

시험성적서

한국시험인증원(주)

서울시 강서구 마곡중앙로 161-8, B동 1216호
Tel:02-6929-1033 Fax:02-6929-1037

성적서번호 : KST-21-049



1. 의뢰자

- 회사(기관)명 : 서울대학교 산학협력단
- 주소 : 서울특별시 관악구 관악로 1 서울대학교 60동 5층

2. 시험대상품목/물질/시료 설명

- 시험대상 : 비디오 튜링 테스트를 통과할 수준의 비디오 스토리 이해 기반 질의응답 기술 과제결과물

3. 시험기간 : 2021년 08월 05일 ~ 2021년 08월 06일

4. 시험장소 : ☐ 고정시험실 ☒ 현장시험

(주소: 서울특별시 관악구 관악로 1 서울대학교 301동 5층)

5. 시험방법 : 의뢰자 제시기준

6. 시험결과 : 시험성적서 '3. 시험항목별 시험방법 및 시험결과' 참조

이 시험결과는 의뢰자가 제시한 시험대상품목/물질/시료에만 한정됩니다.

확 인	시험자 성 명 : 박 한 비  (서명)	기술책임자 성 명 : 차 원 영  (서명)
-----	---	---

2021년 08월 26일

한국인정기구 인정 한국시험인증원(주) 대표이사 (인)



성적서번호 : KST-21-049

한국시험인증원
KOTCA Korea Testing
Certification Agency

시험결과:

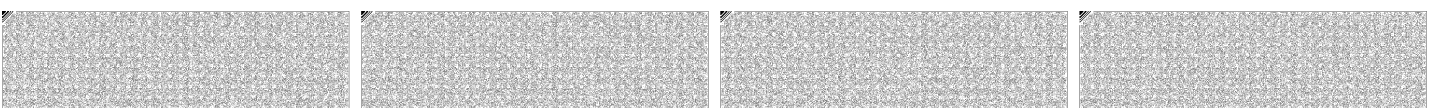
서울대학교 산학협력단

비디오 튜링 테스트를 통과할 수준의 비디오 스토리 이해 기반 질의응답
기술 과제결과물

V&V(Verification & Validation) 시험



본 문서는 KOTCA의 시험성적서로서 누구든지 KOTCA의 사전 승인 없이 문서의 일부분만을 발췌 또는 인용하여 사용하거나 배포할 수 없습니다. 이 성적서의 진위 확인은 기업지원플러스(<http://www.g4b.go.kr>) 웹페이지에서 G4B조회코드로 확인 가능합니다.



성적서번호 : KST-21-049

목 차

1. 시험 개요	4
2. 시험 환경	5
3. 시험항목별 시험방법 및 시험결과	6
3.1 모델 압축 비율	6
3.2 모델 실행시간 비율	7
4. 시험결과에 따른 판정결과	8
<붙임 1> 모델 압축 비율 시험결과 화면	9
<붙임 2> 모델 실행시간 비율 시험결과 화면	10



성적서번호 : KST-21-049

1. 시험 개요

본 문서는 서울대학교 산학협력단이 시험 의뢰한 '비디오 튜링 테스트를 통과할 수준의 비디오 스토리 이해 기반 질의응답 기술 과제결과물'에 대하여 V&V 시험을 수행한 시험 성적서이다. 시험 성적서의 결과는 본 성적서에 국한된다.

서울대학교 산학협력단이 시험 의뢰한 시험항목 및 판정기준은 다음과 같다.

<시험항목 및 판정기준>

구분	시험항목	판정기준
1	모델 압축 비율	10% 이하
2	모델 실행시간 비율	20% 이하

판정결과는 시험항목의 시험결과에 따라 Pass(합격)/Fail(불합격)로 구분하여 기재하였다.

<판정결과 기준>

구분	기준
Pass(합격)	시험항목별 시험결과가 판정기준을 만족하는 경우
Fail(불합격)	시험항목별 시험결과가 판정기준을 만족하지 못하는 경우

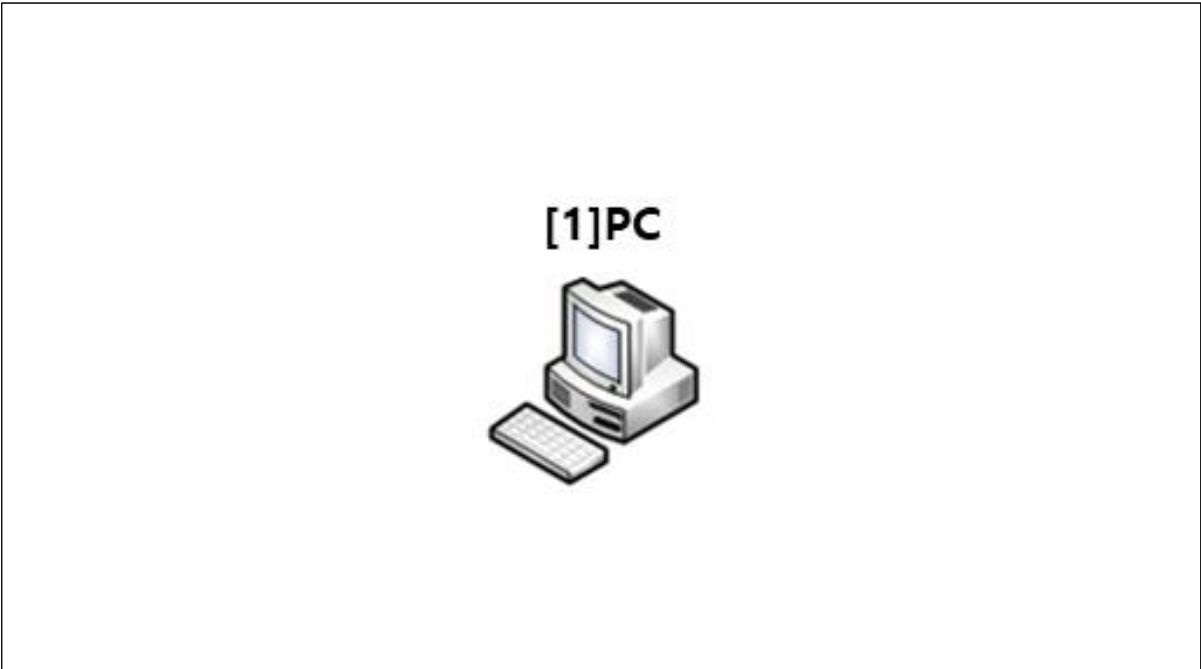
성적서번호 : KST-21-049



2. 시험 환경

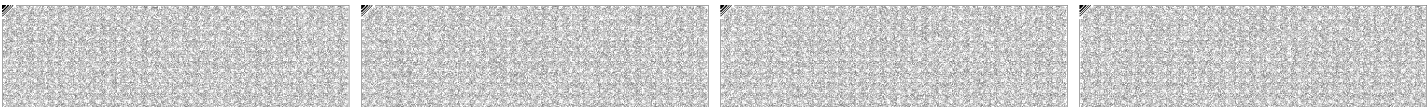
시험환경은 서울대학교 산학협력단에 구축하였고, 시험환경구성 및 세부사양은 아래와 같다.

<시험환경구성도>



<세부사양>

No	Role	OS	CPU	Memory	HDD/SSD	GPU	Pre-Requisite
1	PC	Ubuntu 18.04.5 LTS	Intel® Xeon® Silver 4114 CPU @ 2.20GHz	512GB	HDD 33TB	NVIDIA RTX 2080Ti	<ul style="list-style-type: none">• CUDA 10.1• Pytorch 1.7.1



성적서번호 : KST-21-049

3. 시험항목별 시험방법 및 시험결과

3.1 모델 압축 비율

3.1.1 시험 목적

- 모델 압축 비율 확인

$\text{모델 압축 비율} = \frac{\text{SensiMix 모델 용량}}{\text{BERT 모델 용량}} \cdot 100\%$

- * SensiMix 모델 : BERT 모델을 압축한 모델
- * BERT 모델 : 자연어 처리 모델

3.1.2 시험 방법

- 1) 터미널에 'ft.sh' 명령어를 입력하여 BERT 모델 압축 시작
- 2) Jupyter에 접속하여 './BERT_model' 폴더에서 'pytorch_model.bin' 파일을 통해 BERT 모델 용량 확인
- 3) Jupyter에 'SenxiMix_model_quantized' 폴더에서 'pytorch_model.bin' 파일을 통해 SensiMix 모델 용량 확인
- 4) BERT 모델과 SensiMix 모델의 용량을 비교하여 압축 비율이 10% 이하인지 확인

3.1.3 시험 결과

- 모델 압축 비율 : 9.6%

구분	BERT 모델 용량	SensiMix 모델 용량
1	438.0MB	42.2MB

※ 시험결과 화면은 <붙임 1> 참조

성적서번호 : KST-21-049

3.2 모델 실행시간 비율

3.2.1 시험 목적

- 모델 실행시간 평균 비율 확인

$$\text{모델 실행시간 비율} = \frac{\text{SensiMix 모델 실행시간}}{\text{BERT 모델 실행시간}} \cdot 100\%$$

3.2.2 시험 방법

- 1) 터미널에 'bert_inference.sh' 명령어를 입력하여 BERT 모델 실행
- 2) QNLI dataset이 입력되고 결과로그 생성
 - * QNLI dataset : Question Natural Language Inference
- 3) 결과로그의 'duration' 항목에서 명령어를 입력하는 시각부터 자연어 추론이 완료되는 시각까지의 모델 실행시간 확인
- 4) 터미널에 'sensimix_inference.sh' 명령어를 입력하여 SensiMix 모델 실행
- 5) QNLI dataset이 입력되고 결과로그 생성
- 6) 결과로그의 'duration' 항목에서 명령어를 입력하는 시각부터 자연어 추론이 완료되는 시각까지의 모델 실행시간 확인
 - ※ 3회 시험하여 모델 실행시간 평균 비율 계산

3.2.3 시험 결과

- 모델 실행시간 평균 비율 : 12.1%

시험회차	BERT 모델 실행시간	SensiMix 모델 실행시간	모델 실행시간 비율
1	254,339ms	31,729ms	12.5%
2	254,550ms	29,395ms	11.5%
3	254,698ms	31,566ms	12.4%

※ 시험결과 화면은 <붙임 2> 참조

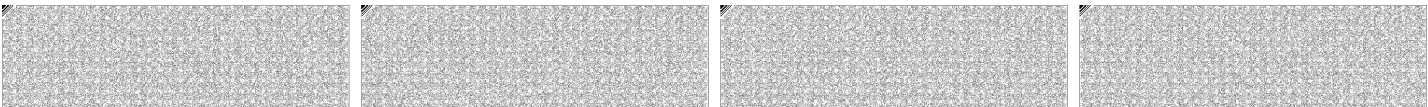
성적서번호 : KST-21-049



4. 시험결과에 따른 판정결과

시험결과에 따른 판정결과는 다음과 같다.

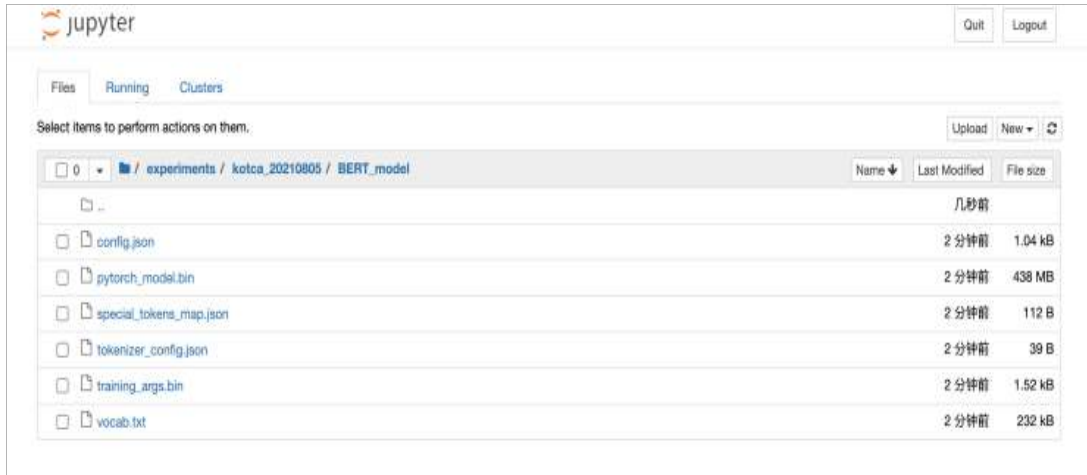
구분	시험항목	시험결과	판정기준	판정결과
1	모델 압축 비율	9.6%	10% 이하	Pass
2	모델 실행시간 비율	12.1%	20% 이하	Pass



성적서번호 : KST-21-049

<붙임 1> 모델 압축 비율 시험결과 화면

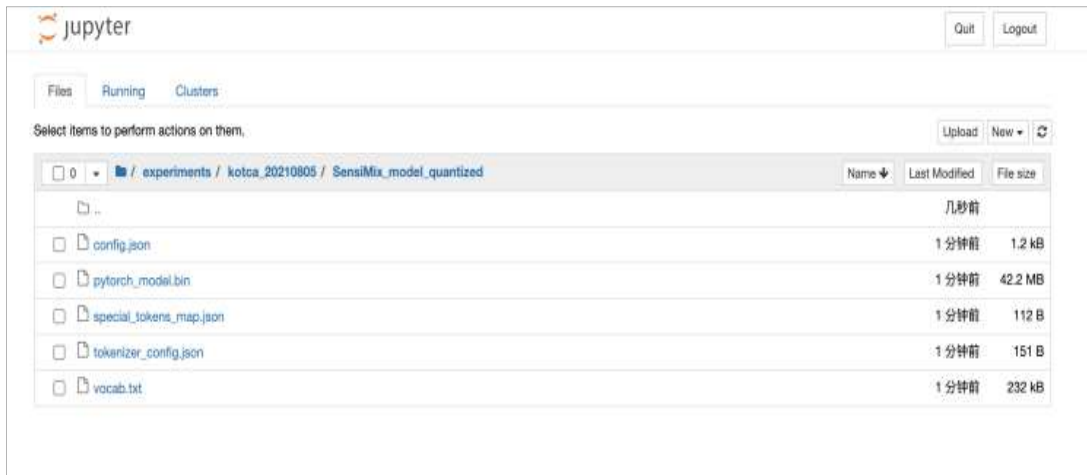
○ BERT 모델 용량 확인



The screenshot shows the JupyterLab file browser interface. The breadcrumb path is `experiments / kotca_20210805 / BERT_model`. The file list contains the following items:

Name	Last Modified	File size
..	几秒钟前	
config.json	2 分钟前	1.04 kB
pytorch_model.bin	2 分钟前	438 MB
special_tokens_map.json	2 分钟前	112 B
tokenizer_config.json	2 分钟前	39 B
training_args.bin	2 分钟前	1.52 kB
vocab.txt	2 分钟前	232 kB

○ SensiMix 모델 용량 확인



The screenshot shows the JupyterLab file browser interface. The breadcrumb path is `experiments / kotca_20210805 / SensiMix_model_quantized`. The file list contains the following items:

Name	Last Modified	File size
..	几秒钟前	
config.json	1 分钟前	1.2 kB
pytorch_model.bin	1 分钟前	42.2 MB
special_tokens_map.json	1 分钟前	112 B
tokenizer_config.json	1 分钟前	151 B
vocab.txt	1 分钟前	232 kB

성적서번호 : KST-21-049

한국시험인증원
KOTCA Korea Testing
 Certification Agency

<붙임 2> 모델 실행시간 비율 시험결과 화면

○ 1회차 시험

- BERT 모델 실행시간 결과로그

```

},
"initializer_range": 0.02,
"intermediate_size": 3072,
"is_decoder": false,
"label2id": {
  "LABEL_0": 0,
  "LABEL_1": 1
},
"layer_norm_eps": 1e-12,
"length_penalty": 1.0,
"max_length": 20,
"max_position_embeddings": 512,
"model_type": "bert",
"num_attention_heads": 12,
"num_beams": 1,
"num_hidden_layers": 12,
"num_labels": 2,
"num_return_sequences": 1,
"output_attentions": false,
"output_hidden_states": false,
"output_past": true,
"pad_token_id": 0,
"pruned_heads": {},
"repetition_penalty": 1.0,
"temperature": 1.0,
"top_k": 50,
"top_p": 1.0,
"torchscript": false,
"type_vocab_size": 2,
"use_bfloat16": false,
"vocab_size": 30522
}

08/05/2021 11:24:47 - INFO - transformers.modeling_utils - loading weights file /home/piaotairen/experiments/kotca_20210805/BERT_model/pytorch_model.bin
08/05/2021 11:24:51 - INFO - __main__ - Loading features from cached file /home/piaotairen/data/glue_data_inference/QNLI/cached_dev_BERT_model_128.qnli
08/05/2021 11:25:00 - INFO - __main__ - ***** Running evaluation *****
08/05/2021 11:25:00 - INFO - __main__ - Num examples = 104743
08/05/2021 11:25:00 - INFO - __main__ - Batch size = 128
Evaluating: 100%
duration: 254330.4751544767 ms
08/05/2021 11:29:15 - INFO - __main__ - ***** Eval results *****
08/05/2021 11:29:15 - INFO - __main__ - acc = 0.5141823319935461

```

- SensiMix 모델 실행시간 결과로그

```

},
"layer_norm_eps": 1e-12,
"length_penalty": 1.0,
"max_length": 20,
"max_position_embeddings": 512,
"min_length": 0,
"model_type": "seq2seq",
"no_repeat_ngram_size": 0,
"num_8bit_layers": 2,
"num_attention_heads": 12,
"num_beams": 1,
"num_hidden_layers": 3,
"num_return_sequences": 0,
"output_attentions": false,
"output_hidden_states": false,
"output_past": true,
"pad_token_id": 0,
"pruned_heads": {},
"repetition_penalty": 1.0,
"temperature": 1.0,
"top_k": 50,
"top_p": 1.0,
"torchscript": false,
"type_vocab_size": 2,
"use_bfloat16": false,
"vocab_size": 30522
}

08/05/2021 11:30:33 - INFO - src.models.modeling_utils - loading weights file /home/piaotairen/experiments/kotca_20210805/SensiMix_model_quantized/pytorch_model.bin
08/05/2021 11:30:34 - INFO - src.models.modeling_utils - Weights from pretrained model not used in MQ(BertForSequenceClassification)inference: ['bert.embeddings.word_embeddings.q_min', 'bert.embeddings.position_embeddings.q_min', 'bert.embeddings.position_embeddings.q_max', 'bert.encoder.layer.0.attention.self.query.q_min', 'bert.encoder.layer.0.attention.self.query.q_max', 'bert.encoder.layer.0.attention.self.value.q_min', 'bert.encoder.layer.0.attention.self.value.q_max', 'bert.encoder.layer.0.attention.output.dense.q_min', 'bert.encoder.layer.0.attention.output.dense.q_max', 'bert.encoder.layer.0.intermediate.dense.q_min', 'bert.encoder.layer.0.intermediate.dense.q_max', 'bert.encoder.layer.0.output.dense.q_min', 'bert.encoder.layer.0.output.dense.q_max', 'bert.encoder.layer.1.attention.self.query.q_min', 'bert.encoder.layer.1.attention.self.query.q_max', 'bert.encoder.layer.1.attention.self.value.q_min', 'bert.encoder.layer.1.attention.self.value.q_max', 'bert.encoder.layer.1.attention.output.dense.q_min', 'bert.encoder.layer.1.attention.output.dense.q_max', 'bert.encoder.layer.1.intermediate.dense.q_min', 'bert.encoder.layer.1.intermediate.dense.q_max', 'bert.encoder.layer.1.output.dense.q_min', 'bert.encoder.layer.1.output.dense.q_max', 'bert.encoder.layer.2.attention.self.query.q_min', 'bert.encoder.layer.2.attention.self.query.q_max', 'bert.encoder.layer.2.attention.self.value.q_min', 'bert.encoder.layer.2.attention.self.value.q_max', 'bert.encoder.layer.2.attention.output.dense.q_min', 'bert.encoder.layer.2.attention.output.dense.q_max']
08/05/2021 11:30:34 - INFO - __main__ - Loading features from cached file /home/piaotairen/data/glue_data_inference/QNLI/cached_dev_SensiMix_model_quantized_128.qnli
08/05/2021 11:31:03 - INFO - __main__ - ***** Running evaluation *****
08/05/2021 11:31:03 - INFO - __main__ - Num examples = 104743
08/05/2021 11:31:03 - INFO - __main__ - Batch size = 128
Evaluating: 100%
duration: 31728.98149403564 ms
08/05/2021 11:31:35 - INFO - __main__ - ***** Eval results *****
08/05/2021 11:31:35 - INFO - __main__ - acc = 0.5157194275512444

```

성적서번호 : KST-21-049

한국시험인증원
KOTCA Korea Testing
Certification Agency

○ 2회차 시험

- BERT 모델 실행시간 결과로그

```

{
  "layer_norm_eps": 1e-12,
  "length_penalty": 1.0,
  "max_length": 20,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 12,
  "num_beams": 1,
  "num_hidden_layers": 12,
  "num_labels": 2,
  "num_return_sequences": 1,
  "output_attentions": false,
  "output_hidden_states": false,
  "output_past": true,
  "pad_token_id": 0,
  "pruned_heads": {},
  "repetition_penalty": 1.0,
  "temperature": 1.0,
  "top_k": 50,
  "top_p": 1.0,
  "torchscript": false,
  "type_vocab_size": 2,
  "use_bfloat16": false,
  "vocab_size": 30522
}

08/05/2021 11:33:01 - INFO - transformers.modeling_utils - loading weights file /home/piotairen/experiments/kotca_20210805/BERT_model/pytorch_model.bin
08/05/2021 11:33:04 - INFO - __main__ - Loading features from cached file /home/piotairen/data/glue_data_inference/QNLI/cached_dev_BERT_model_128_qnli
08/05/2021 11:33:14 - INFO - __main__ - ***** Running evaluation *****
08/05/2021 11:33:14 - INFO - __main__ - Max examples = 104743
08/05/2021 11:33:14 - INFO - __main__ - Batch size = 128
Evaluating: 100%
duration: 25458.14991760254 ms
08/05/2021 11:37:28 - INFO - __main__ - ***** Eval results *****
08/05/2021 11:37:28 - INFO - __main__ - acc = 0.5141823319935461

```

- SensiMix 모델 실행시간 결과로그

```

{
  "layer_norm_eps": 1e-12,
  "length_penalty": 1.0,
  "max_length": 20,
  "max_position_embeddings": 512,
  "min_length": 0,
  "model_type": "mpibert",
  "no_repeat_ngram_size": 0,
  "num_beams": 1,
  "num_attention_heads": 12,
  "num_hidden_layers": 3,
  "num_return_sequences": 0,
  "output_attentions": false,
  "output_hidden_states": false,
  "output_past": true,
  "pad_token_id": 0,
  "pruned_heads": {},
  "repetition_penalty": 1.0,
  "temperature": 1.0,
  "top_k": 50,
  "top_p": 1.0,
  "torchscript": false,
  "type_vocab_size": 2,
  "use_bfloat16": false,
  "vocab_size": 30522
}

08/05/2021 11:40:29 - INFO - src.models.modeling_utils - loading weights file /home/piotairen/experiments/kotca_20210805/SensiMix_model_quantized/pytorch_model.bin
08/05/2021 11:40:30 - INFO - src.models.modeling_utils - Weights from pretrained model not used in MPQbertForSequenceClassificationInference: ['bert.embeddings.word_embeddings.q_min', 'bert.embeddings.position_embeddings.q_min', 'bert.encoder.layer.0.attention.self.query.q_min', 'bert.encoder.layer.0.attention.self.query.q_max', 'bert.encoder.layer.0.attention.output.dense.q_min', 'bert.encoder.layer.0.attention.output.dense.q_max', 'bert.encoder.layer.0.intermediate.dense.q_min', 'bert.encoder.layer.0.intermediate.dense.q_max', 'bert.encoder.layer.0.output.dense.q_min', 'bert.encoder.layer.0.output.dense.q_max', 'bert.encoder.layer.1.attention.self.query.q_min', 'bert.encoder.layer.1.attention.self.query.q_max', 'bert.encoder.layer.1.attention.output.dense.q_min', 'bert.encoder.layer.1.attention.output.dense.q_max', 'bert.encoder.layer.1.intermediate.dense.q_min', 'bert.encoder.layer.1.intermediate.dense.q_max', 'bert.encoder.layer.1.output.dense.q_min', 'bert.encoder.layer.1.output.dense.q_max', 'bert.encoder.layer.2.attention.self.query.q_min', 'bert.encoder.layer.2.attention.self.query.q_max', 'bert.encoder.layer.2.attention.output.dense.q_min', 'bert.encoder.layer.2.attention.output.dense.q_max']
08/05/2021 11:40:30 - INFO - __main__ - Loading features from cached file /home/piotairen/data/glue_data_inference/QNLI/cached_dev_SensiMix_model_quantized_128_qnli
08/05/2021 11:40:39 - INFO - __main__ - ***** Running evaluation *****
08/05/2021 11:40:39 - INFO - __main__ - Max examples = 104743
08/05/2021 11:40:39 - INFO - __main__ - Batch size = 128
Evaluating: 100%
duration: 29394.832649502563 ms
08/05/2021 11:41:09 - INFO - __main__ - ***** Eval results *****
08/05/2021 11:41:09 - INFO - __main__ - acc = 0.515719475512444

```

성적서번호 : KST-21-049

한국시험인증원
KOTCA Korea Testing
 Certification Agency

○ 3회차 시험

- BERT 모델 실행시간 결과로그

```

},
"layer_norm_eps": 1e-12,
"length_penalty": 1.0,
"max_length": 20,
"max_position_embeddings": 512,
"model_type": "bert",
"num_attention_heads": 12,
"num_beams": 1,
"num_hidden_layers": 12,
"num_labels": 2,
"num_return_sequences": 1,
"output_attentions": false,
"output_hidden_states": false,
"output_past": true,
"pad_token_id": 0,
"pruned_heads": {},
"repetition_penalty": 1.0,
"temperature": 1.0,
"top_k": 50,
"top_p": 1.0,
"torchscript": false,
"type_vocab_size": 2,
"use_bfloat16": false,
"vocab_size": 30522
}

08/05/2021 11:41:37 - INFO - transformers.modeling_utils - loading weights file /home/piatairen/experiments/kotca_20210805/BERT_model/pytorch_model.bin
08/05/2021 11:41:40 - INFO - __main__ - Loading features from cached file /home/piatairen/data/glue_data_inference/QNLI/cached_dev_BERT_model_128_qnli
08/05/2021 11:41:50 - INFO - __main__ - ***** Running evaluation *****
08/05/2021 11:41:50 - INFO - __main__ - Num examples = 104743
08/05/2021 11:41:50 - INFO - __main__ - Batch size = 128

Evaluating: 100%
duration: 254697.63898649447 ms
08/05/2021 11:46:05 - INFO - __main__ - ***** Eval results *****
08/05/2021 11:46:05 - INFO - __main__ - acc = 0.5141823319935461

```

- SensiMix 모델 실행시간 결과로그

```

},
"layer_norm_eps": 1e-12,
"length_penalty": 1.0,
"max_length": 20,
"max_position_embeddings": 512,
"min_length": 0,
"model_type": "mpibert",
"no_repeat_ngram_size": 0,
"num_beat_layers": 2,
"num_attention_heads": 12,
"num_beams": 1,
"num_hidden_layers": 3,
"num_return_sequences": 0,
"output_attentions": false,
"output_hidden_states": false,
"output_past": true,
"pad_token_id": 0,
"pruned_heads": {},
"repetition_penalty": 1.0,
"temperature": 1.0,
"top_k": 50,
"top_p": 1.0,
"torchscript": false,
"type_vocab_size": 2,
"use_bfloat16": false,
"vocab_size": 30522
}

08/05/2021 11:47:44 - INFO - src.models.modeling_utils - loading weights file /home/piatairen/experiments/kotca_20210805/SensiMix_model_quantized/pytorch_model.bin
08/05/2021 11:47:44 - INFO - src.models.modeling_utils - Weights from pretrained model not used in MPQbertForSequenceClassificationInference: ['bert.embeddings.word_embeddings.position_embeddings.q_min', 'bert.embeddings.position_embeddings.q_max', 'bert.encoder.layer.0.attention.self.query.q_min', 'bert.encoder.layer.0.attention.self.query.q_max', 'bert.encoder.layer.0.attention.self.value.q_min', 'bert.encoder.layer.0.attention.self.value.q_max', 'bert.encoder.layer.0.attention.output.dense.q_min', 'bert.encoder.layer.0.attention.output.dense.q_max', 'bert.encoder.layer.1.attention.self.query.q_min', 'bert.encoder.layer.1.attention.self.query.q_max', 'bert.encoder.layer.1.attention.self.key.q_min', 'bert.encoder.layer.1.attention.self.key.q_max', 'bert.encoder.layer.1.attention.self.value.q_min', 'bert.encoder.layer.1.attention.self.value.q_max', 'bert.encoder.layer.1.attention.output.dense.q_min', 'bert.encoder.layer.1.attention.output.dense.q_max', 'bert.encoder.layer.2.attention.self.query.q_min', 'bert.encoder.layer.2.attention.self.query.q_max', 'bert.encoder.layer.2.attention.self.key.q_min', 'bert.encoder.layer.2.attention.self.key.q_max', 'bert.encoder.layer.2.attention.self.value.q_min', 'bert.encoder.layer.2.attention.self.value.q_max', 'bert.encoder.layer.2.attention.output.dense.q_min', 'bert.encoder.layer.2.attention.output.dense.q_max']
08/05/2021 11:47:44 - INFO - __main__ - Loading features from cached file /home/piatairen/data/glue_data_inference/QNLI/cached_dev_SensiMix_model_quantized_128_qnli
08/05/2021 11:47:54 - INFO - __main__ - ***** Running evaluation *****
08/05/2021 11:47:54 - INFO - __main__ - Num examples = 104743
08/05/2021 11:47:54 - INFO - __main__ - Batch size = 128

Evaluating: 100%
duration: 31566.34022161391 ms
08/05/2021 11:48:26 - INFO - __main__ - ***** Eval results *****
08/05/2021 11:48:26 - INFO - __main__ - acc = 0.5157194275512444

```