Project Idea List

Hi, KerasNLP community team,

I am Aditya Das, an open-source contributor and a post graduate student in computer science from India. I want to participate in GSoC 2023. I have lots of contributions in the KerasNLP project, I also share the contributions here. And I have some ideas regarding the gsoc KerasNLP project that is also related to official project ideas. I provide all of these ideas below.

List of all the GSoC 2023 ideas :

1) Add longformer-base-4096 model :

longformer-base-4096 is a BERT-like model started from the RoBERTa checkpoint and pretrained for MLM on long documents. It supports sequences of length up to 4,096.

Longformer uses a combination of a sliding window (local) attention and global attention. Global attention is user-configured based on the task to allow the model to learn task-specific representations. Please refer to the examples in modeling_longformer.py and the paper for more details on how to set global attention.

This model has been most downloaded and liked on hugging face almost billion times and also it is mentioned in GSoC 2023 official idea list for new strategic open source backbone model.
https://huggingface.co/allenai/longformer-base-4096

2) Higher priority task model, (HF's task model) :

Right now the community's main focus is on adding the task models that are the Higher priority. If it is not necessary, is there any other model I can add or contribute too, in this library please suggest.

3) Pretranning workflows for models :

The team have models like Bert and all the other models but the team has only one guide to pretranning the models that is keras.IO, which was added many ages ago. This guide only covers all things from scratch.
This guide does not even use Bert for MaskedLM, here basically it builds everything from scratch. Also this needed for gsoc official ideas to pretranning workflows for models.

4) Add training script for generating model like GPT2 :

There are still no issues for generating tasks for training scripts. The team do have training script for seperate training for Bert and stuffs but the team do not have training script for generating tasks like GPT2.
Also this field is mentioned in gsoc official ideas to add training scripts.

5) Add distilGPT2 model

DistilGPT2 is an English-language model pre-trained with the supervision of the 124 million parameter version of GPT-2. DistilGPT2, which has 82 million parameters, was developed using knowledge distillation and was designed to be a faster, lighter version of GPT-2.

DistilGPT2 (short for Distilled-GPT2) is an English-language model pre-trained with the supervision of the smallest version of Generative Pre-trained Transformer 2 (GPT-2). Like GPT-2, DistilGPT2 can be used to generate text. Users of this model card should also consider information about the design, training, and limitations of GPT-2.

This model should be nice to add to the library and it has millions of downloads and likes in huggingface.

https://huggingface.co/distilgpt2

6) Add Task model like NSP(Next sentence prediction)

The community currently focussed on task models because there are lots of

Backbone models, which are right now already present. In this case I want to propose an idea like NSP (Next sentence prediction) task to add as an example to the library.

Also this is needed for gsoc official ideas to support generative modeling.

7) Add sampling API for summarization and machine translation :

The community might not be ready for summarization or machine translation, So I want to work on this idea if maintainers want to add these things to the library. Also this can cover the area of official project ideas as a stand alone sampling API support.

8) plan to ship models that can be deployed to a single mobile device :

Also there is one interesting idea that should be mentioned in the official idea list. Add some guide for TFlite, because of the potential of contributing to an example which can be deployed on a mobile device or something, or an edge device that should be great to add in this project.

List of all my contributions :

Update model and file names for DistilBert #541
https://github.com/keras-team/keras-nlp/pull/541#event-7980214844

Replicate #536 changing GPT2 -> GPT2Backbone #558
https://github.com/keras-team/keras-nlp/pull/558

https://github.com/keras-team/keras-io/issues/1160
https://github.com/keras-team/keras-io/pull/1172

https://github.com/keras-team/keras-nlp/issues/599
https://github.com/keras-team/keras-nlp/pull/617

https://github.com/keras-team/keras-cv/pull/1180 https://github.com/keras-team/keras-cv/issues/1171

https://github.com/keras-team/keras-nlp/issues/573
https://github.com/keras-team/keras-nlp/pull/637

https://github.com/keras-team/keras-nlp/pull/670
https://github.com/keras-team/keras-nlp/issues/650

https://github.com/keras-team/keras-nlp/pull/724
https://github.com/keras-team/keras-nlp/issues/710

https://github.com/keras-team/keras-nlp/issues/810

https://github.com/keras-team/keras-nlp/issues/867
https://github.com/keras-team/keras-nlp/pull/879

https://github.com/keras-team/keras-nlp/pull/821

I want to participate in GSoC(Google Summer Of Code) 2023 and provide these idea lists, is this something maintainers should be interested in adding those ideas to the library !. If the maintainer gets interested in these ideas then I can work on adding these ideas to my gsoc 2023 proposal.
Thank You.

Aditya Das

Email	**adityamca123@gmail.com**
Twitter	**https://twitter.com/ADITYA90546170**
LinkedIn	**https://www.linkedin.com/in/aditya-das-7b2276202**
Github	**https://github.com/ADITYADAS1999**
Portfolio	**https://adityadas1999.github.io/**