CLIPScore:

Abstract

1

2

3 CLIPScore

モデルの紹介

CLIP（Radford et al, 2021）は、ウェブから収集した400M（画像、キャプション）のペアで学習したクロスモーダル検索モデルである。一般的なユニグラム/ビグラム、名前付きエンティティなどからなる500Kの検索クエリを検索エンジンで実行した。各クエリに対して、最大20K（画像とキャプション）のペアが収集された。

私たちが使用しているモデルはViT-B/32バージョンです。Vision Transformer (Vaswani et al., 2017; Dosovitskiy et al., 2021)を介して画像を表現し、畳み込みフィルターを使用せず、224×224ピクセルの入力画像を均等に分割する7×7グリッドの画像パッチ間で計算される自己注意マップを使用する。このモデルには12層のトランスと86Mのパラメーターがある。テキストは同様に、49KのBPEトークン・タイプの語彙（Sennrich et al, 2016）に対して学習された12層の変換器によって表現される（Radford et al (2019)で詳しく説明されている）。これらのベクトルはそれぞれ、入力されたキャプションや画像の内容を表すことを目的としている。ViT-B/32の場合、これらのベクターは512-Dである。モデルの重みは、InfoNCE (Sohn, 2016; Oord et al., 2018)を用いて、真に対応する画像／キャプションのペアのスケールされた余弦類似度を最大化すると同時に、不一致の画像／キャプションのペアの類似度を最小化するように学習される。実験ではこの重みのセットを固定した。

CLIPによるキャプション生成の評価

候補生成の品質を評価するために、画像と候補キャプションの両方をそれぞれの特徴抽出器に通す。次に、埋め込み結果の余弦類似度を計算する。候補の前にプロンプトを付けると相関が少し改善することがわかった：「しかし、Radfordら（2021）の推奨プロンプトである "A photo of "も有効であっ

た。Zhangら（2020）に従い、再スケーリング操作を行う。視覚的なCLIP埋め込み $\mathbf{v}$ を持つ画像と、テキスト的なCLIP埋め込み $\mathbf{c}$ を持つキャプション候補に対して、$w = 2.5$ と設定し、CLIP-Sを次のように計算する.

$$\text{CLIP-S}(\mathbf{c}, \mathbf{v}) = w * \max(\cos(\mathbf{c}, \mathbf{v}), 0)$$

コーパスレベルのCLIP-Sを計算するには、単純に(候補と画像)のペアを平均する。この評価は基本的なリファレンスには依存しないことに注意。ViT-B/32をバックボーンとするCLIP-Sのランタイムは高速だ。私たちのコンシューマー向けGPUとハードディスク・ドライブ1台で、1分間におよそ4Kの画像候補のペアリングを処理できる。

RefCLIPScore

CLIP-Sはさらに、リファレンスがあれば、それを組み込むように拡張することができる。CLIPのテキスト変換器に通すことで、利用可能な各参照のベクトル表現を抽出する。その結果が、すべての参照のベクトル表現の集合、Rである。次に、RefCLIPScoreはCLIP-Sの調和平均として計算され、最大参照余弦類似度、すなわち

$$\text{RefCLIP-S}(\mathbf{c,R,v}) = \text{H-Mean}(\text{CLIP-S}(\mathbf{c,v}), \max(\max_{\mathbf{r} \in \mathbf{R}} \cos(\mathbf{c,r}), 0))$$

となる.