Paper Critique: Perceiver: General Perception with Iterative Attention

Dan Peng
Department of Computer Science
Oct 11, 2024
danpeng@unc.edu

1. Research Problem

1.1. What research problem does the paper address?

The paper addresses the problem how to build a general-purpose model capable of handling arbitrary configurations of different modalities without relying on domain-specific architectural assumptions about their inputs. The Perceiver model won't suffer from scalability issues and doesn't require adjustments for different types of data as the traditional models do.

1.2. What is the motivation of the research work?

The motivation stems from how biological systems perceive the world efficiently by processing inputs from multiple modalities simultaneously. However, current models are designed for specific modalities with strong biases and non generality.

2. Technical Novelty

2.1. What are the key technical challenges identified by the authors?

Key technical challenges are:

- how can we design a single flexible model that can handle high-dimensional inputs from various modalities?
- how can we overcome the quadratic scaling problem and inefficiency of standard transformers?

2.2. How significant is the technical contribution of the paper? If you think that the paper is incremental, please provide references to the most similar work

The paper introduces the Perceiver, which is based on transformers but appeals an asymmetric attention mechanism to iteratively distill inputs into a tight latent bottleneck. Compared to transformers, perceiver applies cross-attention first and then self-attention for several times. It also introduces the latent array which is randomly initialized and its size can be controlled at will. In this case, the computational complexity has gone from the original $O(M * M)$ to the current near-linear $O(NM)$, where N $<<$ M.

2.3. Identify 1-5 main strengths of the proposed approach.

- The Perceiver model achieves SOTA results across multiple downstream tasks like image classification, audio classification, and point cloud classification
- Using Fourier features at input allows the model to retain some spatial structure without relying on architectural biases
- The computational complexity of attention mechanism has gone from the original $O(M * M)$ to the current near-linear $O(NM)$, where N $<<$ M

2.4. Identify 1-5 main weaknesses of the proposed approach.

- The experiments missed the text tasks and forgot to benchmark against the basic transformer at this point
- They're not E2E learning since they still employ modality-specific augmentation and position encoding

3. Empirical Results

3.1. Identify 1-5 key experimental results, and explain what they signify.

- The Perceiver achieves 78.0% top-1 accuracy on ImageNet, which is comparable to models like ResNet-50 (73.5%) and ViT(76.7%). It indicates that the architecture of Perceiver is very powerful and can handle image classification tasks without relying on CNNs.
- The Perceiver achieves a 38.3 mAP with raw audio on AudioSet, outperforming many specialized models. Also, its audio+video fusion results in improved

performance (44.2 mAP). These demonstrate its high ability to process both video and audio data.

- The Perceiver achieves 85.7% accuracy on the ModelNet40 dataset, outperforming many transformer-based models on point clouds. Combining with the previous results, it indicates the Perceiver model can handle multiple modal tasks.

3.2. Are there any weaknesses in the experimental section (i.e., unfair comparisons, missing ablations, etc)?

Yes, there are two weaknesses in the experimental design. First, the experiments missed the text tasks and forgot to benchmark against the basic transformer at this point. Second, the experiments doesn't show detailed ablation studies on the effect of weight sharing across all modalities.

4. Summary

The Perceiver presents a general-purpose model by introducing an asymmetric attention mechanism. that scales well across different types of inputs. This latent bottleneck allows it to outperform other transformer-based models or CNNs on standard datasets without relying on domain-specific architectural assumptions. I appreciate this paper for its idea of reducing time complexity from the original $O(M * M)$ to the current near-linear $O(NM)$, where N $<<$ M.

5. QA Prompt for a Paper Discussion

5.1. Discussion Question

Can the Perceiver model handle text tasks compared to transformer based models?

5.2. Your Answer

Yes, the Perceiver model has the potential to handle text tasks. While it performs very well in this paper on multimodal tasks such as images, audio, and point clouds, its use of cross-attention mechanisms is appropriate for handling sequential data like text.