Resúmen ImageNet Classification with Deep Convolutional Neural Networks

Primero que nada, es necesario entender que este artículo fue revolucionario ya que en el 2012 las redes neuronales convulsiónales no eran las herramientas predilectas para clasificar imágenes. En ese entonces se usaban herramientas efectivas para datasets relativamente pequeños como MNIST y CIFAR-10 pero que generalizaban mal para datasets emergentes más grandes.

Cabe destacar que entrenar una red tradicional era computacionalmente caro, no se solían utilizar las GPUs para entrenar las neuronas y además tenían una tendencia a sobreajuste cuando se tenían redes neuronales largas.

Por tal motivo, se necesitó un modelo robusto que pueda generalizar bien para datasets grandes. Una de las soluciones propuestas fueron los CNN, al tener menos conexiones que las capas totalmente conectadas y su buena generalización por naturaleza era una elección interesante.

Al experimentar se creo una de las redes convulsiónales más largas usando los subconjuntos de imageNet de las competiciones 2010 y del 2012. El tamaño del a red provocó un problema de sobreajuste, pero esto se pudo contrarrestar con distintas técnicas.

En relación con el dataset, este cuenta con 1.2 millones de imágenes de entrenamiento (85%), 50.000 imágenes de validación (5%) y 150.000 (10%) imágenes de prueba. Se redujo la calidad de las imágenes a 256x256 y luego se las normalizo antes de ingresar a la red.

Respecto a la arquitectura, la red contiene 8 capas, de las cuales 5 son convulsiónales y 3 están totalmente conectadas. Además, se utilizó la función de activación ReLu la cual no era una respuesta obvia en ese entonces, su uso permitía un aprendizaje más rápido el cual tenía una influencia en la mejora del rendimiento de modelos con muchas capas con datasets grandes.

Se usaron múltiples GPUs para este experimento, se dividió la red en dos GPUs y estas se comunicarían solo en ciertas capas. Aunque ReLu no necesita hacer uso de la generalización para aprender, se pudo comprobar que la normalización ayudó a la generalización del modelo. Reduciendo el sobreajuste también se hizo uso de una sobreposición de las capas de maxpooling reduciendo el top en 0,4% y 0,3%.

El sobreajuste fue un problema el cual se pudo disminuir a través de distintos mecanismos tales como: augmentación de los datos, esto se hizo generando imágenes rotadas y trasladadas esto hizo que se cuenten con más imágenes en el training set al factor de 2048. Además, se modificó el contraste y el brillo de las imágenes lo cual redujo el error top-1 en un 1%.

Por último, se hizo uso de la técnica dropout para que aprenda los features más robustos los cuales son útiles en conjunción con otros subsets aleatorios de otras neuronas.

Se entrenó usando una gradiente descendiente estocástica con un batch size de 128 ejemplos y un momentum de 0,9 y un weight decay de 0,0005. El weight decay es importante ya que reduce el training error. Se inicializaron los pesos en cada capa con una distribución gaussiana con desviación estándar de 0.01.

Se lograron tasas de error del 37,5% y de 17% comparadas por las ganadoras de la competición ILSVRC-2010 las cuales fueron 47,1% y 28,2% respectivamente.

En conclusión, el artículo revolución la comunidad pudiendo resolver problemas emergentes con soluciones creativas y sentó una base para trabajar con datasets grandes aprovechando el uso de las GPUs.