AT 82.05 NLU

Dr Chaklam Silpasuwanchai

Intention Detection and Question Answering

By
Min Khant Soe (st122277)
Pyae Sone (st122645)

Objectives

- To learn how intent detection and question answering works.

- In intent detection, will the model “ Cross Entropy Loss + Supervised Contrastive Learning Loss” beat the model with “Cross Entropy Loss” alone as the proof in the paper?

- Compare the results of the models

Intent Classification

Dataset

Banking77

Dataset	Data	Label
Train	10003	77
Val	3080	77

Author:
Inigo Casanueva and Tadas Temcinas and Daniela Gerz and Matthew Henderson and Ivan Vulic,

Title:
Efficient Intent Detection with Dual Sentence Encoders
[2003.04807] Efficient Intent Detection with Dual Sentence Encoders

Date: Mar 2020

Book Title:
Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020

Data available at:
PolyAI-LDN/task-specific-datasets

Models:

RoBERTa and XLNET

RoBERTa

A Robustly Optimized BERT Training Approach:
[1907.11692] RoBERTa: A Robustly Optimized BERT Pretraining Approach (arxiv.org)

- Replication study of BERT pretraining
- Hyperparameter choices have significant impact on the final results
- Remove the **Next Sentence Prediction** (NSP) and Train with Full Sentences
- **Dynamic Masking**: dynamically change the masking pattern applied to the training data

RoBERTa

Benchmarking Commercial Intent Detection Services with Practice-Driven Evaluations:
[Paper tables with annotated results for Benchmarking Commercial Intent Detection Services with Practice-Driven Evaluations | Papers With Code]

	CLINC150	HWU64	BANKING77	Average
WA classic	93.9	88.8	90.6	91.1
WA enhanced	95.7	90.5	92.6	92.9
RASA	89.4	84.9	89.9	88.1
Distilbert-base	96.3	91.7	92.1	93.4
BERT-base	96.8	91.6	93.3	93.9
BERT-large	**97.1**	91.9	93.7	94.2
USE-base	94.7	88.9	89.9	91.2
RoBERTa-base	97.0	**92.1**	**94.1**	**94.4**

Table 1: **Accuracy on CLINC150, HWU64 and BANKING77 for Watson Assistant (WA), RASA and pretrained LMs.** Training is performed on the full train sets and evaluation on full test sets.

Algorithm	Resources	CLINC150 Training time	HWU Training time	BANKING77 Training time
WA classic	-	**1.04**	0.85	**0.64**
WA enhanced	-	1.81	**0.82**	1.22
RASA	GPU	13.93	9.43	15.45
Distilbert-base	GPU	35.98	20.35	20.35
BERT-base	GPU	71.08	39.48	38.75
BERT-large	GPU	270	175	175
USE-base	GPU	14.73	8.92	9.47
RoBERTa-base	GPU	90	60	57

Table 5: **Training time (in minutes) and resource requirements** for Watson Assistant (WA), RASA and pretrained LMs. Training is performed on full training sets. All methods except for Watson Assistant are trained using a single NVIDIA K80 GPU.

XLNET

XLNet: Generalized Autoregressive Pretraining for Language Understanding

[1906.08237] XLNet: Generalized Autoregressive Pretraining for Language Understanding

- Retains the benefits of Autoregressive **(AR)** language model while having it learn from bidirectional context as Autoencoding **(AE)** models
- Maximizes the expected log likelihood of a sequence with respect to all possible permutations of the factorization order
- Does not rely on data corruption: **no masks**
- Provides a natural way to use the product rule for factorizing the joint probability of the predicted tokens
- Eliminate the independence assumption

XLNET

XLNet: Generalized Autoregressive Pretraining for Language Understanding

[1906.08237] XLNet: Generalized Autoregressive Pretraining for Language Understanding

Model	IMDB	Yelp-2	Yelp-5	DBpedia	AG	Amazon-2	Amazon-5
CNN [15]	-	2.90	32.39	0.84	6.57	3.79	36.24
DPCNN [15]	-	2.64	30.58	0.88	6.87	3.32	34.81
Mixed VAT [31, 23]	4.32	-	-	0.70	4.95	-	-
ULMFiT [14]	4.6	2.16	29.98	0.80	5.01	-	-
BERT [35]	4.51	1.89	29.32	0.64	-	2.63	34.17
XLNet	**3.20**	**1.37**	**27.05**	**0.60**	**4.45**	**2.11**	**31.67**

Table 4: Comparison with state-of-the-art error rates on the test sets of several text classification datasets. All BERT and XLNet results are obtained with a 24-layer architecture with similar model sizes (aka BERT-Large).

SUPERVISED CONTRASTIVE LEARNING FOR PRE-TRAINED LANGUAGE MODEL FINE-TUNING: https://arxiv.org/pdf/2011.01403.pdf

- In Paper; compared ROBERTA with CE alone and with CE + SCL and proved that CE + SCL is better.
- So, we apply this on both ROBERTA and XLNet and check if it improves the models' accuracy.

Cross Entropy loss

$$\mathcal{L}_{\text{CE}} = -\frac{1}{m}\sum_{i=1}^{m} yi \cdot log(\hat{yi})$$

Supervised Contrastive learning loss

$$\mathcal{L}_{\text{S\_cl}} = -\frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{1}_{yi=yj} \ log\frac{e^{sim(hi,hj)/\tau}}{\sum_{n=1}^{N} e^{sim(hi,hn)/\tau}}$$

- detail
  - ui ~ sentence i
  - hi ~ BERT(ui) in our case using Roberta as a encoder
  - hi : (batch_size,sequence_len,embed_size)
  - hi is the output of model which is last hidden layers before classifier head in the model architecture
  - 1yi=yj ~ we select only the sample that come from the same class to compute in each i and j
  - T ~ the number of pairs that come from the same classes
  - τ ~ temperature parameter
  - Sim(x1,x2) : cosine similarity [-1, 1]
  - λ' is just weighted of cross entropy loss
  - Sim function is the cosine similarity
  - N ~ the number of samples in a batch

$$'sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Loss total

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{s\_cl}} + \lambda' \mathcal{L}_{CE}$$

Model Training

XLNet - **xlnet-base-cased**

12-layer, 768-hidden, 12-heads, 110M parameters.

batch_size = 4
lr= 1e-5
epoch: 5
optimizer: AdamW

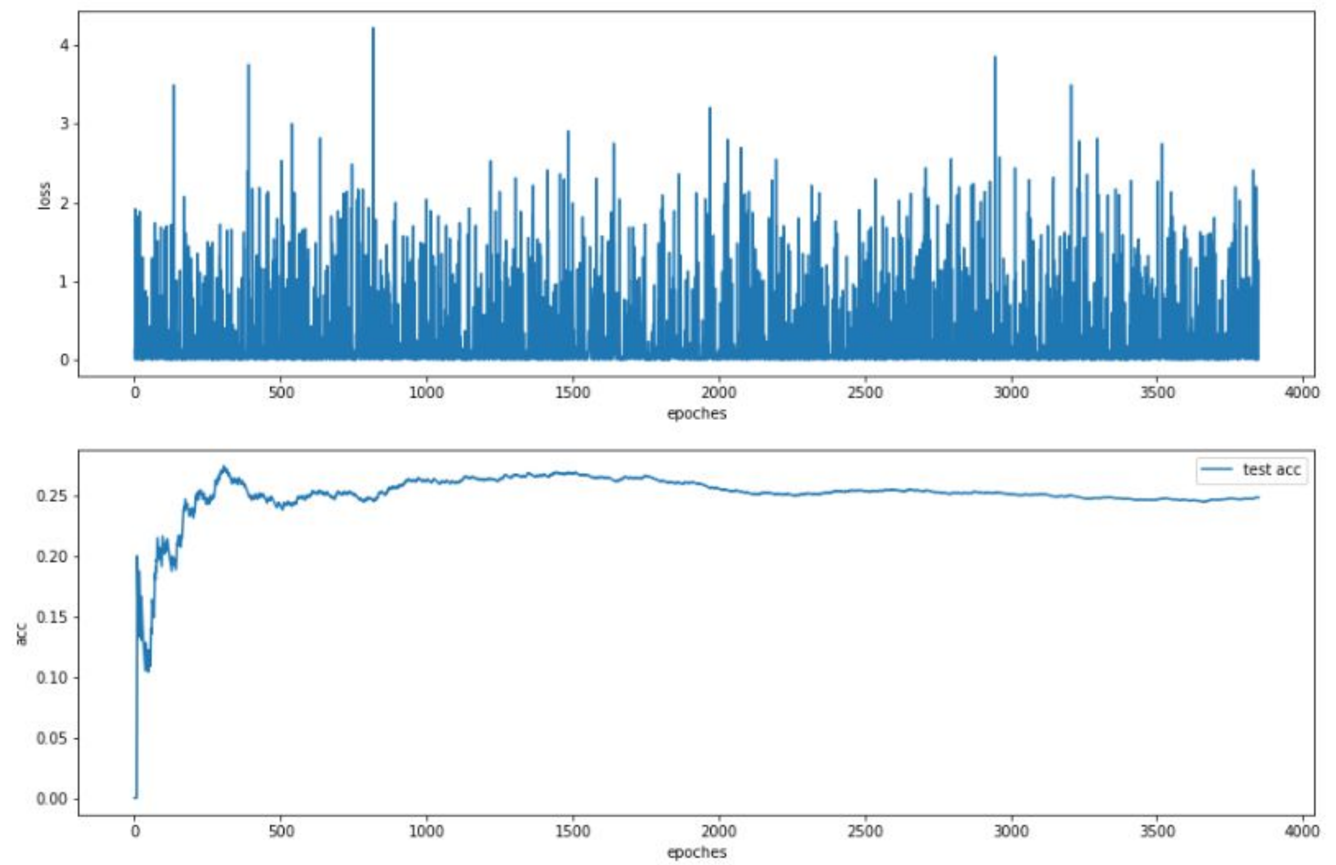
RoBERTa - **roberta-base**

12-layer, 768-hidden, 12-heads, 125M parameters

batch_size = 4
lr= 1e-5
epoch: 5
optimizer: AdamW

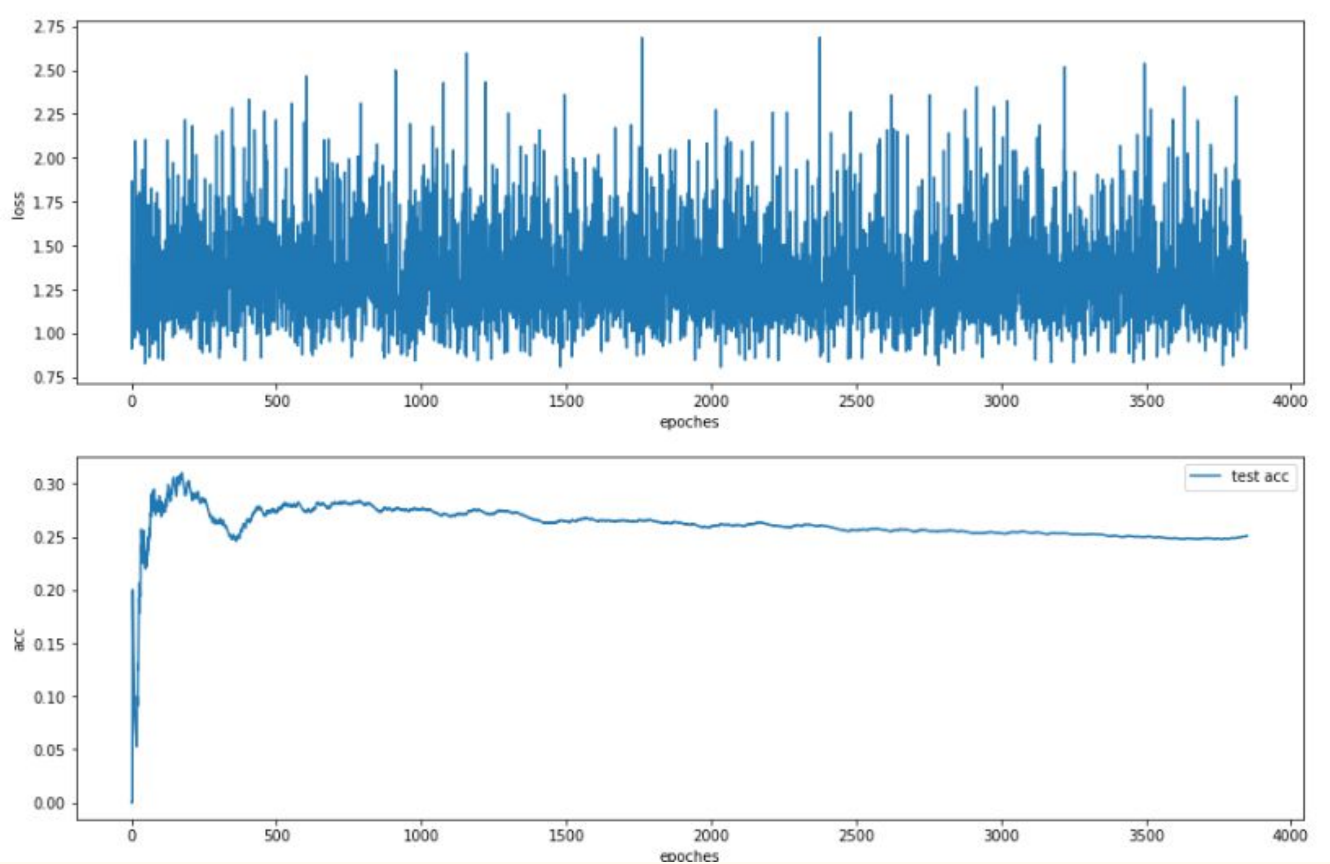
XLNET (Cross-Entropy Loss)

Validation Loss and Accuracy



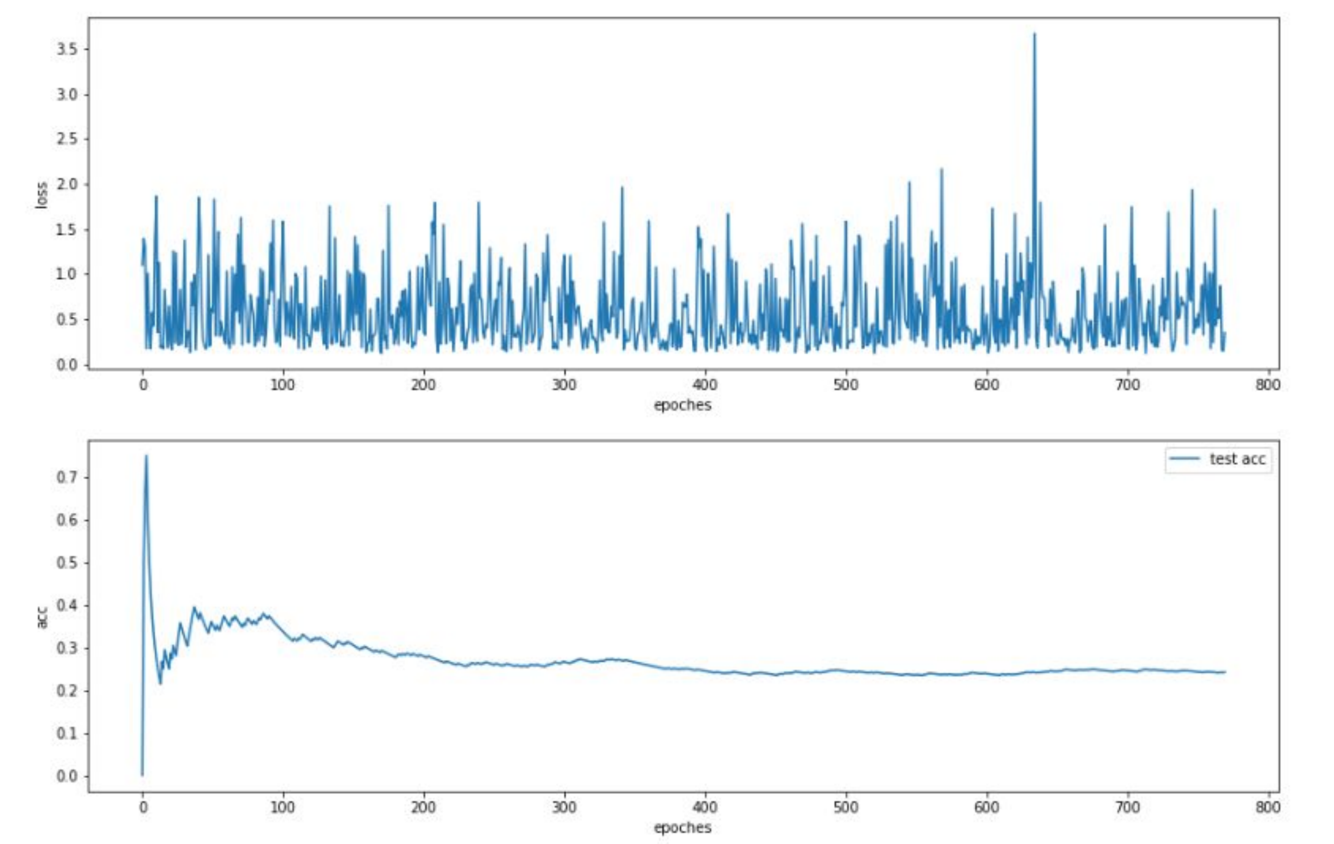
XLNET (Cross-Entropy Loss + Supervised Contrastive Loss)

Validation Loss and Accuracy



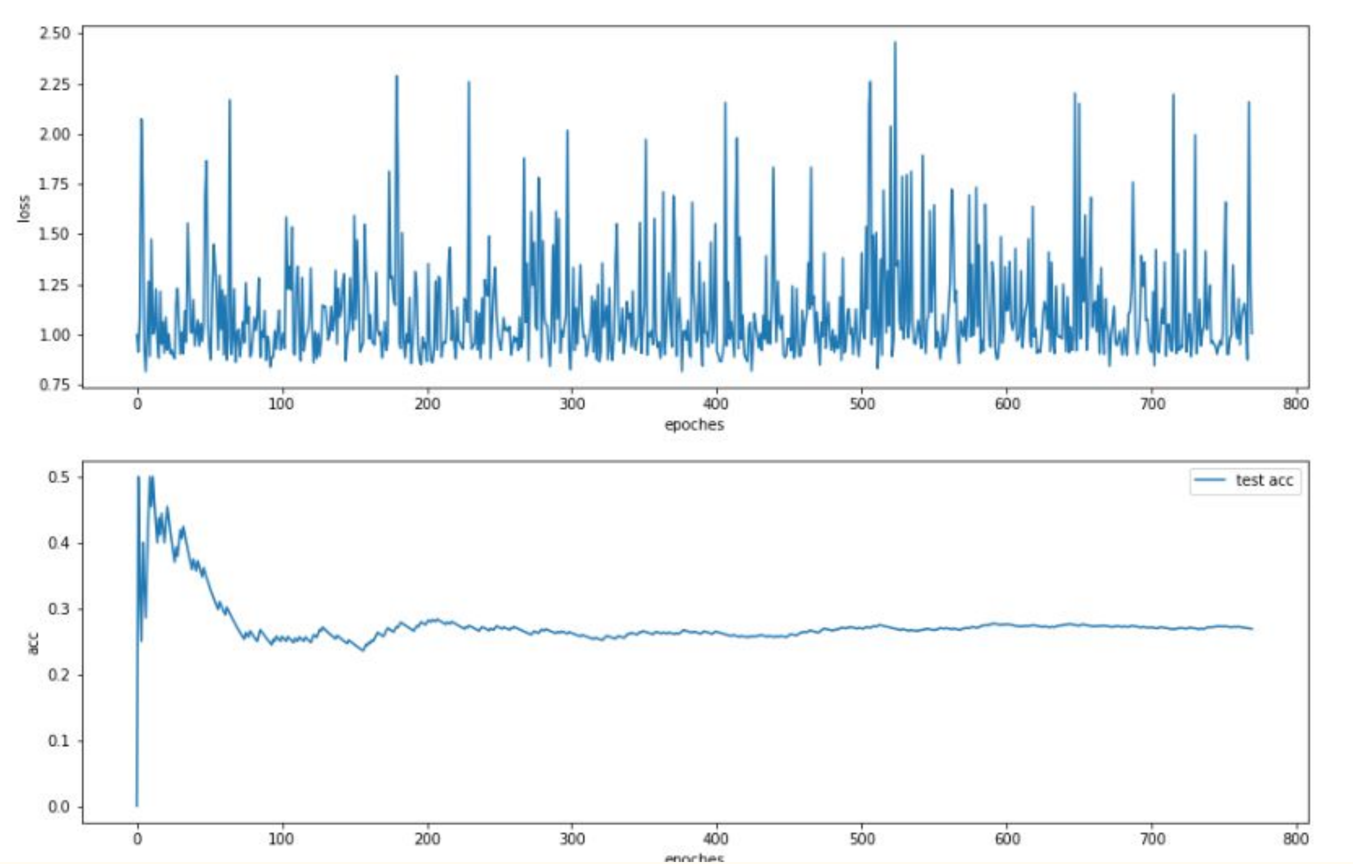
RoBERTa (Cross-Entropy Loss)

Validation Loss and Accuracy



RoBERTa (Cross-Entropy Loss + Supervised Contrastive Loss)

Validation Loss and Accuracy



Results

Transformer	Validation Accuracy (%)
XLNET (Cross Entropy Loss)	24.83 %
XLNET (Cross-Entropy Loss + Supervised Contrastive Loss)	25.09 %
RoBERTa (Cross-Entropy Loss)	24.16 %
RoBERTa (Cross-Entropy Loss + Supervised Contrastive Loss)	**26.88 %**

- Both models increase accuracy with combined loss.
- With cross entropy loss alone, XLNet is slightly higher than RoBERTa while with loss combination, RoBERTa is noticeably higher than XLNet.

Labels in Dataset


```
{'Refund_not_showing_up': 0, 'activate_my_card': 1, 'age_limit': 2, 'apple_pay_or_google_pay': 3,
'atm_support': 4, 'automatic_top_up': 5, 'balance_not_updated_after_bank_transfer': 6,
'balance_not_updated_after_cheque_or_cash_deposit': 7, 'beneficiary_not_allowed': 8, 'cancel_transfer': 9,
'card_about_to_expire': 10, 'card_acceptance': 11, 'card_arrival': 12, 'card_delivery_estimate': 13,
'card_linking': 14, 'card_not_working': 15, 'card_payment_fee_charged': 16, 'card_payment_not_recognised':
17, 'card_payment_wrong_exchange_rate': 18, 'card_swallowed': 19, 'cash_withdrawal_charge': 20,
'cash_withdrawal_not_recognised': 21, 'change_pin': 22, 'compromised_card': 23, 'contactless_not_working':
24, 'country_support': 25, 'declined_card_payment': 26, 'declined_cash_withdrawal': 27, 'declined_transfer':
28, 'direct_debit_payment_not_recognised': 29, 'disposable_card_limits': 30, 'edit_personal_details': 31,
'exchange_charge': 32, 'exchange_rate': 33, 'exchange_via_app': 34, 'extra_charge_on_statement': 35,
'failed_transfer': 36, 'fiat_currency_support': 37, 'get_disposable_virtual_card': 38, 'get_physical_card':
39, 'getting_spare_card': 40, 'getting_virtual_card': 41, 'lost_or_stolen_card': 42, 'lost_or_stolen_phone':
43, 'order_physical_card': 44, 'passcode_forgotten': 45, 'pending_card_payment': 46,
'pending_cash_withdrawal': 47, 'pending_top_up': 48, 'pending_transfer': 49, 'pin_blocked': 50,
'receiving_money': 51, 'request_refund': 52, 'reverted_card_payment?': 53, 'supported_cards_and_currencies':
54, 'terminate_account': 55, 'top_up_by_bank_transfer_charge': 56, 'top_up_by_card_charge': 57,
'top_up_by_cash_or_cheque': 58, 'top_up_failed': 59, 'top_up_limits': 60, 'top_up_reverted': 61,
'topping_up_by_card': 62, 'transaction_charged_twice': 63, 'transfer_fee_charged': 64,
'transfer_into_account': 65, 'transfer_not_received_by_recipient': 66, 'transfer_timing': 67,
'unable_to_verify_identity': 68, 'verify_my_identity': 69, 'verify_source_of_funds': 70, 'verify_top_up': 71,
'virtual_card_not_working': 72, 'visa_or_mastercard': 73, 'why_verify_identity': 74,
'wrong_amount_of_cash_received': 75, 'wrong_exchange_rate_for_cash_withdrawal': 76}
```

Testing XLNET Model

(Cross Entropy Loss)

```
Text  1 :  I want to deposit
predicted intent: transfer_into_account
Text  2 :  I want to withdraw
predicted intent: declined_cash_withdrawal
Text  3 :  I want to check account
predicted intent: why_verify_identity
Text  4 :  I want to transfer money
predicted intent: transfer_into_account
```

(Cross-Entropy Loss + Supervised Contrastive Loss)

```
Text  1 :  I want to deposit
predicted intent: transfer_into_account
Text  2 :  I want to withdraw
predicted intent: pending_cash_withdrawal
Text  3 :  I want to check account
predicted intent: why_verify_identity
Text  4 :  I want to transfer money
predicted intent: transfer_into_account
```

Testing ROBERTA Model

(Cross Entropy Loss)

```
Text  1 :  I want to deposit
predicted intent: top_up_by_cash_or_cheque
Text  2 :  I want to withdraw
predicted intent: declined_cash_withdrawal
Text  3 :  I want to check account
predicted intent: why_verify_identity
Text  4 :  I want to transfer money
predicted intent: receiving_money
```

(Cross-Entropy Loss + Supervised Contrastive Loss)

```
Text  1 :  I want to deposit
predicted intent: top_up_by_cash_or_cheque
Text  2 :  I want to withdraw
predicted intent: declined_cash_withdrawal
Text  3 :  I want to check account
predicted intent: why_verify_identity
Text  4 :  I want to transfer money
predicted intent: receiving_money
```

Question Answering

Question Answering Method We Used

- Closed Domain
  - Closed-domain question answering **deals with questions under a specific domain** (for example, medicine or automotive maintenance), and can exploit domain-specific knowledge frequently formalized in ontologies.

- Extractive
  - extract the answer from the given context.

Dataset

Health_QA @AEOP RESEARCH

Dataset	Question	Context	Answers
Train	1615	1615	1615
Val	404	404	404

https://github.com/Akomand/AEOP_Research_2021/tree/main/Week4/datasets/covidqa

Data External Source (For Testing)

HIV: https://www.health.ny.gov/publications/0213.pdf

Covid19: https://www.cdc.gov/coronavirus/2019-ncov/faq.html

Context

['COVID-19 is a disease caused by a virus called SARS-CoV-2. Most people with COVID-19 have mild symptoms, but some people become severely ill. Older adults and people who have certain underlying medical conditions are more likely to get severely ill. Post-COVID conditions are a wide range of health problems people can experience four or more weeks after first getting COVID-19. ', 'HIV (Human Immunodeficiency Virus) is a virus that only affects human beings. AIDS (Acquired Immune Deficiency Syndrome) is a late stage of HIV disease. A person develops HIV if the virus gets into his or her bloodstream and begins making more and more of itself, or reproducing. People living with HIV may have no symptoms for ten or more years. ']

Question

['What is Coivd19?', 'What is AIDS?']

Answers

[{'text': 'COVID-19 is a disease caused by a virus called SARS-CoV-2.', 'answer_start': 0, 'answer_end': 58}, {'text': 'AIDS (Acquired Immune Deficiency Syndrome) is a late stage of HIV disease.', 'answer_start': 78, 'answer_end': 152}]

Models:

RoBERTa and DistilBERT

DistilBERT

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter:
https://arxiv.org/pdf/1910.01108.pdf

- DistilBERT is distilled on very large batches leveraging gradient accumulation (up to 4K examples per batch) using dynamic masking and without the next sentence prediction objective.

- Most of the operations used in the Transformer architecture (linear layer and layer normalisation) are highly optimized in modern linear algebra frameworks.

- Variations on the last dimension of the tensor (hidden size dimension) have a smaller impact on computation efficiency (for a fixed parameters budget) than variations on other factors like the number of layers.

Model Training

RoBERTa - **roberta-base**

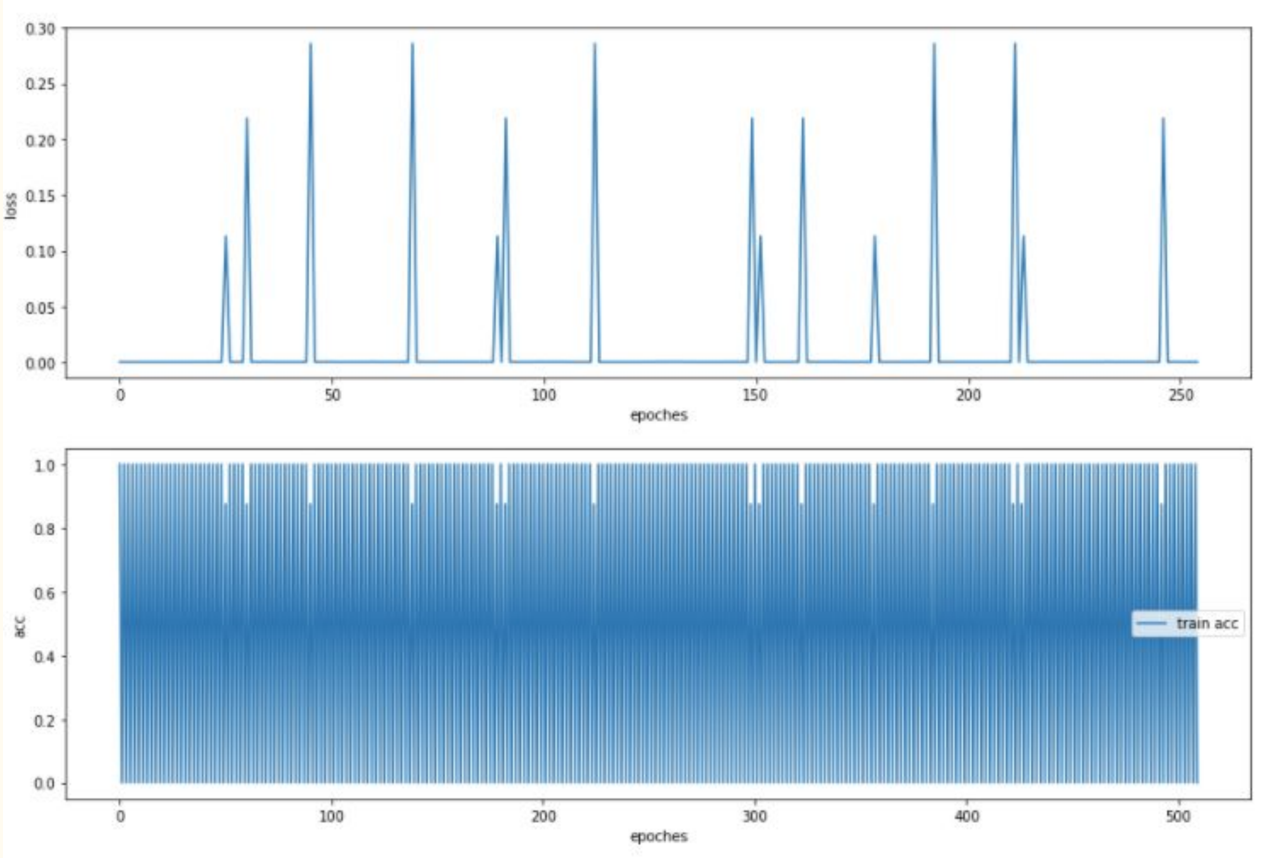
12-layer, 768-hidden, 12-heads, 125M parameters
lr= 1e-5
epoch: 5
optimizer: AdamW

DistilBERT - **distilbert-base-uncased**

6-layer, 768-hidden, 12-heads, 66M parameters
lr= 1e-5
epoch: 5
optimizer: AdamW

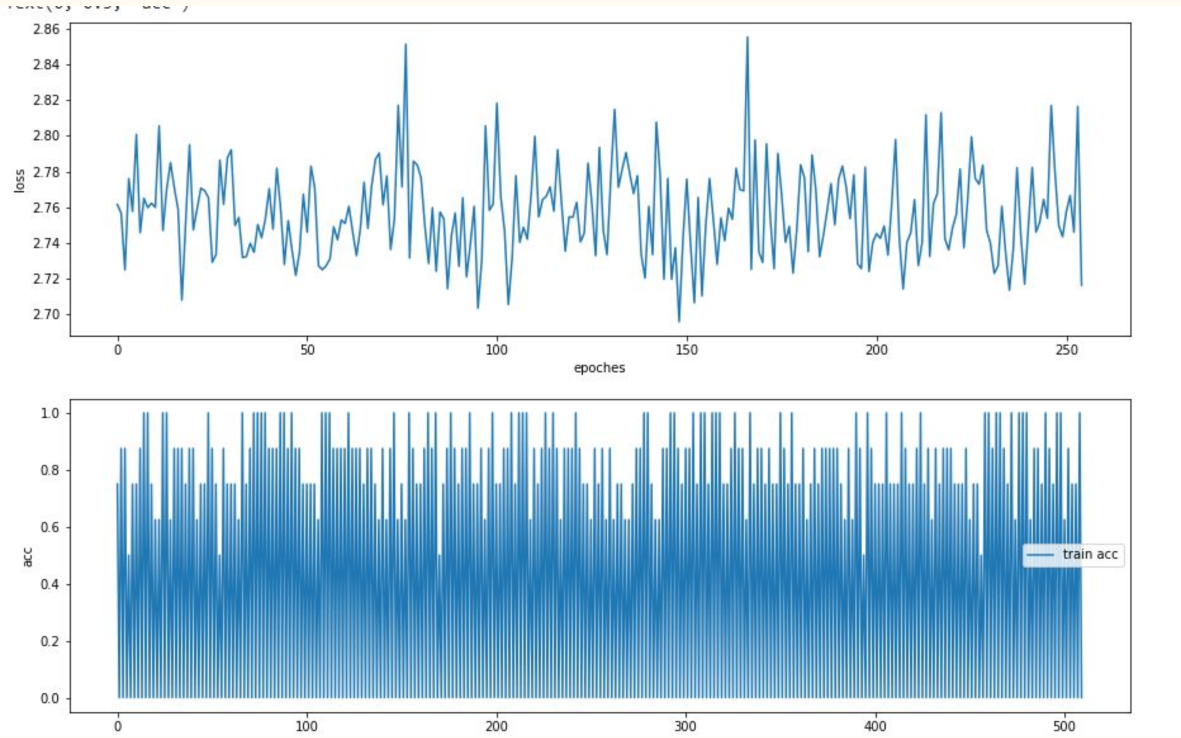
RoBERTa

Validation Loss and Accuracy



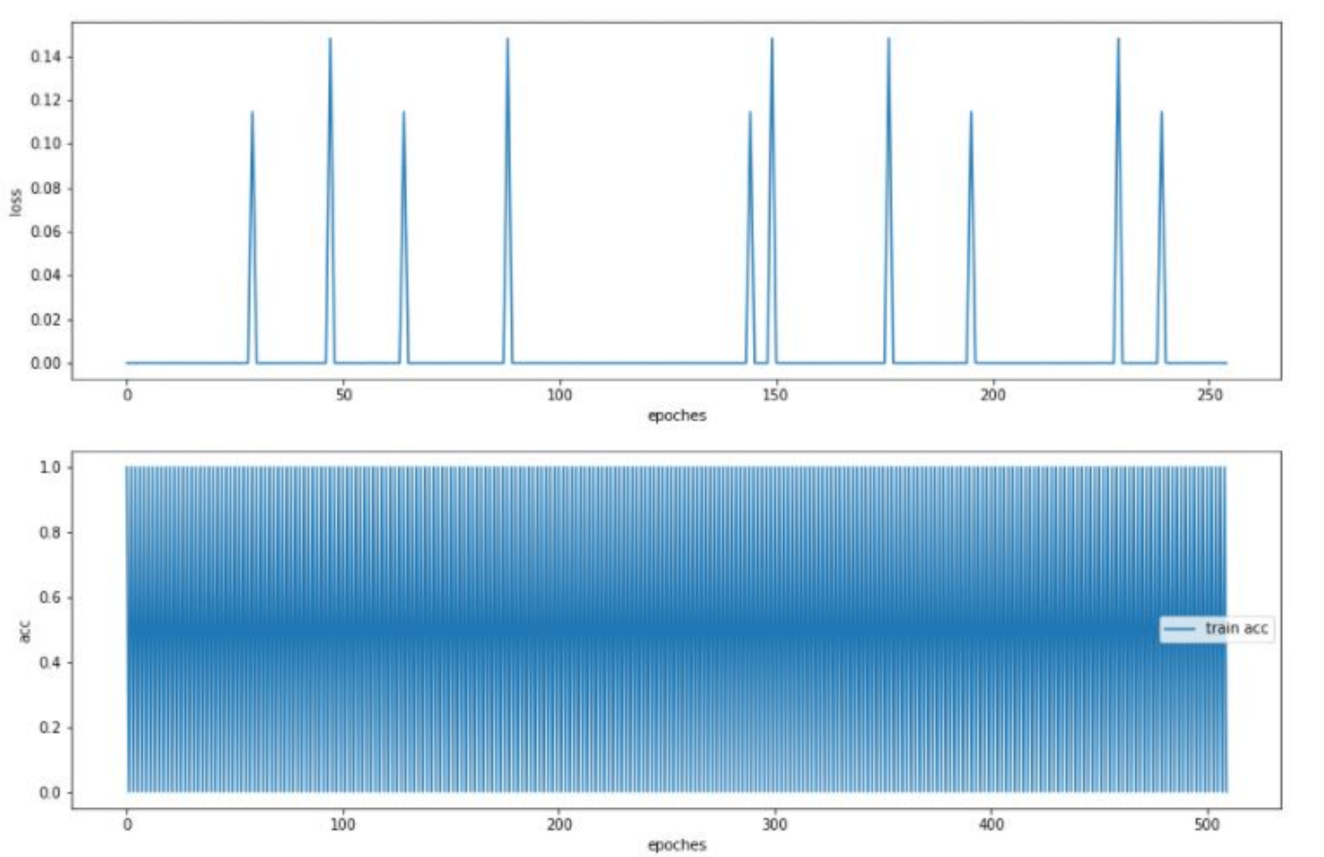
ROBERTA (Freeze All Layers)

Validation Loss and Accuracy



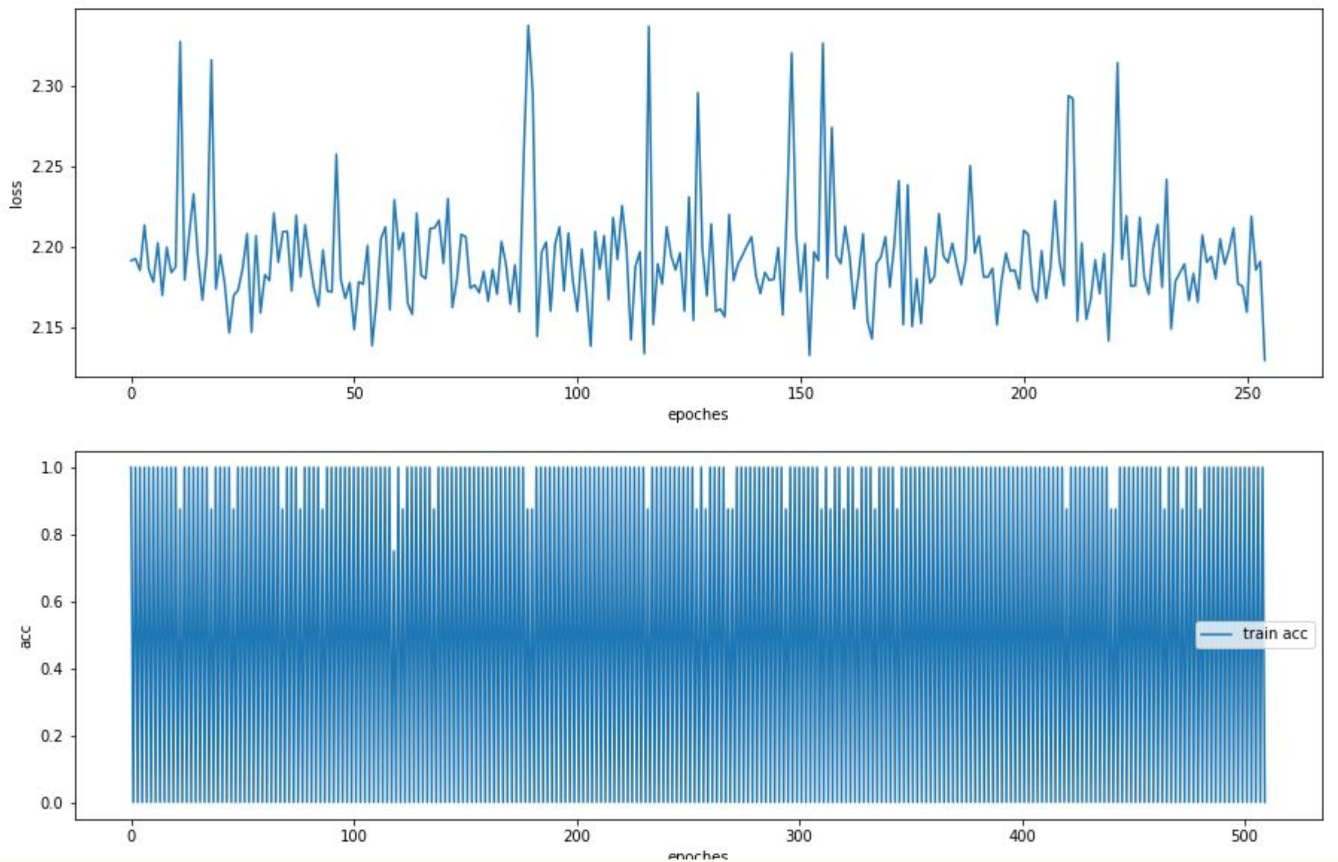
DistilBERT

Validation Loss and Accuracy



DistilBERT (Freeze All Layers)

Validation Loss and Accuracy



Results

Transformer	Validation Accuracy (%)
RoBERTa	49.75%
RoBERTa (Freeze all layers)	41.74 %
DistilBERT	50.00 %
DistilBERT (Freeze all layers)	49.26 %

- In this case, DistilBERT has slightly higher performance than RoBERTa.

Testing ROBERTA Model

Testing data from validation set

```
Question  1 :
 What should be investigated in the future?

Predicted Answer  1 :
 </s>Respir

Question  2 :
 What further can viral persistence lead to?

Predicted Answer  2 :
  components remaining in
```

Testing data from external source

```
Question  1 :
 What is Coivd19?

Predicted Answer  1 :
 ivd19

Question  2 :
 What is AIDS?

Predicted Answer  2 :
 HIV (
```

Testing ROBERTA (Freeze All Layers) Model

Testing data from validation set

```
Question  1 :
 What should be investigated in the future?

Predicted Answer  1 :
 </s>Respiratory Viral Infections in Exacerbation of Chronic Airway Inflammatory Diseases: Novel Mechanisms and Insights From the Upper
Airway Epithelium
```

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7052386/

```
SHA: 45a566c71056ba4faab425b4f7e9edee6320e4a4

Authors: Tan, Kai Sen; Lim, Rachel Liyu; Liu, Jing; Ong, Hsiao Hui; Tan, Vivian Jiayi; Lim, Hui Fang; Chung, Kian Fan; Adcock, Ian M.;
Chow, Vincent T.; Wang, De Yun
Date: 2020-02-25
DOI: 10.3389/fcell.

Question  2 :
 What further can viral persistence lead to?
Predicted Answer  2 :
  components remaining in the airway?</s></s>Respiratory Viral Infections in Exacerbation of Chronic Airway Inflammatory Diseases: Novel
Mechanisms and Insights From the Upper Airway Epithelium
```

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7052386/

```
SHA: 45a566c71056ba4faab425b4f7e9edee6320e4a4

Authors: Tan, Kai Sen; Lim, Rachel Liyu; Liu, Jing; Ong, Hsiao Hui; Tan, Vivian Jiayi; Lim, Hui Fang; Chung, Kian Fan; Adcock, Ian M.;
Chow, Vincent T.; Wang, De Yun
Date: 2020-02-25
DOI:
```

Testing ROBERTA (Freeze All Layers) Model

Testing data from external source

```
Question  1 :
 What is Coivd19?

Predicted Answer  1 :
 ?</s></s>COVID-19 is a disease caused by a virus called SARS-CoV-2. Most people with COVID-19 have
mild symptoms, but some people become severely ill. Older adults and people who have certain

Question  2 :
 What is AIDS?

Predicted Answer  2 :
 ?</s></s>HIV (Human Immunodeficiency Virus) is a virus that only affects human beings. AIDS
(Acquired Immune Deficiency Syndrome) is a late stage of HIV disease. A person develops HIV if the
virus gets into his or her bloodstream and begins making more and more of itself, or reproducing.
People living with HIV may have no symptoms for ten or more years.
```

Testing DISTILBERT Model

Testing data from validation set

```
Question  1 :
 What should be investigated in the future?

Predicted Answer  1 :
 viral

Question  2 :
 What further can viral persistence lead to?

Predicted Answer  2 :
 remaining
```

Testing data from external source

```
Question  1 :
 What is Coivd19?

Predicted Answer  1 :
 ##iv

Question  2 :
 What is AIDS?

Predicted Answer  2 :
 (
```

Testing DISTILBERT (Freeze All Layers) Model

Testing data from validation set

```
Question  1 :
 What should be investigated in the future?

Predicted Answer  1 :
 viral infections in exacerbation of chronic airway inflammatory diseases : novel mechanisms and insights from the upper airway epithelium https : / / www. ncbi. nlm. nih. gov / pmc / articles / pmc7052386 / sha : 45a566c71056ba4faab425b4f7e9edee6320e4a4 authors : tan, kai sen ; lim, rachel liyu ; liu, jing ; ong, hsiao hui ; tan, vivian jiayi ; lim, hui fang ; chung, kian fan ; adcock, ian m. ; chow, vincent t. ; wang, de yun date : 2020 - 02 - 25 doi : 10. 3389 / fcell. 2020. 00099 license : cc - by abstract : respiratory virus infection is one of the major sources of exacerbation of chronic airway inflammatory diseases. these exacerbations are associated with high morbidity and even mortality worldwide. the current understanding on viral - induced exacerbations is that viral infection increases airway inflammation which aggravates disease symptoms. recent advances in in vitro air - liquid interface 3d cultures, organoid cultures and the use of novel human and animal challenge models have evoked new understandings as to the mechanisms of viral exace

Question  2 :
 What further can viral persistence lead to?

Predicted Answer  2 :
 remaining in the airway? [SEP] respiratory viral infections in exacerbation of chronic airway inflammatory diseases : novel mechanisms and insights from the upper airway epithelium https : / / www. ncbi. nlm. nih. gov / pmc / articles / pmc7052386 / sha : 45a566c71056ba4faab425b4f7e9edee6320e4a4 authors : tan, kai sen ; lim, rachel liyu ; liu, jing ; ong, hsiao hui ; tan, vivian jiayi ; lim, hui fang ; chung, kian fan ; adcock, ian m. ; chow, vincent t. ; wang, de yun date : 2020 - 02 - 25 doi : 10. 3389 / fcell. 2020. 00099 license : cc - by abstract : respiratory virus infection is one of the major sources of exacerbation of chronic airway inflammatory diseases. these exacerbations are associated with high morbidity and even mortality worldwide. the current understanding on viral - induced exacerbations is that viral infection increases airway inflammation which aggravates disease symptoms. recent advances in in vitro air - liquid interface 3d cultures, organoid cultures and the use of novel human and animal challenge models have evoked new understandings
```

Testing DISTILBERT (Freeze All Layers) Model

Testing data from external source

```
Question  1 :
 What is Coivd19?

Predicted Answer  1 :
 ? [SEP] covid - 19 is a disease caused by a virus called sars - cov - 2. most people with covid - 19
have mild symptoms, but some people become severely ill. older adults and people who have certain
underlying medical conditions are more likely to get severely ill. post - covid conditions are a wide
range of health problems people can experience four or more weeks after first getting

Question  2 :
 What is AIDS?

Predicted Answer  2 :
 ? [SEP] hiv ( human immunodeficiency virus ) is a virus that only affects human beings. aids (
acquired immune deficiency syndrome ) is a late stage of hiv disease. a person develops hiv if the
virus gets into his or
```

Conclusion

In intent detection, we proved that combining Supervised Contrastive Loss with Cross-Entropy Loss improves the accuracy of the models.

In Question Answering, we built a chatbot system with two different pretrained transformers. Although accuracy is not good, it can be improved when we have deeper knowledge about dialogue system.

From this project, we understood more about pre-trained transformers and their architecture, and learnt how to implement pretrained models for intent detection and question answering tasks.

We also learnt that Reinforcement Learning can be applied to improve the performance of the dialogue system.

Although this project is challenging and difficult for us, it is fun to learn and create a chatbot.

Reference

A Comparative Study of Transformer-Based Language Models on Extractive Question Answering:
https://arxiv.org/pdf/2110.03142.pdf

A Robustly Optimized BERT Training Approach:
[1907.11692] RoBERTa: A Robustly Optimized BERT Pretraining Approach (arxiv.org)

Benchmarking Commercial Intent Detection Services with Practice-Driven Evaluations:
Paper tables with annotated results for Benchmarking Commercial Intent Detection Services with Practice-Driven Evaluations | Papers With Code

XLNet: Generalized Autoregressive Pretraining for Language Understanding:
[1906.08237] XLNet: Generalized Autoregressive Pretraining for Language Understanding

Reference

SUPERVISED CONTRASTIVE LEARNING FOR PRE-TRAINED LANGUAGE MODEL FINE-TUNING: https://arxiv.org/pdf/2011.01403.pdf

SUPERVISED CONTRASTIVE LEARNING FOR PRE-TRAINED LANGUAGE MODEL FINE-TUNING: https://openreview.net/pdf?id=cu7IUiOhujH

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter: https://arxiv.org/pdf/1910.01108.pdf

A SIMPLE BUT EFFECTIVE BERT MODEL FOR DIALOG STATE TRACKING ON RESOURCE-LIMITED SYSTEMS: https://arxiv.org/pdf/1910.12995.pdf

Thank you