NOMBRE: Benjamín Farías Valdés

N.ALUMNO: 22102671

Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC3692 — Tópicos Avanzados en Inteligencia Artificial — 2' 2022

Lectura 22

Crítica

Masked Autoencoders Are Scalable Vision Learners

El paper propone una arquitectura para aprender representaciones de imágenes, denominada *MAE* (*Masked Autoencoder*). Este modelo se basa en entrenar un *encoder* para aprender representaciones de imágenes en donde la mayoría de la información está oculta (enmascarada), y además entrenar un *decoder* para reconstruir la imagen de salida a partir de las representaciones internas y la información que se ocultó. Ambos componentes se basan en *visual transformers*.

La principal idea detrás de esta propuesta está en que el *encoder* tenga acceso a poca información de la imagen, de forma que aprenda a caracterizar la constitución visual de forma inteligente en vez de simplemente memorizar patrones visuales. A su vez, otro aspecto importante es que el *decoder* sí tiene acceso a toda la información que se le ocultó al *encoder*, de forma que se pueda evaluar el resultado final por medio de comparación de píxeles entre imagen generada y *ground truth*. Esto es algo bastante innovador, ya que se tiene una arquitectura asimétrica en vez de la típica simetría *encoder-decoder*, y además, es una buena estrategia para ahorrar computo y memoria al hacer que las capas sean lo más livianas posibles.

A mi parecer, algo muy importante que se desprende del éxito de este modelo, es el hecho de que mejora al esconder gran parte de la imagen. Esto se debe a que, en general, las imágenes contienen mucha información redundante, a diferencia del lenguaje natural. Esto a su vez surge de que el lenguaje natural fue creado y diseñado por humanos para poder comunicarse eficientemente, mientras que las imágenes se producen al capturar la luz reflejada sobre objetos de forma natural, lo que corresponde a un fenómeno completamente agnóstico a los humanos (refiriéndome a la luz y sus propiedades).

Otro tema interesante que me gustaría mencionar es que, a pesar de que el reconstruir imágenes a partir de poca información pueda parecer poco relacionado a la inteligencia visual humana, lo cierto es que eso es justamente lo que nuestros reflejos usan (al menos en mi opinión). Cuando el cuerpo humano debe reaccionar ante algún evento de forma rápida (posiblemente para evitar peligro), me parece que en ese instante de tiempo, el sistema nervioso es capaz de darse cuenta de lo que sucede a partir de información visual muy escasa. Esto último debido a que no se alcanza a procesar completamente la información visual (los ojos ni siquiera alcanzan a enfocar) y uno actúa en base a una intuición visual (por ejemplo, ver algo acercarse a gran velocidad y esquivarlo, sin saber realmente lo que era).

En términos de desventajas de este *approach*, la principal me parece que serían los problemas de escala. Cuando los objetos aparecen cerca o lejos en la imagen, este método podría presentar problemas, dado que

asume un tamaño fijo para las subdivisiones de la imagen que se utilizan para interpretarla. En estos casos, las redes *CNN* serán netamente superiores, dada su invarianza ante transformaciones de escala.

Como conclusión, me parece un enfoque muy interesante, que sin duda abre las puertas a nuevos esfuerzos para adaptar la idea de los *transformers* al campo de la visión por computador. Si bien no logra superar a las redes convolucionales actuales, encuentro que tiene potencial, al tratarse de un modelo que busca la relación entre zonas de la imagen, en vez de simplemente memorizar patrones de características.