

제목 짓기 어려워서 만든 제목 생성 모델

#비전공자의반란 #생성요약 #LongBART #BigBART

NLP-02
(메타몽팀)

목차

- 1 Introduction
- 2 학습 파이프라인
- 3 성능 개선
- 4 경량화
- 5 더 나아가기
- 6 회고

1. Introduction

#team 소개
#프로젝트 협업
#주제_선택_배경
#시스템 아키텍처
#시연영상

Introduction

Team 소개



데이터 수집
모델 개발 및 튜닝
Text infilling
Teacher forcing scheduler

고창용_T2008
#PM(Main)+코드검사 담당



R-drop
시각화
Serving

이예빈_T2165
#PM(Sub) 담당



모델 개발 및 튜닝
Serving 프로토타입 제작

이기성_T2149
#코드검사 담당



Baseline 구축
데이터 수집
모델 개발 및 튜닝

정유석_T2204
#리서치 담당



모델 개발
모델 경량화
ElasticSearch

안명철_T2126
#리서치 담당



Baseline 구축
한국어 평가지표 개발
시각화
데이터 후처리

박범진_T2081
#기록 담당



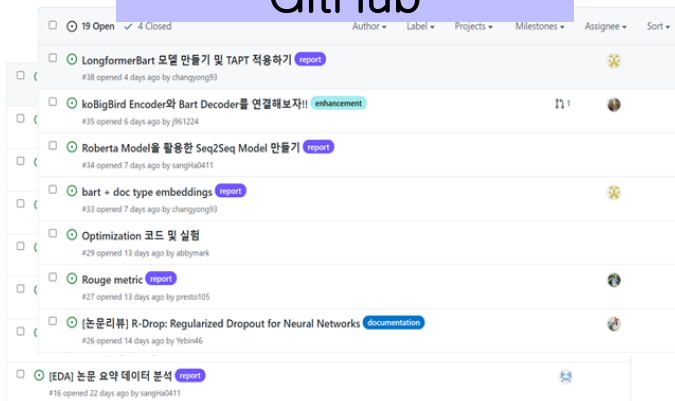
데이터 전처리
모델 개발
Text infilling
Teacher forcing scheduler

박상하_T2084
#기록 담당

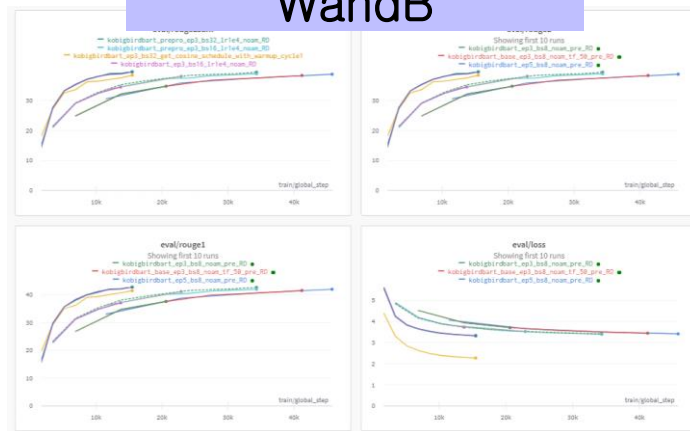
Introduction

프로젝트 협업

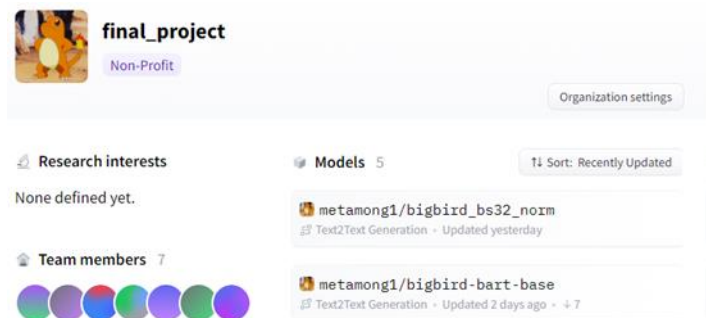
GitHub



WandB



HuggingFace Hub



Google Sheet

Task	Assigned to	Progress	Start Date	End Date
Data	모두	100%	2021-11-16	2021-11-22
프로젝트 계획 및 관리 정리	모두	100%	2021-11-22	2021-11-25
데이터 수집 EDA	그림용 작성자, 박정민, 장유석	100%	2021-11-25	2021-11-27
데이터 전처리 및 Testset 분리	박정민	100%	2021-11-28	2021-11-29
모델 자료 공유	박정민	100%	2021-12-14	2021-12-15
Baseline			2021-11-30	2021-12-23
train, dataloader 구축	박정민, 장유석	100%	2021-11-30	2021-12-22
model argument 구축	박정민, 장유석	100%	2021-12-02	2021-12-23
predict 및 코드 리팩	박정민, 장유석	100%	2021-12-03	2021-12-26
Baseline 코드 리팩	그림용, 이기성	100%	2021-12-04	2021-12-25
Baseline Train 종료 및 업로드	장유석	100%	2021-12-05	2021-12-23
Model			2021-11-30	2021-12-24
longformer 논문 등 source code 리뷰	그림용	100%	2021-11-30	2021-12-23
LongBART	그림용, 박정민	100%	2021-12-03	2021-12-16
BigBirdSeq2Seq	이기성, 장유석	100%	2021-12-05	2021-12-10
bart + R-drop	이예민	100%	2021-12-03	2021-12-26
BigBART 모델 학습	장유석	100%	2021-12-06	2021-12-24
BigBART	이기성, 장유석	100%	2021-12-07	2021-12-14
bart + Doc_type_embedding	그림용, 박정민	100%	2021-12-10	2021-12-11
robusta-wc2Vec	그림용, 박정민	100%	2021-12-10	2021-12-11
distilBertEncoder + distilBertDecoder	장유석	100%	2021-12-11	2021-12-17
TAPT(Text Infilling, LongformerBart 7종)	그림용, 박정민, 장유석, 이기성	100%	2021-12-14	2021-12-15
Optimizing			2021-12-08	2021-12-16
performance eval 구축	장유석	100%	2021-12-04	2021-12-26
Quantization	장유석	100%	2021-12-04	2021-12-26
Structured Pruning	장유석	100%	2021-12-05	2021-12-27
Knowledge Distillation	장유석	100%	2021-12-06	2021-12-26
Try Distillation	장유석	100%	2021-12-10	2021-12-15
Serving			2021-12-05	2021-12-26
Visualization			2021-12-05	2021-12-26
시각화 - attention text heatmap	박정민, 이예민	100%	2021-12-05	2021-12-14
시각화 - network graph	박정민, 이예민	100%	2021-12-16	2021-12-20
etc			2021-12-16	2021-12-24
자료 준비	장유석	100%	2021-12-18	2021-12-24
elastic search	장유석	100%	2021-12-21	2021-12-23

다양한 툴을 활용하여 협업

논문, 보고서... 제목 짓기 너무 어렵지 않나요?



.....?

Introduction

주제 선정 배경

“생성 요약을 활용해 제목을 만들어주는 모델 개발”

- 논문, 보고서… 제목 짓기 너무 어렵지 않나요?

영어 제목 생성기는 있는데 한국어는 없네?

- 우리가 만들자!

Bidirectional Encoder Representations from Transformers: A New Language Processing Model

ABSTRACT

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

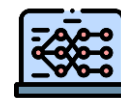
(RE)GENERATE TITLE

Introduction

주제 선정 배경

1

다양한 시도를 통해
우리의 성장을
가장 잘 보여줄 수 있는가?



부스트캠프에서 배운 내용들을 모두 녹아낼 수 있는 프로젝트
배운 내용보다 더 나아갈 수 있는 프로젝트

2

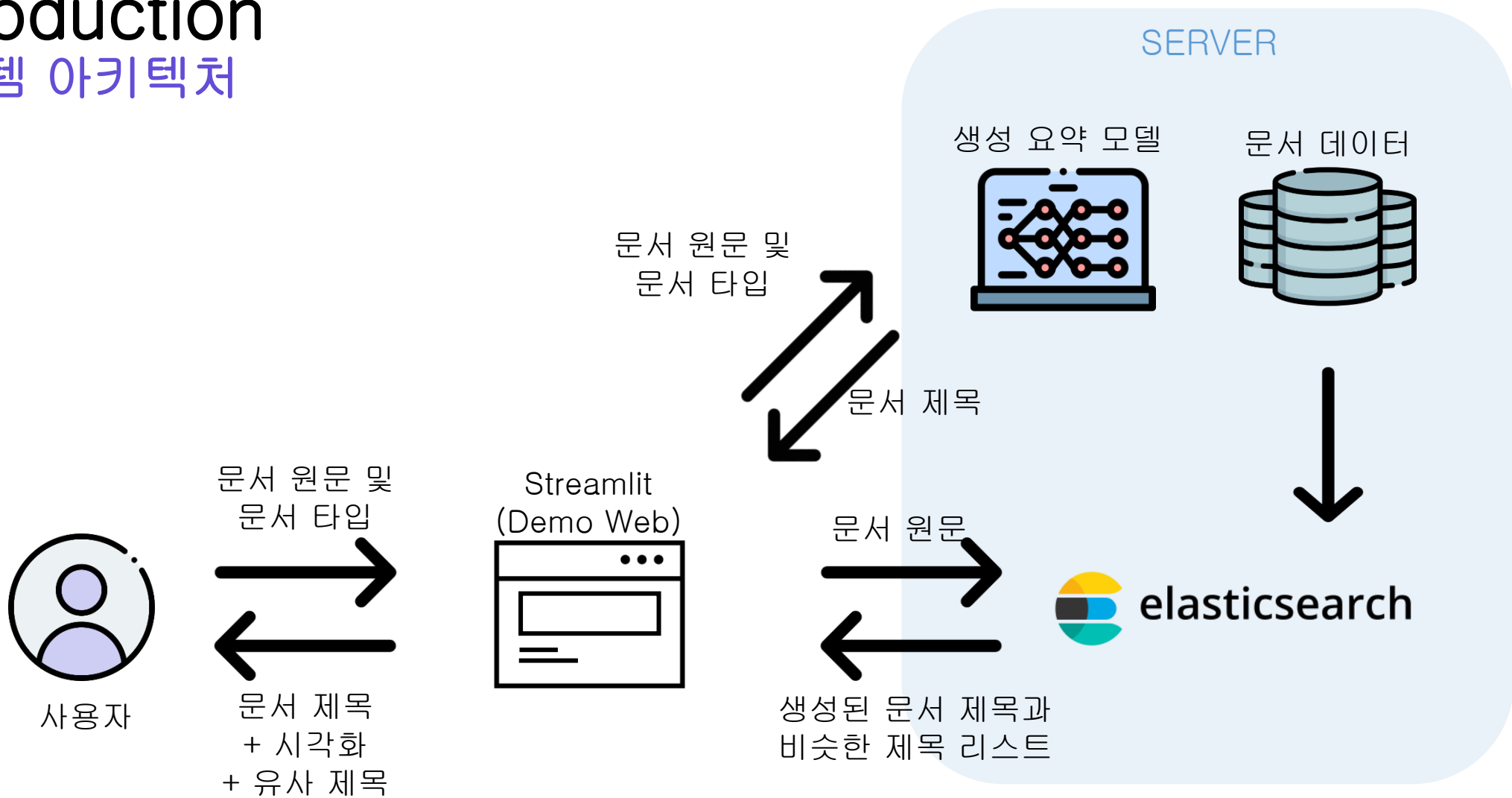
주어진 환경에서
완성할 수 있는가?



개발자로서 주어진 리소스 내에서
최선의 결과물을 만들 수 있는 프로젝트

Introduction

시스템 아키텍처



Introduction

시연 영상

✕

문서 타입을 선택해주세요!

해당없음 ▾

문서 내용을 입력해주세요!

I

Welcome in text generation website

좌측에 본문 내용을 넣어주세요!

Visualization!

Made with Streamlit

2. 학습 파이프라인

#Model Pipeline
#EDA
#Pre/Post-Processing
#Model
#평가 방법

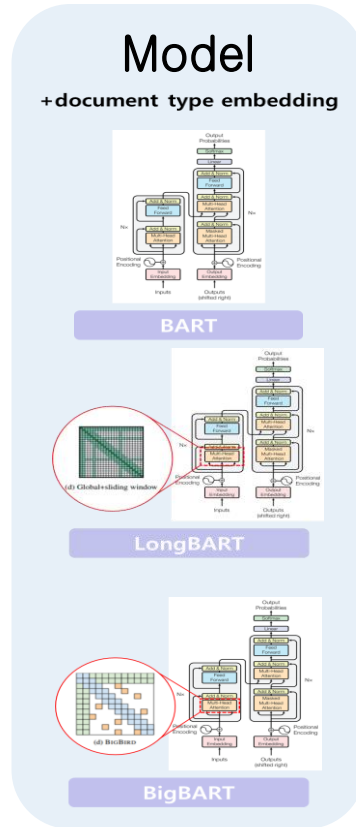
학습 파이프라인

Model Pipeline



Train 50%

Preprocessing

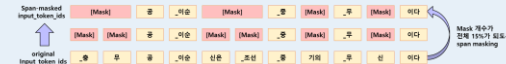


Yes

doc type
embedding

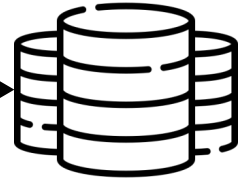
No

Pretraining / TAPT



Finetuning

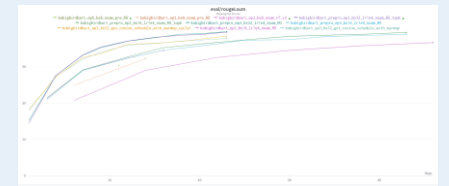
Noam Scheduler with warmup
Linear Scheduler with Warmup
Regularized Dropout
Teacher Forcing Scheduler
Hyperparameter
Rouge Score specialized in Korean.



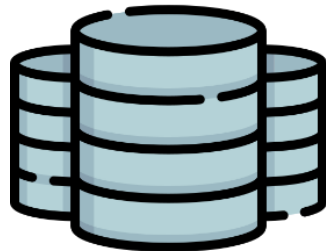
Eval 50%



Selecting Models

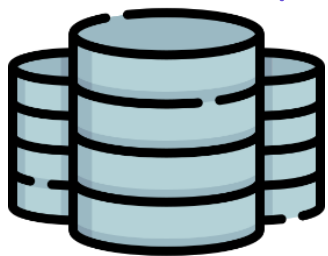


Train 50%로 학습한 모델 중 성능이 좋은 후보 모델을 전체 데이터로 재학습



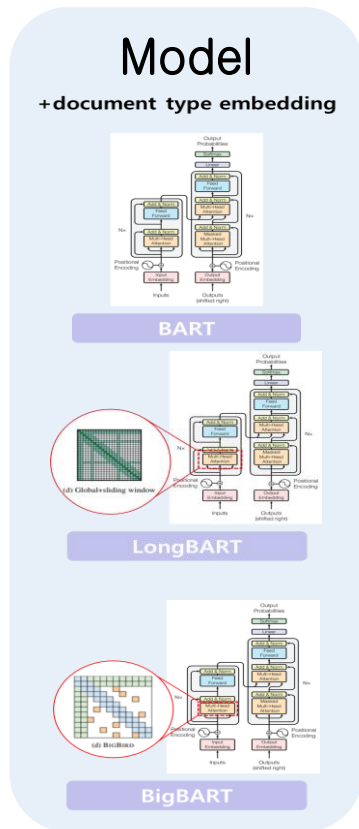
Train 100%

Model Pipeline



Train 100%

Preprocessing

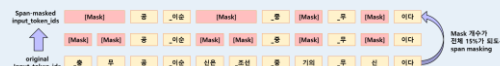


Yes

**doc type
embedding**

No

Pretraining / TAPT

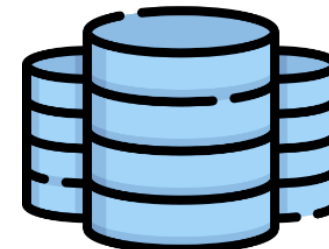


Finetuning

Noam Scheduler with warmup
Linear Scheduler with Warmup
Regularized Dropout
Teacher Forcing Scheduler
Hyperparameter
Rouge Score specialized in Korean.



Eval 100%



Test 100%



최종 모델 선정

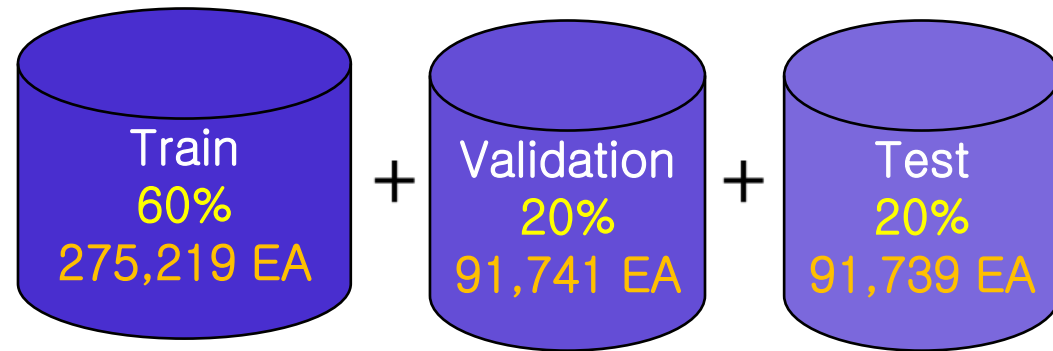


Optimization (optional)

Quantization
Knowledge Distillation

학습 파이프라인

데이터 분석(EDA)



데이터 본문 문체 및 특성 분석 (max length : 1887)

1. 논문 → 목적, 증거

본 연구의 목적은 청소년의 스마트폰 중독에 대한 학교부적응의 영향과 이에 대한 우울의 매개효과를 검증하는데 있다. 첫째, ~~~~. 둘째, ~~~~.

2. 뉴스 → 발언 인용

"재판관 관심사항과 논리를 보강한 보충서면을 준비하고 대법원 현장검증 시 대응에 철저를 기해 당진땅을 꼭 수호하자"고 강조했다.

3. 사설 잡지 → 주장, 의견 제시

핵담판이 난기류에 빠진 만큼 우리 정부는 북·미의 움직임을 예의 주시하면서 한·미 간 대북 공조를 강화해야 할 것이다.

데이터 제목 문체 및 특성 분석 (max length : 128)

1. 논문 → 연구, 분석

불변시장점유율모형과 BCG 매트릭스를 이용한 향만의 철강 수출변동 분석

2. 뉴스 → 지역, 단체

GS홈쇼핑, '벤처투자'에 사활 걸었지만... '지분법손실' 확대

3. 사설 잡지 → 주장, 의견 제시

개각으로 들쭉인 공직사회, 이젠 개혁에 전념하라
한국판 라인-야후재팬 빅뱅' 막는 지주사 규제, 완화해야

학습 파이프라인

Post/PreProcessing

Pre-processing

Model이 이해하지 못하는 문자를 최대한 제거

- 괄호 안에 있는 한자, 영어로 된 부연 설명 제거
- 한자, 한글, 영어 및 구두점을 제외한 특수 문자 제거

도애 홍석모의 금강산 유기, 『간
관록』 일고 (陶厓 洪錫謨의 금강
산 유기, 『艮觀錄』 一考)



도애 홍석모의 금강산 유기,
간관록 일고

Post-processing

마무리가 제대로 지어지지 않은 부분 제거

- 쌍을 이루지 못한 문장기호와 마무리 되지 못한 내용 제거: (), [], {}, ‘ ’, “ ” 등
- 개행문자, 문장 끝 특수문자 제거

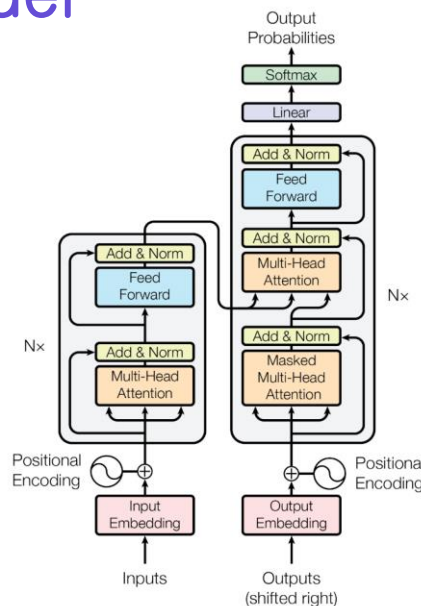
물도 껌질이 있는 걸까? - [물에
관한



물도 껌질이 있는 걸까?

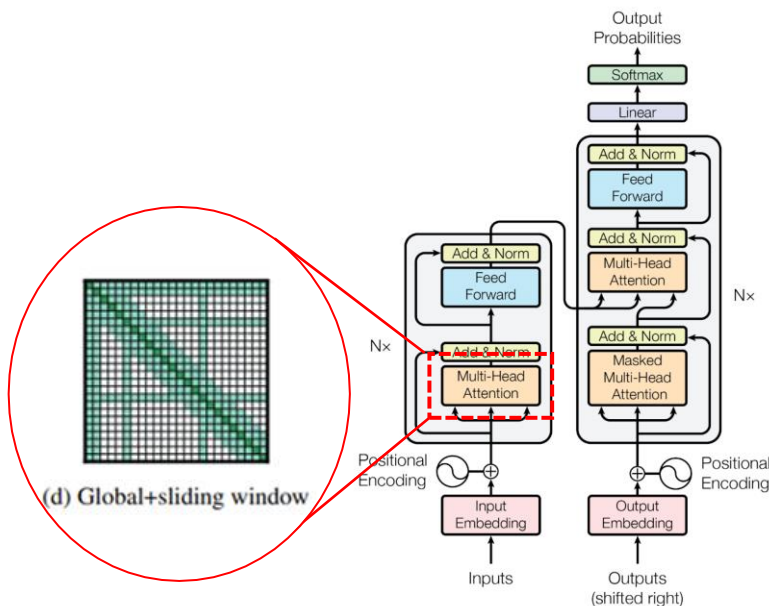
학습 파이프라인

Model



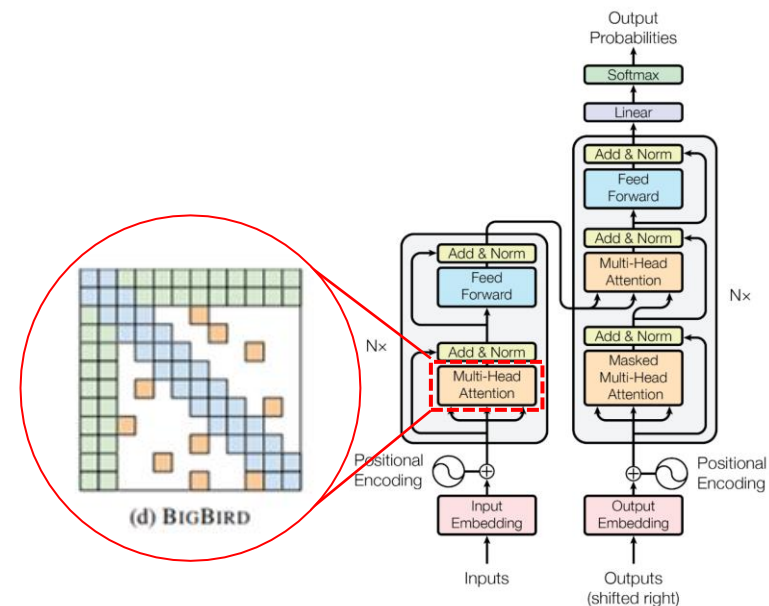
BART

CNN/Daily Mail Dataset
Vanilla 모델로 8위



LongBART

Longformer + BART



BigBART

BigBird + BART

8	BART	44.16	21.28	40.90
---	------	-------	-------	-------

9	ERNIE-GENLARGE	44.02	21.17	41.26
---	----------------	-------	-------	-------

10	T5	43.52	21.55	40.69
----	----	-------	-------	-------

학습 파이프라인

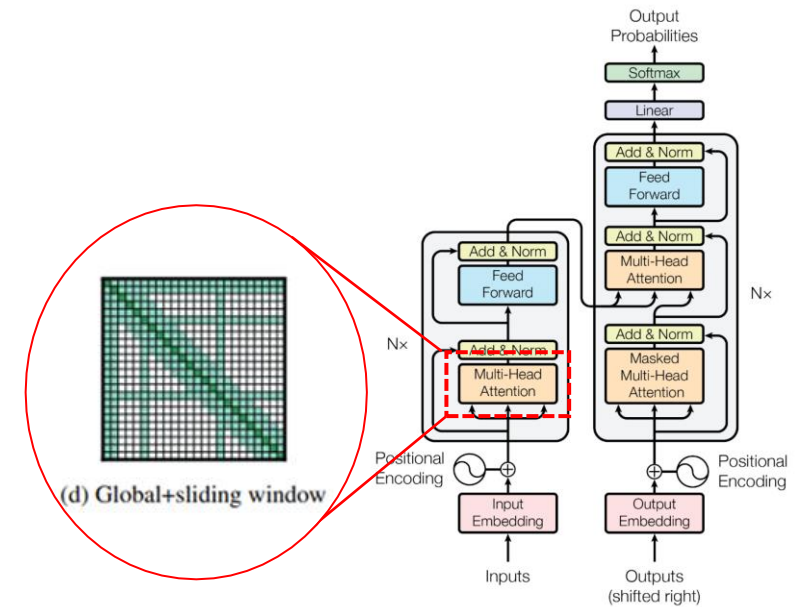
Model – LongBART (Longformer + BART)

· 모델 선정 이유

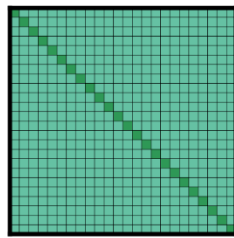
- Time Complexity(Self Attention)
 - Bart $\rightarrow O(N^2)$, Longformer $\rightarrow O(N)$
- Long Sequence를 효율적으로 학습할 수 있는 모델

· 모델 구현 과정 중 제한 사항

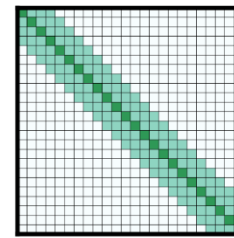
- Longformer 모델은 한국어 데이터로 pretrained 된 모델이 없음
- Pretraining 과정이 필요



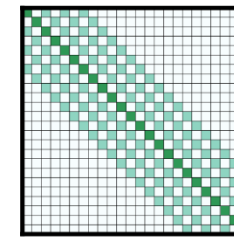
Bart



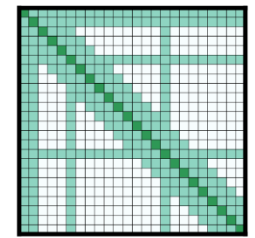
(a) Full n^2 attention



(b) Sliding window attention



(c) Dilated sliding window



(d) Global+sliding window

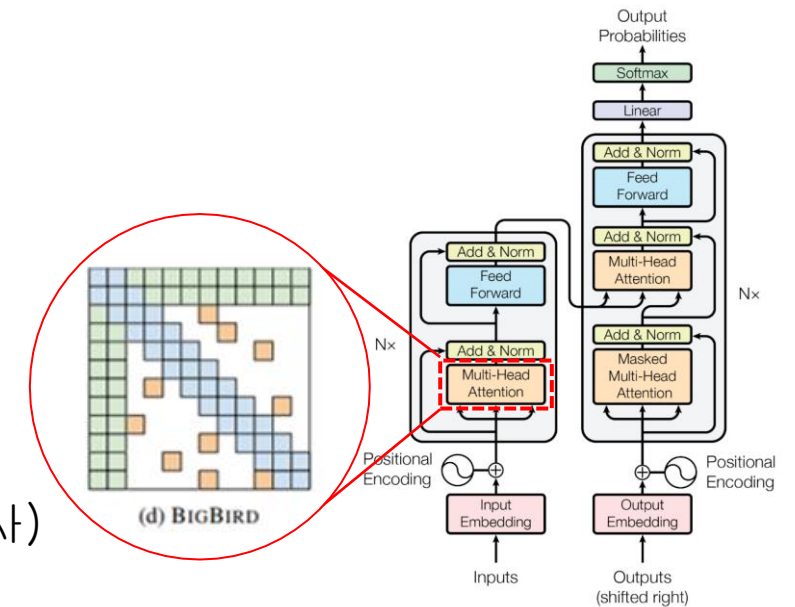
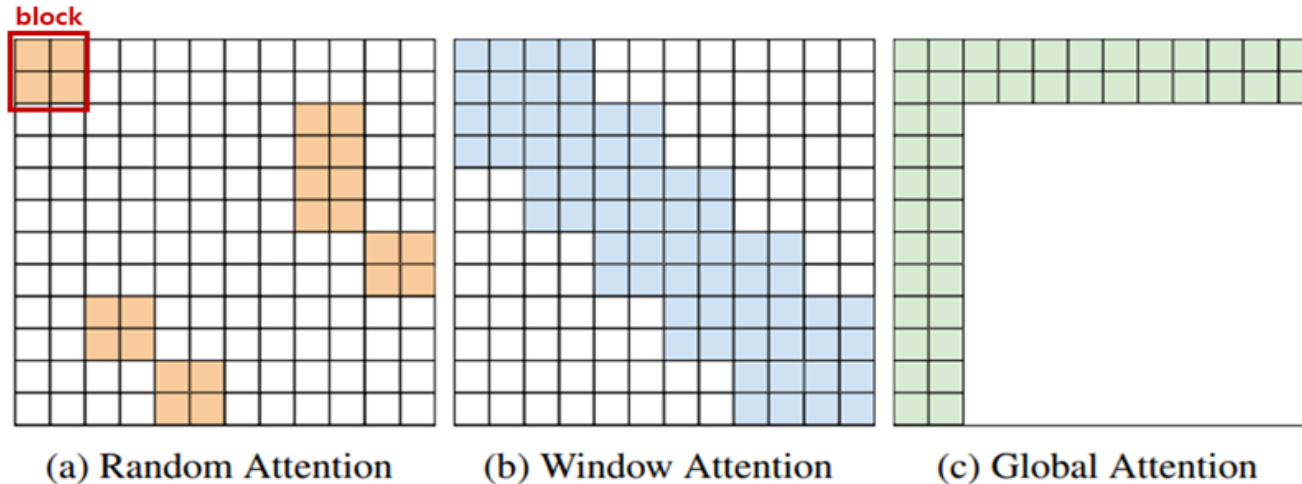
Longformer

학습 파이프라인

Model – BigBART (Bigbird + BART)

모델 선정 이유

- 한국어 사전 학습 모델이 존재
- Sparse Attention
 - Longformer → sliding, global attention
 - BigBird → sliding, global, **random attention** (full attention에 근사)
- Block 단위의 sparse attention 사용 → 연산 최적화



학습 파이프라인

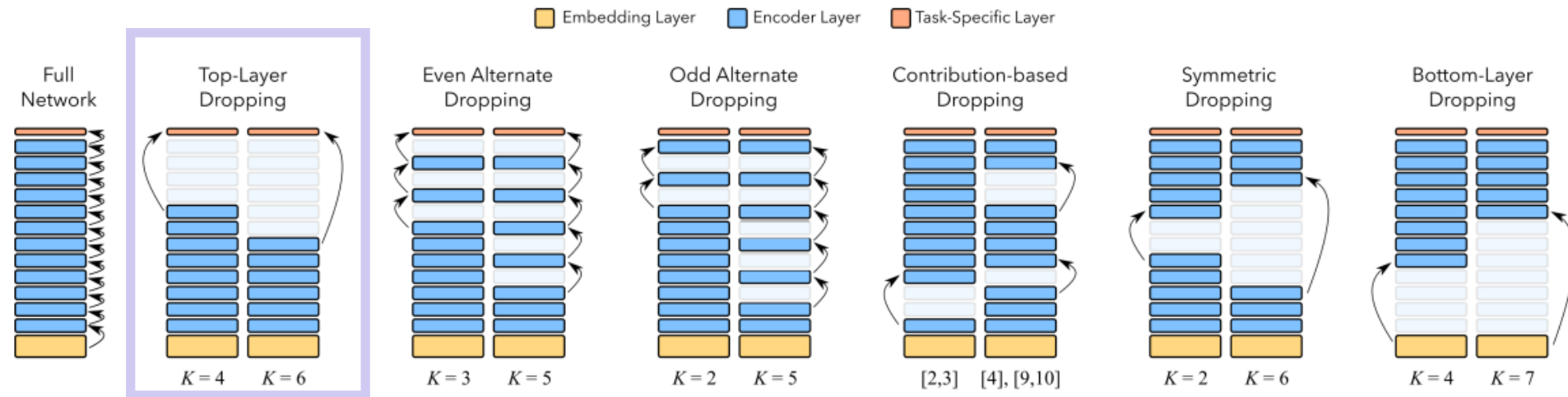
Model – BigBART (Bigbird + BART)

두 사전 학습 모델의 Word embedding 공유

Vocab size of word embedding

- Encoder(BigBird) : 32500 ← shared
- Decoder(Bart) : 30000

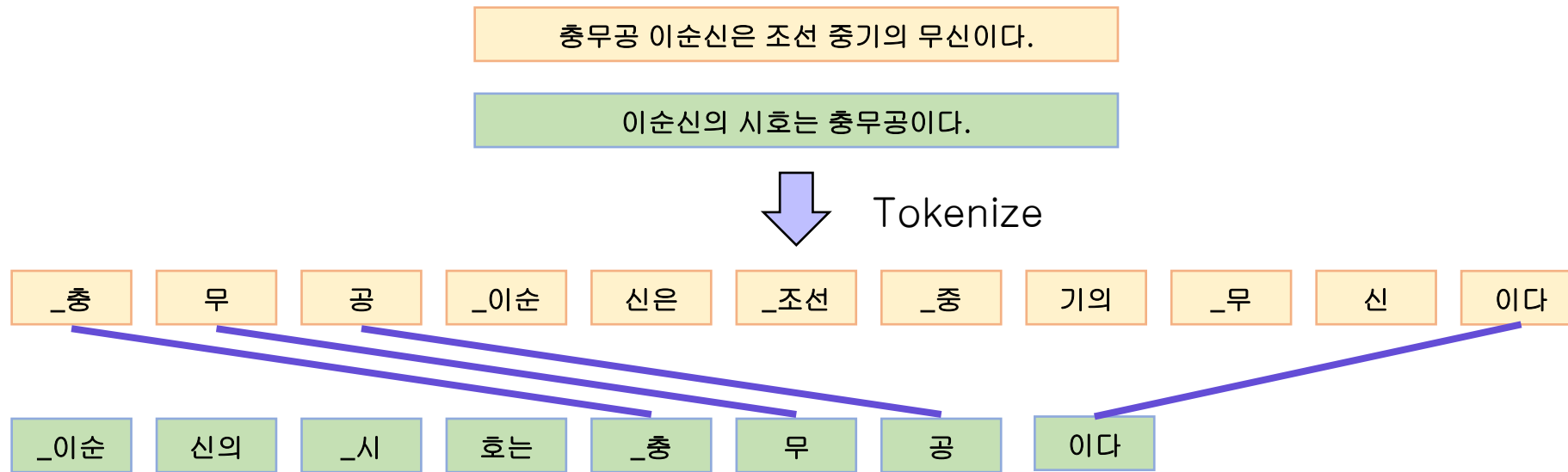
모델 사이즈 축소(파라미터 1억9천개 → 1억5천개)



학습 파이프라인

평가 방법: Rouge-L

Rouge-L: Longest common subsequence(LCS), 단일 문장 비교



Tokenizer의 한계를 인식하고 **Mecab POS**(part of speech) tagging, 명사 등 usable token만 평가에 사용하는 방식으로 업데이트

3. 성능 개선

#Document_type

#Text_infilling

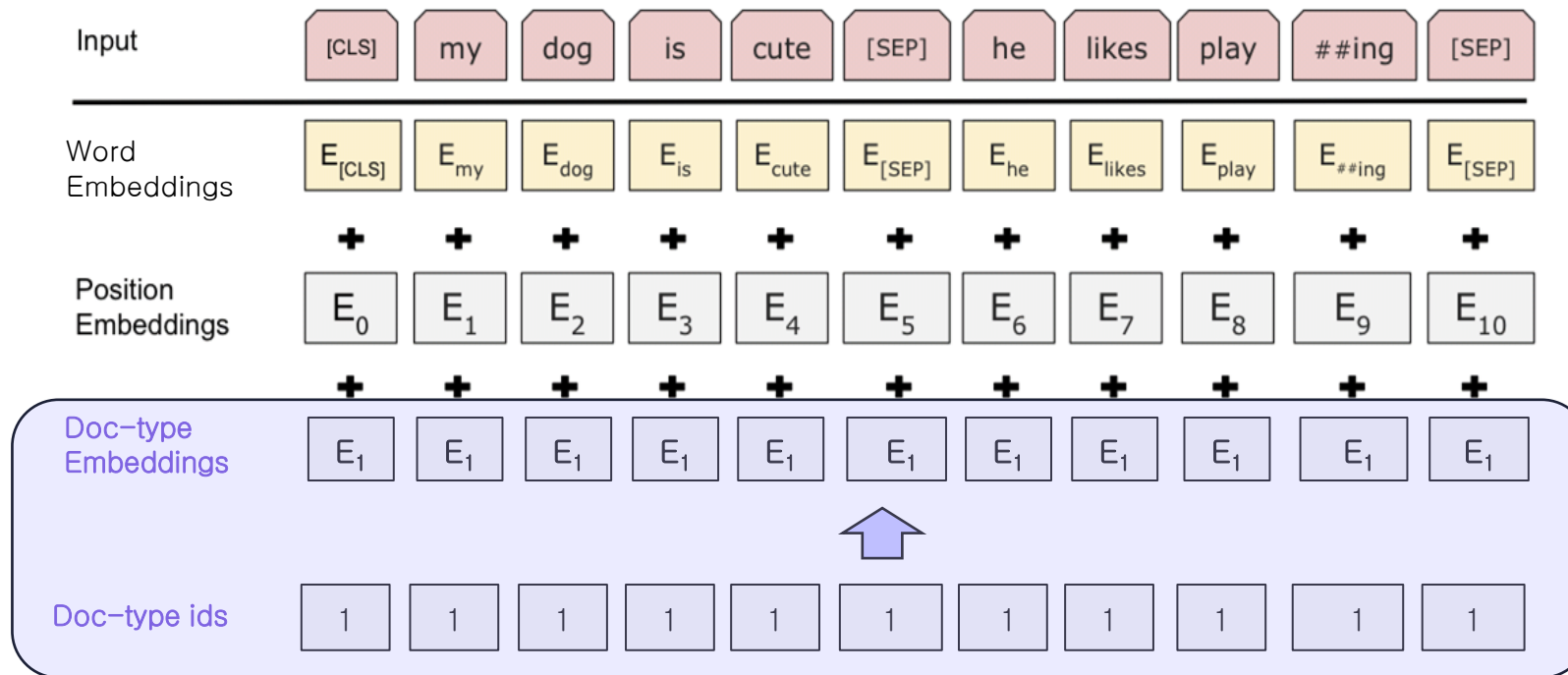
#R-drop

#Teacher_forcing

#결과

성능 개선

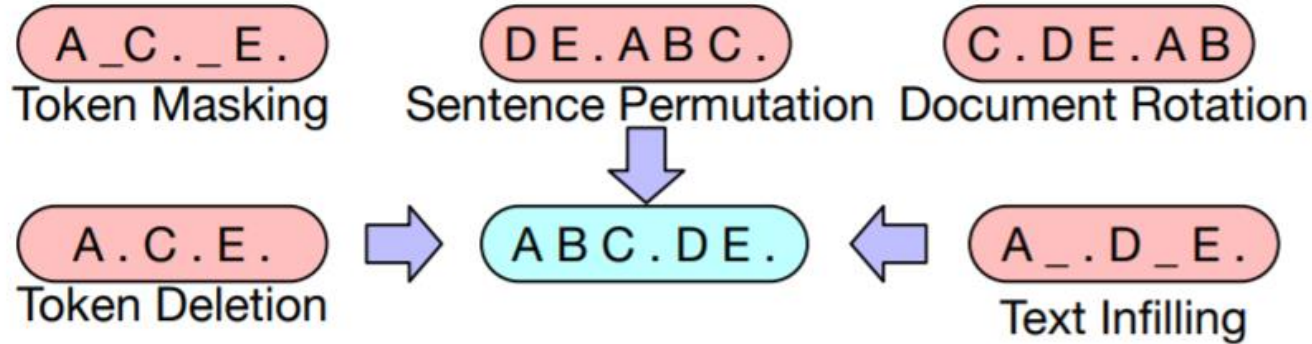
Document Type Embeddings



Document type Embedding + pretraining/Task-adaptive pretraining(Text Infilling)

성능 개선

Text Infilling



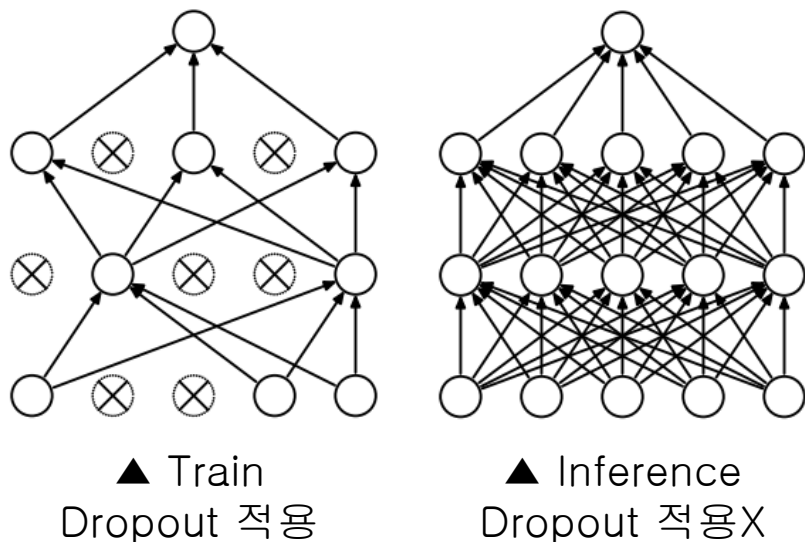
Model	SQuAD 1.1 F1	MNLI Acc	ELI5 PPL	XSum PPL	ConvAI2 PPL	CNN/DM PPL
BERT Base (Devlin et al., 2019)	88.5	84.3	-	-	-	-
Masked Language Model	90.0	83.5	24.77	7.87	12.59	7.06
Masked Seq2seq Language Model	87.0	82.1	23.40	6.80	11.43	6.19
Permuted Language Model	76.7	80.1	21.40	7.00	11.51	6.56
Permuted Language Model	89.1	83.7	24.03	7.69	12.23	6.96
Multitask Masked Language Model	89.2	82.4	23.73	7.50	12.39	6.74
BART Base						
w/ Token Masking	90.4	84.1	25.05	7.08	11.73	6.10
w/ Token Deletion	90.4	84.1	24.61	6.90	11.46	5.87
w/ Text Infilling	90.8	84.0	24.26	6.61	11.05	5.83
w/ Document Rotation	77.2	75.3	53.69	17.14	19.87	10.59
w/ Sentence Shuffling	85.4	81.5	41.87	10.93	16.67	7.89
w/ Text Infilling + Sentence Shuffling	90.8	83.8	24.17	6.62	11.12	5.41



Document type Embedding + pretraining/Task-adaptive pretraining(Text Infilling)

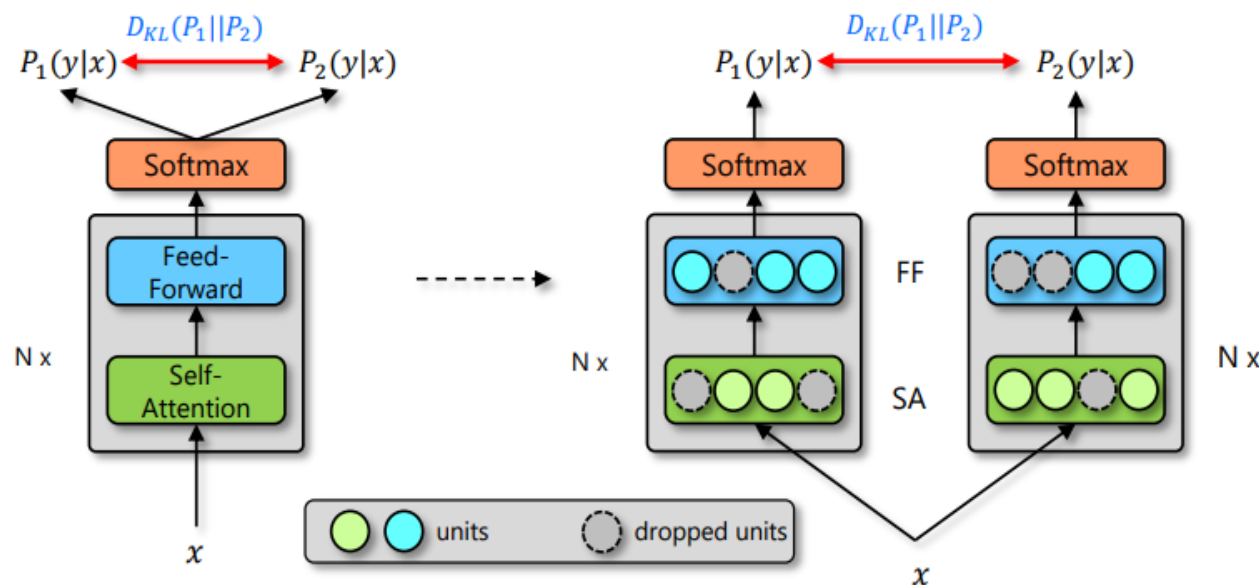
성능 개선

Regularized Dropout(R-Drop)



Rank	Model	ROUGE-1	ROUGE-2	ROUGE-L	Extra Training Data
1	GLM-XXLarge	44.7	21.4	41.4	×
2	BART + R-Drop	44.51	21.58	41.24	×

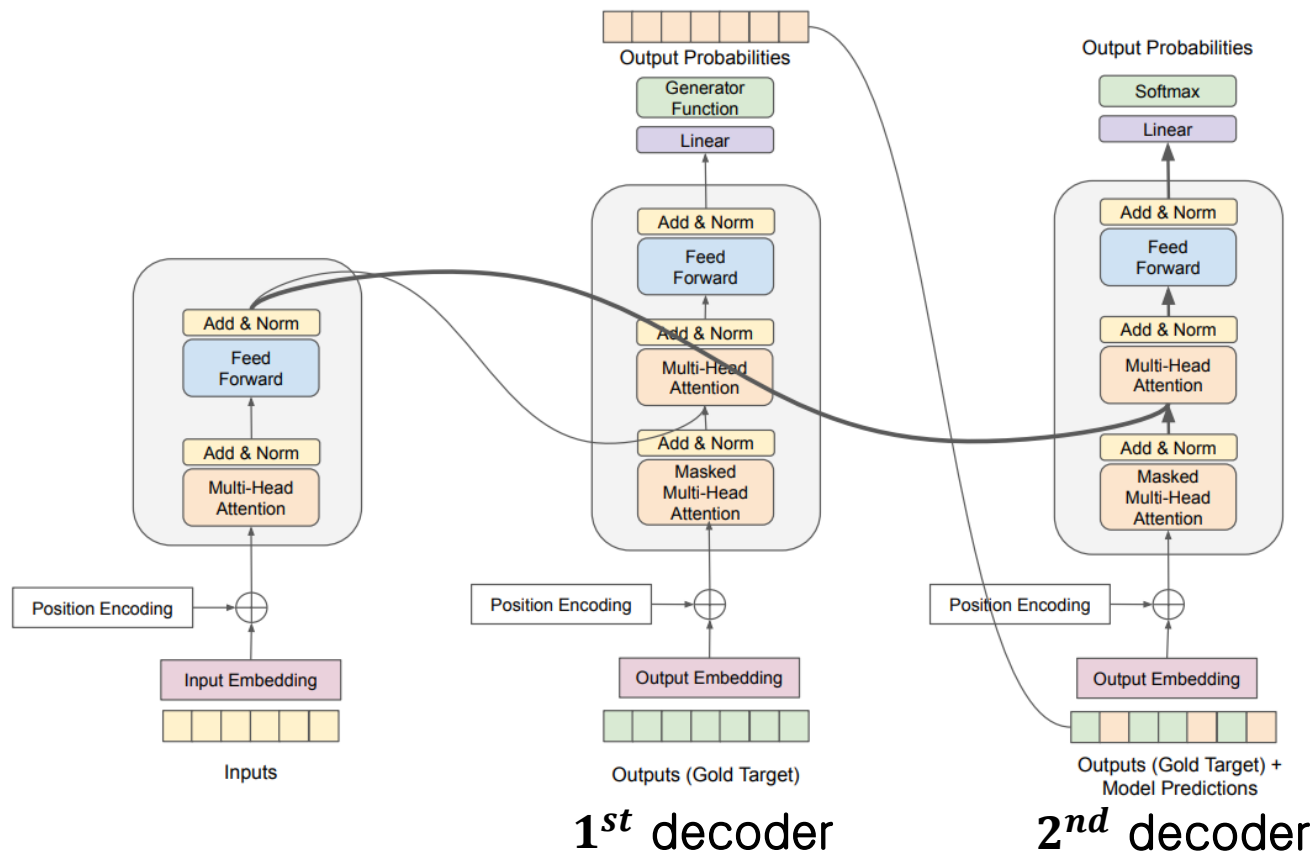
◀ CNN/Daily Mail 2위



같은 input에 대해 다른 output이 발생하는 것을 줄이기 위해
두 output 분포 사이의 거리에 대한 loss를 추가하여 학습

성능 개선

Teacher Forcing Scheduler



Experiment	DE-EN		JA-EN	
	Dev	Test	Dev	Test
Teacher Forcing Baseline	35.05	29.62	18.00	19.46
No backprop				
Argmax	23.99	20.57	12.88	15.13
Top-k mix	35.19	29.42	18.46	20.24
Softmax mix $\alpha = 1$	35.07	29.32	17.98	20.03
Softmax mix $\alpha = 10$	35.30	29.25	17.79	19.67
Gumbel Softmax mix $\alpha = 1$	35.36	29.48	18.31	20.21
Gumbel Softmax mix $\alpha = 10$	35.32	29.58	17.94	20.87
Sparsemax mix	35.22	29.28	18.14	20.15
Backprop through model decisions				
Softmax mix $\alpha = 1$	33.25	27.60	15.67	17.93
Softmax mix $\alpha = 10$	27.06	23.29	13.49	16.02
Gumbel Softmax mix $\alpha = 1$	30.57	25.71	15.86	18.76
Gumbel Softmax mix $\alpha = 10$	12.79	10.62	13.98	17.09
Sparsemax mix	24.65	20.15	12.44	16.23

Step0이 증가할 수록 stage-1 decoder을 통해 구한 prediction 사용 비율 증가 → **모델의 추론 능력 향상**

성능 개선 결과

	Model	Rouge-L	비고
Model	BART → 146M	34.747	
	LongBART(not pretrained) → 67M	27.037(-7.71)	Pretraining with Text Infilling
	LongBART(pretrained) → 67M	32.327(-2.42)	
	BigBART(doc type and TAPT) → 153M	40.987(+6.21)	Pretraining with Text Infilling
	BigBART(pretrained) → 153M	36.457(+1.71)	

- LongBART 모델 사이즈가 가장 작지만,
리소스 내에서 최대의 성능을 발휘하는 BigBART(doc type and TAPT) Model로 선정
- 최종 모델: BigBART + 전처리 + R-Drop → Rouge-L :41.687

4. 경량화

#Dynamic_Quantization

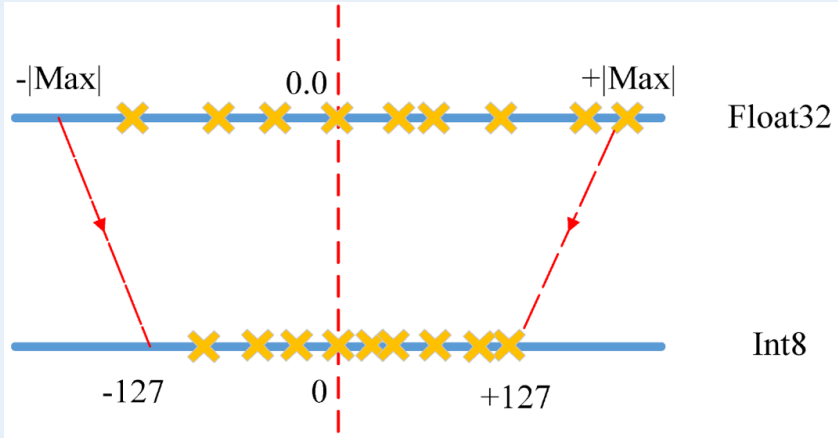
#Knowledge_Distillation

#경량화 결과

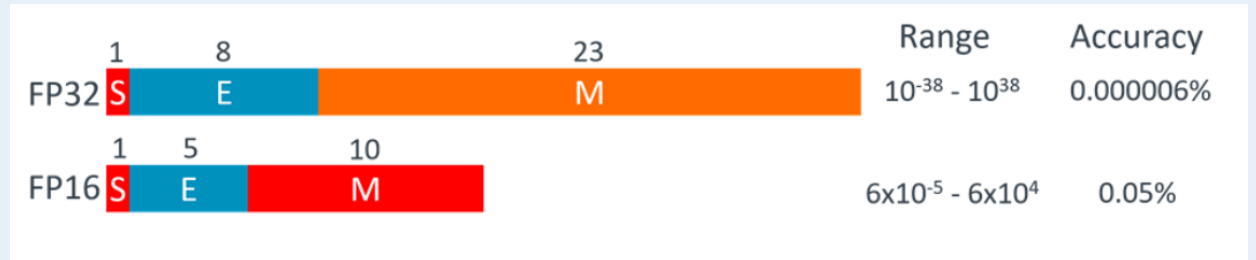
경량화

Dynamic Quantization

Quantization → Int8



Quantization → Float16

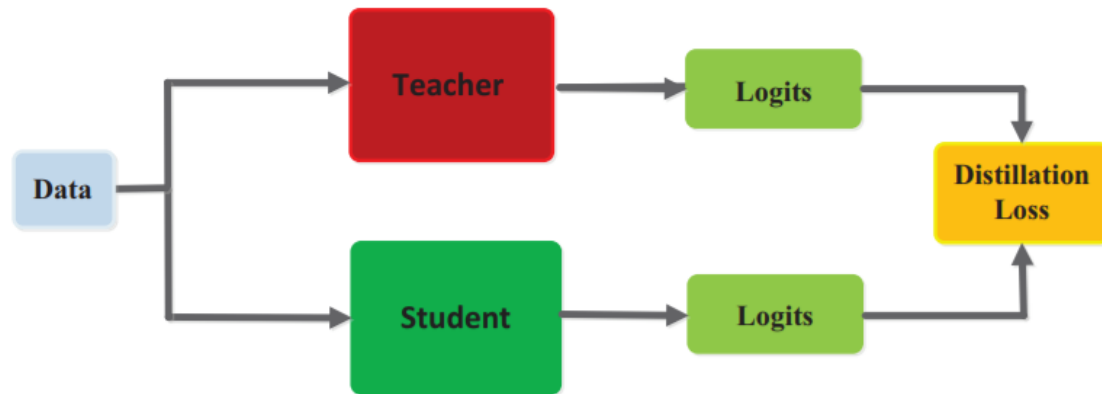


적용방법	개선사항	적용가능 Hardware
Int8	589.40MB → 223.58MB	CPU
Half	589.40MB → 294.76MB	GPU

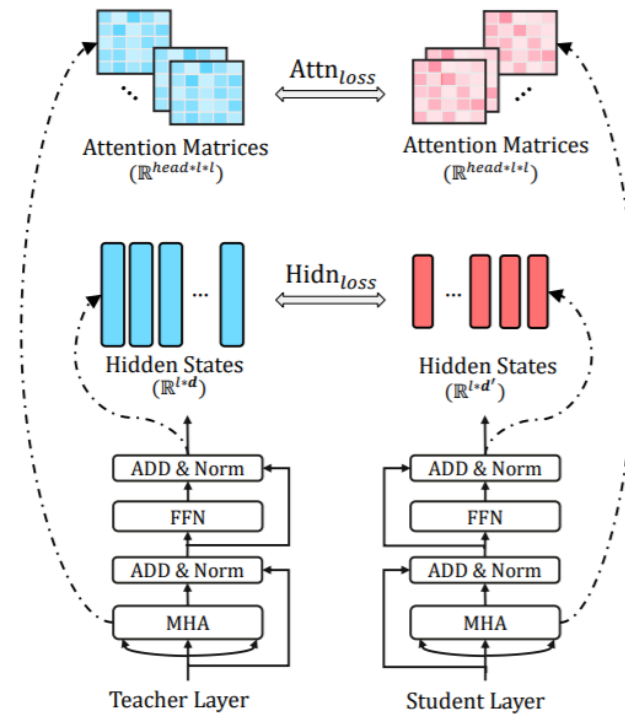
Rouge score은 유지하였으나, latency에서의 개선사항은 없었음

경량화

Knowledge Distillation



Basic Distillation



Tiny Distillation

성능이 좋은 teacher 모델로부터 student 모델이 학습

경량화 결과

Pruning(Student Model)

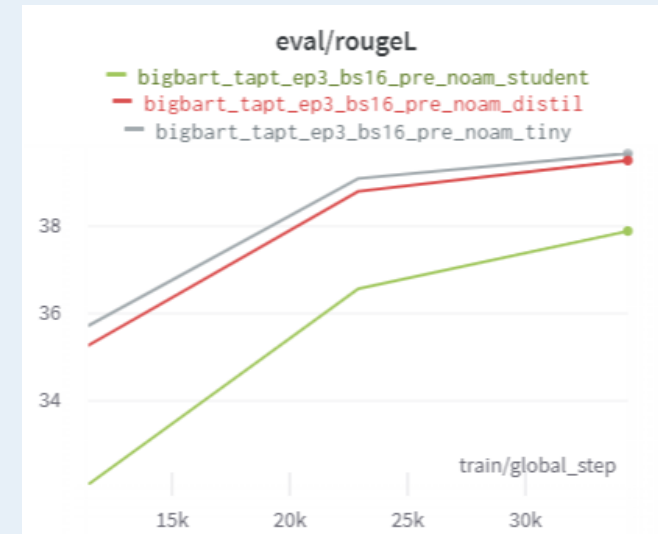
589.40MB → 400.07MB(32% 경량화)
13ms → 6.8ms(48% 감축)

Dynamic Quantization

400.07MB → 200.08MB(50% 경량화)

Distillation

적용방법	결과(Rouge-L)
Student Model	37.889
Basic Distillation	+ 1.621
Tiny Distillation	+ 1.783



5. 더 나아가기

#시각화
#ElasticSearch

더 나아가기

시각화

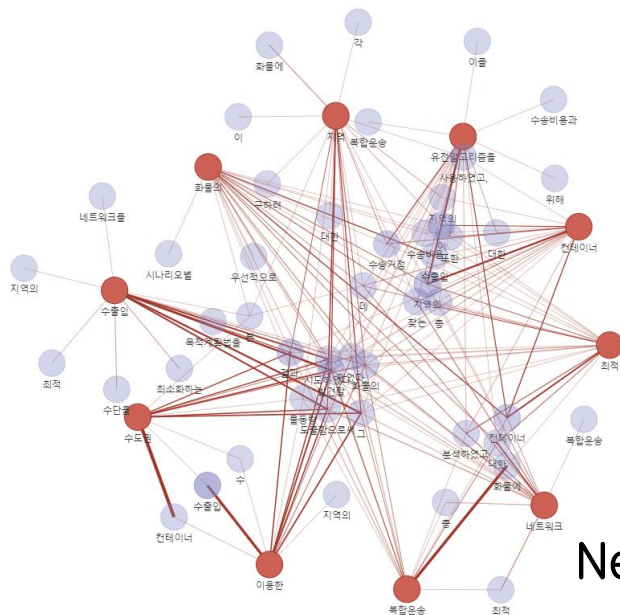
원문 제목: 유전알고리즘을 이용한 복합운송최적화모형에 관한 연구

생성된 제목: 수도권 지역의 수출입 컨테이너 화물에 대한 최적 복합운송 네트워크 탐색에 관한 연구

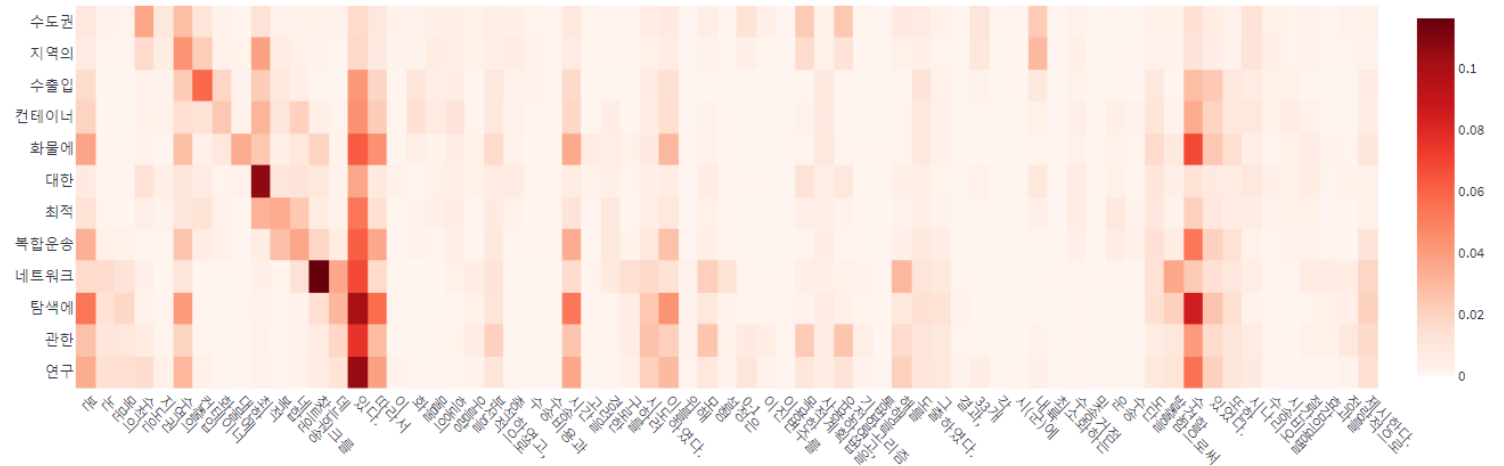
Text highlight

본 논문의 목적은 수도권 지역의 수출입 컨테이너 화물에 대한 최적 복합운송 네트워크를 찾는 데 있다. 따라서 이 지역의 컨테이너 화물의 물동량 흐름을 우선적으로 분석하였고, 총 수송비용과 총 수송 시간을 고려한 최적 경로를 구하려 시도하였다. 이를 위해 모형 설정은 0-1 이진변수를 이용한 목적 계획법을 사용하였고, 유전알고리즘 기법을 통해 해를 도출하였다. 그 결과, 수도권 지역의 33개 각 시(군)에 대한 내륙 수송비용과 수송 시간을 최소화 하는 수송거점 및 운송 수단을 도출함으로써 이 지역의 수출입 컨테이너 화물에 대한 최적 복합운송 네트워크를 발견할 수 있었다. 또한 시나리오별 수송비용 및 수송 시간이 적당 증가를 전라적으로 제시한다.

2D heatmap



Network graph



더 나아가기

ElasticSearch

문서 내용을 입력해주세요!

정부 방역지침에 반발한 자영업자들이 집회를 열고 방역패스와 영업시간 제한 등 조치에 대한 철폐를 촉구했다.

PC방업계와 호프업계, 공간대여업체 등으로 구성된 '코로나19 대응 전국자영업자비상대책위원회(비대위)'는 22일 오후 3시부터 서울 중로구 광화문 시민열린마당 인근에 모여 집회를 열었다. 앞서 **입력 문서** 집회·시위만 허가하는 정부 방역 수칙에 따라 299명 규모로 집회를 사전 신청했다.

현장에는 방역당국과 경찰, 주위 측 협조로 질서 유지선과 방역점검소가 설치됐다. 집회 참석자들은 체온 측정과 명부 작성 등 방역수칙을 준수한 채 자리를 잡았다. 다만 집회 참가자 간 거리



생성된 제목

Titles: 자영업자 299명 집회... "방역패스 허용해야"

[generate...] done in 1.365 s

유사 문서의 제목

☒ 유사제목 1: 코로나19 행정명령 비슷한 교회·클럽들 사회적 책임 물어야

본문내용: 서울시가 23일 방역수칙도 무시하고 예배를 강행해온 성북구 사랑제일교회에 대해 집회금지 명령을 내렸다. 다음달 5일까지 예배·찬양이 밀집해 예배 보면서 신도 간 1~2m 거리 두기를 하지 않은 것으로 확인됐다. 참석자 명단 작성 수칙도 어겼고, 바깥 붙어 기도·찬송·울동을 전광훈 담임목사(한국기독교총연합회 회장)의 석방을 요구하는 기도회를 평일에도 계속해왔다. 집회엔 전 목사가 이끌던 태극기집회 참석자들이 "며 폭언과 욕설을 퍼붓기도 했다. 서울시는 집회금지 명령을 어긴 사람은 1인당 300만원의 벌금을 부과하고, 확진자 발생 시 구상권도 청구할 성 행사를 해온 교회로선 자업자득이다. 23일 교회 3185 **유사 문서 내용** 를 받았다. 전국 교회 4만5420곳 중 7%가 방역수칙이 미배로 전환했고, 나머지는 현장예배 시 방역수칙을 지켰다고 평가됐다. 정부가 22일부터 2주간 종교·실내체육·유흥시설 운영을 중단토록 요구한 금까지 집단감염이 주로 일어난 중·소교회들이다. 정부는 감독의 고삐를 늦출 수 없는 큰 숫자라는 것을 명심해야 한다. 정부가 현장조사를 교회장 때 발열과 마스크 착용 여부를 확인했지만 내부에선 방역수칙이 무시됐다. 내달 6일 각급 학교 개학 전까지 펼쳐질 고강도 거리 두기는 코로나 준비도 마쳐야 한다. 언제까지 코로나19에 끌려갈 것인지도 여기에 달려 있다. 미꾸라지 한 마리가 물을 흐리게 해서도 안된다. 시민들은 마지막

☐ 유사제목 2: 광화문 원천봉쇄, 과도한 행정편의주의 걱정된다

☐ 유사제목 3: 집회의 자유와 감염 예방 조화시킨 판결에 주목한다

보유 데이터 중 유사한 문서의 제목과 모델을 통해 생성된 제목 비교 가능

6. 회고

#회고

회고

회고

잘한 점

- 유사 데이터셋에서 좋은 성능을 보인 성능 개선 방법론에 대해 다양한 시도를 함
- 개개인의 역량 파악을 통해 각 태스크 별로 정확한 일정 관리가 이뤄짐
- 프로젝트 전 과정에 대한 기록 및 github/notion/google sheet/wandb를 통한 협업이 원활히 이뤄짐
- 이전 대회에서 각자 다뤄보지 못한 태스크(model,data,visualization 등)들을 충분히 다뤄봄으로써 개인의 역량이 강화됨

아쉬운 점

- Korean Rouge Score의 평가 방법 오류를 늦게 파악하여 전체 성능 파악이 늦어짐
- Unstructured Pruning을 BART모델에 적용해보지 못함.

개선 방향

- 데이터 증량 시 성능 개선 및 LongBART의 Pretraining이 가능할 것으로 판단됨.
- BART 이외에 다양한 모델(T5,PEGASUS, switch transformers 등) 시도를 통해 성능 개선 및 적용 범위를 확대시킬 수 있을 것이라 판단됨
- 생성 요약만 진행했지만, 추출요약을 수행 후 시도를 했더라면 더 좋은 모델이 구현됐을 것으로 판단됨

Title : 부스트캠프, 지속 가능한 개발자를 꿈

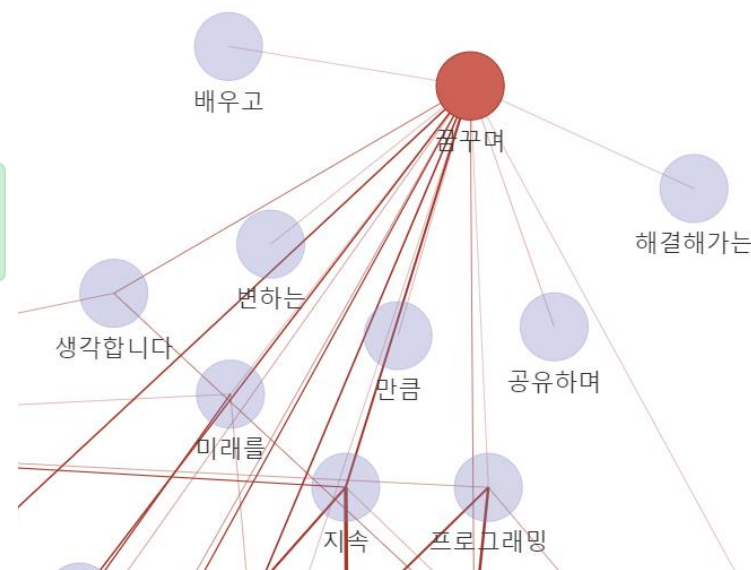
Titles: 부스트캠프, 지속가능한 개발자를 꿈꾸며

꾸며

[generate...] done in 1.227 s

Visualization!

부스트캠프는 단순히 뛰어난 **프로그래밍** 스킬을 가진 사람이 아닌, 사람들과 끊임없이 커뮤니케이션하고 서로에게 배우고 공유하며 문제를 해결해가는 사람, 그렇게 더 큰 **미래를 그리는** 사람이 좋은 **개발자**이고 **지속 가능한 개발자**라고 생각합니다. 좋은 경험과 습관을 가진 **지속 가능한 개발자** 양성을 목표로, 최고의 전문가들이 모여 만들어진 **소프트웨어** 교육 커뮤니티. 누구에게나 열려 있는 기회, 그러나 강도 높은 주도적인 문제 해결의 경험을 제공합니다. 세상이 빠르게 변하는 만큼 **부스트캠프**는 매년 새로운 시도들로 새로운 가능성을 찾습니다.



감사합니다



Q&A

