Discussion points/Reasons for not training a DPR model separately:

In the dataset we have z|x. So the probability p(z|x) for the given x is 1.

Now when have the z available for a given x, we can just pick up that z and search through the entire dataset for similar documents.

This can be using FAISS (which is much more optimized), or any similarity measure like cosine similarity.

So for a given x, the given z can act as the query to FAISS and rest of the documents act as the search space.

Now the way we have designed our dataset, we have mulitple questions on a single context (document). So the search space would retrieve the same documents (very-likely).

Now let's say treat the documents independently to be sent to the encoder and questions independently, to send the questions to a second encoder and try to minimize the negative log-likelihood of probability p(z|x) (or maximize p(z|x).

We are in way wasting the resources to do the job which has been already done while designing the dataset.

So I guess, all we need to do is to train a generator model, where given x, z we train it to predict the y.

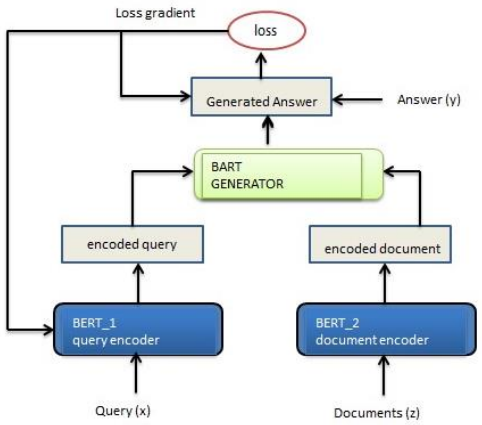
The question remains what do we do to answer a new query which does not come with the context.

So to answer that, I dont think we should train the retriever independent of generator. We should train the whole set of bi-encoders, (encoder for query (to be fine-tuned), encoder for documents (to be used as it is)) and the generator model to minimize negative log-likelihood of p(y|x)).

So all in all, to train a retriver model, independent of generator model, does not make sense to me and is a futile exercise.

Further to this, if we are looking at the RAG model and trying to do something similar, then there we do not have context attached to the questions (as I have understood). The authors have taken wikipedia pages divided into chunks of size 100 and 21 million such chunks, and I guess independent questions. Now here DPR (or any retriever model) in between makes sense.

So guess, the way I plan to work this out is this:



Please let me know if you have any reservations/ counter points about this.