01.11.2023
FIAP - 2TIAR - Visão Computacional

Henrico Nardelli | RM 95985
Emilly Gabrielly | RM 94437
Felype Nunes | RM 96232
Sara Leal | RM 96302
Daniel Faria | RM 94026

Visual Transformers - Vantagens e Desvantagens

Entendendo as vantagens e desvantagens do uso de Visual Transformers

Introdução

Os Visual Transformers, ou Transformers Visuais, representam uma inovação significativa no campo da visão computacional e processamento de imagens. Inspirados na arquitetura dos modelos de linguagem baseados em Transformers, esses modelos foram projetados para capturar relações complexas e de longo alcance em dados visuais, tornando-se uma abordagem revolucionária para tarefas como classificação de imagens, detecção de objetos, segmentação semântica e muito mais.

Eles introduzem a capacidade de processar imagens em sua totalidade, em vez de depender de técnicas tradicionais de extração de características, abrindo caminho para uma nova era de eficiência e desempenho em visão computacional. Nesta era dos Visual Transformers, a análise de imagens está se beneficiando do poder da atenção, paralelismo e escalabilidade inerentes a essa arquitetura, redefinindo os padrões em uma ampla gama de aplicações visuais.

Vantagens e desvantagens

Os Visual Transformers, uma extensão dos Transformers para processamento de imagens, oferecem vantagens e desvantagens importantes. Entre as vantagens, destaca-se a capacidade de modelar relações de longo alcance em imagens, permitindo uma compreensão global da cena, o que os torna eficazes em tarefas como segmentação, detecção de objetos e classificação de imagens.

Além disso, podem aprender automaticamente representações em múltiplas escalas, sendo úteis na identificação de objetos de diferentes tamanhos e contextos visuais variados. Modelos pré treinados, como o ViT, podem ser ajustados para tarefas específicas, economizando recursos de treinamento. Também são adaptáveis para processar informações multimodais, como texto e imagem, sendo aplicáveis em descrições de imagens.

Por outro lado, as desvantagens incluem a complexidade computacional, pois os Visual Transformers têm um grande número de parâmetros, tornando o treinamento e inferência computacionalmente intensivos. Além disso, exigem conjuntos de dados rotulados em grande escala para treinamento, o que pode ser desafiador de se obter. A interpretabilidade é limitada em comparação com métodos tradicionais de visão computacional, o que pode ser uma desvantagem em cenários onde a explicabilidade é crucial. Também podem não ser ideais para dados sequenciais, como vídeos, e podem herdar viés dos dados de treinamento.

Portanto, a escolha de usar Visual Transformers depende da natureza da tarefa, dos recursos disponíveis e das necessidades específicas do projeto, considerando cuidadosamente suas vantagens e desvantagens.

CNNs e ResNets x Transformers

Os Transformers têm superado técnicas convencionais, como CNNs e ResNets, em problemas de visão computacional. Sua capacidade de capturar relações de longo alcance e processar imagens de forma mais eficiente tem proporcionado melhorias significativas em tarefas desafiadoras, apontando para um futuro promissor na resolução de problemas visuais complexos.

A arquitetura de atenção dos Transformers, que permite a captura de relações de longo alcance entre pixels ou regiões de uma imagem, tem demonstrado vantagens notáveis em tarefas desafiadoras. Os modelos baseados em Transformers, incluindo os Visual Transformers, conseguem processar imagens em sua totalidade, eliminando a necessidade de etapas de extração de características manuais. Isso resultou em melhorias significativas na eficiência e na capacidade de generalização para diferentes domínios, tornando-os competitivos e, em muitos casos, superiores às técnicas tradicionais.

No entanto, seu rápido progresso e capacidade de lidar com informações contextuais complexas estão impulsionando a evolução da visão computacional e apontando para um futuro onde os Transformers irão desempenhar um papel cada vez mais proeminente na resolução de problemas visuais complexos.

CNNs e Transformers em detecção de objetos

Convolutional Neural Networks (CNNs) têm sido a pedra angular da visão computacional por muitos anos e demonstraram notável eficácia em tarefas como classificação de imagens e detecção de objetos. Sua capacidade de extrair características espaciais e locais em imagens torna-os ideais para detectar padrões visuais em dados visuais.

No entanto, modelos baseados em Transformers, como o Vision Transformer (ViT), surgiram como concorrentes promissores. Eles têm a capacidade de capturar relações de longo alcance em imagens e aprender representações em várias escalas espaciais. Em alguns casos, os Transformers têm superado as CNNs, especialmente em tarefas que envolvem conjuntos de dados com contextos visuais complexos ou relações espaciais difíceis de modelar com CNNs.

A escolha entre CNNs e Transformers depende de fatores como a natureza da tarefa, o tamanho do conjunto de dados, os recursos computacionais disponíveis e os objetivos do projeto. Em muitos cenários, a combinação de ambas as arquiteturas ou o ajuste fino de modelos pré-treinados pode ser a estratégia mais eficaz para obter o melhor desempenho em tarefas de visão computacional.

É importante notar que a pesquisa continua evoluindo nessa área, e novos desenvolvimentos podem levar a avanços significativos tanto em CNNs quanto em modelos baseados em Transformers, tornando a escolha da arquitetura uma consideração crítica para qualquer projeto de visão computacional.

Uso de Visual Transformers

Os Visual Transformers têm demonstrado eficácia em uma ampla gama de problemas além da visão computacional, incluindo processamento de linguagem natural (NLP), tais como tradução automática e geração de texto. Eles também mostram potencial em áreas como reconhecimento de fala, sequenciamento de DNA e até mesmo na otimização de problemas complexos. Suas vantagens na captura de relações de longo alcance e no processamento de dados sequenciais têm impactado positivamente várias disciplinas. Um exemplo é o GPT-3, um modelo baseado em Transformers que se destacou em tarefas de PLN, incluindo questionamento e resposta, geração de texto e muito mais.

Conclusão

Em conclusão, os Visual Transformers representam uma inovação impressionante no campo da visão computacional e do processamento de imagens. Inspirados na arquitetura dos modelos baseados em Transformers, esses modelos têm o potencial de transformar a maneira como compreendemos e interagimos com dados visuais complexos. Eles superaram desafios, como a captura de relações de longo alcance em imagens e a eficiência no processamento de dados visuais, oferecendo melhorias notáveis em tarefas desafiadoras, como classificação de imagens, detecção de objetos e segmentação semântica.

No entanto, é importante lembrar que a escolha entre Visual Transformers e outras arquiteturas, como CNNs e ResNets, depende das necessidades específicas do projeto, recursos disponíveis e natureza da tarefa. A pesquisa continua evoluindo nessa área, e a combinação de diferentes arquiteturas ou o ajuste fino de modelos pré-treinados pode ser a estratégia mais eficaz para obter o melhor desempenho em tarefas de visão computacional.

Além disso, os Visual Transformers não se limitam à visão computacional e têm se mostrado promissores em diversas outras disciplinas, ampliando seu impacto e contribuindo para avanços em áreas como processamento de linguagem natural, reconhecimento de fala e muito mais. Assim, eles representam não apenas uma inovação no processamento de imagens, mas também uma tendência que está moldando o futuro da inteligência artificial e da computação visual.

Referências

- [2005.14165] Language Models are Few-Shot Learners (arxiv.org)
- Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm | Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (acm.org)
- Visual Transformer Meets CutMix for Improved Accuracy,Communication Efficiency, and Data Privacy in Split Learning
- End-to-End Object Detection with Transformers