MANOJ KUMAR MAHARANA
M20AIE252

① The simplification that computes the $P(y|x)$ independently for each label in the vocabulary and picks the top 5 labels with the maximum likelihood would work under the following assumptions / conditions as per my understanding are →

① Given the input features, the labels are conditionally independent (Presence or the absence of one label does not affect the Probability of another label appearing in the predicted labels set).

② The input features are sufficient to capture all the information needed to predict the labels accurately.

③ The dataset is big to estimate the probabilities of each labels accurately/correctly.

④ The no. of labels in the vocabulary is not too much large (The computational complexity of the model would increase significantly as the number of labels increases).

What I think that while this simplification can make the model more computationally efficient but it may not always result in the best performance. If the above assumptions are not correct then we need more sophisticated models that will capture the dependency between the labels will achieve better results.

So, In Summary the simplification that computes $P(y|x)$ independently for each label in the vocabulary and picks the top 5 labels with the maximum likelihood assumes that the labels are conditionally independent given the input features. This assumption is known as "Independence assumption" and is made to reduce the computational complexity of the model.

② In class we discussed Network Partitioning, Layer-wise Partitioning and data Parallelism. (Multiple GPUs).

One new approach is using hybrid of network and layer-wise partitioning. Here, neural network will be divided into multiple sub-networks and each subnetwork is further partitioned layer-wise to enable data parallelism. I read one article last week and I found this is also possible.

Here suppose we assume →

→ The dataset is large to divide into multiple batch for processing.

→ Also think, the subnets are of equal number or size depending upon available GPUs.

→ Each subnet is assigned to different GPU.

→ During training each GPU processes assigned using data parallesim. (Each GPU trains with different dataset).

This approach has giving advantage of combining the benefits of both network partitioning and layer-wise also, I think by doing this memory requirement of each GPUs are reduced, so that help to train larger models and for faster training layer-wise (data parallelism).