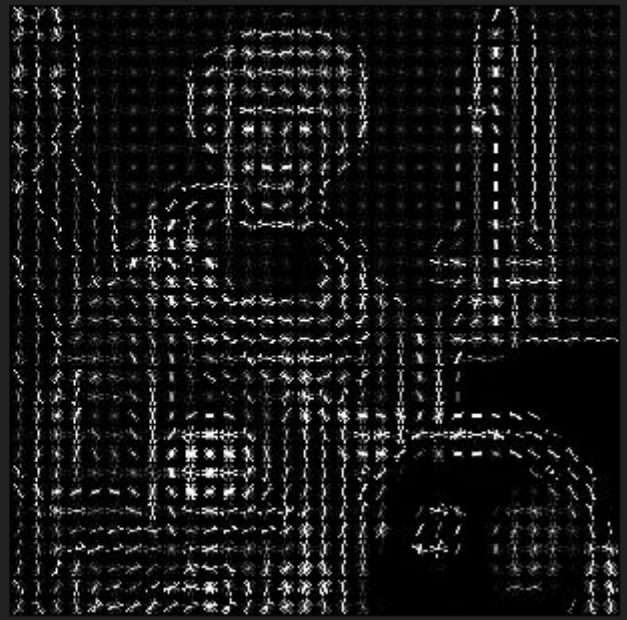
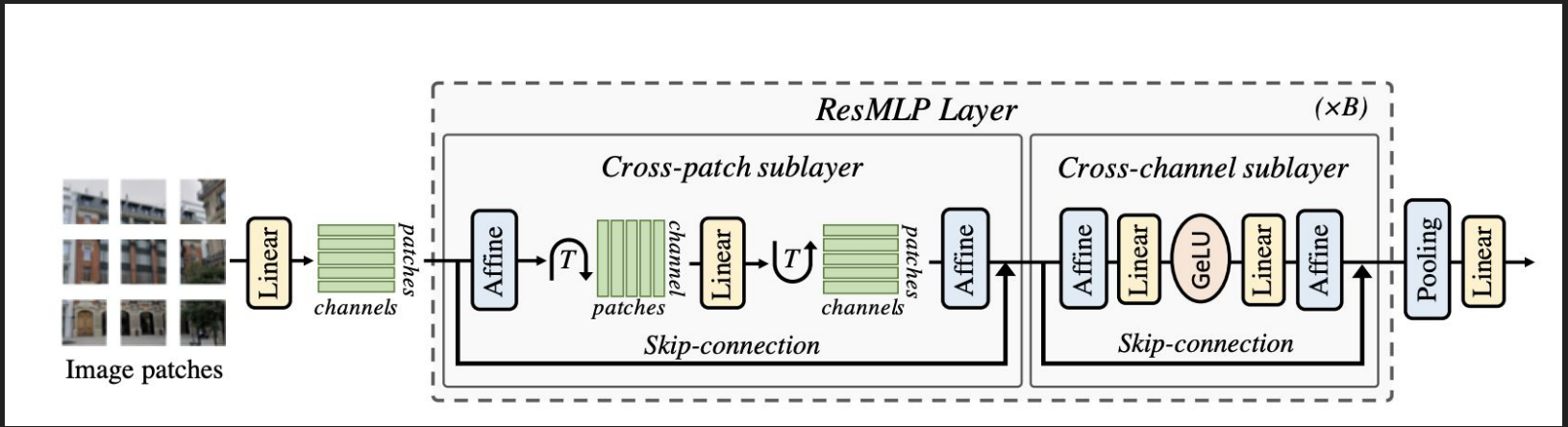
ResMLP

From Hardcoded Feature to non-hardcoded



Architecture



Mathematical Representation of Architecture

$$\texttt{Aff}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{x}) = \texttt{Diag}(\boldsymbol{\alpha})\mathbf{x} + \boldsymbol{\beta},$$

$$\begin{aligned} \mathbf{Z} &= \mathbf{X} + \texttt{Aff}\left((\mathbf{A}\,\texttt{Aff}\,(\mathbf{X})^\top)^\top\right), \\ \mathbf{Y} &= \mathbf{Z} + \texttt{Aff}\,(\mathbf{C}\,\texttt{GELU}(\mathbf{B}\,\texttt{Aff}(\mathbf{Z}))), \end{aligned}$$

ResMlp Unique Features

- No batch norm or Layer norm
- Use Linear layer instead of Attention layer Hence claim to have more stability.
- No positional encoding to encode patches position

Findings During Implementation

Dataset: Cifar 20 Model: ResMlp-s12 patches=16 Embed dim=768

- Direct paper implementation output loss: Nan
- Overfitting at some point

Thanks