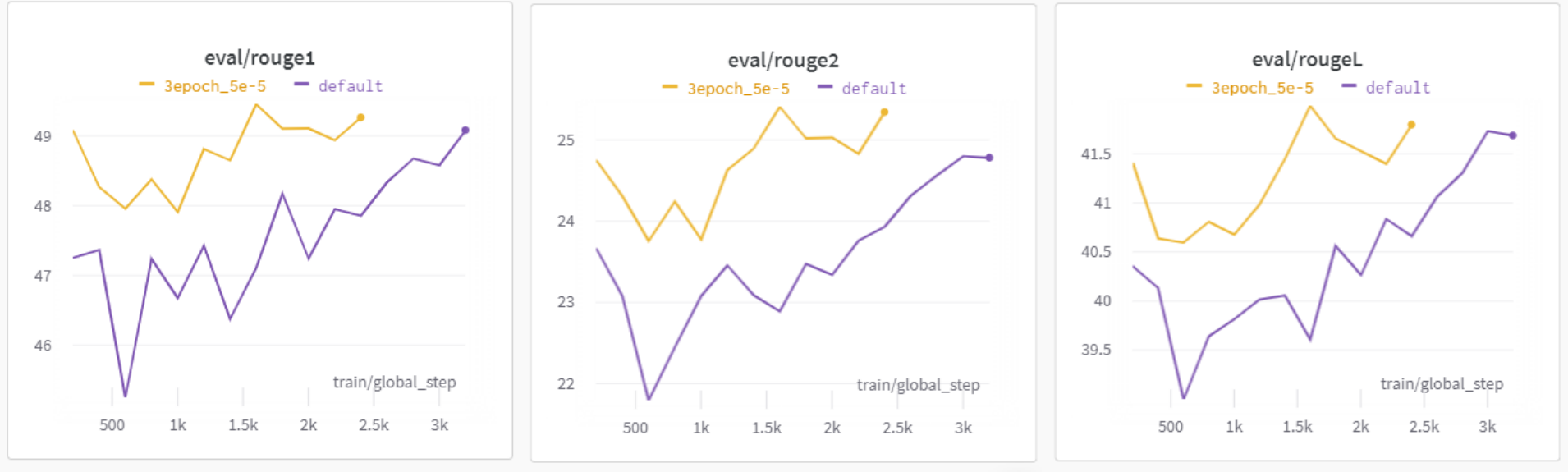
VS

BART

Feasibility Check

Dialogue Summarization

SAMSum dataset



Metric

Dialogue Summarization

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

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

- RDASS (More details in Appendix.)
  - < 문서, 정답 요약 문장, 예측 요약 문장 > 사이의 관계 고려
  - Dialogue-Golden Summary간의 유사도 정확하게 평가 불가

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

- Cosine Similarity (Golden Summary – Prediction)
  - RDASS에서 아이디어 착안

$$s(p,r)$$

Experiments

Dialogue Summarization

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

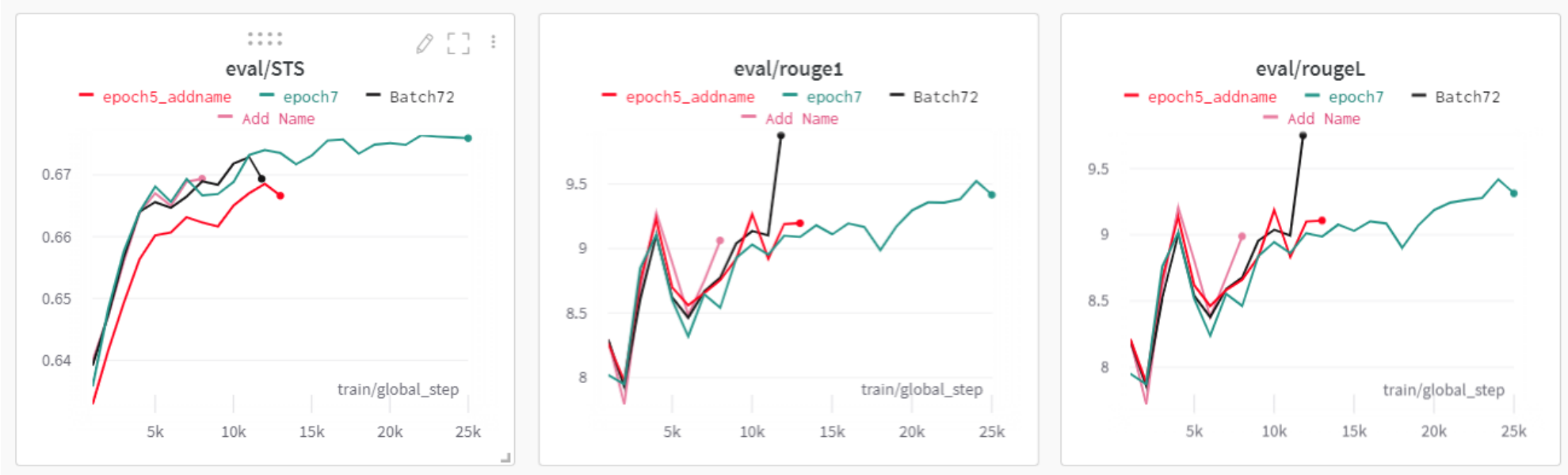
```
P01: 너~~무 앞인데
P02: ㅋㄲㅋㅋ
P02: #@이름#가
P02: 겉돈단 생각 없지않아 조금하고있어서 조금 예민할수도
P02: 너랑나처럼 #@기타#편하고 막 잘놀고그런건아니자너
P02: 그와중에 놀아달라할때 씹으믄 약간의 소외감느낄껄ㅋㅋ
P02: 근데오긴올듯
P02: 근데 헤어졌는데
P02: 너무 무시해셤ㅋ맘상했것다..
P01: 아아아아
P01: ㅇㅇ갑자기
P01: 이해 팍 된다
P01: 확 와닿앗음
P01: ㅠㅠㅠㅠㅠ
P01: 겉돈단 느낌 지금도 들려나
P01: 이제 대놓고 괴롭히는 좋아ㅋㅋ이없는뎅
P01: 운다 울어운다 울어근데 소외감 느꼇을만하네
P01: #@기타# 멘탈 나갓을때니까
P01: 아무도안들어주는거같거
P02: 그러게
P02: 하씨 친구랑놀다 얼핏본거여서
P01: 나도 ncs하다가 내얘기하느<unk>웃기다ㅋ#@기타#웃기다ㅋㅋㅋ
P02: 별거아니겠지해쏜디
P01: 십어버림
```

Experiments

Dialogue Summarization

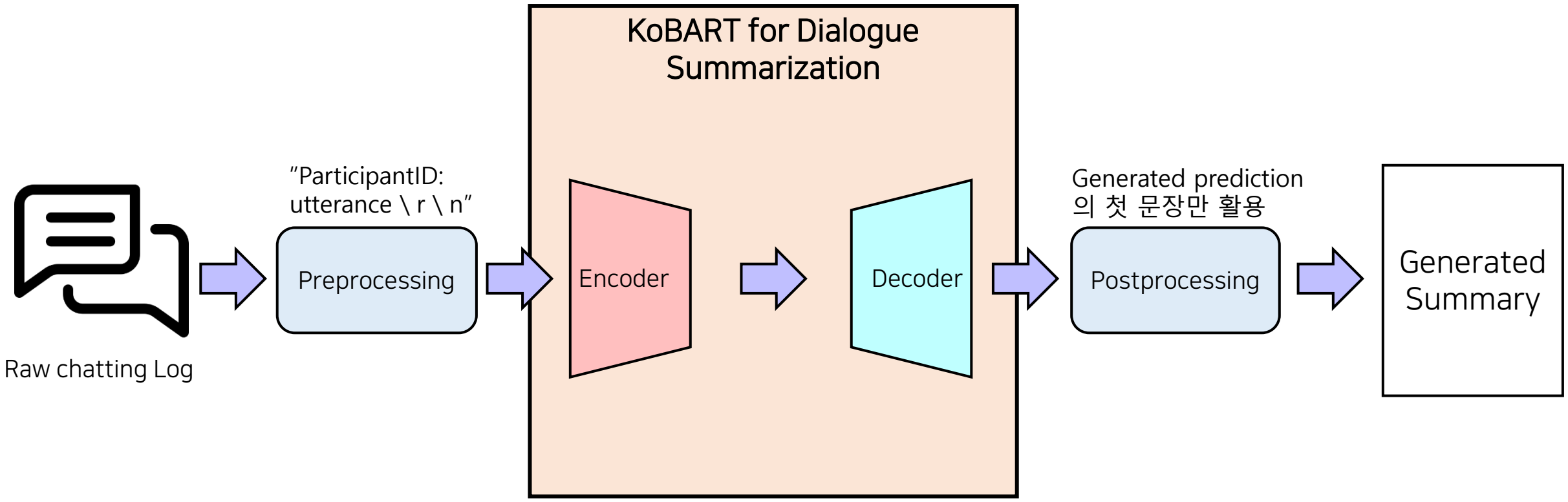
- Hyperparameter Tuning

- **Change Input Format** (Add Name)
  - {ParticipantIDs} + Dialogue



Architecture

Dialogue Summarization



Experiments

Dialogue Summarization

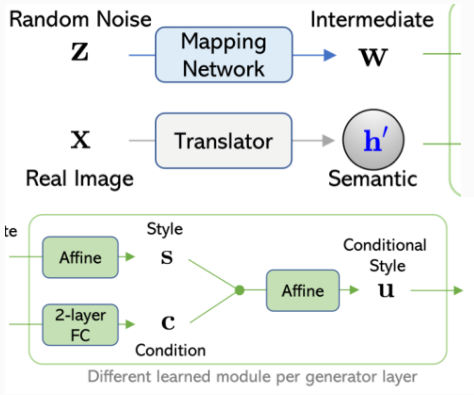
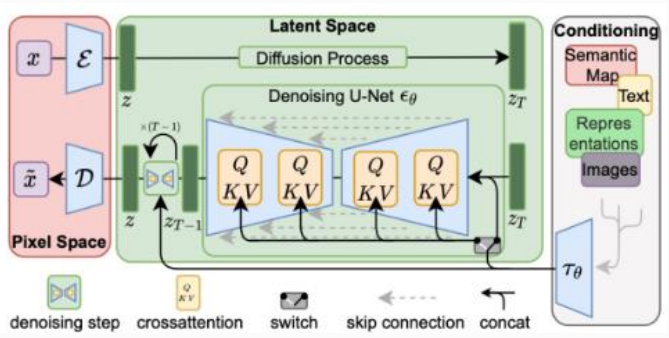
- Ablation Study of Decoding Methods
  - Top_k: 확률 상위 50개만
  - Top_p: 확률 0.95 이상만
  - No_repeat_ngram: trigram은 최대 1회 반복
  - Temperature: 모델의 randomness 약하게

```
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	1. Model Size ↓ (75M) → 속도 ↑ 2. Fine-tuning 하기 좋음 → 다양한 pretrained 모델 제공 3. Language-Free training으로 이미지로만 학습 가능	1. 4억 개의 image-text 쌍을 학습한 모델 → 높은 성능 기대 2. GAN을 앞선 높은 image 합성 성능	1. Zero-shot 평가에도 높은 성능을 가진 DALL-E의 구조를 가져옴 2. V100에서 학습 가능
모델 구조			DALL-E의 축소

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

Model Search

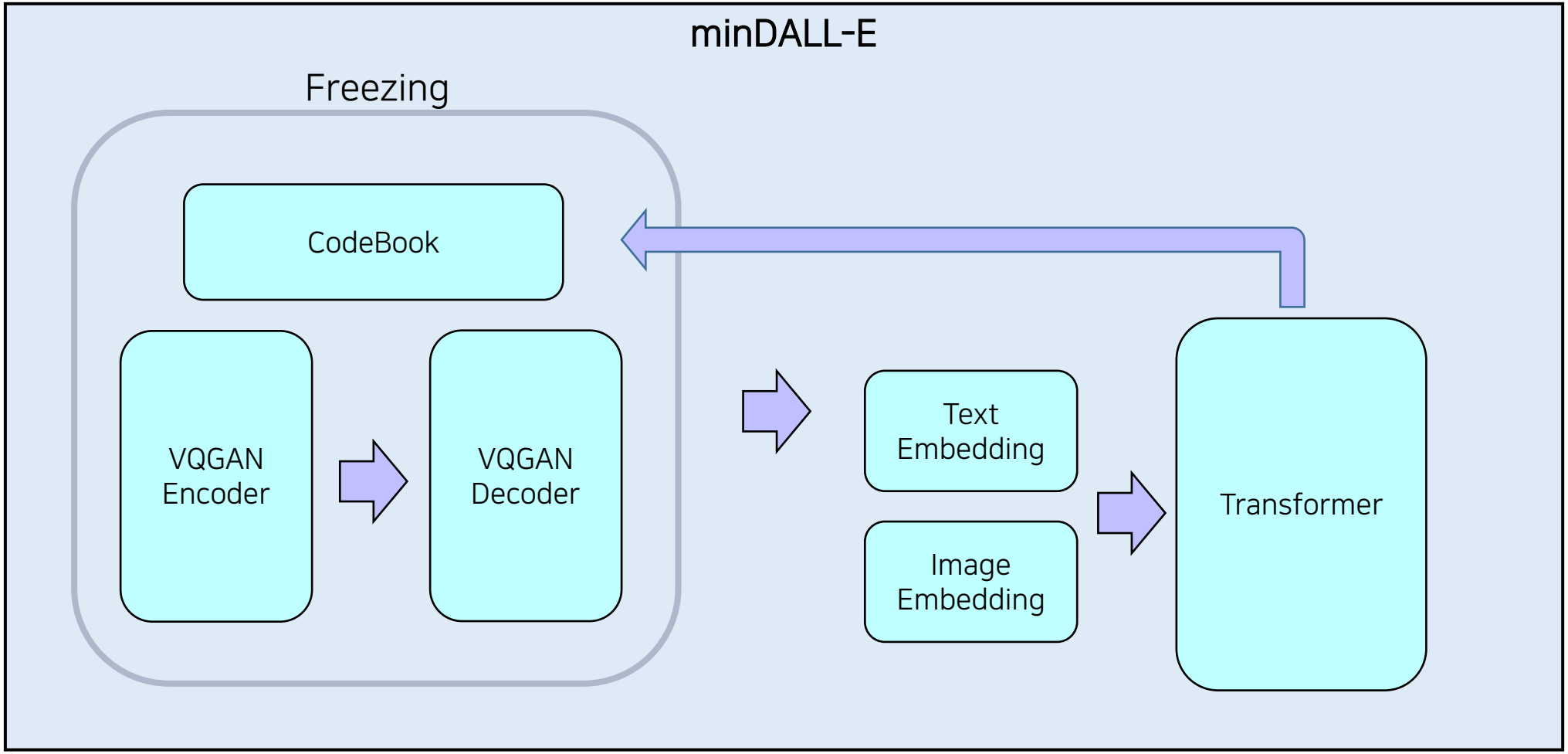
Text-to-Image

- **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

Architecture

Text-to-Image



Experiments

Text-to-Image

- Image GPT였던 stage2를 Transformer1D로 변경
  - Image GPT 는 클래스와 이미지를 학습
  - 텍스트와 이미지 임베딩을 concat하여 Input으로 취하는 Transformer1D 사용

```
# sos token embedding
if self.use_cls_cond:
    self.sos = nn.Embedding(hparams.n_classes, hparams.embed_dim)
else:
    self.sos = nn.Parameter(torch.randn(1, 1, hparams.embed_dim))

# input embedding
self.tok_emb_img = nn.Embedding(vocab_size_img, hparams.embed_dim)
self.pos_emb_img = nn.Embedding(hparams.ctx_len_img, hparams.embed_dim)

self.drop = nn.Dropout(hparams.embd_pdrop)
```

< stage2 – image GPT>

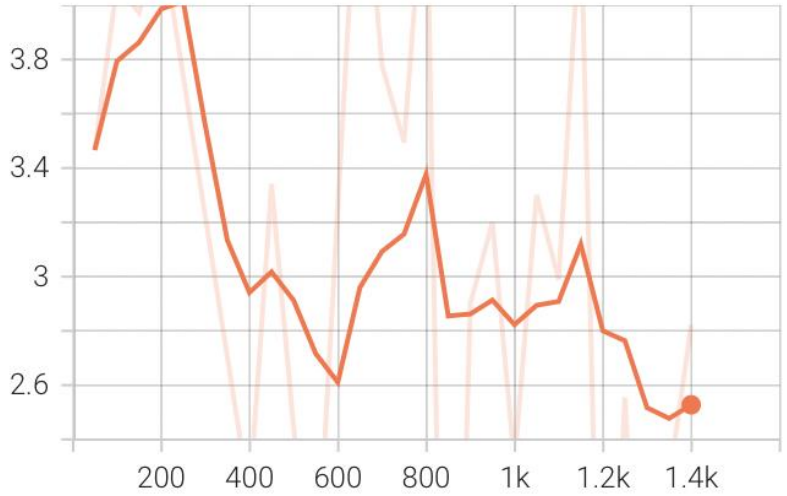
```
# input embedding for image and text
self.tok_emb_img = nn.Embedding(vocab_size_img, hparams.embed_dim)
self.tok_emb_txt = nn.Embedding(vocab_size_txt, hparams.embed_dim)

self.pos_emb_img = nn.Embedding(hparams.ctx_len_img, hparams.embed_dim)
self.pos_emb_txt = nn.Embedding(hparams.ctx_len_txt, hparams.embed_dim)

self.drop = nn.Dropout(hparams.embd_pdrop)
```

< stage2 - Transformer1D>

Experiments

Text-to-Image

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



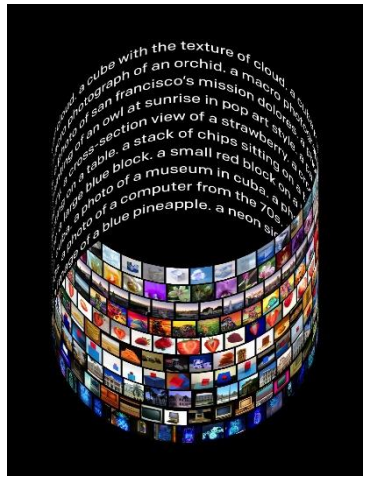
Tensor board CE loss graph

- Early stopping

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

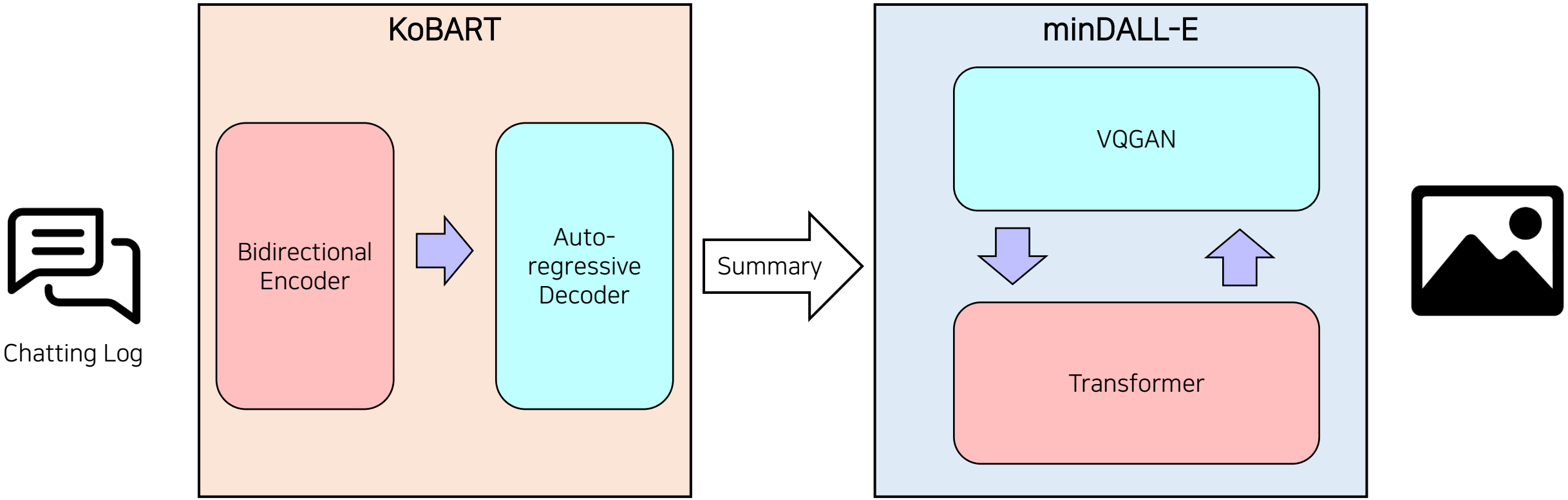
Architecture

KoBART



minDALL-E

Architecture

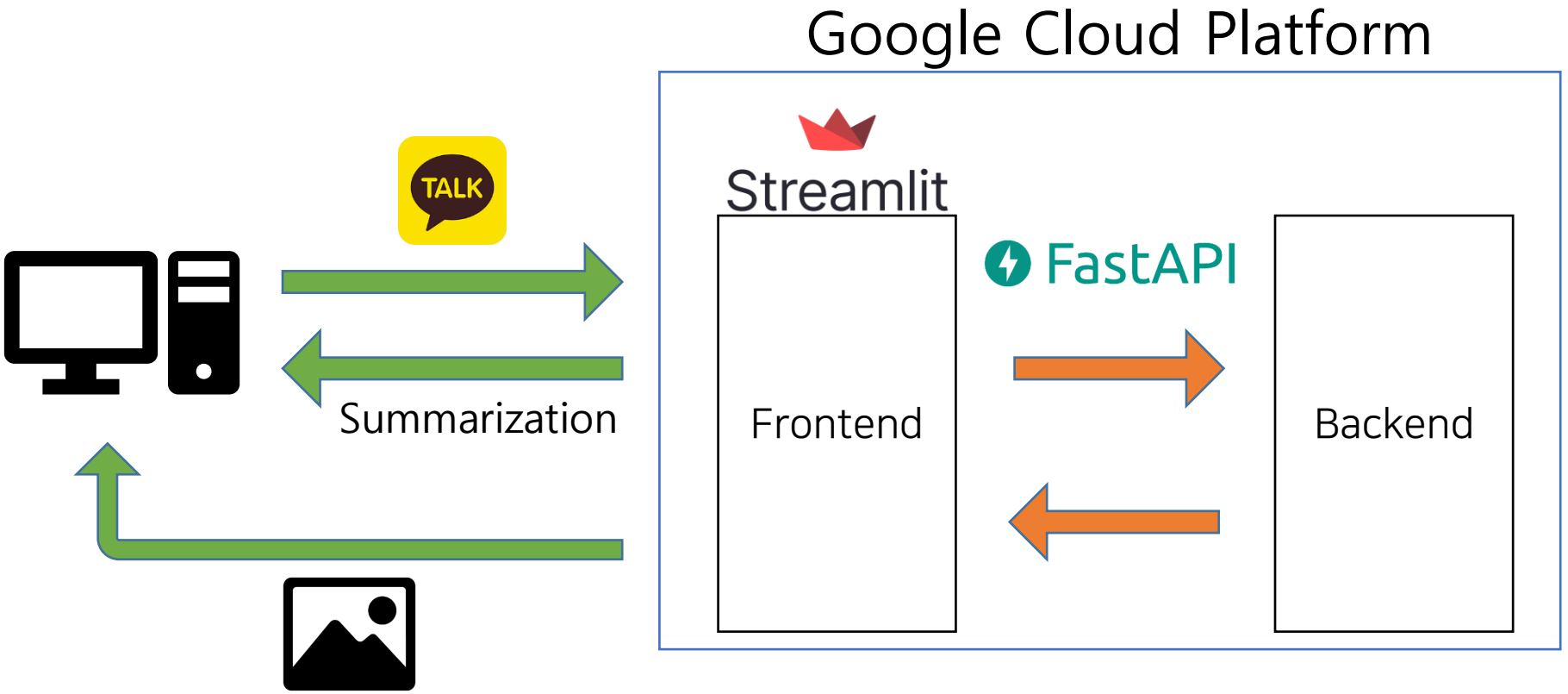


Serving

Cloud

VS

Service Architecture



Result

영상자리

Future Works

- Translation 단계 생략
- 모델 경량화
- Summarization이 복잡할 수록 성능 하락
- ROUGE, RDASS나 Cosine-similarity보다 더 설득력있는 metric
- 모델의 재학습
- 완전한 개인정보 전처리위한 룰
- 추상적 개념의 이미지화

Appendix

Input text preprocessing

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 preprocessing

Text-to-Image

- 12가지 문장 변형 방법에 대한 모델 성능 비교

  BEST : A painting of + extracting POS tag/removing stopwords + 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.	0.210205
① + ⑤ + ⑦	went home right away today n't rest well	0.211792
① + ⑤ + ⑧	went, home, right, away, today, n't, rest, well	0.222656
① + ⑥ + ⑦	went home right away today , could n't rest well .	0.199951
① + ⑥ + ⑧	went, home, right, away, today, ,, could, n't, rest, well, .	0.206420
② + ④ + ⑦	A painting of I went home right away today, but I couldn't rest well.	0.247558
② + ⑤ + ⑦	A painting of went home right away today n't rest well	0.252197
② + ⑤ + ⑧	A painting of went, home, right, away, today, n't, rest, well	0.271972
② + ⑥ + ⑦	A painting of went home right away today , could n't rest well .	0.257080
② + ⑥ + ⑧	A painting of went, home, right, away, today, ,, could, n't, rest, well, .	0.272705
③ + ⑤ + ⑨	A painting = went + home + right + away + today + n't + rest + well	0.253662
③ + ⑥ + ⑨	A painting = went + home + right + away + today + , + could + n't + rest + well + .	0.252197

Lesson Learn

- AI 서비스 개발에서 문제정의와 그에 맞는 양질의 데이터셋 수집 및 구축의 중요성
- Subproblem의 해결보다 product의 흐름에 더 집중해야한다.
- 논문의 목적과 활용 가능성을 고려하여 아이디어를 얻어오는 것에 대한 중요성의 인지
- 주제 선정 시 실현 가능성 확인이 우선되어야 함
- 정량적 평가 지표 선택 및 실험 기록의 중요성
- Clean Code & Modularization의 중요성
- 실행 속도 향상을 위한 모델 경량화
- 어려운 문제 해결 시 세분화의 필요

Contribution

임동진

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

정재윤

- 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

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

- Cross-Entropy Loss가 FID score를 낮춘다고 판단할 수 있는 이유는 무엇인가요?
  - FID: 정답이미지의 분포와 생성된 이미지의 분포와의 거리 계산
  - Cross-Entropy는 모델 예측 확률값과 실제값의 차이를 구함
  - Cross-Entropy Loss를 줄이는 방향으로 쉽게 FID score를 낮출 수 있음

예상 Q&A

- RDASS가 아닌 Sentence BERT를 최종적인 Metric으로 선택한 이유가 있을까요?
  - Dialogue – Golden summary의 낮은 cosine similarity
  - Golden summary-Prediction의 cosine similarity만을 활용

- 생성 요약 문장 평가 기준을 0.6으로 골랐던데, 혹시 기준이 있을까요?
  - “한 비행기가 착륙하고 있다. 애니메이션화된 비행기 하나가 착륙하고 있다.” (korSTS label 2.8)
  - KorSTS 기준 3.0을 넘기면 매우 유사한 문장이라고 판단
  - Cosine similarity는 -1~1사이의 값 도출해서 5로 나눈 0.6을 기준으로 삼음

- 생성 문장의 가장 첫 문장만 선택한 이유가 있을까요?
  - 생성 결과에서 첫 문장 이후 ㅋㅋㅋ등의 무의미한 단어나 보완 설명이 생성
  - Text-to-Image의 입력은 한 문장만 들어가는 것이 좋음
  - 가장 핵심 요약을 담고 있는 첫 문장만을 활용하기로 결정

RDASS

Dialogue Summarization

- Model
  - Sentence-BERT

- Pre-training
  - NLI, STS 데이터셋을 이용해 학습

- Fine-tuning
  - <$\boldsymbol{H_p}$, $\boldsymbol{V_p^r}$, $\boldsymbol{V_n^r}$> 각각의 벡터로 triplet 구성
  - d($\boldsymbol{H_p}$, $\boldsymbol{V_p^r}$)와 d($\boldsymbol{H_p}$, $\boldsymbol{V_n^r}$) 계산
  - d($\boldsymbol{H_p}$, $\boldsymbol{V_p^r}$)는 0으로 수렴하도록, d($\boldsymbol{H_p}$, $\boldsymbol{V_n^r}$)는 1에 수렴하도록 학습

$$H_p : \text{Generated Summary vector}$$
$$v_p^r : \text{Golden Summary vector}$$
$$v_n^r : \text{Incorrect Summary vector}$$

RDASS

Dialogue Summarization

- Sentence embedding을 활용한 요약 모델 평가 지표

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

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

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

- $\boldsymbol{v_d}$: Document
- $\boldsymbol{v_r}$: Golden Summary
- $\boldsymbol{v_p}$: Generated Summary

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/