

Summary Results

Joel Pointon

July 27, 2023

1 Fine-Tuned T5

Fine-tuning a model using a question answering dataset achieves the following results:

Train Loss	Val Loss	Rouge1	Rouge2	RougeL	RougeLsum	Bleu	Gen Len	Meteor	Non-ans acc
1.3849	1.0848	41.527	33.324	38.4866	38.4856	29.906	17.1296	0.377	1.9%

As can be seen above, after 19 epochs of training, both the training and validation loss are very low. Furthermore, the model achieves excellent results using Rouge, Bleu and Meteor scores. Rouge scores are by nature difficult to compare, as the vocabulary and context vary each application. However, we can say that around 42% words in the target answer were included in the generated answer (rouge1). Additionally, model still performs well when comparing bigram overlap (rouge2). The model also scores highly using the Bleu score and Meteor score, which are more complex than a Rouge score, but similarly compare the accuracy of the generated text to the target answer.

However, even with 30% unanswerable questions, the model still struggles to identify unanswerable questions, with only 2% being correctly identified. Additionally, the model appears to be underfitting, as opposed to overfitting to the data. This is shown by the training loss being higher than the validation loss. Therefore, there is scope to use regularisation to improve model performance. This is currently being attempted and hopefully be implemented in the final model.

1.1 Usage

The above performance metrics are promising, however, these are difficult to interpret, and so the output from testing the model will provide the greatest insights.

Below is the output from asking the model: What is PCA?

The PCA is used to approximate the face.

To construct this answer, I used the following documents:

1. *Fundamentals_of_Digital_Image_Processing_Textbook->Page(s)278/279/280/281: When we carry out a PCA on this ensemble of faces, we find that the eigenvalue spectrum dies off qu...*

While the framework has correctly identified one of the most important sections in the textbook, the model has failed to understand the intention of the question (to give a meaning of the acronym and a short description) and extract the correct answer.

Rewriting the question in alternative manner yields an improved but still unsatisfactory answer: What is Principal Component Analysis?

The Principal Component Analysis is calculated on a pixel-to- pixel basis.

To construct this answer, I used the following documents :

1. *Fundamentals_of_Digital_Image_Processing_Textbook->Page(s)274/275: Depending on the precise context, there may be several closely related reasons for doing PCA, but t...*

This time the model has extracted an accurate statement within the text on pages 274/275: "one possible approach to PCA the covariance is calculated on a pixel-to-pixel basis...". However, it has failed to provide a satisfactory description of what Principal Component Analysis actually is.

A more simple question can be asked: What are true-colour images?

The true-colour images are three distinct 2-D planes.

To construct this answer, I used the following documents :

1. *Fundamentals_of_Digital_Image_Processing_Textbook->Page(s)28/29: RGB (or true colour) images are 3-D arrays that we may consider conceptually as three distinct 2-D ...*

This answer demonstrates the potential of such a model - it has correctly identified the answer and provided a succinct response. Such answers are possible by the model, but aren't consistently provided.

Finally, we can also see the model's inability to detect unanswerable questions. For example, when asked: Who is Joel?, it responds:

The Joel is Chris Solomon.

Unfortunately, it has identified that both Joel and Chris Solomon are people, but has not understood that there is not an answer to such a question in the textbook provided.

2 GPT-3.5-turbo

To benchmark these responses against an alternative, state-of-the-art model, we can ask the same questions to GPT-3.5-turbo and provide the same context to the model. Asking the model, What is PCA? yields the following response:

PCA stands for Principal Component Analysis. It is a statistical technique used for dimensionality reduction and data compression. PCA identifies the most important features or components in a dataset and represents the data in a new coordinate system defined by these components. It is commonly used in computer vision for tasks such as face recognition and image compression.

To construct this answer, I used the following documents :

1. *Fundamentals_of_Digital_Image_Processing_Textbook->Page(s)278/279/280/281: When we carry out a PCA on this ensemble of faces, we find that the eigenvalue spectrum dies off qu...*
2. *Fundamentals_of_Digital_Image_Processing_Textbook->Page(s)274/275: Depending on the precise context, there may be several closely related reasons for doing PCA, but t...*

As can be seen, the model is given the similar sources (2 rather than 1, as it can process more tokens), and succinctly provides an accurate answer using the sources, and is grounded in the context of computer vision. Additionally, when asked "Who is Joel?", the model responds appropriately:

I could not find an answer in the text I've been provided, sorry! Please try again.

3 Summary

Clearly the state-of-the-art API managed by OpenAI can produce far superior answers. While the fine-tuned T5 model produces excellent numerical outcomes, actual output from the model lacks substance and demonstrates a failure to fully understand the question. These outcomes are not inconsistent and

show that there is a need to use a similarity measure which compares the semantic meaning of sentences (e.g. cosine similarity) rather than a simple overlap of words (unigram, bigram etc).

However, there are ways to improve the fine-tuned T5 model further. Firstly, a confidence level can be implemented so that the model is only able to respond if the textbook has a highly relevant section. This can be achieved by providing a cosine similarity cutoff. Secondly, the model can be improved by using regularisation techniques and by providing more training examples. This can reduce the impact of underfitting and enable the model to learn further.

The framework presented by this paper offers promising results and potential for training a long-form, closed-domain question-answering chatbot which has not yet been directly attempted by the literature. I offer a novel dataset of naturally worded questions with succinct answers written in natural-language to contribute to the literature. By improving and adding to this dataset, such models can be improved further. This framework can very easily be used in a variety of contexts, as the model is trained to answer questions from a context, rather than be an expert in a specific domain, and is open-source so is completely free.