Dominio emergente de los Transfomers sobre CNN

Cristhian Sanchez
UNI
Lima, Peru
csanchez@uni.pe

Fernando Zambrano
UNI
Lima, Peru
fzambranoa@uni.pe

Angel Larreategui
UNI
Lima, Peru
alarreateguic@uni.pe

Abstract

A pesar de que las Redes Convolucionales (CNN) son el estándar actual en reconocimiento de imágenes, la arquitectura Transformer (que se ha convertido en el estándar de facto para tareas de procesamiento del lenguaje natural), aplicado directamente a secuencias de parches de una imagen, puede realizar muy bien varias de estas tareas. Cuando se pre-entrena de forma supervisada en grandes cantidades de datos, Vision Transformer (ViT) logra excelentes resultados (88.55% en ImageNet) en comparación con las CNN de última generación, al tiempo que requieren sustancialmente menos recursos computacionales para entrenarse[1]. Asimismo, cuando se pre-entrena de manera autosupervisada bajo el enfoque de DINO, ViT muestra propiedades emergentes de fácil interpretabilidad que podrían usarse en tareas de segmentación, copy detection, entre otros.

1 Introducción

Inicialmente presentados en el 2017, los Transformers (Vaswani et al. [1], 2017) han sido usados en muchas tareas de NLP por su beneficiosa paralelización en el entrenamiento. Gracias a la escalabilidad y eficiencia computacional de los Transformers, se ha hecho posible entrenar modelos de tamaño sin precedentes, con varios billones de parámetros. Con el crecimiento de modelos y conjuntos de datos, todavía no hay señales de saturación del rendimiento.

En el campo de la visión por computadora, sin embargo, las arquitecturas convolucionales siguen siendo dominantes (LeCun et al. [2], 1989; Krizhevsky y col. [3], 2012; He et al. [4], 2016). Inspirados por los éxitos en NLP, varios trabajos intentan combinar arquitecturas similares a las de CNN con algunas componentes de los Transformers (Wang et al. [5], 2018. Si bien estos modelos son teóricamente eficientes, aún no se han escalado de manera efectiva en aceleradores de hardware modernos (GPU's y TPU's) debido al uso de complejos patrones de diseño. Por lo tanto, en el reconocimiento de imágenes a gran escala, las arquitecturas clásicas como ResNet (en la figura 1 se muestra la arquitectura ResNet-50) siguen siendo de vanguardia (Mahajan et al. [6], 2018; Xie et al. [7], 2020; Kolesnikov et al. [8], 2020).

Los autores de este trabajo se inspiraron en los éxitos de Transformer en NLP, aplicaron un Transformer estándar directamente a las imágenes, dividiendo una imagen en parches (o sea en porciones de la imagen original) y proporcionaron la secuencia de incrustaciones (o embeddings) lineales de estos parches como entrada al Transformer. Los parches de imagen se tratan de la misma manera que los tokens (palabras) en una aplicación de NLP. Ellos entrenaron el modelo de clasificación de imágenes de forma supervisada. Cuando se entrena en conjuntos de datos de un tamaño mediano como ImageNet, dichos modelos producen modestos accuracy's de unos pocos

[1] Sin embargo, siguen siendo recursos muy poco accesibles para principiantes

puntos porcentuales por debajo de ResNets, el cual es de tamaño comparable.

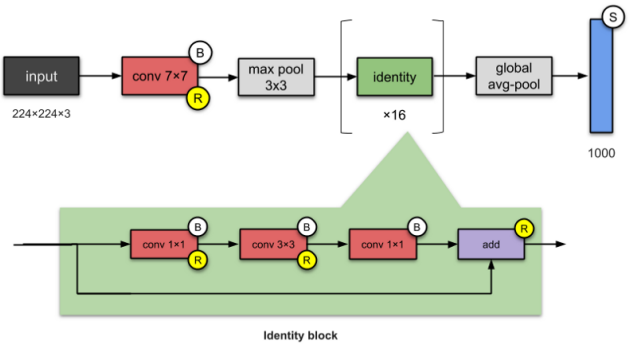

Figure 1: Arquitectura convolucional ResNet-50 [4]

El resultado parece negativo, y puede deberse a que los Transformers carecen de algunos de los sesgos inductivos que son esenciales a las CNN, como la invarianza traslacional y de localidad, por lo que no se generaliza bien cuando se entrenan con pequeñas cantidades de datos. Sin embargo, el panorama cambia si los modelos se entrenan en conjuntos de datos más grandes (entre 14M-300M imágenes). Se encontró que el entrenamiento a gran escala triunfa sobre el sesgo inductivo. Vision Transformer (ViT) logra excelentes resultados cuando se pre-entrena a una escala suficiente y transfiriendo el aprendizaje a tareas con menos datapoints. En particular, el mejor modelo alcanza el accuracy del 88,55% en ImageNet, el 90,72% en ImageNet-ReaL, el 94,55% en CIFAR-100, y 77,63% en el conjunto de 19 tareas de VTAB.

Los avances relacionados con ViT no se quedaron allí, por lo que en la búsqueda de más eficiencia e interpretabilidad del modelo, ViT fue entrenado bajo el enfoque auto-supervisado de DINO, consiguiendo de esa manera resultados muy alentadores en la visualización de características, las cuales pueden ser manejadas con el algoritmo K-NN. En particular, entrenando DINO con ViT con solo 8-GPU's en 3 días se logró un accuracy del 76.1% en ImageNet.

References

[1] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[2] Y. LeCun et al. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1 (1989), pp. 541–551.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012, pp. 1097–1105. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[4] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[5] Xiaolong Wang et al. *Non-local Neural Networks*. 2018. arXiv: 1711.07971 [cs.CV].

[6] Dhruv Mahajan et al. *Exploring the Limits of Weakly Supervised Pretraining*. 2018. arXiv: 1805.00932 [cs.CV].

[7] Qizhe Xie et al. *Self-training with Noisy Student improves ImageNet classification*. 2020. arXiv: 1911.04252 [cs.LG].

[8] Alexander Kolesnikov et al. *Big Transfer (BiT): General Visual Representation Learning*. 2020. arXiv: 1912.11370 [cs.CV].