Sigmoid Loss for Language Image Pre-Training

- https://arxiv.org/pdf/2303.15343

概要

- vision and languageのpre-trainingの話
- SigLIPを提案
- CLIPはsoftmax lossを使うけどSigLIPはsigmoid lossを使う

CLIP

- 比較のためにCLIPから説明
- データセットからとってきたmini-batchを$\mathcal{B}=\{(I_1,T_1),(I_2,T_2),\cdots\}$とする
- image modelを$f(\cdot)$, text modelを$g(\cdot)$とする
- $\mathbf{x}_i=\frac{f(I_i)}{\|f(I_i)\|_2},\mathbf{y}_i=\frac{g(I_i)}{\|g(I_i)\|_2}$とする
- このときCLIPはpairとなっている(i,i)の確率を大きくするためlossは以下

$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\to\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\to\text{image softmax}}\right)$$

- tは学習可能なパラメータ

SigLIP

- SigLIPでは二値分類タスクを考えて、画像とテキストが与えられたときにペアならpositive, ペアでないならnegativeとする
- このときlossは以下

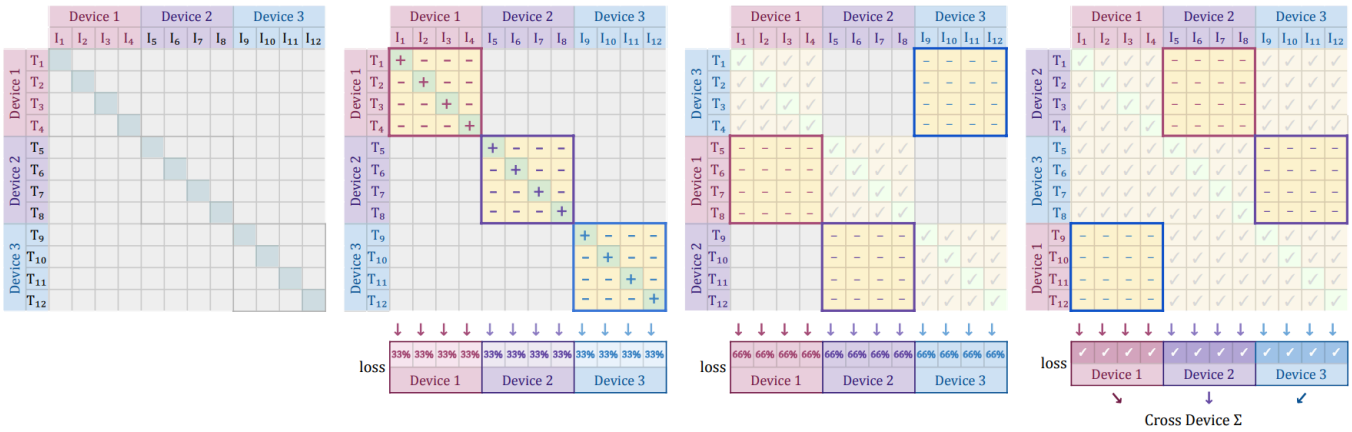
$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

- bは学習可能なパラメータ

Efficient chunked implementation of siglip

- デバイスの個数をDとして、$b=\frac{\|\mathcal{B}\|}{D}$とする
- このとき以下のようにして計算可能(未理解)

$$-\frac{1}{|\mathcal{B}|}\underbrace{\sum_{d_i=1}^{D}}_{\textbf{A: }\forall\text{ device }d_i}\ \underbrace{\sum_{d_j=1}^{D}}_{\textbf{B: swap negs across devices}}\ \underbrace{\underbrace{\sum_{i=bd_i}^{b(d_i+1)}}_{\text{all local positives}}\ \underbrace{\sum_{j=bd_j}^{b(d_j+1)}}_{\text{negs from next device}}\mathcal{L}_{ij}}_{\textbf{C: per device loss}}$$



(論文より引用)

英語

- dissimilar : 異なる
- unstablle : 不安定な
- outdate : 時代遅れの