DEVIEW 2019

Dialog-BERT: 100억 건의 메신저 대화로 일상대화 인공지능 서비스하기

이주홍 (@roomylee)
Pingpong, Scatter Lab

SCATTER LAB

CONTENTS

DEVIEW 2019

1. 일상대화와 BERT 이해하기

2. 일상대화를 위한 Dialog-BERT 학습시키기

3. Dialog-BERT로 일상대화 태스크 해결하기

4. 서비스를 위한 BERT 경량화

CONTENTS

DEVIEW 2019

1. 일상대화와 BERT 이해하기 → 도메인 이해
2. 일상대화를 위한 Dialog-BERT 학습시키기 → Pre-training
3. Dialog-BERT로 일상대화 태스크 해결하기 → Fine-tuning
4. 서비스를 위한 BERT 경량화 → 모델 경량화

1. 일상대화와 BERT 이해하기

어제 말이죠…

DEVIEW 2019

내일 데뷰에서 발표하는데 너무 떨려

어제 말이죠…

DEVIEW 2019

내일 데뷰에서 발표하는데 너무 떨려

???

어제 말이죠...

DEVIEW 2019

내일 데뷰에서 발표하는데 너무 떨려

이해를 하지 못했어요
제가 할 수 없는 일이에요

아... 그래

위로하고 공감해주는 편안한 일상대화 능력 부족

왜 일상대화를 잘 못하지?

DEVIEW 2019

1. 대화 주제가 무한하다

사람이 얘기하는 모든 주제를 커버해야 함

2. 필요한 지식과 상식이 무한하다

"사과는 빨갛다", "동물은 숨을 쉰다", "택시는 타는 것이다"

3. 의도나 목적이 불분명하다 (=정답이 불분명하다)

“오늘 날씨 어때?” → “(아하 날씨를 알려달라는 거구나!)”

"아 커피 너무 맛있다” → "(??? 어쩌라고...)"

어휘의 변형이 심하다.
생략이 빈번하다.
일상대화는 데이터가 부족하다.

이러한 점들이 모델링을 굉장히 어렵게 한다.

저희 핑퐁이는요…

DEVIEW 2019

내일 데뷰에서 발표하는데 너무 떨려

잘할 수 있을 거에요!
화이팅!

고마워! 열심히 하고 올게!

대용량 데이터와 BERT를 통해 일상대화를 많이 이해하게 되었어요!

도대체 BERT가 뭐길래 (Devlin et al., 2018)

Bidirectional **E**ncoder **R**epresentation from **T**ransformer

도대체 BERT가 뭐길래 (Devlin et al., 2018)

Transformer(Vaswani et al. 2017)의 Encoder를 사용!

도대체 BERT가 뭐길래 (Devlin et al., 2018)

GLUE Test Results of BERT

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

11개의 다양한 NLP 태스크에서 state-of-the-art 성능을 보였으며,
이중 일부에서는 **사람보다도 뛰어난 결과를 얻음**

BERT 학습시키기

Pre-training

Fine-Tuning

Pre-training: 아주 많은 데이터. 오랜 시간 학습.

언어 전반에 대해 **깊게 이해**하는 단계

Fine-tuning: 특정 문제에 적응하는 단계.

깊은 언어의 이해를 바탕으로 **특정 문제에 맞춰 적응**하는 단계

BERT Pre-training

이순신은 그 즉시 조정에 장계를 올렸고 아울러 경상, 전라, 충청도에도 왜의 침략을 알리는 파발을 보냈다. 그 뒤 이순신은 휘하의 병력 700여명을 비상 소집하여 방비를 갖추도록 하였다.

- Wikipedia 이순신 中

Next Sentence Prediction (NSP)

Input: 이순신은 그 즉시 ... 파발을 보냈다. | 그 뒤 이순신은 ... 갖추도록 하였다

Output: True

Masked Language Modeling (Masked LM)

Input: 이순신은 그 즉시 조정에 [MASK] 를 올렸고 아울러 ... 파발을 보냈다.

Output: 장계

NSP 와
Masked LM 문제를 푸는 것이
언어를 이해하는 길이다
라고 BERT 저자가 설명함.

BERT Fine-tuning

(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

Question Answering (Machine Reading Comprehension)

Input: 이순신이 태어난 년도는 언제인가?

Output: 1545년

Sentiment Analysis

Input: 스토리면 스토리 액션이면 액션 정말 너무 재밌네요!!

Output: Positive

2. 일상대화를 위한 Dialog-BERT 학습시키기 (Pre-training)

일상대화 데이터

100억 건의 한국어 카카오톡 데이터, 2억 건의 일본어 라인 데이터

모든 데이터는 사용자의 동의를 받고 **사용자에게 직접 제공 받았으며,**
개인정보 식별이 불가능한 데이터만 연구 개발 목적으로 사용하고 있습니다.

일상대화 데이터 전처리: Tokenization

DEVIEW 2019

1. 형태소 분석 기반

   MeCab, Khaiii 등

2. Subword 기반

   SentencePiece, WordPiece 등

BERT에 기본적으로 장착되어 있는 것이 WordPiece 토크나이저이다.

3. Combined Approach

   Mecab으로 먼저 자르고 SentencePiece로 또 자르고

일상대화 데이터 전처리: Tokenization

영어와 다르게 한글은 공백 분할은 성능이 좋지 못하다.

한국어에서 음절(char) 분할은 괜찮게 된다.

MeCab이 LM에서 좋은 이유: 조사 분할을 잘해서.

Method	**LM** (Perplexity)	**NSMC** (Accuracy)	**Intent** (Accuracy)
Space	72.9	71.2	53.4
Char	35.4	83.0	70.1
MeCab	61.5	85.6	76.6
SentencePiece	295.7	85.3	77.1
MeCab + SentencePiece	**58.2**	**86.1**	**81.5**

* Size of Vocabulary: 30000
* LM (Language Modeling): 2-layer Bi-LSTM
* NSMC (Naver Sentiment Movie Corpus): Attention-based Bi-LSTM
* Intent (Intent Classification): Attention-based Bi-LSTM

일상대화 데이터 전처리

DEVIEW 2019

'안녕 데뷰야 나는 핑퐁이야 ㅎㅎ'

Tokenization
(MeCab + SentencePiece)

['_안녕', '_데', '뷰', '야', '_나', '는', '_핑퐁', '이', '야', '_ㅎㅎ']

일상대화 데이터 전처리

65억 Tokens

50 GB

30000 Vocabulary

사전 크기는
OOV 비율과
테스크 성능을 기준으로 결정함.

Pre-training, 그냥 하면 될까?

이순신(李舜臣, 1545년 4월 28일 ~ 1598년 12월 16일 (음력 11월 19일))은 조선 중기의 무신이다. 본관은 덕수(德水), 자는 여해(汝諧), 시호는 충무(忠武)이며, 한성 출신이다. 문반 가문 출신으로 1576년(선조 9년) 무과(武科)에 급제[2]하여 그 관직이 동구비보 권관, 훈련원 봉사, 발포진 수군만호, 조산보 만호, 전라좌도 수군절도사를 거쳐 정헌대부 삼도수군통제사에 이르렀다.

위키피디아

A: 아 너무 배고프다

B: 이따 머 먹으러 갈려?

A: 옼ㅋㅋ 타이밍 쩌네 ㅇㅇ ㄱㄱ

B: 떡볶이?

A: ㅇㅇㅋ 이따 봥

B: 그랭 ㅋㅋ 늦지 말고 오셈

카카오톡

대화체와 문어체는 매우 다른데...?

대화에 알맞은 Pre-training 전략

고민1. 대화체로 학습시켜도 괜찮을까?

고민1. 대화체로 학습시켜도 괜찮을까?

DEVIEW 2019

이순신(李舜臣, 1545년 4월 28일 ~ 1598년 12월 16일 (음력 11월 19일))은 조선 중기의 무신이다. 본관은 덕수(德水), 자는 여해(汝諧), 시호는 충무(忠武)이며, 한성 출신이다. 문반 가문 출신으로 1576년(선조 9년) 무과(武科)에 급제[2]하여 그 관직이 동구비보 권관, 훈련원 봉사, 발포진 수군만호, 조산보 만호, 전라좌도 수군절도사를 거쳐 정헌대부 삼도수군통제사에 이르렀다.

위키피디아

A: 아 너무 배고프다

B: 이따 머 먹으러 갈려?

A: 옼ㅋㅋ 타이밍 쩌네 ㅇㅇ ㄱㄱ

B: 떡볶이?

A: ㅇㅇㅋ 이따 봥

B: 그랭 ㅋㅋ 늦지 말고 오셈

카카오톡

대화체.
문장 길이가 짧다.
단어의 변형과 생략이 잦다.

고민1. 대화체로 학습시켜도 괜찮을까?

DEVIEW 2019

이순신(李舜臣, 1545년 4월 28일 ~ 1598년 12월 16일 (음력 11월 19일))은 조선 중기의 무신이다. 본관은 덕수(德水), 자는 여해(汝諧), 시호는 충무(忠武)이며, 한성 출신이다. 문반 가문 출신으로 1576년(선조 9년) 무과(武科)에 급제[2]하여 그 관직이 동구비보 권관, 훈련원 봉사, 발포진 수군만호, 조산보 만호, 전라좌도 수군절도사를 거쳐 정헌대부 삼도수군통제사에 이르렀다.

위키피디아

A: 아 너무 배고프다

B: 이따 **머** 먹으러 **갈려?**

A: **올**ㅋㅋ 타이밍 쩌네 **ㅇㅇ ㄱㄱ**

B: **떡볶이?**

A: ㅇㅇㅋ 이따 **봉**

B: **그랭** ㅋㅋ 늦지 말고 **오셈**

카카오톡

고민1. 대화체로 학습시켜도 괜찮을까?

Model	문어체 (Wiki)		대화체 (Dialog)		대화체
	Masked LM	**NSP**	**Masked LM**	**NSP**	**NSMC**
Wiki-BERT	**55.7**	**95.3**	36.2	44.6	87.7
Dialog-BERT	24.6	46.4	**53.8**	**84.1**	**88.8**

고민1. 대화체로 학습시켜도 괜찮을까?

Model	문어체 (Wiki)		대화체 (Dialog)		대화체
	Masked LM	**NSP**	**Masked LM**	**NSP**	**NSMC**
Wiki-BERT	**55.7**	**95.3**	36.2	44.6	87.7
Dialog-BERT	24.6	46.4	**53.8**	**84.1**	**88.8**

1. 대화체 데이터로도 충분히 잘 학습된다

고민1. 대화체로 학습시켜도 괜찮을까?

Model	문어체 (Wiki)		대화체 (Dialog)		대화체
	Masked LM	**NSP**	**Masked LM**	**NSP**	**NSMC**
Wiki-BERT	**55.7**	**95.3**	36.2	44.6	87.7
Dialog-BERT	24.6	46.4	**53.6**	**84.1**	**88.8**

1. 대화체 데이터로도 충분히 잘 학습된다

2. 서로 다른 도메인에서는 성능이 매우 떨어진다

문어체 vs. 대화체

이와 같은 이유로 pre - training도
대화체로 학습하기로 결정!

대화에 알맞은 Pre-training 전략

고민2. Turn에 대한 구분이 있어야 하지 않을까?

고민2. Turn에 대한 구분이 필요하지 않을까?

DEVIEW 2019

이순신(李舜臣, 1545년 4월 28일 ~ 1598년 12월 16일 (음력 11월 19일))은 조선 중기의 무신이다. 본관은 덕수(德水), 자는 여해(汝諧), 시호는 충무(忠武)이며, 한성 출신이다. 문반 가문 출신으로 1576년(선조 9년) 무과(武科)에 급제[2]하여 그 관직이 동구비보 권관, 훈련원 봉사, 발포진 수군만호, 조산보 만호, 전라좌도 수군절도사를 거쳐 정헌대부 삼도수군통제사에 이르렀다.

위키피디아

A: 아 너무 배고프다

B: 이따 머 먹으러 갈려?

A: 올ㅋㅋ 타이밍 쩌네 ㅇㅇ ㄱㄱ

B: 떡볶이 ㄱ?

A: ㅇㅇㅋ 이따 봉

B: 그랭 ㅋㅋ 늦지 말고 오셈

카카오톡

대화에는 turn 이라는 게 존재함.
대화는 화자 A와 화자 B가 번갈아가면서 이야기를 한다. 즉, 두 개 이상의 흐름이 서로 상호작용한다.

이러한 turn 정보를 모델에게 줘야 하지 않을까?

고민2. Turn에 대한 구분이 필요하지 않을까?

DEVIEW 2019

고민2. Turn에 대한 구분이 필요하지 않을까?

DEVIEW 2019

SEPT: 각 turn을 구분.

고민2. Turn에 대한 구분이 필요하지 않을까?

DEVIEW 2019

Turn Embedding: 각 화자에 대한 임베딩

고민2. Turn에 대한 구분이 필요하지 않을까?

DEVIEW 2019

Model	Masked LM	NSP	NSMC
Dialog-BERT	53.6	84.1	88.8
+ Turn SEP & EMB	**55.3**	**88.4**	**90.4**

각 발화 문장과 대화의 흐름을 더 잘 이해할 수 있다

Dialog-BERT Configuration

for Pre - training.

DEVIEW 2019

Hyperparameters	Value
BERT Config.	
hidden_size (H)	768
intermediate_size	3072
num_attention_heads (A)	12
num_hidden_layers (L)	12
vocab_size	30006
Training Config.	
num_steps	**10,000,000**
batch_size	**1024**
learning_rate	5e-5
max_sequence_length	**48**
max_predictionas_per_seq	**8**
masked_lm_prob	15

대화체가 문어체보다 짧기 때문에 128에서 48로 축소.

Google Cloud

TPU v3-8로 약 30일간 학습

1~2일 학습에도 성능이 꽤 괜찮았음.
그 이후로는 2~3% 성능 향상.
시간과 비용이 없다면, 짧은 시간으로 학습해도 괜찮을 것!.

3. Dialog-BERT로 일상대화 태스크 해결하기 (Fine-tuning)

일상대화 태스크

오늘 저녁 뭐 먹을까?

음 짜장면 어떰??

User Input

어떤 일련의 과정

System Output

일상대화 태스크

일상대화라는 큰 문제를
다음과 같이 작은 3가지 문제로
나눠서 풀려고 했음.

Task #1. Semantic Textual Similarity:

주어진 두 문장이 의미적으로 유사한가?

Task 1과 Task 2는
전형적인 Dialog 테스크.

Task 3는 핑퐁에서 만듦.

Task #2. Query-Reply Matching:

주어진 문장 다음에 올 문장으로 적절한가?

Task #3. Reaction Classification:

주어진 문장 다음에 올 리액션은 무엇일까?

Task #1. Semantic Textual Similarity

Sentence A	Sentence B	Is Similar ?
어떤 실수 했는데?	무슨 실수요?	1
네 보통 몇시에 자요?	응응 평소엔 보통 언제쯤 자?	1
나한테 관심도 없고	난 이미 관심 없어서	0
자다가 깼어요?	졸린데 깨웠어요?	0
...	...	...
오늘 저녁 뭐 먹을까?	저녁 뭐 먹징?	1

두 문장이 의미적으로 유사한지를 분류하는 문제

Task #1. Semantic Textual Similarity

DEVIEW 2019

Semantic Textual Similarity 데이터 셋을 기준으로 Fine - tuning을 한다.

Fine-tuning!!

Task #1. Semantic Textual Similarity

Query	Reply
저녁 뭐 먹징?	음 짜장면 어떰??
잘 잤어요?	간만에 푹잤다 ㅎㅎ
...	...
와 진짜 배고파	헐 나도 ㅋㅋ

Paired Data

Task #1. Semantic Textual Similarity

오늘 저녁 뭐 먹을까?

Query	Reply
저녁 뭐 먹징?	음 짜장면 어떰??
잘 잤어요?	간만에 푹잤다 ㅎㅎ
...	...
와 진짜 배고파	헐 나도 ㅋㅋ

User Input

Paired Data

Task #1. Semantic Textual Similarity

DEVIEW 2019

어떤 Query가 가장 유사할까?

오늘 저녁 뭐 먹을까?

Query	Reply
저녁 뭐 먹징?	음 짜장면 어떰??
잘 잤어요?	간만에 푹잤다 ㅎㅎ
...	...
와 진짜 배고파	헐 나도 ㅋㅋ

User Input

Paired Data

Task #1. Semantic Textual Similarity

DEVIEW 2019

Inference!!

Task #1. Semantic Textual Similarity

DEVIEW 2019

오늘 저녁 뭐 먹을까?

	Query	Reply
0.9	저녁 뭐 먹징?	음 짜장면 어떰??
0.1	잘 잤어요?	간만에 푹잤다 ㅎㅎ
...	...	...
0.6	와 진짜 배고파	헐 나도 ㅋㅋ

User Input

Paired Data

Task #1. Semantic Textual Similarity

DEVIEW 2019

오늘 저녁 뭐 먹을까?

Query	Reply
저녁 뭐 먹징?	**음 짜장면 어떰??**
잘 잤어요?	간만에 푹잤다 ㅎㅎ
...	...
와 진짜 배고파	헐 나도 ㅋㅋ

음 짜장면 어떰??

User Input

Paired Data

System Output

Task #2. Query-Reply Matching

DEVIEW 2019

Sentence A	Sentence B	Is Next ?
나 너무 살찐 것 같아	그정도면 날씬하죠	1
나 너무 살찐 것 같아	그런 것 같아요	0
...	...	...
오늘 저녁 뭐 먹을까?	음 짜장면 어떰?	1

논리적으로는 맞지만, 올바른 대화 방향이 아니라서 0이다.

단순히 다음에 올 수 있는지(NSP)보다 "얼마나 좋은 답인지"를 분류하는 문제

Task #2. Query-Reply Matching

Fine-tuning!!

Task #2. Query-Reply Matching

Reply
음 짜장면 어떰??
간만에 푹잤다 ㅎㅎ
...
헐 나도 ㅋㅋ

Unpaired Data

Task #2. Query-Reply Matching

오늘 저녁 뭐 먹을까?

Reply
음 짜장면 어떰??
간만에 푹잤다 ㅎㅎ
...
헐 나도 ㅋㅋ

User Input

Unpaired Data

Task #2. Query-Reply Matching

DEVIEW 2019

다음에 올 Reply로 뭐가 좋을까?

오늘 저녁 뭐 먹을까?

Reply
음 짜장면 어떰??
간만에 푹잤다 ㅎㅎ
...
헐 나도 ㅋㅋ

User Input

Unpaired Data

Task #2. Query-Reply Matching

Inference!!

Task #2. Query-Reply Matching

오늘 저녁 뭐 먹을까?

	Reply
0.9	음 짜장면 어떰??
0.1	간만에 푹잤다 ㅎㅎ
...	...
0.6	헐 나도 ㅋㅋ

User Input

Unpaired Data

Task #2. Query-Reply Matching

오늘 저녁 뭐 먹을까?

Reply
음 짜장면 어떰??
간만에 푹잤다 ㅎㅎ
...
헐 나도 ㅋㅋ

음 짜장면 어떰??

User Input

Unpaired Data

System Output

Task #3. Reaction Classification

DEVIEW 2019

Sentence	Reaction Class (1384)
출근하기 너무 싫다	화이팅! (id=13)
오늘 짜장면 먹을까?	좋아요 (id=5)
아 뭐 잘못먹었나 배가 아프네 ㅠㅠ	괜찮아요? (id=78)
퇴근이다!	수고했어요 (id=324)
...	...
벌써 12시네 ㅠㅠ 나 자야겠다	잘자요 (id=1016)

다양한 리액션을 class로 정의.

하나의 문장을 <u>미리 정의된 Reaction Class</u> 중 하나로 분류하는 문제

Task #3. Reaction Classification: Motivation

DEVIEW 2019

우리가 하는 말 중에 대부분은 "Reaction"에 가까움

빈도수 기준 상위 0.01%의 문장이 전체 문장 중 20%를 차지함

리액션만 잘해도..
왠만한 대화를 커버할 수 있겠구나라고 생각.

리액션
총 1384개의 리액션을 정의.

알겠어	귀여워	거짓말	잘자	어딘데요?	잘했어
대박!	수고했어	맛있게 먹어	미안해	아니에요	좋아!
부러워	졸려	밥 먹었어?	진짜요?	사랑해	괜찮아
누구랑?	언제?	글쎄...	나도나도	아직도?	안녕

Task #3. Reaction Classification

Fine-tuning!!

Task #3. Reaction Classification

Reaction Class
좋아요
고마워요
...
수고했어요

Predefined Class Set

(1384)

Task #3. Reaction Classification

DEVIEW 2019

오늘 짜장면 먹을까?

Reaction Class
좋아요
고마워요
...
수고했어요

User Input

Predefined Class Set
(1384)

Task #3. Reaction Classification

DEVIEW 2019

다음에 올 적절한 Reaction은 뭘까?

오늘 짜장면 먹을까?

Reaction Class
좋아요
고마워요
...
수고했어요

User Input

Predefined Class Set
(1384)

Task #3. Reaction Classification

DEVIEW 2019

Inference!!

Task #3. Reaction Classification

오늘 짜장면 먹을까?

	Reaction Class
0.6	좋아요
0.01	고마워요
...	...
0.01	수고했어요

User Input

Predefined Class Set
(1384)

Task #3. Reaction Classification

DEVIEW 2019

오늘 짜장면 먹을까?

좋아요

User Input

Predefined Class Set
(1384)

System Output

일상대화 태스크

Task	Quality	Coverage
1. Query Similarity	상	하
2. Reply Matching	중	중
3. Reaction	하	상

서로 상호보완적으로 답변을 만들어낼 수 있다

일상대화 태스크 성능

Model	**Query Similarity** (Accuracy)	**Reply Matching** (NDCG@10)	**Reaction** (Accuracy)
RNNs (Pingpong)	85.0	83.1	19.3
Multilingual BERT (Google)	87.8	70.9	18.6
KorBERT (ETRI)	90.5	78.2	21.8
KoBERT (SK T-Brain)	89.5	67.6	19.7
Dialog-BERT (Pingpong)	**93.3**	**87.0**	**25.7**

Google, ETRI, T - Brain 에서는 모두 문어체를 사용.
핑퐁은 대화체.
그냥 데이터셋만 다르게 세팅한 것이지 않나?

대화 관련 NLP 태스크 성능

Model	**NSMC** (Accuracy)	**Intent** (Accuracy)
RNNs (Pingpong)	86.1	81.5
Multilingual BERT (Google)	87.5	82.6
KorBERT (ETRI)	**90.4**	87.6
KoBERT (SK T-Brain)	90.1	83.4
Dialog-BERT (Pingpong)	**90.4**	**88.9**

NSMC는 문장 길이 100이 넘어갈 정도로 길다. 우리는 48 길이밖에 되지 않는다. 이런 핸디캡이 있음에도 가장 좋은 성능을 보였다.

인공지능과 대화해보기

DEVIEW 2019

긴 문장도 이해를 잘하고 적절한 대답을 해요

상대방 발화를 명확히 파악해서 상황에 적절한 답변을 할 수 있음

어제 엄청 늦게 잤더니 늦게 일어나서 아침도 못 먹고 나왔어

멀티턴 리액션 😃

- 몇시에 잤는데요? (0.2168)
- 아이고 (0.2058)
- 으이구 (0.2043)

여러 개 문장이 중첩되서 들어와도,
의미를 잘 파악해서.
적절한 답변을 만들어낸다.

헐.. 나 데뷰 발표자료 거의 다 만들었는데 갑자기 컴퓨터 꺼져써.. 다 날아간건 아니겠지ㅠㅠ?

멀티턴 리액션 😃

- 설마요 (0.2280)
- 아닐거예요 (0.2234)
- 아니겠죠 (0.2121)

인공지능과 대화해보기

DEVIEW 2019

사회적인 개념에 대한 이해도 하고 있어요

수많은 데이터 속에서 자연스럽게 학습하면서 사회적 개념을 이해함

오늘 월요일이다

멀티턴 리액션 😃

알아요 (0.2191)

아 맞다 (0.2127)

으악 (0.2017)

오늘 금요일이다

멀티턴 리액션 😃

좋아요? (0.2169)

오예 (0.2147)

신나요 (0.2108)

인공지능과 대화해보기: Multi-turn 학습

DEVIEW 2019

문맥을 이해하면서 대화할 수 있어요

BERT를 Single-turn이 아닌 Multi-turn Context로 학습시키면 문맥을 이해함

인공지능과 대화해보기: Multi-turn 학습

DEVIEW 2019

문맥을 이해하고 있기 때문에 짧게 끊어지는 발화에도 대답할 수 있어요

문맥을 이해해서 생략된 부분을 잘 유추해서 적절한 대답을 함

대화 문맥 속에
녹아든 발화의 의미 역시 잘
유추하는 것을 확인.

4. 서비스를 위한 BERT 경량화

서비스... 쉽지 않네

Number of Parameters
110,000,000
82,500,000
55,000,000
27,500,000
0
26M
109M
RNNs
BERT

Inference Time
160
120
80
40
0
50ms
160ms
RNNs
BERT

모델 경량화

Knowledge Distillation

(Hinton et al., 2015)

모델 경량화: Knowledge Distillation

모델 경량화: Knowledge Distillation

DEVIEW 2019

Hard Label

dog	cat	tiger		car
0	1	0	···	0

Train

Model

모델 경량화: Knowledge Distillation

DEVIEW 2019

Hard Label

dog	cat	tiger		car
0	1	0	···	0

Train

Large Model

Hard Label

dog	cat	tiger		car
0	1	0	···	0

Train

Small Model

모델 경량화: Knowledge Distillation

DEVIEW 2019

Hard Label

dog	cat	tiger		car
0	1	0	···	0

Train

Large Model

Knowledge = Softmax Output

softmax output은 large model이 오랜시간 동안 학습한 결과물이다.

Soft Label

dog	cat	tiger		car
0.1	0.8	0.05	···	0.01

Train

Small Model

BERT Distillation

Distillation은
pre - training에 적용하는 것이 아닌,
fine - tuning에 적용했다.

BERT-Base

num_layer = 12

num_head = 12

hidden_size = 768

BERT-Small

num_layer = 10

num_head = 8

hidden_size = 512

BERT-XSmall

num_layer = 8

num_head = 8

hidden_size = 512

Result of Distillation

DEVIEW 2019

Performance of Reaction Classification

Model Size and Inference Time

Model	# of Parameters	Inference Time
BERT-Base	109M (1x)	160ms (1x)
BERT-Small	47M (2.3x)	90ms (1.8x)
BERT-XSmall	41M (2.7x)	60ms (2.7x)
RNNs	26M (4.2x)	50ms (3.2x)

* Intel Xeon CPU E5-2698 v4 @ 2.20GHz

Multi-task Distillation (Clark et al., 2019)

Result of Multi-task Distillation

Model	Average Score	Reply Matching	Reaction
BERT-Base	56.4	87.0	25.7
BERT-XSmall	55.3	85.8	24.7
BERT-XSmall + Single-task	55.8	86.4	25.1
BERT-XSmall + Multi-task	**56.0**	**86.7**	**25.3**

5. Conclusion

3줄 요약

1. BERT, 이제는 선택이 아닌 필수!

2. 태스크의 문체에 맞춰서 BERT를 사용하면 더 좋아요!

3. Distillation 꼭 해보세요!

Future Work

1. Neural Generation: 세상에 없는 기가 막힌 문장을 만들어내기

2. Text Style Transfer: 존댓말, 성향 등에 대한 스타일 변환

3. Knowledge / Topic: 지식을 기반으로한 대화, 다양한 주제의 대화

4. Representation: 더 좋은 Language Understanding

5. Computer Vision: 이미지에 대해서 대화 나누기

(이외에도 너무 너무 많아요 ㅠㅠ)

일상대화 인공지능 같이 만들어요!

DEVIEW 2019

우리만큼 열심히 하고 / 진짜 잘 놀고 / 재미있게 일하는 곳 없을 거에요! 😻

저희 부스도 하고 있어요! 채용상담하시고 "스타벅스 쿠폰" 받아가세요!

지금 바로 GO GO! 👉 https://scatterlab.co.kr/recruiting

Thank You

References

Paper

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding: https://arxiv.org/abs/1810.04805
- Attention Is All You Need: https://arxiv.org/abs/1706.03762
- Distilling the Knowledge in a Neural Network: https://arxiv.org/abs/1503.02531
- BAM! Born-Again Multi-Task Networks for Natural Language Understanding: https://arxiv.org/abs/1907.04829

Article

- The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning): http://jalammar.github.io/illustrated-bert/
- Lunit Tech Blog - Distilling the Knowledge in a Neural Network (NIPS 2014 Workshop): https://blog.lunit.io/2018/03/22/distilling-the-knowledge-in-a-neural-network-nips-2014-workshop/

Third-party Material

- Pingpong Blog: https://blog.pingpong.us/
- Pingpong Demo: https://demo.pingpong.us/

Q & A