Learning Transferable Visual Models From Natural Language Supervision

Indo contra o estado da arte, que atualmente, no campo de visão computacional concentra-se em treinar para predizer um conjunto de categorias, este artigopropõe aprender diretamente através de textos sobre as imagens, o que torna-se uma ótima alternativa, devido a dificuldade de dados com rótulos em visão computacional. Demonstra-se que com um pré-treino simples é possível predizer qual título refere-se a imagem. Também é possível pré-treinar modelos de linguagem natural para representar conceitos visuais, permitindo assim o uso de abordagens “zero-shot”.

Há algum tempo métodos de linguagem natural vêm revolucionando a ciência, sistemas importantes como o GPT-3 são muito competitivos através de várias tarefas. Porém para áreas de visão computacional, ainda é comum realizar pré-treinos com um conjunto de dados gigantescos. Vários pesquisadores investigam como representar imagens através de textos. Porém provas de conceitos utilizando linguagem natural para representação de imagens ainda é algo complicado e raro de ser utilizado. Esta escassez de trabalhos nesta linha deve-se aos baixos resultados obtidos em alguns trabalhos em comparação com técnicas já bem estabelecidas como estado da arte ou até mesmo técnicas mais clássicas. Porém alguns problemas podem ser observados, entre eles a utilização de uma função softmax estática para predizer e utilizar mecanismos dinâmicos.

Observa-se que aprender através de linguagens naturais têm grandes pontos positivos, uma vez que é mais fácil escalar se comparado com modelos tradicionais, visto que não necessitam de rótulos, ao contrário, podem aprender passivamente através da vasta quantidade de texto. Também não é necessário aprender uma representação, deve-se aprender como conectar a representação em relação à linguagem.

Em relação aos dados disponíveis, se comparado com conjunto de dados de visão computacional, com bilhões de imagens, quando filtrado e organizado, os dados passam a ter poucos exemplos, de 100 milhões passam para 6 à 15 milhões de conjuntos imagem, texto. Para isto, no trabalho, construiu-se um novo conjunto de dados com 400 milhões de pares (imagem, texto), coletados em através da internet. Chamado então de WIP.

Para seleção de um modelo para pré-treino, focou-se principalmente na eficiência do treinamento. Observou-se que alguns trabalhos focaram em predizer a exata palavra que acompanha a imagem, o que é muito complexo, neste trabalho focou-se em predizer qual texto como um todo corresponde a uma imagem, em seu sentido. Para isto, utilizou-se na abordagem CLIP um modelo com espaço multimodal treinando um codificador para imagem e outro para texto, procura-se maximizar a similaridade do cosseno das imagens e textos dos pares reais e diminuir as dos pares incorretos. Não se preocupou com o sobreajuste, o que simplificou o modelo, devido a grande quantidade de dados. Também, modificou-se o modelo, não utilizando uma projeção não linear entre a representação e o espaço, ao invés, uma projeção linear foi utilizada, não houve diferenças de eficiência entre estes dois modos nos treinos. Outra simplificação em relação aos trabalhos relacionados foi a remoção da função t_u que realiza uma amostra de uma única frase uniformemente, visto que no dado usado os valores são de apenas uma frase. Também simplificou-se a função t_v. Como aumento de dados, utilizou-se apenas um corte quadrado aleatório. A Temperatura t controla o alcance dos logitos e é um hiperparâmetro a ser otimizado.

Considerou-se múltiplas arquiteturas, ResNet com várias modificações, e Vision Transformeres, com pequenas alterações. Foram considerados também, para verificação de robustez do modelo, um treinamento zero-shot, onde verifica-se o vetor de características da imagem e dos textos através do método de similaridade dos cossenos escalados por uma temperatura t e finalmente normalizado via uma distribuição softmax.