MSc Project Demonstration

Commonsense Validation and Explanation in Natural Language Processing

Supervisor : Dr Mark Lee
Student : Letian Li (2214560)

- **Task Description**
- **Methodology**
- **Implementation & Experiment**
- **Demonstration**

Task Description

MSc Project: Commonsense Validation and Explanation in NLP

- **Objective Task: SemEval-2020 Task 4: Commonsense Validation and Explanation [1]**

➢ Subtask 1: Validation

Task: Which of the two similar statements is against common sense?

Statement 1: He put a turkey into the fridge.

Statement 2: He put an elephant into the fridge.

➢ Subtask 2: Explanation (Multi-Choice)

Task: Select the most appropriate reason as to why this statement is against common sense.

Statement: He put an elephant into the fridge.

A: An elephant is much bigger than a fridge.

B: Elephants are usually white while fridges are usually white.

C: An elephant cannot eat a fridge.

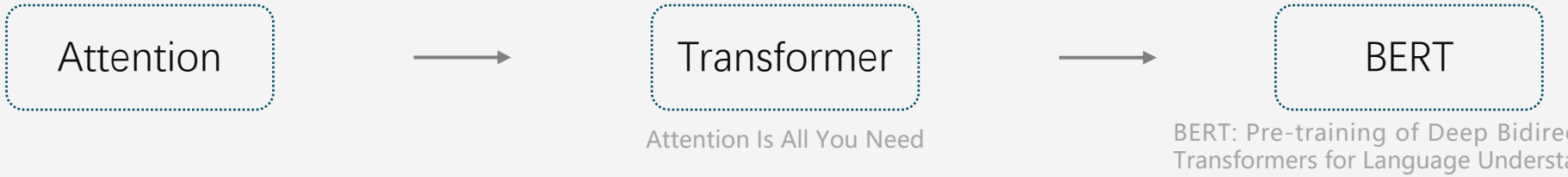
[1] Wang, C., Liang, S., Jin, Y., Wang, Y., Zhu, X. & Zhang, Y. SemEval-2020 Task 4: Commonsense Validation and Explanation in Proceedings of the Fourteenth Workshop on Semantic Evaluation (International Committee for Computational Linguistics, Barcelona (online), Dec. 2020), 307–321. https://aclanthology.org/2020.semeval- 1.39.

Methodology

MSc Project: Commonsense Validation and Explanation in NLP

- **Architecture Choices:**
  - ☐ **CNN (Convolutional Neural Network)**
  - ☐ **RNN (Recurrent Neural Network)**
  - ✓ **BERT (Bidirectional Encoder Representations from Transformers)**
- **Why BERT ?**
  - ➢ **Pretrained on a large corpus of English data.**
  - ➢ **Self-supervised learning with two objectives:**

    Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).
  - ➢ **Achieved state-of-the-art performance on big-name datasets like GLUE, MultiNLI, and SQuAD in 2019 [1].**
- **Core Mechanism:**
  - ➢ **Self-Attention**

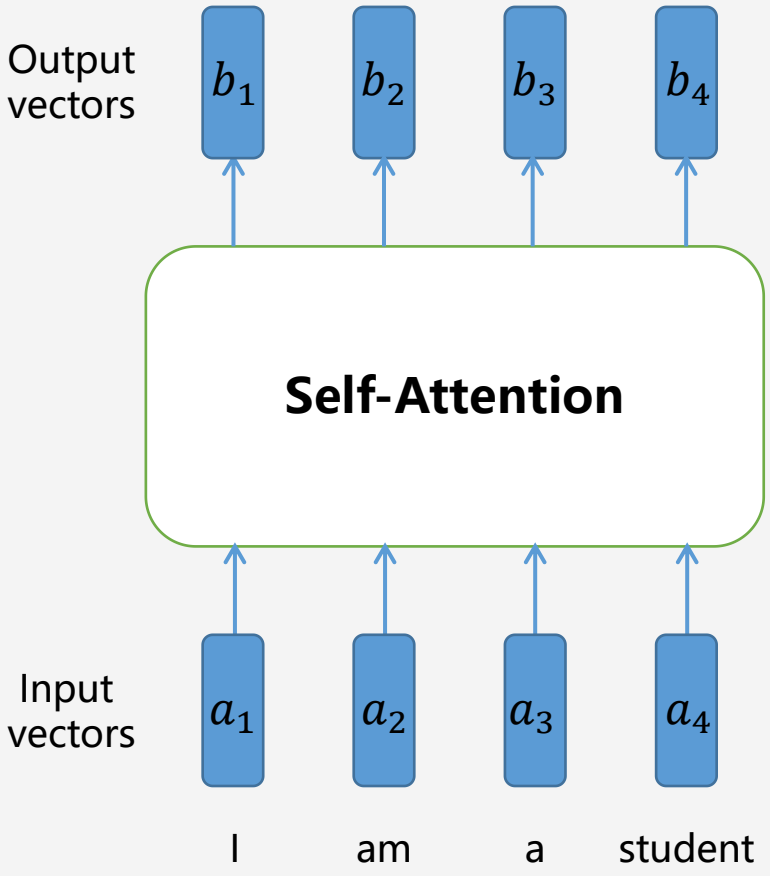


[1] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Association for Computational Linguistics, Minneapolis, Minnesota, June 2019), 4171–4186. https://aclanthology.org/N19-1423.

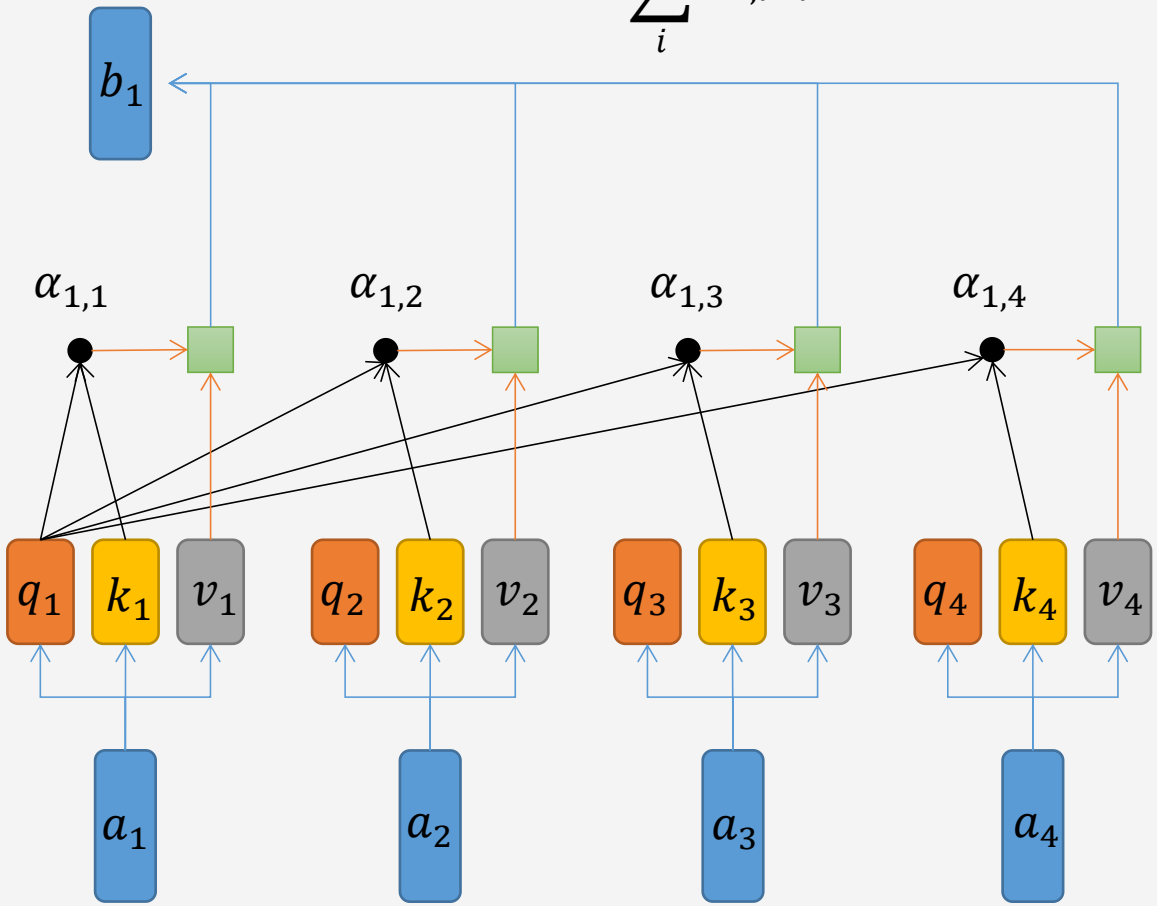
Methodology

MSc Project: Commonsense Validation and Explanation in NLP

Briefly Introduction to Self-Attention:



$$b_1 = \sum_i \alpha_{1,i} v_i$$



a: input vector
b: output vector
α: attention score
q: query matrix
k: key matrix
v: value matrix

Implementation

MSc Project: Commonsense Validation and Explanation in NLP

- **Models:**
  - BERT and its variant
- **Programming Language:**
  - Python
- **Deep Learning Framework:**
  - PyTorch
- **Integrated Development Environment (IDE):**
  - Jupyter
- **Visualization:**
  - TensorBoard

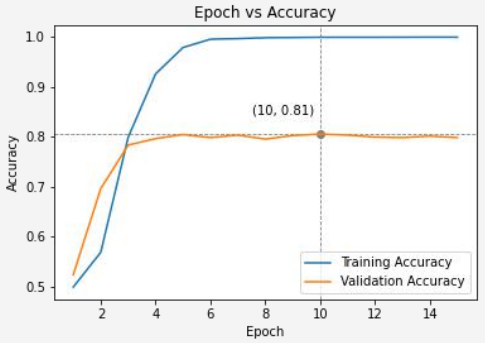
- **Python Libraries:**
  - **Main Library to Implement BERT:**
    - Transformers
  - **Data Processing:**
    - NumPy
    - Pandas
  - **Plotting:**
    - Matplotlib
    - Seaborn

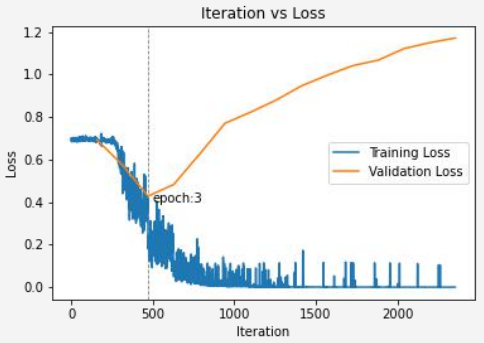
Dataset (for both Subtask A and Subtask B):

Training Data: 10,000

Validation Data: 997

Test Data: 1,000





Experiment

MSc Project: Commonsense Validation and Explanation in NLP

Experiment Objectives:

- Find the best performance model and hyperparameters for the task.
- Research on how hyperparameters influence model training.

Experiment Workflow:

- **Round 1: Grid Search on Different Models**
  - Compare the performance of different models and choose the one with the best performance to be the final model.
  - Locate a good range of learning rate and batch size and select a baseline for subsequent experiments.
- **Round 2: Research on Epoch**
  - Determine an appropriate epoch for the final model.
- **Round 3: Research on Learning Rate**
  - Compare how different learning rates affect model training.
  - Determine an appropriate learning rate for the final model.
- **Round 4: Research on Batch Size**
  - Compare how different batch sizes affect model training.
  - Determine an appropriate batch size for the final model.

Experiment

MSc Project: Commonsense Validation and Explanation in NLP

Round 1: Grid Search on Different Models

Grid search is a process of exhaustive search in a certain hyperparameter space.

It tries all possible combinations of hyperparameters.

Aim:

- Choose the best performance model to be the final model.
- Locate a good range of hyperparameters.
- Select appropriate hyperparameters as the baseline for the model.

Experiment Models:

Bert-Base-Uncased (12-layer, 768-hidden, 12-heads, 110M parameters)

Distilbert-Base-Uncased (6-layer, 768-hidden, 12-heads, 66M parameters)

Roberta-Base (12-layer, 768-hidden, 12-heads, 125M parameters)

Roberta-Large (24-layer, 1024-hidden, 16-heads, 355M parameters)

Hyperparameters Search List:

Learning rate list: [1e-7,1e-6,1e-5,1e-4,1e-3,1e-2]

Batch size list: [8,16,32,64,128,256]

Epoch: 5

Experiment Workflow:

Round 1: Grid Search ⟶ Round 2: Research on Epoch ⟶ Round 3: Research on Learning Rate ⟶ Round 4: Research on Batch Size

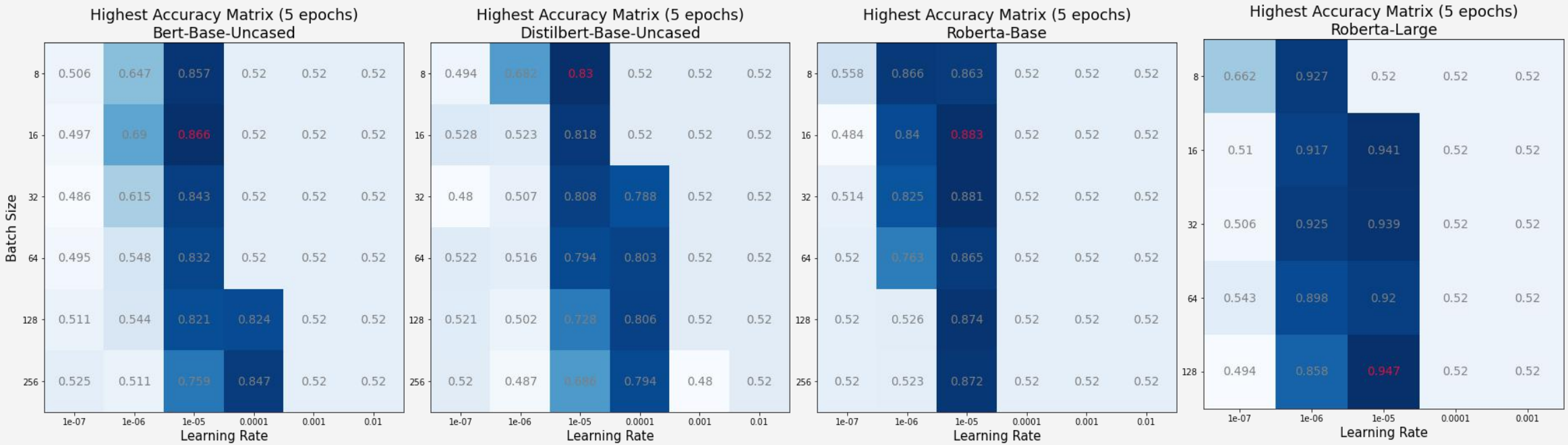
Experiment

MSc Project: Commonsense Validation and Explanation in NLP

Round 1: Grid Search on Different Models

Choose a baseline for subsequent experiments：

Model: Roberta-Large, Learning Rate: 1e-5, Batch Size: 128



Experiment Workflow:

Round 1: Grid Search ⟶ Round 2: Research on Epoch ⟶ Round 3: Research on Learning Rate ⟶ Round 4: Research on Batch Size

Experiment

MSc Project: Commonsense Validation and Explanation in NLP

Round 2: Research on Epoch

Epoch is the value of how many times the dataset has been used to train the model.

Aim:

Determine an appropriate epoch for the final model.

Eexperiment Model and Hyperparameters:

Model: Roberta-Large

Learning Rate: 1e-5

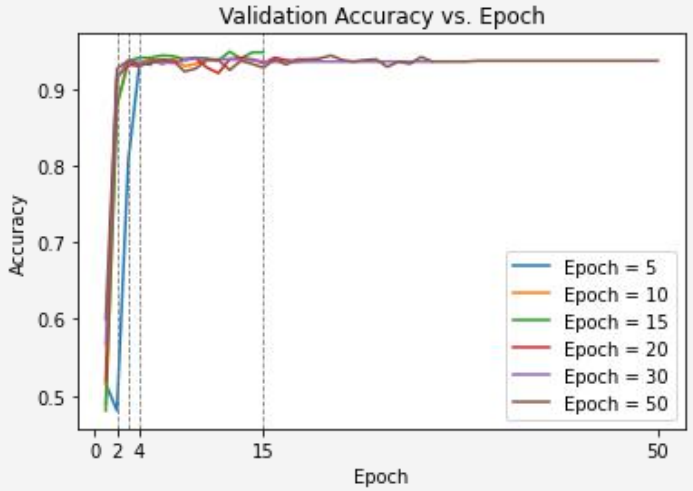
Batch Size: 128

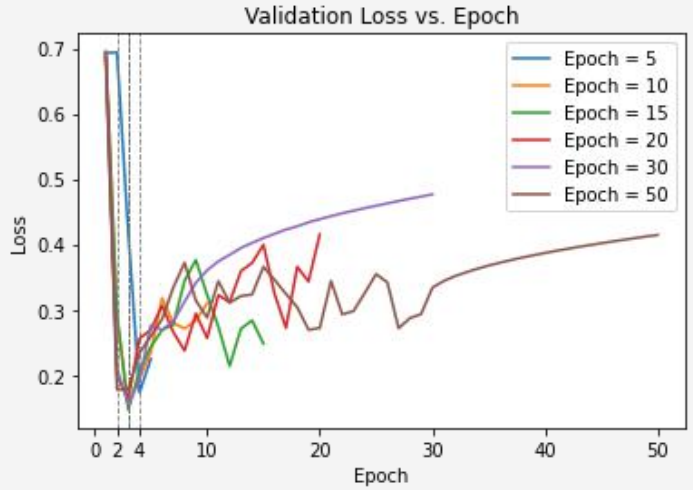
Search List of Epoch:

[5, 10, 15, 20, 30, 50]

Final Choice:

Epoch = 15





Experiment Workflow:

Round 1: Grid Search ⟶ Round 2: Research on Epoch ⟶ Round 3: Research on Learning Rate ⟶ Round 4: Research on Batch Size

Experiment

MSc Project: Commonsense Validation and Explanation in NLP

Round 3: Research on Learning Rate

Learning rate controls how quickly the model learns a problem.

Aim:

Compare how different learning rates affect model training.

Determine an appropriate learning rate for the final model.

Eexperiment Model and Hyperparameters:

Model: Roberta-Large

Batch Size: 128

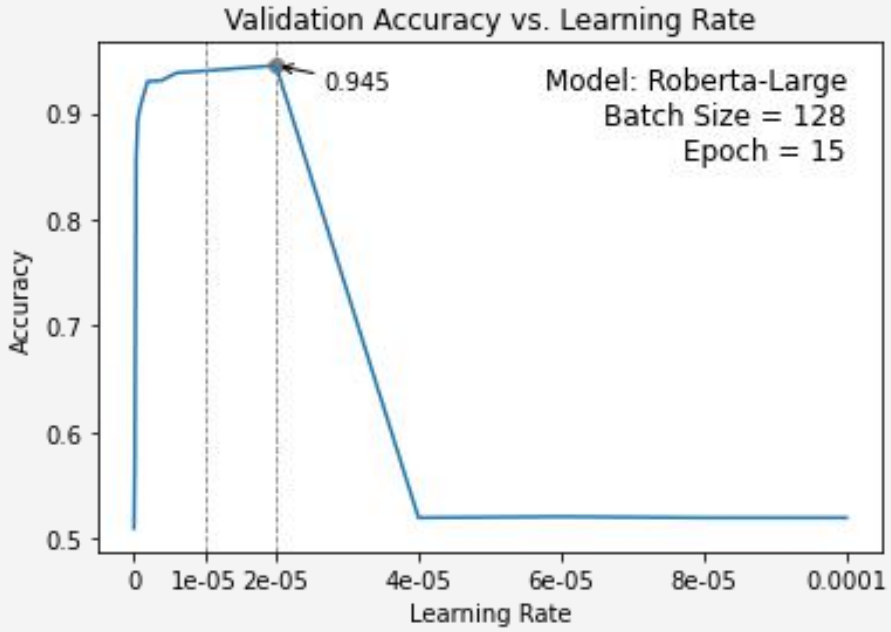
Epoch: 15

Search List of Learning Rate:

[1E-7, 2E-7, 4E-7, 6E-7, 8E-7]

[1E-6, 2E-6, 4E-6, 6E-6, 8E-6]

[1E-5, 2E-5, 4E-5, 6E-5, 8E-5, 1E-4]



Experiment Workflow:

Round 1: Grid Search ⟶ Round 2: Research on Epoch ⟶ Round 3: Research on Learning Rate ⟶ Round 4: Research on Batch Size

Experiment

MSc Project: Commonsense Validation and Explanation in NLP

Round 3: Research on Learning Rate

Learning rate controls how quickly the model learns a problem.

Aim:

Compare how different learning rates affect model training.

Determine an appropriate learning rate for the final model.

Eexperiment Model and Hyperparameters:

Model: Roberta-Large

Batch Size: 128

Epoch: 15

Search List of Learning Rate:

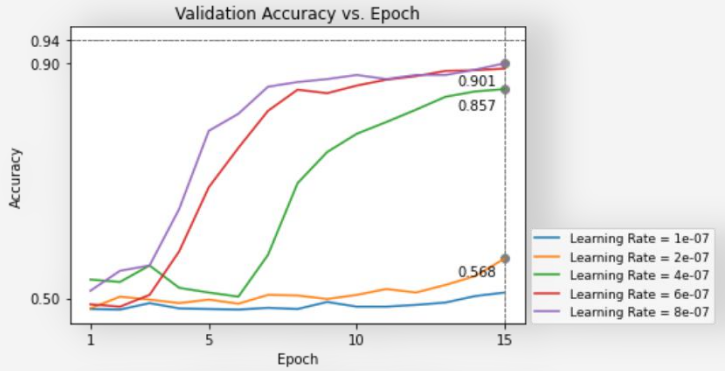
[1E-7, 2E-7, 4E-7, 6E-7, 8E-7]

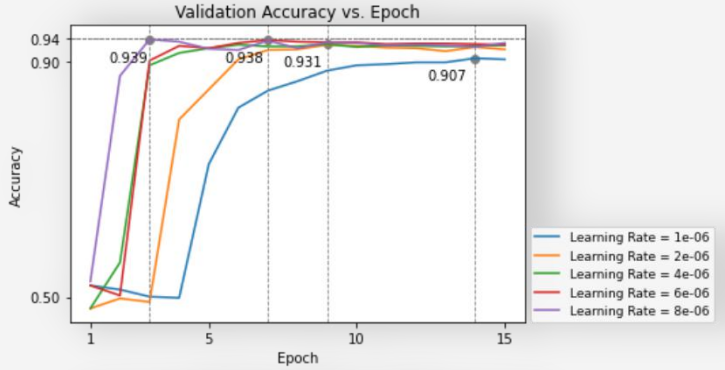
[1E-6, 2E-6, 4E-6, 6E-6, 8E-6]

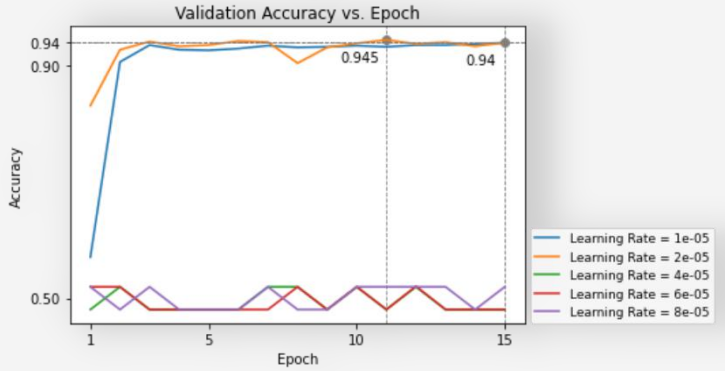
[1E-5, 2E-5, 4E-5, 6E-5, 8E-5, 1E-4]

Final Choice:

Learning Rate = 1E-5







Experiment Workflow:

Round 1: Grid Search ⟶ Round 2: Research on Epoch ⟶ Round 3: Research on Learning Rate ⟶ Round 4: Research on Batch Size

Experiment

MSc Project: Commonsense Validation and Explanation in NLP

Round 4: Research on Batch Size

Batch size is the number of training examples utilized in one iteration.

Aim:

Compare how different batch sizes affect model training.

Determine an appropriate batch size for the final model.

Eexperiment Model and Hyperparameters:

Model: Roberta-Large

Learning Rate: 1E-5

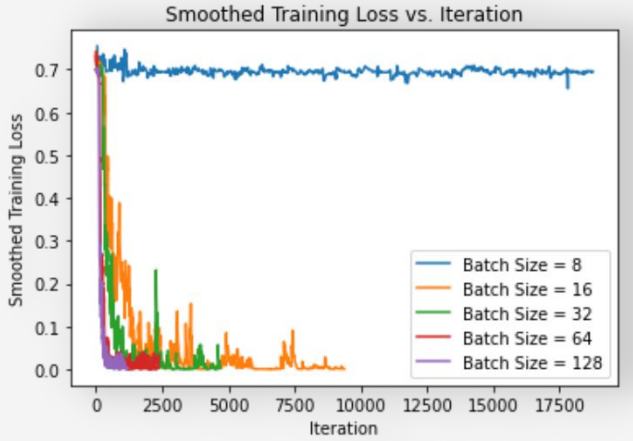
Epoch: 15

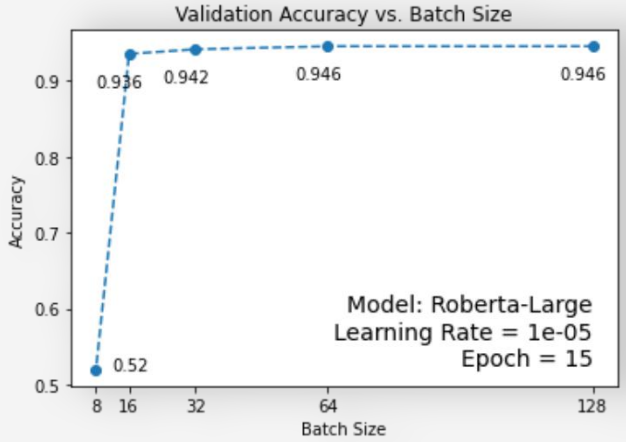
Search List of Batch Size:

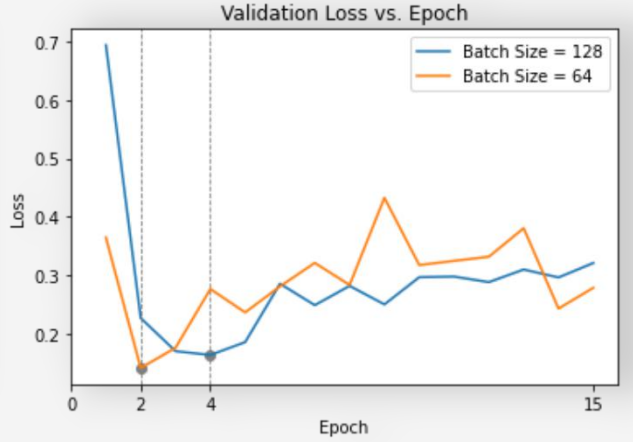
[8, 16, 32, 64, 128]

Final Choice:

Batch Size = 128







Experiment Workflow:

Round 1: Grid Search ⟶ Round 2: Research on Epoch ⟶ Round 3: Research on Learning Rate ⟶ Round 4: Research on Batch Size

Demonstration

MSc Project: Commonsense Validation and Explanation in NLP

Performance of the Final Model

Best Implementation for Subtask A:

Model: Roberta-Large,

Learning Rate: 1E-5,

Batch Size: 128,

Epoch: 15,

Achieved Accuracy: 94.3%

Best Implementation for Subtask B:

Model: Roberta-Large,

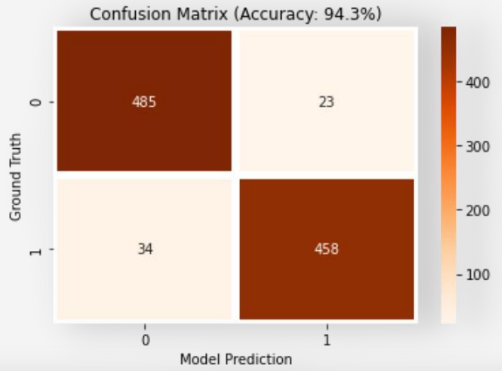
Learning Rate: 1E-5,

Batch Size: 16,

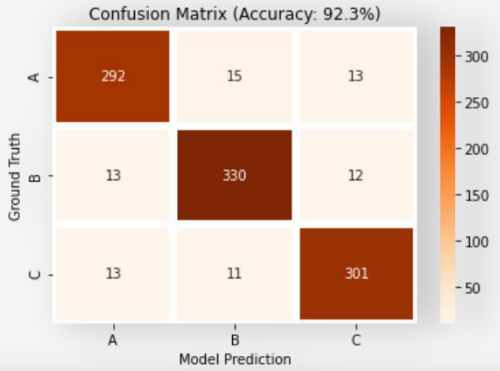
Epoch: 10,

Achieved Accuracy: 92.3%

A:



B:



	ID	Statement 0	Statement 1	Ground Truth	Model Prediction
0	1175	He loves to stroll at the park with his bed	He loves to stroll at the park with his dog.	0	0
1	452	The inverter was able to power the continent.	The inverter was able to power the house	0	0
2	275	The chef put extra lemons on the pizza.	The chef put extra mushrooms on the pizza.	0	0
3	869	sugar is used to make coffee sour	sugar is used to make coffee sweet	0	0
4	50	There are beautiful flowers here and there in ...	There are beautiful planes here and there in t...	1	1

	ID	False Statement	Option A	Option B	Option C	Ground Truth	Model Prediction
0	1175	He loves to stroll at the park with his bed	A bed is too heavy to carry with when strollin...	walking at a park is good for health	Some beds are big while some are smaller	A	A
1	452	The inverter was able to power the continent.	An inverter is smaller than a car	An inverter is incapable of powering an entire...	An inverter is rechargeable.	B	B
2	275	The chef put extra lemons on the pizza.	Many types of lemons are to sour to eat.	Lemons and pizzas are both usually round.	Lemons are not a pizza topping.	C	C
3	869	sugar is used to make coffee sour	sugar is white while coffee is brown	sugar can dissolve in the coffee	sugar usually is used as a sweetener	C	C
4	50	There are beautiful planes here and there in t...	A plane flies upon the garden	You can have a small garden in your private plane	A plane can never be seen in garden	C	C

Manual Test