FINE-TUNING [facebook/nllb-200-distilled-600M] FOR MACHINE TRANSLATION ON LOW RESOURCE LANGUAGES

Translation language:	Hindi -> Kinnauri	Hindi -> Kangri
Dataset used:	Kinnauri-Pahari	Kangri_corpus
No. of sentences:	20307	26785
Framework used:	Pytorch	Pytorch

For this implementation two Tesla T4 GPUs (15GB) were utilized.

Implementation details:

1. The dataset present in the original .txt format in both cases was zipped together and converted to a list format. This was followed by a conversion to the .csv format.
2. To get more data for the LLM to train on, the dataframe was essentially repeated. The number of unknown tokens were identified and ultimately expanding the tokenizer's vocabulary wasn't required.
3. The LLM chosen for this implementation is the "facebook/nllb-200-distilled-600M". No Language Left Behind (NLLB) is a first-of-its-kind, AI breakthrough project that open-sources models capable of delivering high-quality translations directly between any pair of 200+ languages — including low-resource languages like Asturian, Luganda, Urdu and more. It aims to help people communicate with anyone, anywhere, regardless of their language preferences.
4. The NLLB tokenizer consists of predefined language codes for the corresponding source and target languages. Since, Kangri & Kinnauri aren't available codes in the FLORES-200 code index, new language codes were created (kangri_Deva and kan_Deva respectively) and inserted into the nllb tokenizer.
5. This was followed by resizing the model embeddings based on the length of the new tokenizer. Finally, the dataset was tokenized.
6. For evaluation the sacrebleu metric was utilized and the results for the two datasets after fine tuning are as follows:

Kinnauri	**Kangri**
{'score': 15.669726436489944, 'counts': [17796, 7979, 4025, 2089], 'totals': [46021, 40538, 35080, 29727], 'precisions': [38.66930314421677, 19.682766786718634, 11.473774230330672, 7.027281595855619], 'bp': 0.9955770512173886, 'sys_len': 46021, 'ref_len': 46225}	{'score': 0.09152850478601893, 'counts': [1192, 74, 19, 3], 'totals': [55718, 50361, 45044, 39894], 'precisions': [2.139344556516745, 0.1469390997001648, 0.04218097859870349, 0.007519927808693036], 'bp': 0.9159447113479372, 'sys_len': 55718, 'ref_len': 60610}