Title: Perceiver

URL: https://en.wikipedia.org/wiki/Perceiver

PageID: 68379149

Categories: Category:Machine learning

Source: Wikipedia (CC BY-SA 4.0).

Perceiver is a variant of the Transformer architecture, adapted for processing arbitrary forms of data, such as images, sounds and video, and spatial data . Unlike previous notable Transformer systems such as BERT and GPT-3 , which were designed for text processing, the Perceiver is designed as a general architecture that can learn from large amounts of heterogeneous data. It accomplishes this with an asymmetric attention mechanism to distill inputs into a latent bottleneck.

Perceiver matches or outperforms specialized models on classification tasks. [1]

Perceiver was introduced in June 2021 by DeepMind . [1] It was followed by Perceiver IO in August 2021. [2]

Design

Perceiver is designed without modality -specific elements. For example, it does not have elements specialized to handle images, or text, or audio. Further it can handle multiple correlated input streams of heterogeneous types. It uses a small set of latent units that forms an attention bottleneck through which the inputs must pass. One benefit is to eliminate the quadratic scaling problem found in early transformers. Earlier work used custom feature extractors for each modality. [1]

It associates position and modality-specific features with every input element (e.g. every pixel, or audio sample). These features can be learned or constructed using high-fidelity Fourier features. [1]

Perceiver uses cross-attention to produce linear complexity layers and to detach network depth from input size. This decoupling allows deeper architectures. [1]

Components

A cross-attention module maps a (larger) byte array (e.g., a pixel array) and a latent array (smaller) to another latent array, reducing dimensionality . A transformer tower maps one latent array to another latent array, which is used to query the input again. The two components alternate. Both components use query-key-value (QKV) attention. QKV attention applies query, key, and value networks, which are typically multilayer perceptrons – to each element of an input array, producing three arrays that preserve the index dimensionality (or sequence length) of their inputs.

Perceiver IO

Perceiver IO can flexibly query the model's latent space to produce outputs of arbitrary size and semantics. It achieves results on tasks with structured output spaces, such as natural language and visual understanding, StarCraft II , and multi-tasking. Perceiver IO matches a Transformer-based BERT baseline on the GLUE language benchmark without the need for input tokenization and achieves state-of-the-art performance on Sintel optical flow estimation. [2]

Outputs are produced by attending to the latent array using a specific output query associated with that particular output. For example to predict optical flow on one pixel a query would attend using the pixel’s xy coordinates plus an optical flow task embedding to produce a single flow vector. It is a variation on the encoder/decoder architecture used in other designs. [2]

Performance

Perceiver's performance is comparable to ResNet -50 and ViT on ImageNet without 2D convolutions . It attends to 50,000 pixels . It is competitive in all modalities in AudioSet . [1]

See also

Convolutional neural network

Transformer (machine learning model)

References

External links

DeepMind Perceiver and Perceiver IO | Paper Explained on YouTube

Perceiver: General Perception with Iterative Attention (Google DeepMind Research Paper Explained) on YouTube , with the Fourier features explained in more detail