CLIP

論文ソース

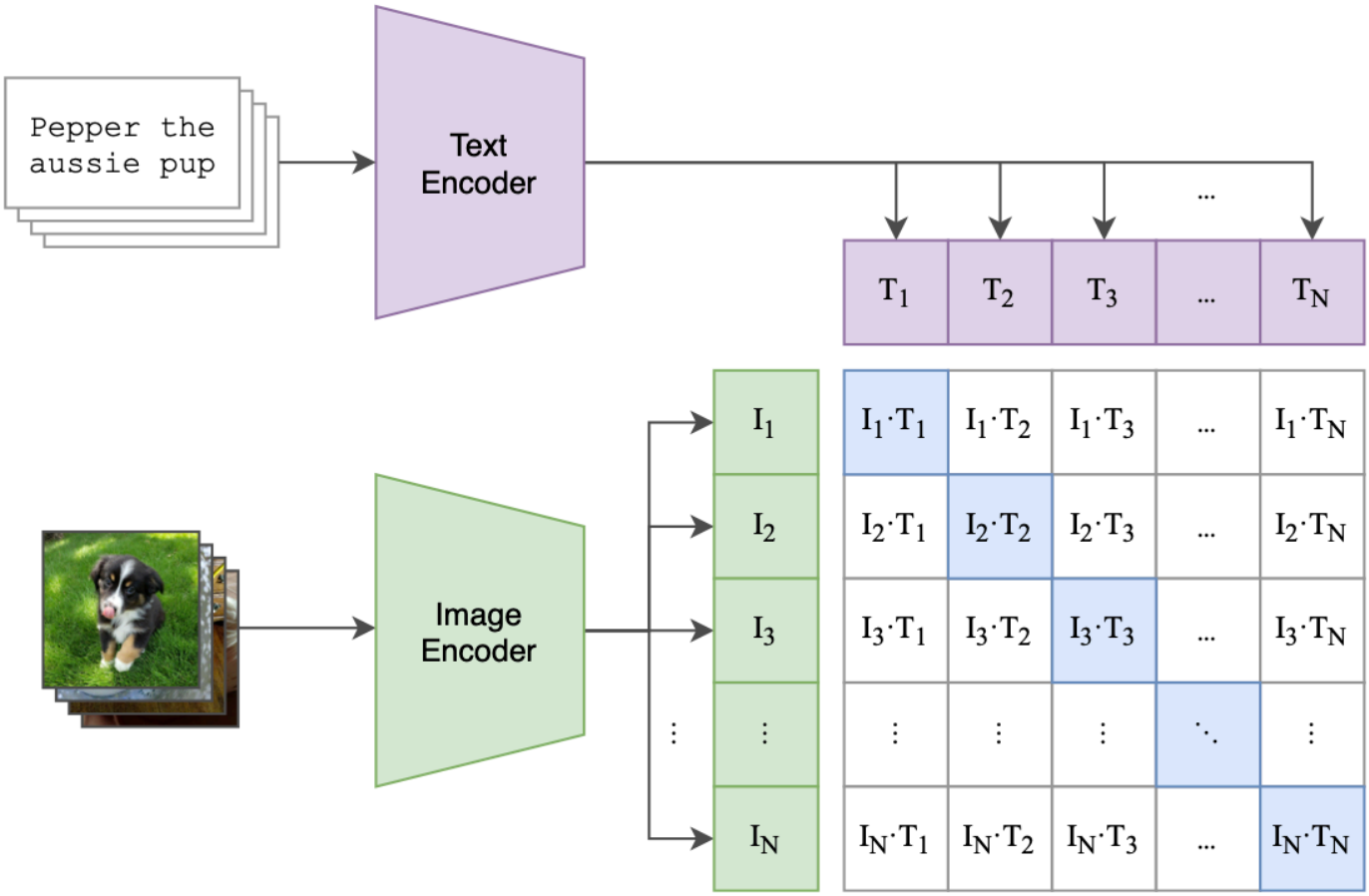
- Learning Transferable Visual Models From Natural Language Supervision

概要

- Contrasive Language Image Pre-training
- 事前学習方法として提案
  - モデルは既存のEncoderをいろいろ使用して実験されている
- 目的:ゼロショットでも画像分類をうまくできるようにする
  - 初めてみるデータセットでも分類したいという意味
  - OpenAIがこれを重視しているらしい(汎用的AIを作りたいから？)
  - 以前の学習手法では以下のような問題がある
    - ラベル付けにコストかかる
    - ラベルの種類が限定的だから初めてみる対象については分類精度が低い
  - CLIPはWebから大量の画像とテキストのペアを取得するためラベル付けいらない
  - ゼロショットを可能にするためにContrasive objectiveを目的関数にした事前学習を行う

対照学習

- 以下が提案学習手法の学習時全体図



- n 個の(画像、テキスト)のペアに対してそれぞれEncoderに入力して潜在表現ベクトルを得る
- 内積とって $n \times n$ の行列を得る
- 得られた潜在空間上でペアになる画像とテキストの類似度が高くなるようにする
- i 番目の画像に対してはi番目のテキストが正解になる
  - 正例が n 個(対角成分)
  - 負例が $n^2 - n$ 個(対角成分以外)
- 以下が論文にある数式

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

- 画像EncoderはResNetかViT
- テキストEncoderはCBOWかText Transformer
- 以下数式の説明
- 1行目:画像 $I \in \mathbb{R}^{n \times h \times w \times c}$ を画像Encoderに入れて $I_f \in \mathbb{R}^{n \times d_i}$ を得る
- 2行目:テキスト $T \in \mathbb{R}^{n \times l}$ をテキストEncoderに入れて $T_F \in \mathbb{R}^{n \times d_t}$ を得る
- 3行目:I_f に対して $W_i \in \mathbb{R}^{d_i \times d_e}$ で埋め込みして正規化して $I_e \in \mathbb{R}^{n \times d_e}$ を得る
- 4行目:T_f に対して $W_t \in \mathbb{R}^{d_t \times d_e}$ で埋め込みして正規化して $T_e \in \mathbb{R}^{n \times d_e}$ を得る
- 5行目:内積とって $n \times n$ の行列 $logits$ を得る
  - 値の大きさは e^t で調整
- 6行目:$labels = (0, 1, \cdots, n-1)$ 作成
- 7行目:画像を基準にした誤差 $loss_i$ を計算する
  - 行方向にクロス・エントロピー誤差を計算

    - 各行 i に対して i が正解となるように先程の $labels$ を使用して記述している
- 8行目:テキストを基準にした誤差 $loss_t$ を計算する
    - 列方向にクロス・エントロピー誤差を計算
- 9行目:得られた2つのロスの平均をとって全体のロスとする。これを小さくするように学習する