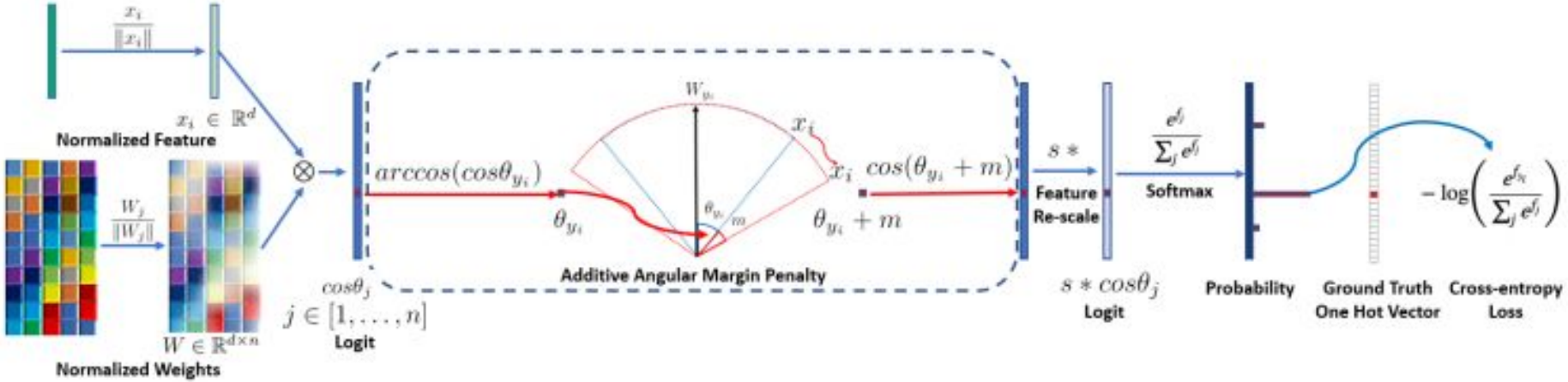
DL Bonus Assignment

Submitted by:
Chiranjeev
P21CS007

Arcface Loss



- We have studied about Softmax with categorical cross entropy loss.
  - Here the probability distribution difference is calculated and minimized while backpropagation.
  - But, this loss doesn't differentiate between classes with a margin.
  - As we want a similar class to be highly similar and different classes to be far-ly dissimilar to each other. But Categorical cross entropy doesn't differentiate with a great margin so the boundary is blurry for class differentiation.
  - So the Arcface loss was introduced which come up with the angular margin in-order to separate different classes with a margin for clear separability between classes.
- ArcFace loss: $$L_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}}$$
  - This loss calculates the Geodesic distance of embeddings.
  - Steps to use/calculate this Arcface loss:
    - The angle(theta) is calculated between normalizedΘ weight vector and feature vector by using formula: $\boldsymbol{\theta}$ **= arccos($W^T X$)**
    - Margin "**m**" is added to the angle: $\boldsymbol{\theta}$**+m**
    - Now we get the arface logits: **s*cos($\boldsymbol{\theta}$+m)** for each class and this is further fed to the softmax function in order to get the probabilities followed by cross entropy loss.

Transformer

- Transformer is a deep learning model in which attention mechanism/approach is introduced in the neural network.
- Attention mechanism helps in focusing much on various regions/parts of the input in order to improve the capability of the model to learn significantly.
- The focus on various parts of input is done by giving weightage to each part of the input on the basis of future prediction to be made.
- Transformer came into the picture through one of the most cited paper published in NIPS - “Attention is All you need”.
- Transformer is basically used to boost the speed of learning while training as transformer lends itself to parallelization.
- It is majorly used in the field of NLP, Language task. But now it is the upcoming growing research of transformer in the field of vision.

Explaining Concept of Transformer in task of Language Translation

- Suppose we want to translate from one language (let's say Hindi) to another language (let's say English).
- There are encoder and decoder modules for this translation purpose.
- In encoder the concept of self-attention is introduced which helps in paying attention to input words(previous position words) which will help in encoding the word at the current position.
- Similarly in decoder the concept of self-attention is used But, here is a little bit change in this attention module i.e., it has multi-head attention module which gives the attention layer multiple representation subspaces.
- Basic steps:
  1. Take Input sentence word-wise as input
  2. Create word embedding for each word
  3. Split word embeddings into multiple heads and get multiplied with weight matrices
  4. Calculate attention using matrices (Query(Q), Key(K), Value(V))
  5. Calculate z scores for each attention head using $\text{Softmax}((Q \times K^T)/\text{sqrt}(d_K)).V$
  6. Use the word now predicted till now with their positional encodings to generate next future translated word. With similar process as above.

* Attention Is All You Need Ashish et.al

Applications of Transformer

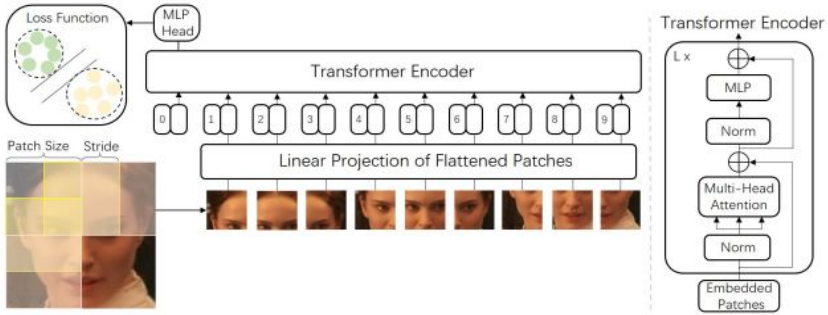
- Majorly it is used in the NLP, Machine Translation task.
- But now it has been observed that Transformer is extended to Image related task. Namely Vision Transformers.
- For Face Recognition task - Face Transformer for Recognition

<u>Steps for Face Transformer:</u>

1. Patches of Face is created and are passed to the Transformer's encoding module which contains Multi-Head Attention part.

2. The embedded patches and output of Multi-Head Attention module are concatenated and passed to Fully Connected Neural Network. Then further general training process is followed like loss calculation and backward propagation.

3. But this paper showed that Normal Convolutional Neural Network performs much better than Face Transformer.



* Face Transformer for Recognition Yaoyao Zhong, Weihong Deng

Thank You !!