Resource-Efficient NLP

KHIPU 2025

What is Low-resource NLP?

- **Low-resource Data:** Dataset sizes which are smaller than usually available to train a given model.
    - This is going to be simulated in this tutorial using subsamples from Spanish/Portuguese datasets.

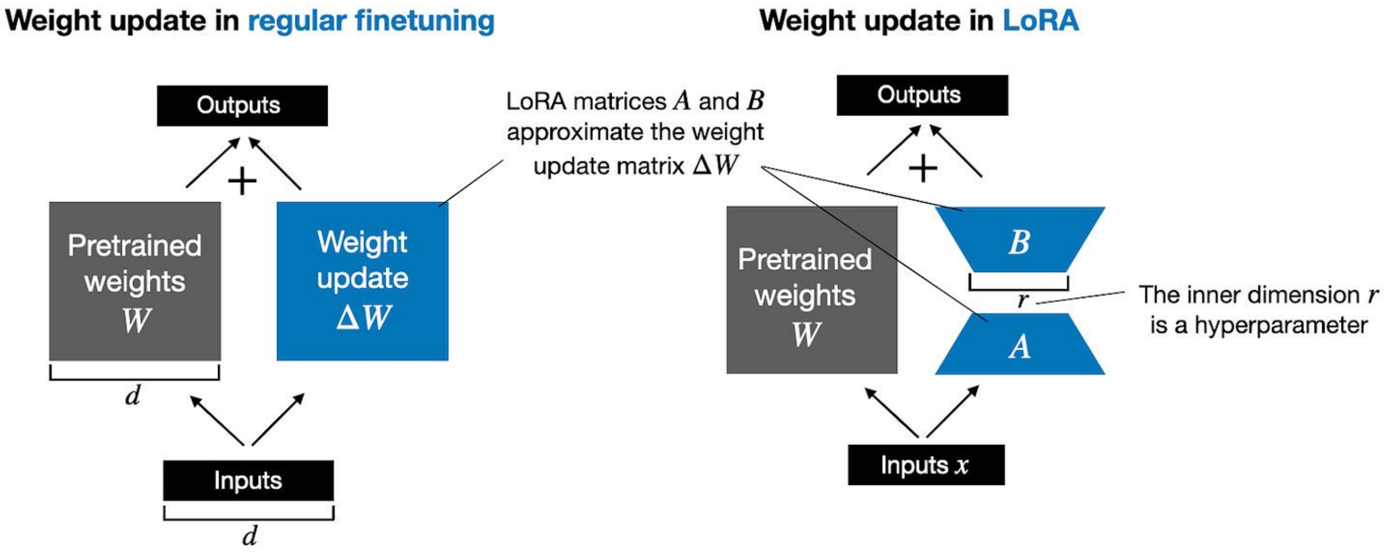
- **Low-resource Compute:** Constrained computing resources than usually required by standard training/inference methods.

Today: Sampling of Resource-efficient NLP Toolkit

Today we'll cover a sample of different approaches which may be used to circumvent problems related to settings with low-resource data and/or compute!

Parameter-Efficient Fine-tuning (PEFT)

A family of methods which allow for modifying a subset of parameters during fine-tuning, reducing computational requirements.


Weight update in regular finetuning
Outputs
+
Pretrained weights W
Weight update ΔW
d
Inputs
d
LoRA matrices A and B approximate the weight update matrix ΔW
Weight update in LoRA
Outputs
+
Pretrained weights W
B
r
A
The inner dimension r is a hyperparameter
Inputs x

Img source: https://sebastianraschka.com/blog/2024/lora-dora.html

Pre-trained Models

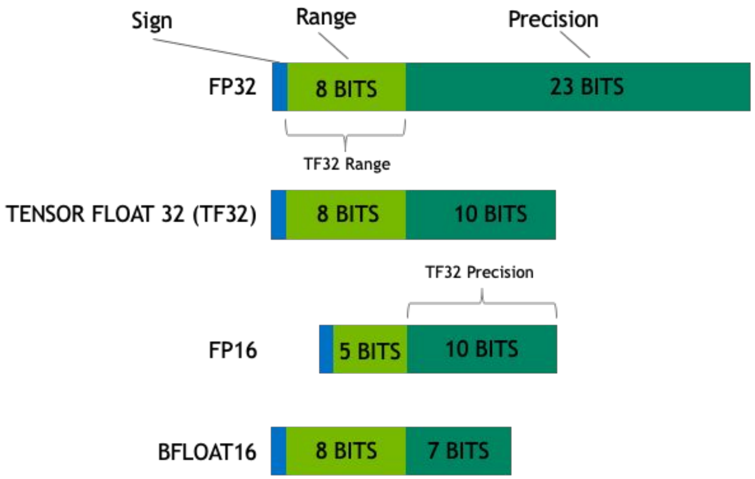
Plenty of pre-trained models available open source in hubs like Hugging Face! This can save you compute resources that would otherwise be needed to pre-train a base model.

- Can use models off-the-shelf
- Can use models as a base to further fine-tune on your own data
- **Note:** Pre-trained models may not be readily available for low-resource data domains.

huggingface_hub

Model Quantization

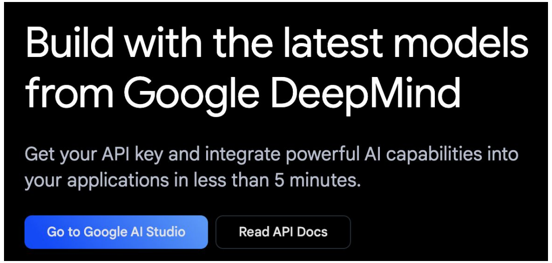
Model quantization methods reduce the precision of numerical representations of model weights and activations, yielding reductions in memory costs and potentially computational speedups as well.



Source: https://huggingface.co/blog/hf-bitsandbytes-integration

API Usage

Many state-of-the-art models are also available for use through APIs. Although APIs generally incur a cost to use, the cost is significantly lower than what would be required to train and house such models.



Build with the latest models from Google DeepMind

Get your API key and integrate powerful AI capabilities into your applications in less than 5 minutes.