Taming Transformers for High-Resolution Image Synthesis

Ronaldo Lopez
FC-UNI
Lima, Peru
ronaldo.lopez.c@uni.pe

Cristhian Sánchez Sauñe
FC-UNI
Lima, Peru
csanchez@uni.pe

Davis Alderete
FC-UNI
Lima, Peru
davis.alderete.v@uni.pe

Abstract

Diseñado para aprender interacciones de largo alcance en datos secuenciales, los Transformers continúan mostrando resultados de vanguardia en una amplia variedad de tareas. A diferencia de las CNN, no contienen sesgos inductivos que prioricen las interacciones locales. Esto los hace expresivos, pero también computacionalmente inviables para secuencias largas, como imágenes de alta resolución. Demostramos cómo la combinación de la efectividad del sesgo inductivo de las CNN con la expresividad de los Transformers les permite modelar y, por lo tanto, sintetizar imágenes de alta resolución. En este trabajo se muestra cómo (i) usar CNN para aprender un vocabulario rico en contexto de los componentes de la imagen y, a su vez, (ii) utilizar Transformers para modelar de manera eficiente su composición dentro de imágenes de alta resolución. El enfoque propuesto se aplica fácilmente a las tareas de síntesis condicional, donde tanto la información no espacial, como las clases de objetos, como la información espacial, como las segmentaciones, pueden controlar la imagen generada. En particular, se presentan los primeros resultados sobre la síntesis guiada semánticamente de imágenes megapíxeles con Transformers y se obtiene el estado del arte entre los modelos autorregresivos en ImageNet condicional de la clase. El código y los modelos previamente entrenados se pueden encontrar en github.com/HiroForYou/Image-Synthesis-with-Transformer

1 Introducción

Los Transformers van en aumento: ahora son la arquitectura estándar de facto para las tareas de lenguaje [1, 2, 3, 4] y se adaptan cada vez más en otras áreas como el audio [5] y la visión [6, 7]. En contraste con la arquitectura de visión predominante, las redes neuronales Convolucionales (CNN), la arquitectura del Transformer no contiene un previo inductivo incorporado en la localidad de interacciones y, por lo tanto, es libre de aprender relaciones complejas entre sus entradas. Sin embargo, esta generalidad también implica que tiene que aprender todas las relaciones, mientras que las CNN se han diseñado para explotar el conocimiento previo sobre las fuertes correlaciones locales dentro de las imágenes. Por lo tanto, la mayor expresividad de los Transformers viene con costos computacionales que aumentan cuadráticamente, porque se tienen en cuenta todas las interacciones por pares. Los requisitos de energía y tiempo resultantes de los modelos de Transformers de última generación plantean problemas fundamentales para escalarlos a imágenes de alta resolución con millones de píxeles.

Las observaciones de que los Transformers tienden a aprender estructuras Convolucionales [7] plantean, por tanto, la pregunta: ¿Tenemos que volver a aprender todo lo que sabemos sobre la estructura local y la regularidad de las imágenes desde cero cada vez que entrenamos un modelo de visión, o podemos codificar de manera eficiente el sesgo inductivo de la imagen mientras se conserva la flexibilidad de los Transformers?

Figure 1: El enfoque propuesto permite a los Transformers sintetizar imágenes de alta resolución como esta, que contiene 1280x460 píxeles.

Se presume que la estructura de la imagen de bajo nivel está bien descrita por una conectividad local, es decir, una arquitectura Convolucional, mientras que esta suposición estructural deja de ser efectiva en niveles semánticos superiores. Además, las CNN no solo exhiben un fuerte sesgo de localidad, sino también un sesgo hacia la invariancia espacial mediante el uso de pesos compartidos en todas las posiciones. Esto los hace ineficaces si se requiere una comprensión más holística de la entrada. La idea clave planteada para obtener un modelo eficaz y expresivo es que, en conjunto, las arquitecturas Convolucional y del Transformer pueden modelar la naturaleza compositiva de nuestro mundo visual [8]: se utilizó un enfoque Convolucional para aprender de manera eficiente un libro de códigos de partes visuales ricas en contexto y, posteriormente, aprenderá un modelo de sus composiciones globales. Las interacciones de largo alcance dentro de estas composiciones requieren una arquitectura Transformer expresiva para modelar distribuciones sobre sus partes visuales constituyentes. Además, se utilizó un enfoque adversario para garantizar que el diccionario de partes locales capture la estructura local perceptualmente importante para aliviar la necesidad de modelar estadísticas de bajo nivel con la arquitectura del Transformer. Permitir que los Transformers se concentren en modelar relaciones de largo alcance les permite generar imágenes de alta resolución como en la Fig. 1, una hazaña que anteriormente estaba fuera de su alcance.

La formulación planteada da control sobre las imágenes generadas mediante el condicionamiento de la información sobre las clases de objetos deseados o los diseños espaciales. Finalmente, los experimentos demuestran que este enfoque conserva las ventajas de los Transformers al superar los enfoques de vanguardia basados en libros de códigos anteriores basados en arquitecturas Convolucionales.

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] Alec Radford et al. “Improving language understanding by generative pre-training”. In: (2018).

[3] Alec Radford and Jeffrey Wu. “Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019”. In: *Language models are unsupervised multitask learners. OpenAI Blog* 1.8 (2019), p. 9.

[4] Tom B Brown et al. “Language models are few-shot learners”. In: *arXiv preprint arXiv:2005.14165* (2020).

[5] Rewon Child et al. “Generating long sequences with sparse transformers”. In: *arXiv preprint arXiv:1904.10509* (2019).

[6] Mark Chen et al. “Generative pretraining from pixels”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1691–1703.

[7] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[8] Bjorn Ommer and Joachim M Buhmann. “Learning the compositional nature of visual objects”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2007, pp. 1–8.