Perceiver-Architecture-Study

Perceiver: General Perception with Iterative Attention (2021)

Paper Link: https://arxiv.org/abs/2103.03206 (Author: Jaegle et al., DeepMind)

---

1. Problem Addressed

Transformers scale poorly with high-dimensional inputs because self-attention has quadratic $O(N^2)$ complexity, making them inefficient for images, audio, video, and multi-modal data. The paper addresses the problem of building a domain-agnostic neural architecture that can ingest very large inputs (pixels, audio samples, point clouds) without being restricted by input size. The goal is to decouple model depth from input dimensionality, allowing a single architecture to handle many different modalities efficiently.

2. How the Transformer Model Is Applied

Perceiver modifies the Transformer architecture using two key mechanisms:

- **Cross-Attention Bottleneck**
- **Latent Transformer Stack**
- **Fourier Positional Encodings**

In my reproduction, I implemented:
a Fourier encoder → input projection → cross-attention → 4× latent self-attention blocks → mean-pooled classifier. This captures the core ideas of the original architecture using a lightweight PyTorch model.

3. What I Learned From the Paper & Insights From My Reproduction

The paper showed how Transformers can be redesigned to handle very large inputs by using asymmetric cross-attention to extract information into a compact latent bottleneck, drastically reducing computation. I learned how this latent space enables a fixed-cost Transformer stack that is independent of input size, and how Fourier positional encodings replace convolutional priors while still preserving spatial structure. Overall, the work demonstrated a clean, scalable way to build a general-purpose perception model across images, audio, and other modalities.

Training a small Perceiver for 10 epochs on CIFAR-10 reached ~52% validation accuracy, showing that even a lightweight latent-bottleneck model can learn useful representations from raw pixels. The curves were smooth and stable, and the early trend of validation > training accuracy reflected strong generalization under augmentations. These results support the paper's claim that Perceiver remains effective without convolutional priors.

4. Limitations & Interesting Observations

The model's accuracy is naturally capped by the small latent size, shallow depth, and limited training time, along with the absence of heavy augmentations. Because it lacks the spatial biases of CNNs, it learns more slowly and relies heavily on positional encodings to organize raw pixels. Deeper latent stacks and additional cross-attention iterations, as used in the original paper, would further improve performance. Despite these constraints, the reproduction behaved exactly as expected for a lightweight Perceiver.
