Pathfinder-X2: A Challenging Dataset for Evaluating Large Language Models on Long-Range Dependencies

Tyler Suard

April 2023

1 Abstract

The rapid progress of large language models has led to impressive results in a wide array of tasks. However, there remains a need for increasingly challenging datasets to evaluate these models' ability to handle long-range dependencies. In this paper, we present Pathfinder-X2, a novel dataset that builds upon the Pathfinder and Pathfinder-X datasets. Pathfinder-X2 comprises 512x512 pixel images, designed to test large language models' capacity to segment a specific white line dash "snake" with a circle at its tip among a collection of similar, distractor snakes. We demonstrate that the increased image resolution and complexity of Pathfinder-X2 present a substantially more challenging task for large language models, contributing to the ongoing development and assessment of such models.

2 Introduction

Large language models, such as those developed by OpenAI and Meta, have demonstrated remarkable performance in various tasks, including translation, summarization, and natural language understanding. However, their capacity to handle long-range dependencies remains a crucial aspect of model evaluation. In this context, the Pathfinder and Pathfinder-X datasets have played an essential role in testing model performance on long-range dependency tasks.

Recently, Meta released the Moving Average Equipped Gated Attention model[2] that scored 97% on the Pathfinder-X dataset, highlighting the need for more challenging datasets to push the limits of large language models. This paper introduces Pathfinder-X2, a dataset comprising 512x512 and pixel images designed to evaluate models on long-range dependency tasks.

1

3   Related Work

Pathfinder [1] is a dataset containing 32x32 pixel images, designed to test large language models on their ability to segment a specific white line dash "snake" with a circle at its tip from a collection of distractor snakes. The Pathfinder-X dataset is also built on this premise, using larger 256x256 pixel images, making it more challenging for models to identify the target snake.

4   Pathfinder-X2 Dataset

The Pathfinder-X2 dataset expands on the original Pathfinder and Pathfinder-X datasets by increasing image size to 512x512 pixels. These larger images introduce more complex long-range dependencies, requiring models to process and understand longer sequences of pixels (262,144 pixels for the 512x512 images). In addition, the increased resolution allows for more intricate snake patterns, further challenging the models' ability to discern the target snake from distractors. The dataset features a Segmentation task. The segmentation task requires a model to predict a mask over only the dashed line ending with a circle, and to exclude all other lines (fig 1). The dataset includes 200,000 images for the input, and 200,000 images for the target, or 400,000 images in total.
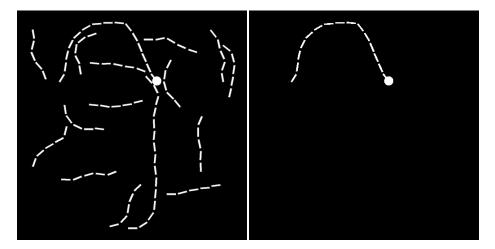

Figure 1: Segmentation Task, Input and Label

The release of the Pathfinder-X2 dataset is intended to stimulate research in this area by providing a new benchmark for evaluating the performance of language models on long-range dependency tasks. By doing so, we hope to inspire the development of new techniques and architectures that can better handle long-range dependencies in natural language processing and other domains.

Overall, the release of the Pathfinder-X2 dataset represents a step forward in the development of language models that can handle long-range dependencies,

and we hope that it will inspire continued progress towards more advanced and powerful language models that can tackle even more complex tasks in the future.

References

[1] Drew Linsley, Alekh K Ashok, Lakshmi N Govindarajan, Rex Liu, Thomas Serre. *Stable and Expressive Recurrent Vision Models.* Brown University, 22 Oct 2020. www.github.com/drewlinsley/pathfinder

[2] Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, Luke Zettlemoyer *Mega: Moving Average Equipped Gated Attention.* arXiv:2209.10655 [cs.LG] 23 Jan 2023.