Finetuning on Unimodal Tasks

What is often lost sight of in multimodal models is the performance on unimodal downstream tasks. While the main goal of multimodal models is to learn a joint representation of text and images, a multimodal model should also excel at unimodal tasks. In our case: image classification and text classification.