LUAN MANTEGAZINE

FERRAMENTA PARA BUSCA DE IMAGENS USANDO LINGUAGEM NATURAL

Desenvolvimento de aplicações BigData e FoG/EDGE

2023

INSTITUTO DE INFORMÁTICA UFRGS

Sumário

Introdução

- **Trabalhos relacionados:**
  - **Métodos de pré-treinamento que aprendem diretamente do texto** (Dai&le,2015; Perter et al.,2018; Howard e Ruder, 2018; Radford et al., 2018; Devlin et al., 2018; Raffel et al., 2019)
  - **O desenvolvimento do “texto para texto” como uma interface de entrada-saída padronizada** (McCann et al.,2018; Radford et al., 2019; Raffel et al. 2019)
  - **Sistema GPT-3, que exige pouco ou nenhum conjunto de dados de treinamento especifico** (Brown et al.,2020)

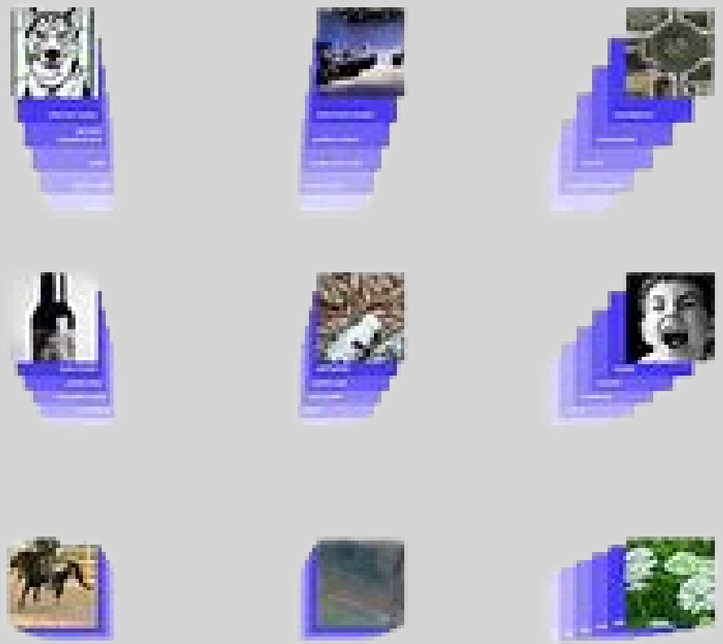
Introdução

- Trabalhos recentes que utilizam arquiteturas profundas para recuperação de imagens se limitam principalmente ao uso de uma rede pré-treinada como extrator de recursos locais.

- O projeto usa o modelo CLIP (Contrastive Language-Image Pre-Training) da OpenAI para gerar incorporações para as imagens no conjunto de dados e para a consulta do usuário. As incorporações são usadas para calcular a semelhança entre as imagens e a consulta

Introdução

- **Modelo CLIP**
  - É um avançado modelo de aprendizado de máquina Ele foi projetado para entender e correlacionar informações de texto e imagens, permitindo que o modelo compreenda o contexto visual e linguístico de maneira conjunta



Introdução

- **Modelo CLIP**

  - **Aprendizado de Máquina Contrastivo**: O CLIP é baseado em aprendizado de máquina contrastivo, o que significa que ele é treinado para encontrar semelhanças e diferenças entre pares de dados. Isso é feito treinando o modelo em tarefas que exigem que ele relacione informações visuais e textuais.

  - **Compreensão Multimodal:** É capaz de compreender texto e imagens simultaneamente. Isso o torna versátil para uma ampla gama de tarefas que envolvem a compreensão de informações em diferentes modalidades, como classificação de imagens, busca de imagens por descrições textuais, tradução de texto para imagens e vice-versa, entre outras.

Introdução

- **Modelo CLIP**

  - **Pré-treinamento**: Antes de ser usado em tarefas específicas, o modelo é pré-treinado em um grande conjunto de dados que contém pares de texto e imagens de toda a Internet. Isso permite que o modelo adquira um conhecimento geral do mundo e desenvolva representações semânticas para diferentes conceitos.

  - **Transferência de Aprendizado**: Após o pré-treinamento, o modelo pode ser ajustado (fine-tuning) para tarefas específicas, como classificação de imagens, geração de texto a partir de imagens, resolução de quebra-cabeças e muito mais. Ele demonstrou desempenho impressionante em várias dessas tarefas, muitas vezes superando modelos anteriores.

Metodologia

- **DATASET**

kaggle → Unsplash

Plataforma online de compartilhamento de imagens e fotografias de alta qualidade.

- Pacote com 25 mil imagens
- 13GB

Metodologia

- **Codificadores**

  - **Codificador Visual:** É uma rede neural convolucional (CNN) que processa as imagens de entrada. Sua função é extrair representações semânticas das imagens, convertendo-as em vetores numéricos de alta dimensão. Esses vetores capturam informações visuais das imagens, permitindo que o modelo compreenda seu conteúdo visual.

  - **Codificador de Linguagem**: É uma rede neural que converte texto em vetores numéricos de alta dimensão. Ele codifica as descrições de texto, títulos ou outras informações textuais associadas às imagens em representações numéricas que capturam o significado semântico do texto.

Metodologia

- **Treinamento (Constrative Learning)**

Metodologia

- **Classificação (Fine-Turning)**

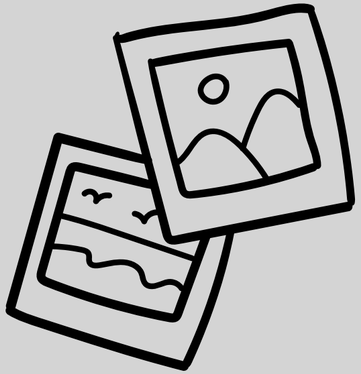
Entrada das labels

Análise das legendas das imagens do banco de dados

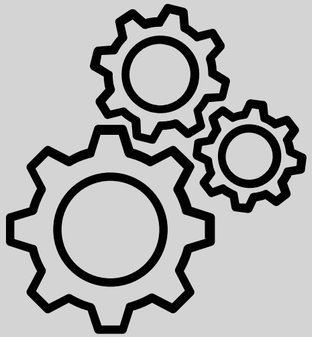
Codificador de texto

Metodologia

- **Predição (zero-shot prediction)**



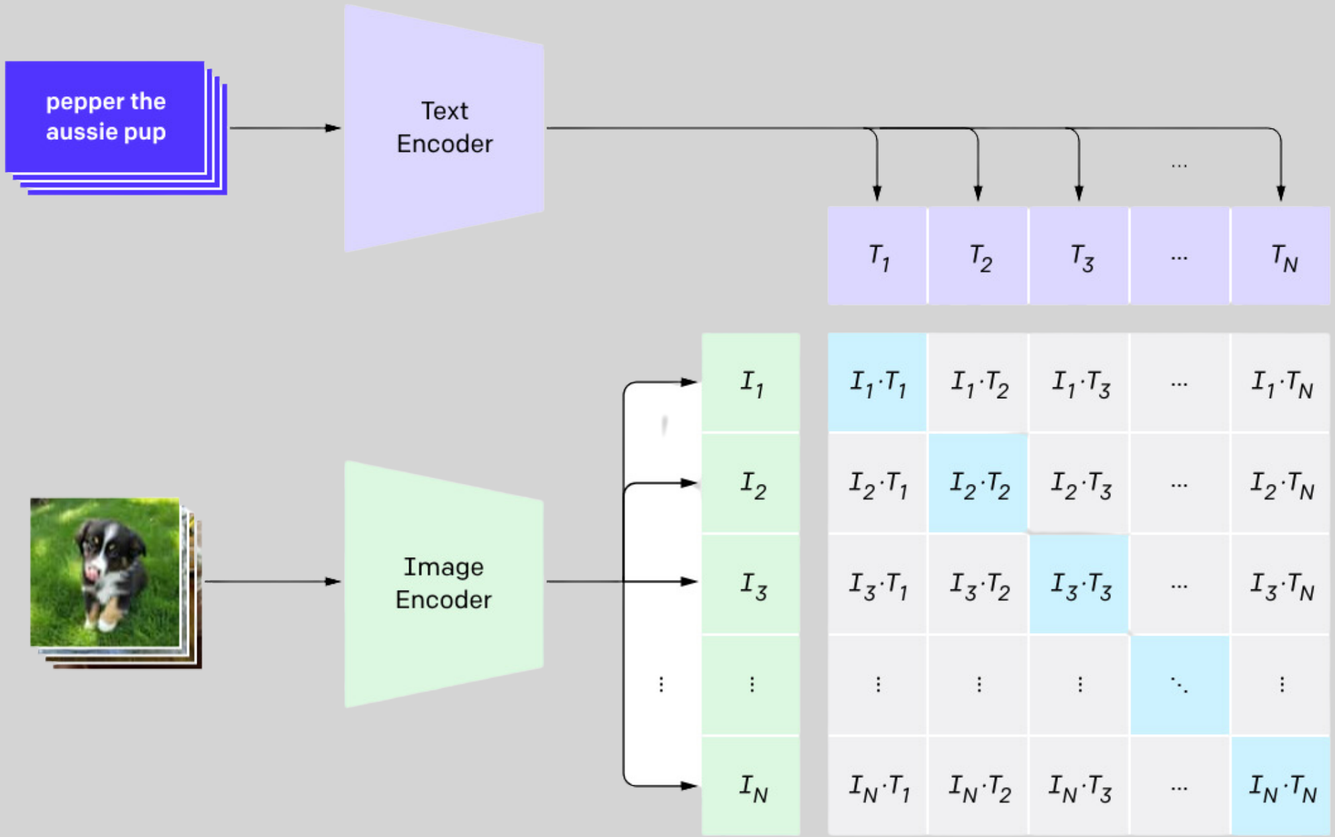
Imagens do banco de dados



Codificador de imagem

Metodologia

- **Esquema final**



Metodologia

- **Aplicação Web**

Metodologia

- **Avaliação do algoritmo**

Top-K accuracy

Para avaliar o modelo de codificador duplo, foi utilizado as imagens e legendas de amostras fora de treinamento para avaliar a qualidade da recuperação.

"Top-K accuracy"- é uma métrica de avaliação usada em tarefas de classificação multiclasse, especialmente em problemas onde não é suficiente considerar apenas a previsão correta mais provável. Em vez disso, essa métrica avalia se a classe correta está entre as K principais previsões feitas pelo modelo.

Metodologia

- **Avaliação do algoritmo**

Por que Top-K accuracy?

Modelos que integram informações visuais e textuais, podem fazer previsões com base em múltiplos modais de dados (texto e imagem). Nesse contexto, Top-K Accuracy pode ser relevante para avaliar se o modelo é capaz de relacionar adequadamente texto e imagens, identificando não apenas a previsão mais provável, mas também previsões alternativas que são semanticamente relevantes.

Resultados

```
  0%|

Scoring training data...

100%|██████████████████████████████████████████████████████████████████████████████
  0%|

Train accuracy: 13.373%
Scoring evaluation data...

100%|██████████████████████████████████████████████████████████████████████████████

Eval accuracy: 6.235%
```

Considerações Finais

- Aplicar em sistemas de grande porte, esta etapa é realizada utilizando uma estrutura paralela de processamento de dados, como Apache Spark ou Apache Beam;

- Realizar PLN em português;

- Treinar a CNN para leituras de diagnósticos médicos sobre imagens de CT/MRI

OBRIGADO

LUAN.MANTEGAZINE@INF.UFRGS.BR