Resumo – Learning Transferable Visual Models From Natural Language Supervision

O artigo estuda o comportamento de classificadores de imagem treinados com linguagem natural auto-supervisionada de larga escala.

O aprendizado de linguagem natural oferece algumas vantagens sobre outros métodos: é muito mais fácil escalá-lo que bancos de imagens classificadas coletivamente, já que não requer anotações em formato compatível com rotinas de machine learning. Ao contrário, métodos que funcionam em linguagem natural podem aprender passivamente de massas de textos disponíveis na internet. Outra vantagem desse tipo de aprendizado é que não apenas aprendem uma representação do conhecimento mas também conecta essa representação à linguagem, o que permite transferência do zero.

Favorecidos pela larga disponibilidade de dados disponibilizados publicamente nesta forma na Internet, os autores criaram um conjunto de dados de 400 milhões de pares (imagem, texto), e 500.000 consultas, cada qual balanceada com aproximadamente 20.000 pares associados.

Criaram um método eficiente de pré-treinamento a partir de linguagem natural auto-supervisionada, chamado de Contrastive Language-Image Pre-training (CLIP. Ele aprende um espaço de embedding multi-modal que treina simultaneamente um codificador de imagem e outro de texto de forma a maximizar a similaridade de cosseno de um texto e uma imagem e, ao mesmo tempo, minimizar a similaridade de cosseno entre embeddings de pares incorretos, depois aplica função de entropia cruzada a ser otimizada. Para codificador de imagem, usou duas arquiteturas: uma ResNet-50 modificada /9ou similares), e um Vision Transformer (ViT). Para codificador de texto, empregou um Transformer modificado.

Os experimentos realizados incluem transferência do zero, que é a capacidade de generalização de um modelo para bases de dados nunca vistas. O artigo estuda a escalabilidade do CLIP pelo treinamento de uma série de 8 modelos distribuídos em quase 2 ordens de magnitude de computação e observa que o desempenho da transferência é uma função aproximada da computação. O CLIP, à família GPT, aprende a executar um conjunto abrangente de tarefas durante pré-treinamento incluindo geo-localização, reconhecimento ótico de caracteres, reconhecimento de emoções faciais, e reconhecimento de ação.

A medição é realizada pela comparação do desempenho zero-shot do CLIP sobre mais de 30 bancos de dados e ele pode ser competitivo com modelos supervisionados treinados por tarefas específicas existentes.

Outros experimentos medem a robustez com relação a mudanças na distribuição natural. Modelos ImageNet com alta acurácia ainda cometem erros em bases modificadas. É difícil generalizar muito a partir de achados em modelos treinados sobre o ImageNet. Modelos CLIP, por outro lado, alcançam escores relativamente maiores em relação a modelos ImageNet com acurácia original equivalente, o que sugere que a avaliação de modelos zero-shot agnósticos é muito mais representativa da capacidade de um modelo. Modelos CLIP também são mais eficientes computacionalmente em relação a modelos ImageNet equivalentes.

O trabalho inclui ainda experimentos acerca de vieses sociais. Os modelos CLIP são especialmente vulneráveis a decisões de design que gerem esses vieses, dada a flexibilidade que oferecem. Os experimentos ilustram problemas potenciais decorrentes do design de classe e outras fontes de preconceito, e destinam-se a estimular a investigação.