

Automated Flashcard Generation Using LoRA-Enhanced Transformer Models

16:954:577:01 Statistical Software
Final Project

Team 12: Master Blasters
Adish Golechha
Sarvesh Kharche
Sneh Bhandari

Problem Statement and Dataset Overview

Students often struggle to retain key concepts from extensive academic notes. Traditional methods of creating flashcards are time-consuming and may not cover all critical areas, leading to inefficient study practices.

Objective: To automate the generation of high-quality, contextually accurate flashcards from academic notes, enabling students to efficiently test and reinforce their understanding.

SQuAD Dataset - Wikipedia's extractive Q&A pairs for generating context-aware questions and answers. (Extractive)

RACE Dataset - Abstractive Q&A pairs from middle and high school English exams, enhancing question diversity and complexity. (Abstractive)

SciQ Dataset - Fine-tuned on 13,679 science multiple-choice questions with supporting evidence for improved accuracy and relevance.

RNN: Baseline Model

Recurrent Neural Networks (RNNs) capture sequential dependencies in data using internal memory. The RNN model in this project includes an Embedding Layer, Simple RNN Layer, and Dense Layer for output prediction.

RNNs were chosen as a baseline to compare against more complex transformer models like T5, highlighting the performance improvements transformers offer for generating question-answer pairs from educational content.

Parameters

- Optimizer: Adam
- Loss function: Sparse Categorical Cross Entropy
- Epochs = 10
- Batch Size = 64

Challenges

- Capturing Long-Range Dependencies
- Model Simplicity
- Overfitting

T5: Transformer Model

T5 (Text-to-Text Transfer Transformer) is a pre-trained transformer model designed to handle a wide range of NLP tasks by converting them into a text generation problem. It treats all tasks as a text-to-text problem, making it highly flexible across domains.

Model was pre-trained on massive datasets like SQuAD (extractive Q&A) and RACE (abstractive Q&A), allowing it to capture diverse linguistic and contextual nuances.

For our project, we fine-tuned the pretrained T5 model on a science-specific dataset to generate meaningful question-answer pairs directly from educational texts

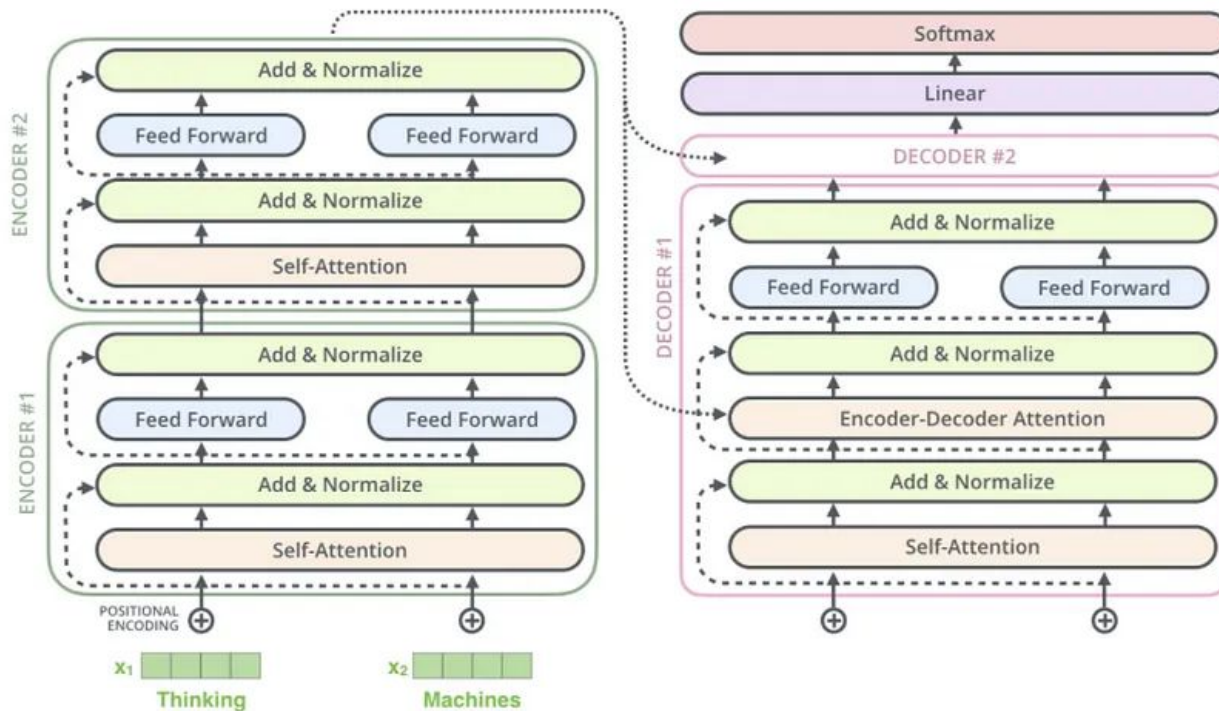
Parameters

- T5 Large: 770 million parameters
- Model Layers: 24 transformer layers
- Attention Heads: 16 attention heads per layer

Challenges

- Computational Cost
- Input Length Limitation
- Time Intensive

T5: Transformer Model Architecture



Vanilla Encoder Decoder Transformer [Jay Alammar's blog](#)

Fine-Tuning the T5 Model

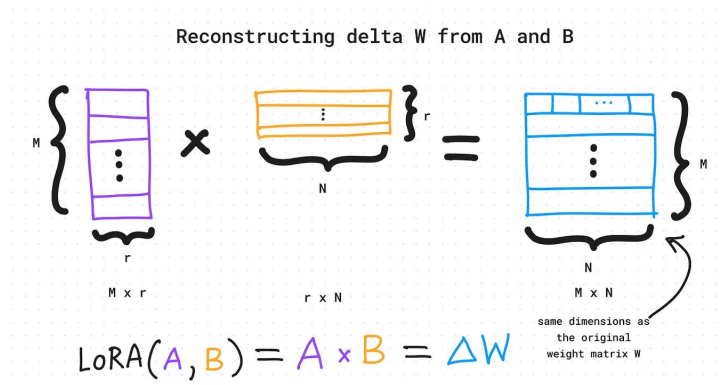
By fine-tuning T5 on science-specific educational datasets, we aimed to enhance its ability to generate meaningful question-answer pairs relevant to academic content, especially for science education.

Parameter Efficient Fine-tuning - LoRA (Low Rank Adaption) was used as it reduces the number of parameters to train, making fine-tuning more memory and computationally efficient.

The model generated high-quality, diverse, and domain-specific question-answer pairs with improved contextual understanding and accuracy.

Parameters

- Rank = 16
- Train Test Split = 80/20
- Total Parameters = 770M
- Trainable Parameters = 4M (0.6% of Total)



Pipeline, Preprocessing, and Enhancements with Tagging

Pipeline and Preprocessing

To handle input token limits, we split large documents into smaller chunks. The chunk size is adjustable via the Streamlit web app. Overlapping chunks were implemented to preserve context between sections.

Each chunk is summarized using a BART Large CNN model, ensuring the summary stays within the 512-token limit. Beam search is used to generate the best possible question-answer pairs. This approach efficiently processes long PDFs (20+ pages).

Tagging

We experimented with Named Entity Recognition (NER) and Part-of-Speech (POS) tagging to enhance the generated output.

Re-finetuning the T5 model on tagged SciQ datasets revealed that POS tagging contributed more significantly to improving question-answer quality than NER, making it a better choice for our approach.

Results

We used ROUGE-L and BERTScore to evaluate the model's performance:

- **ROUGE-L:** Measures the longest common subsequence between generated and reference text, assessing syntactic similarity and content overlap.
- **BERTScore:** Utilizes BERT embeddings to evaluate semantic similarity, ensuring contextual accuracy and meaning alignment.

These metrics provide a comprehensive evaluation of the model's syntactic and semantic quality.


	RNN	T5-Race Fine-tuned	T5-Squad Fine-tuned	T5-Squad Fine-tuned and POS tagging
ROUGE-L	0.222	0.633	0.699	0.733
BERT Score	0.766	0.917	0.928	0.931

User Interface

Flash Card Generator

Enter your academic notes, PowerPoint presentations, or reading material to generate Flash Cards on the material to help you prepare better.

Upload New File

 Drag and drop file here
Limit 200MB per file • PDF

Browse files



transformers.pdf 2.0MB



Number of Questions



[Open Flashcard App](#)

Flashcard App

What mechanism in transformers helps the model focus on relevant parts of the input when generating an output?

Attention

Next

Show Answer

Built on: Streamlit, Dash