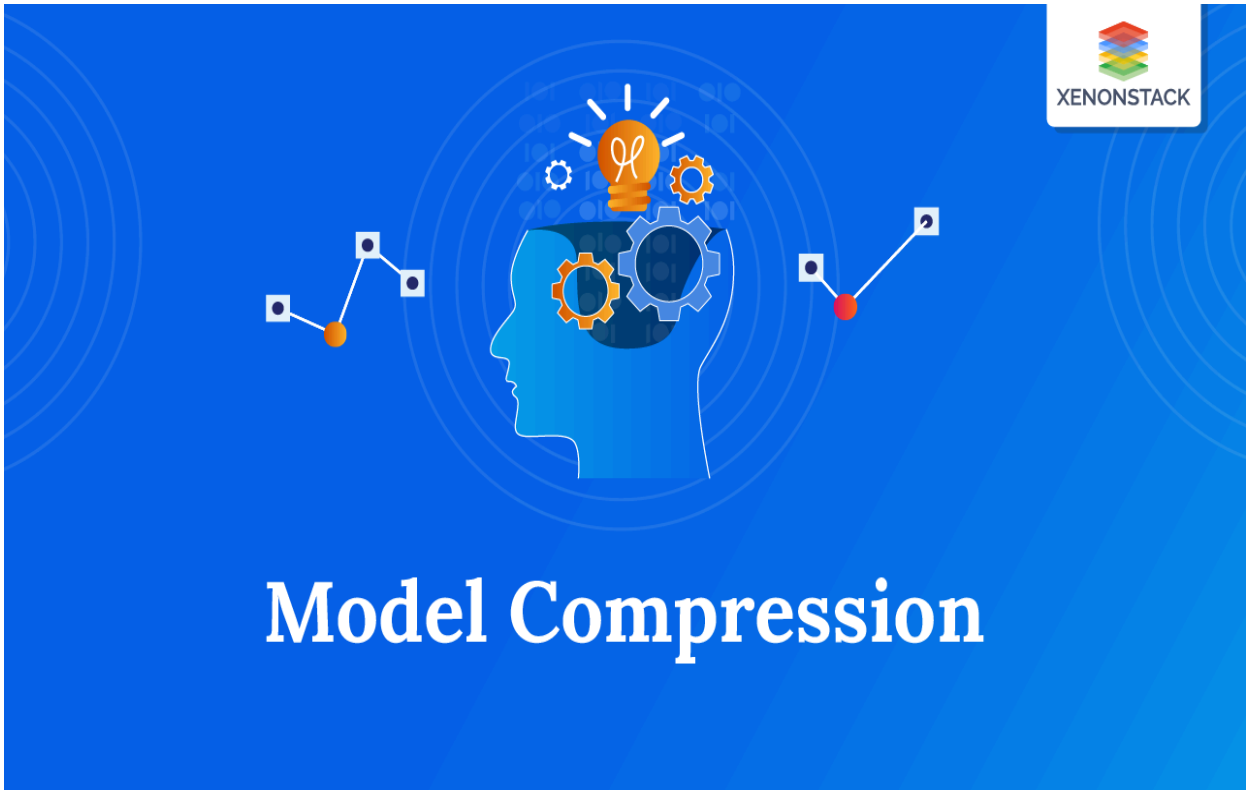


Model Compression and Safe API

18-04-2024

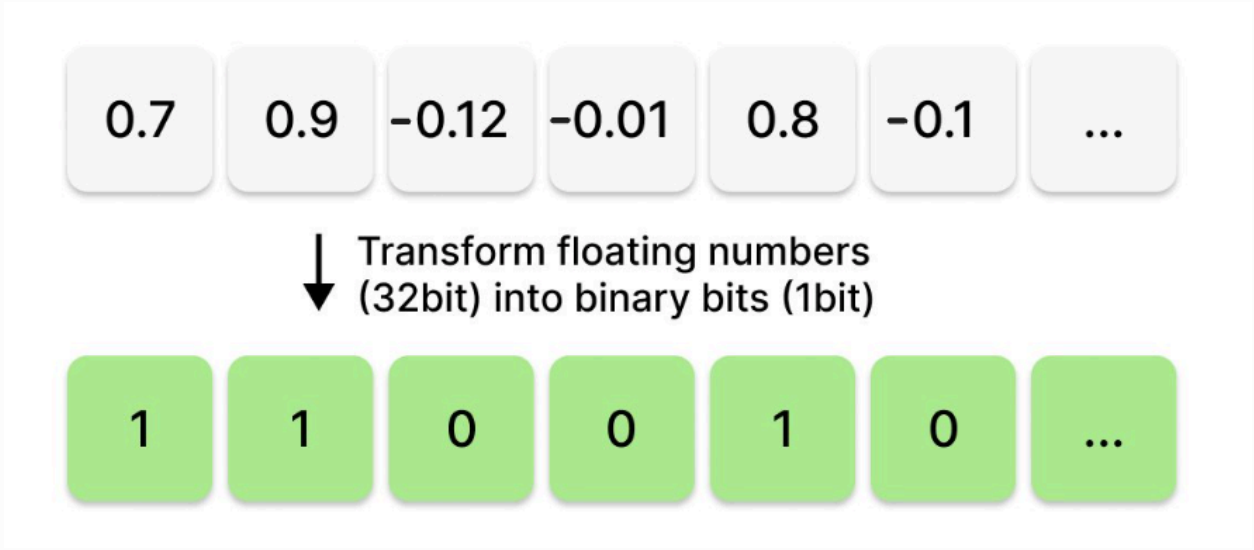
Mohd Saqib
Grid Dynamics
Data Science Intern

Embedding vector dimension reduction :

Our embedding model has 768 vector dimensions with storage of 350 MB. We have reduced the embedding dimension to any size by using Principle Component Analysis (PCA). We have reduced 768 dimensions to 128 dimensions, reducing the storage requirement by factor 6.
This technique neither improves the runtime, nor the memory requirement for running the model. It only reduces the needed space to store embeddings.

Binary Quantization :

Binary quantization refers to the conversion of the float32 values in an embedding to 1-bit values, resulting in a 5x reduction in memory and storage usage. To quantize float32 embeddings to binary, we simply threshold normalized embeddings at 0: if the value is larger than 0, we make it 1, otherwise we convert it to 0.
After reducing the size of embeddings the MAP(mean average precision) on the testing data is 63 % which is 4% lower than the actual model
This technique does not improve the runtime to generate the embeddings. It just simply reduces the size of embedding where it is getting stored.



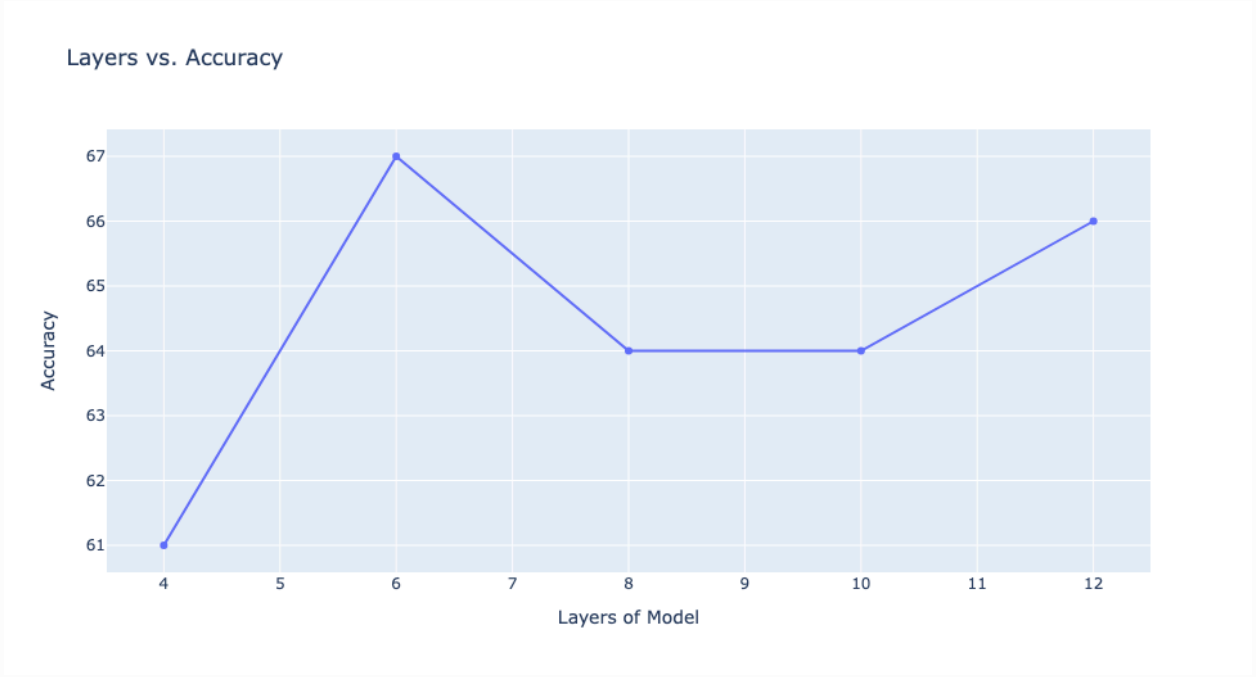
Pruning :

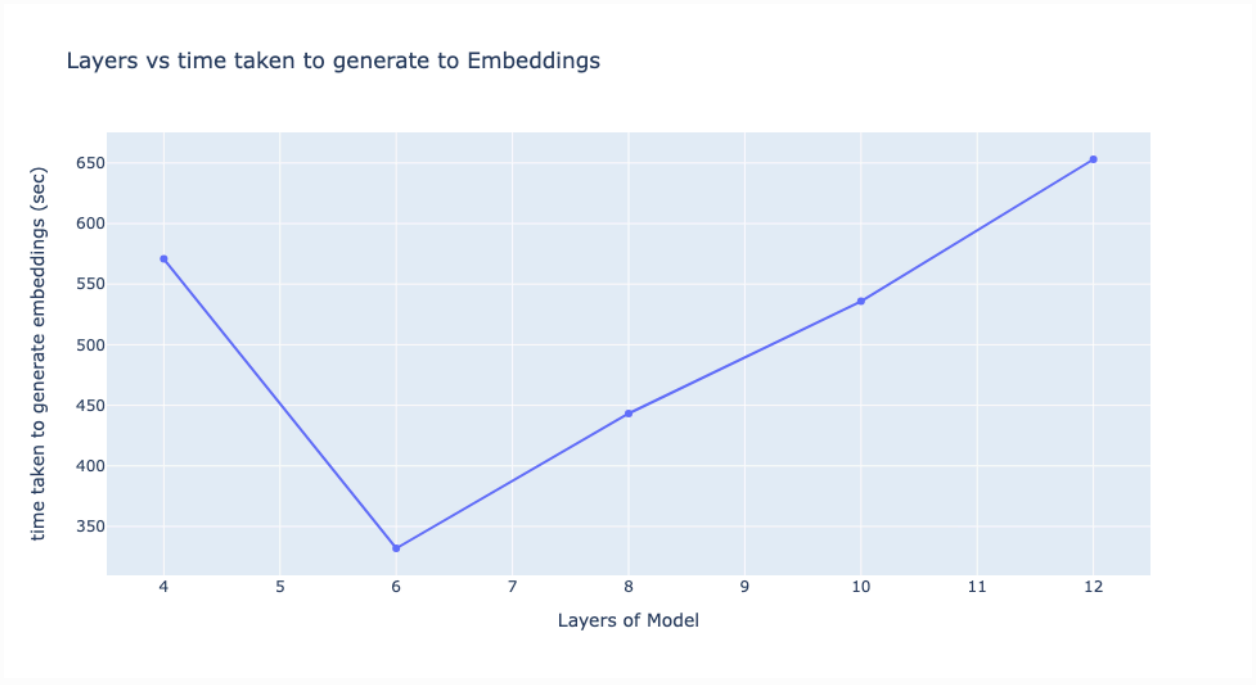
In pruning we are keeping some layers from the actual model and now with these layers , training the reduced layer model to generate the embeddings same as base model with the help of back propagation .
With the help pruning the size of the model decreases and speed to generate the embeddings decreases.
For different layers , speed is decreased without much loss in MAP(mean average precision).

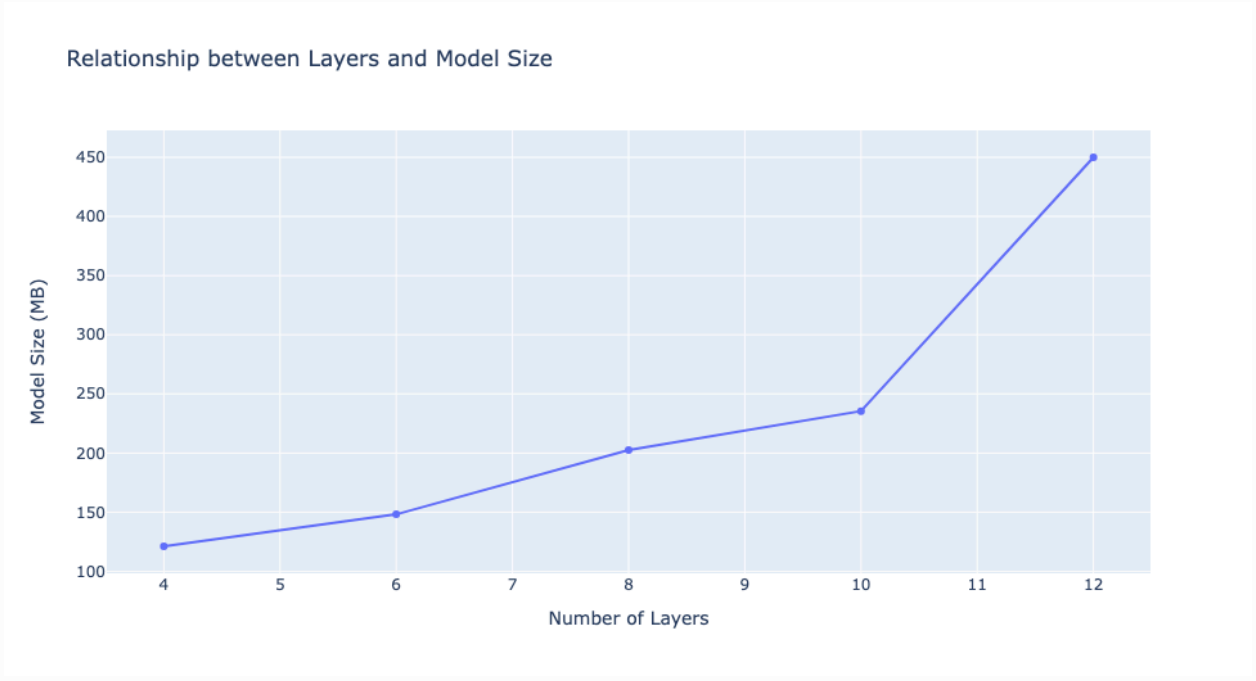
Layers	Accuracy	Time to generate Embeddings (seconds)	Model size (MB)
4	61.40	570.79	121.15
6	67.72	331.94	148.19
8	64.38	443.23	202.27
10	64.74	535.87	235.48
12 (original model)	66.14	653.05	450

The best performing model among comes out to be 6 Layer models .
It generates the embeddings fastest with less model size .
All the Embedding generated by models are of same dimension 768 and same memory usage

Removal of the layer makes the model work faster but at an extent of decreasing the accuracy .





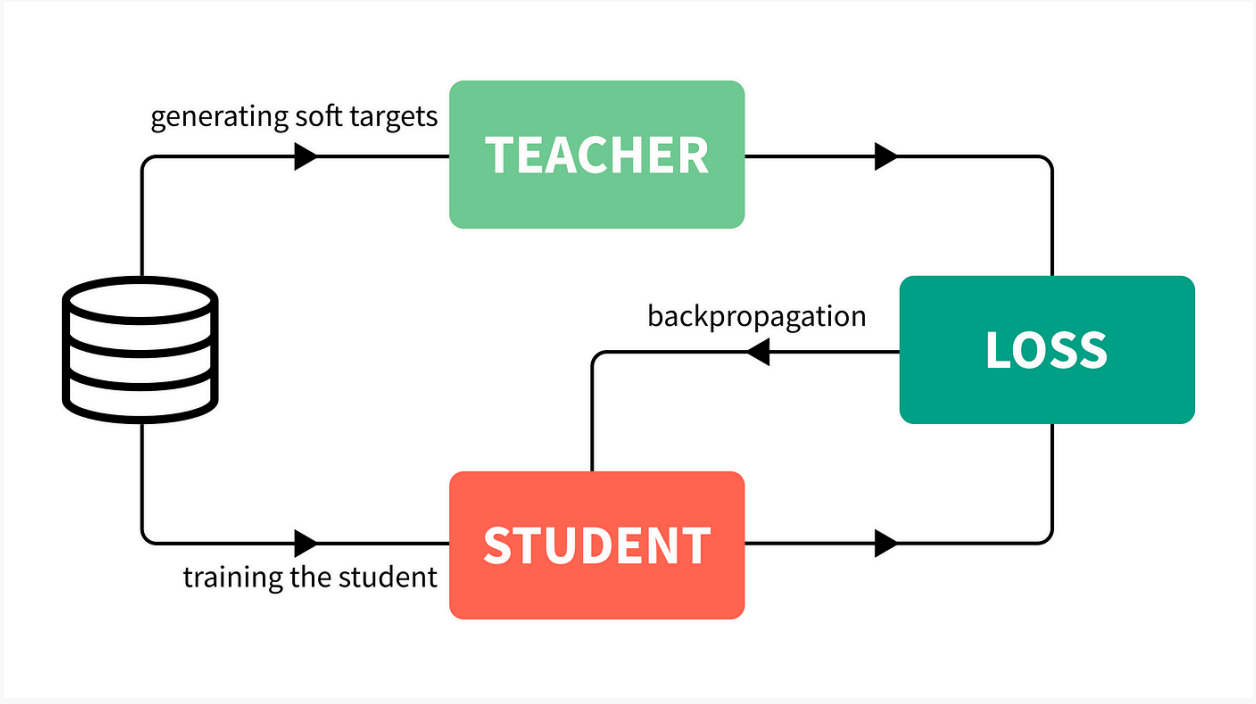


API :

In an api call the input is from the user and it will fetch the most relevant code from data that can be possible for this input query .
For invalid input like “Hello World” our model will somehow give the code for this query but in reality it is an invalid input so our model should not give this type of input .
To overcome this scenario what we have done is while fetching the docstring we will check the cosine similarity of top 10 docstring should be greater than 0.5 .So by doing this we have overcome the case of invalid input.

Knowledge Distillation :

Knowledge distillation describes the process to transfer knowledge from a teacher model to a student model.
In this student model is “sentence-transformers/all-MiniLM-L12-v2” and Teacher model is “sentence-transformers/all-mpnet-base-v2”



Student model has 33 million parameters, 384 dimension embedding vectors, 12 layers and the teacher model has 110 million parameters ,768 dimension embedding vectors, 12 layers .

For mimicking the teacher model, the embedding dimension generated by student model and teacher model should be the same but in our case it is not . To make the embedding of the same dimension for the teacher model we have applied the pca to reduce the dimension of the teacher model in the last layer . Now the embedding vector generated by student and teacher is the same.

Now we can compare the loss between the embeddings and with this loss we back propagate to do some changes in weight and then again train the student model.

With this procedure the student model after distillation has a size of 150 MB and time taken to generate the embeddings is 327 second which is 50 % less time than the teacher model.

Model	Model size (MB)	Time taken to generate embeddings (sec)	Accuracy
Student model(after knowledge distillation)	148	327	62.6 %
Teacher Model	450	653	66.14 %

With knowledge distillation the generation of embeddings is reduced to half time and size of model is also 1/4th time of teacher model