Common sense inference using Macaw Encoder
End Semester Project Presentation

Guide: Dr. B Sivaselvan

Asst Prof. IIITDM

Sreepathy Jayanand

CED17I038

Problem Statement:

- Common sense inference is a task in natural language processing involving understanding the context and suggesting possible logical extensions from the context.
- E.g. :“How do you make a house conduct electricity?”
- Macaw’s answer: Paint the house with metallic paint
- The answer given by Macaw isn’t obtained from a part of the question but rather it was able to associate house with paint and electricity with metal.
- The problem that we are trying to solve here is – Given a context and a list of possible extensions from the context, select the most probable option.

Literature Review:

- Here are 2 of some of important papers leading to the proposed solution.
  - Multi angle question answering - MACAW
  - EncT5: Fine-tuning T5 Encoder for non-autoregressive Tasks

MACAW:

- Macaw is a generative question answering model that is built on T5, and only has 11billion parameters, as compared to GPT-3, a leading generative model which contains 175billion parameters but still produces similar benchmarks.
- T5 is an encoder-decoder generative model that has achieved the state-of-the-art benchmarks in a variety of generation tasks such as language translation, summarization, question answering etc.
- Macaw is trained on different “angles” – For example, Given a question, it can generate the answer; Given an answer it can generate possible options; Given the option it can generate a question etc.

MACAW:

- It is fine-tuned on the ARC dataset, which is a school level science-based reasoning dataset with a question with multiple options.
- Hence, we can expect it to learnt a good context in general common-sense based questions.
- Macaw's syntax:
  - "$question$ = What is the color of a cloudy sky? ; $answer$; $mcoptions$"
  - Macaw's output: '$answer$ = gray ; $mcoptions$ = (A) blue (B) white (C) grey (D) white'

EncT5:

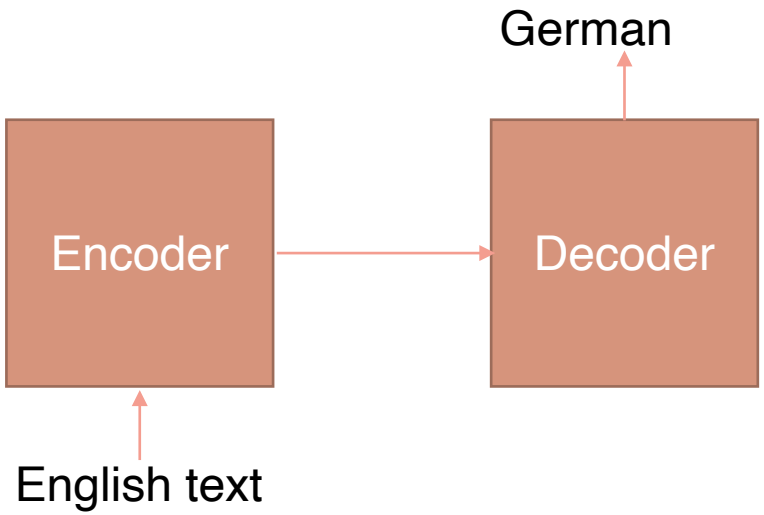
- T5 is an encoder-decoder generative model.
- This paper sheds light on the fact that the encoder of a trained T5 model can be used to obtain embeddings of the input sentence which would have more representational power.
- These encoded input vectors would have good representational power of the input sentence and can be used further.

SWAG dataset:

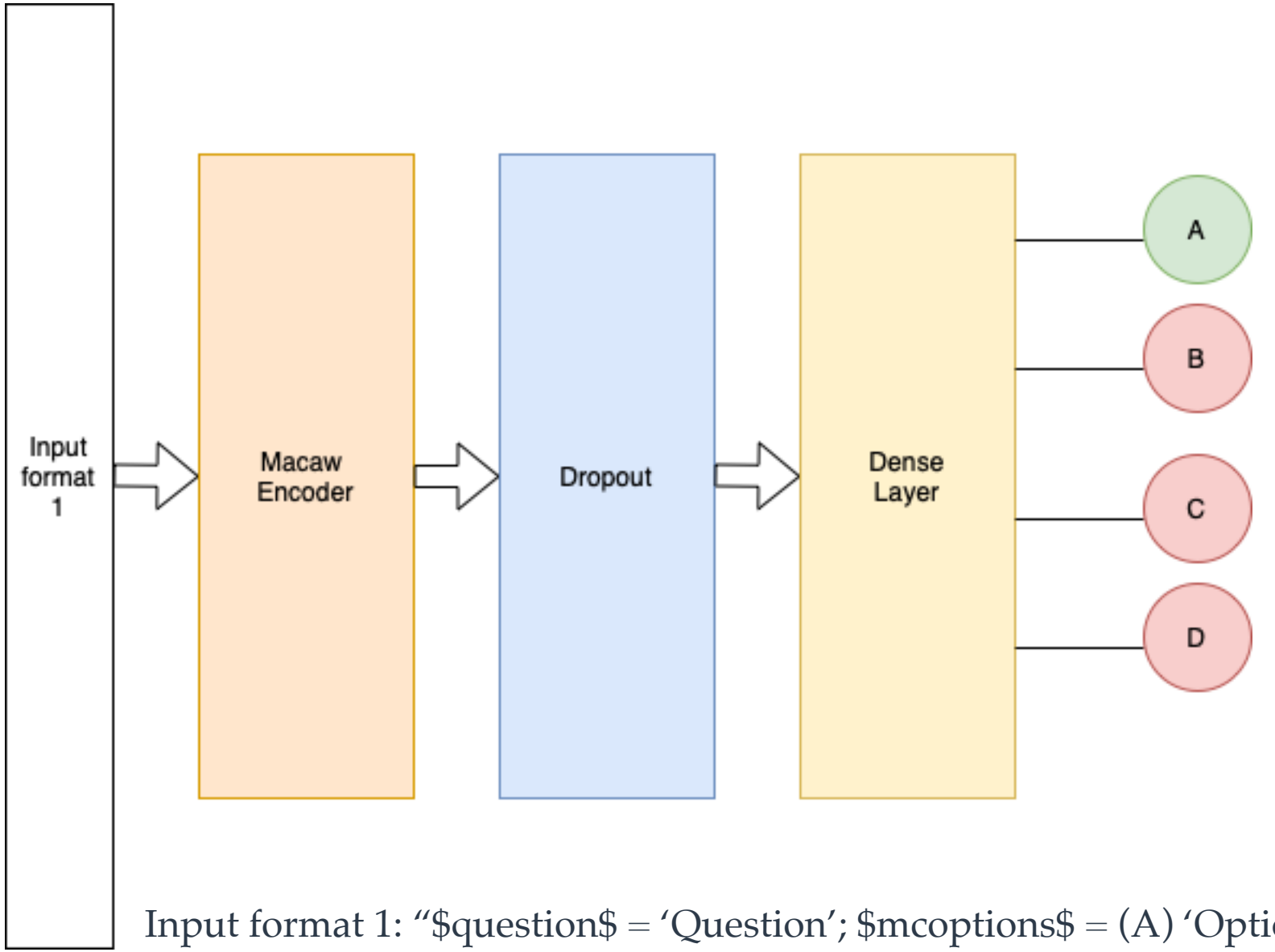
- SWAG / Situations With Adversarial Generations dataset contains instances where in a situation is present and a good amount of “common-sense” is required for predicting the next phase of the situation.
- E.g. “On stage, a woman takes a seat at the piano. She,
  - A. sits on a bench as her sister plays with the doll.
  - B. smiles with someone as the music plays.
  - C. is in the crowd, watching the dancers.
  - D. nervously sets her fingers on the keys – Correct option

Architecture of Macaw:

- Macaw is essentially and encoder-decoder architecture, like the T5 which itself is based of the transformer.



Model 1:

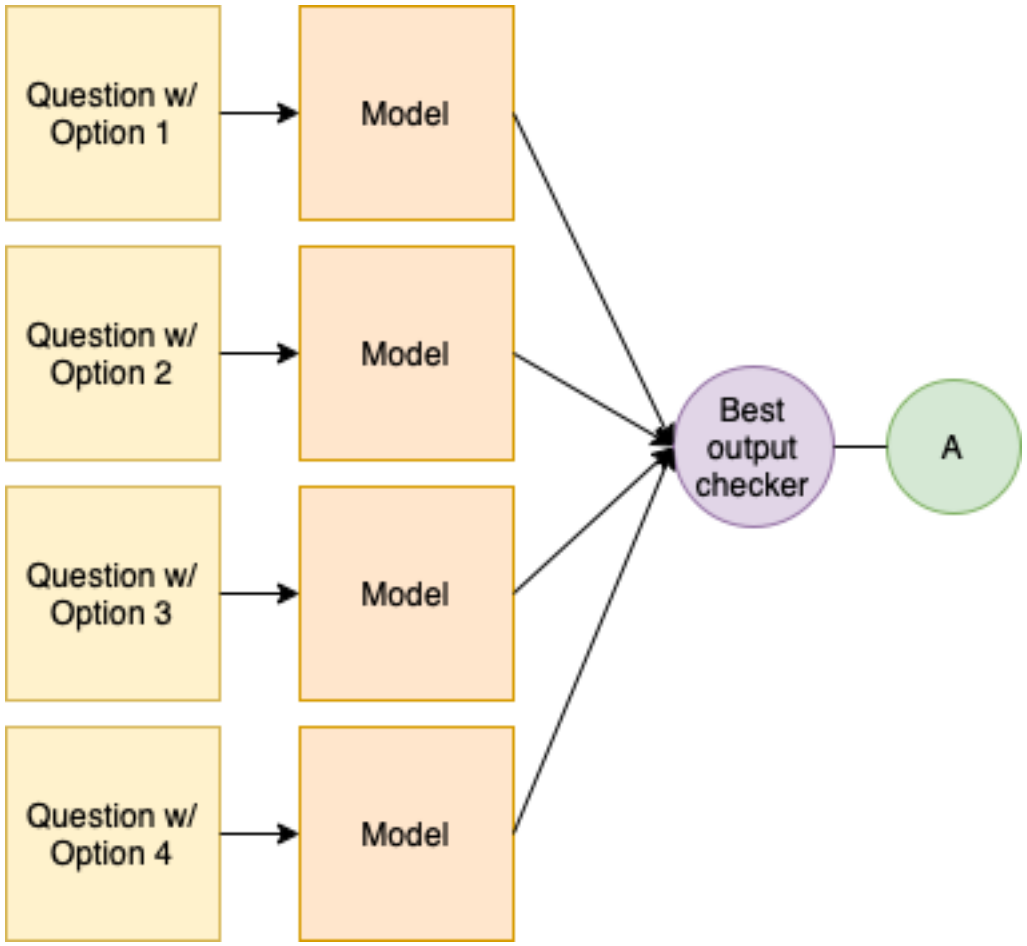


Input format 1: “$question$ = ‘Question’; $mcoptions$ = (A) ‘Option 1’ (B) ‘Option2’ (C) ‘Option 3’ (D) ‘Option 4’ ”

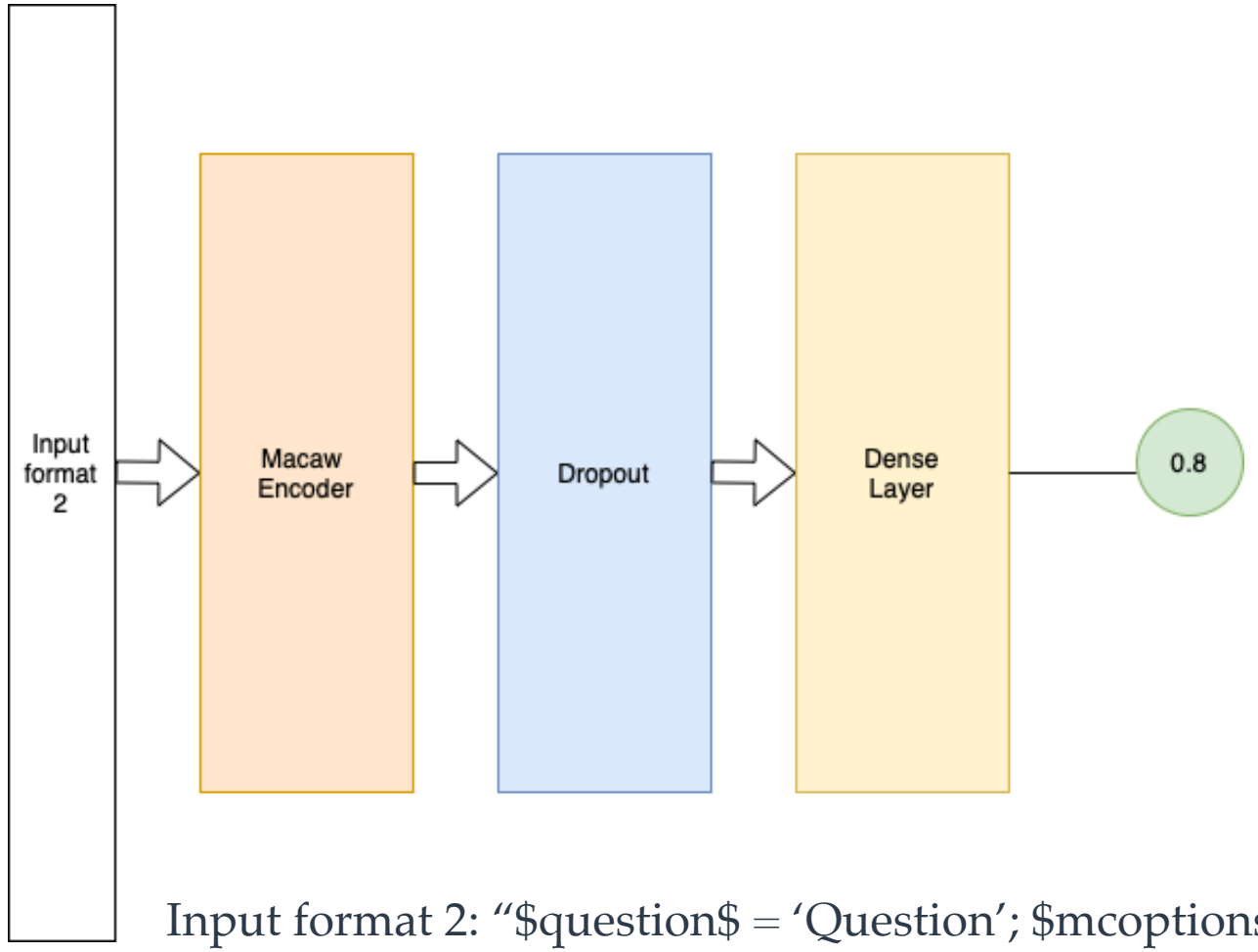
Model 1:

- Trained for 7 epochs
- Learning rate - 0.00001
- Batch size - 8
- Equal distribution of correct option choice (~25% each)
- Total testing instances - 20,008. Correctly classified - 15,974. Accuracy - 79.8
- Average F1 score per class - 0.797
- MCC score - 0.731

Model 2:


Question w/ Option 1
Model
Question w/ Option 2
Model
Question w/ Option 3
Model
Question w/ Option 4
Model
Best output checker
A

Model 2:



Input format 2: “$question$ = ‘Question’; $mcoptions$ = (A) ‘Option 1’ ”

Model 2:

- Trained for 7 epochs
- Learning rate - 0.00001
- Batch size - 8
- Equal distribution of correct option choice (~25% each)
- Total testing instances - 20,008. Correctly classified - 16,561. Accuracy - 82.8
- Average F1 score per class - 0.825
- MCC score - 0.816

Results:

- Human Performance - 88.0
- Model 2 - 82.8
- T5 large - 72.6
- RoBERTa base - 82.6

Conclusions and Future Scope:

- Model 1 achieved 79.8 % and model 2 achieved 82.8 % accuracy
- Both models have higher accuracy than the baseline T5 model from which this encoder for this model is taken and has only less than half the parameters of that of T5.
- Rather than using encoder based architectures like variants of BERT which lead the SWAG benchmarks, if we can compromise the accuracies a little we can use an encoder of a generative model so that a new architecture needn't be improvised every-time we have to perform a new classification task.
- This model can be extended for various other non regressive tasks like sentiment analysis, sentence similarity, answer span prediction etc and expect it to deliver good results.

THANK YOU