Transformers

CS 4277 Deep Learning

Kennesaw State University

KENNESAW STATE UNIVERSITY

Transformers

- Convolutional networks take advantage of the relatedness of nearby pixels in images to deal with the high dimensionality of image data.
- Text data is also high-dimensional with much redundancy that can be exploited, e.g., most instance of `dog` have the same meaning.

Problems: text sequences vary in length and cannot easily be resized.

Consider:

- Please do not cut all my hair off.

shortened to:

Transformers

- Convolutional networks take advantage of the relatedness of nearby pixels in images to deal with the high dimensionality of image data.
- Text data is also high-dimensional with much redundancy that can be exploited, e.g., most instance of `dog` have the same meaning.

Problems: text sequences vary in length and cannot easily be resized.

Consider:

- Please do not cut all my hair off.

shortened to:

- Cut all my hair off.

Processing Text Data

The restaurant refused to serve me a ham sandwich because it only cooks vegetarian food. In the end, they just gave me two slices of bread. Their ambiance was just as good as the food and service.

Dot-Product Self-Attention

Recall that NN layer $\boldsymbol{f}(\boldsymbol{x})$ takes $D \times 1$ input $\boldsymbol{x}$, applies linear transformation to input then passes result to activation function:

$$\boldsymbol{f}(\boldsymbol{x}) = \text{ReLU}(\boldsymbol{\beta} + \boldsymbol{\Omega}\boldsymbol{x})$$

A self-attention block $\boldsymbol{sa}(\cdot)$ takes N inputs.