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 17

Crítica

FiLM: Visual Reasoning with a General Conditioning Layer

El paper introduce una arquitectura llamada *FiLM* (*Feature-Wise Linear Modulation*), la que permite influenciar el aprendizaje de una red neuronal mediante capas que actúan sobre los estados internos de la red para lograr modular la salida con una transformación lineal, dependiendo de la entrada recibida.

El enfoque específico de esta propuesta está en mejorar el rendimiento actual en tareas de razonamiento visual, para lo que utilizan el conocido *dataset CLEVR*, que hemos revisado anteriormente en el curso en la parte de razonamiento y arquitecturas composicionales. Además, en la sección de experimentos se muestran pruebas realizadas sobre variantes del *dataset*, de forma que sea posible probar con contextos más complejos y parecidos al razonamiento humano, además de validar la capacidad de aprender reglas lógicas sobre los datos en vez de simplemente memorizar combinaciones. Esto último me parece una muy buena decisión, ya que el objetivo de este tipo de modelos es demostrar que son capaces de aprender de forma más racional que las arquitecturas clásicas de redes neuronales que utilizan un enfoque estilo fuerza bruta para memorizar.

En mi opinión, uno de los puntos más fuertes de esta propuesta es que es bastante liviana de implementar (al ser una transformación lineal), lo que la hace muy escalable y eficiente de entrenar. Además, aplicar estas capas *FiLM* sobre distintos bloques internos de la red principal permite una gran granularidad al momento de condicionar la salida de la red, permitiendo atacar de forma flexible varios tipos de tareas de razonamiento distintas (lo que se ve evidenciado claramente por los positivos resultados de los experimentos).

Dentro del estudio de sensibilidad de los parámetros, se encontraron resultados interesantes, tales como el hecho de que el parámetro que modula el efecto de los mapas de características de la red es el más relevante dentro de la transformación aprendida. Esto era de esperarse, puesto que la información importante proviene justamente de estos mapas de características aprendidos por la red principal, mientras que el otro parámetro de la transformación lineal permite realizar un ajuste más específico pero menos importante. Otro resultado que me gustaría destacar es que mediante los experimentos se observó que estas capas *FiLM* no están muy relacionadas con el efecto de las capas de normalización, lo que implica que es posible aplicar esta técnica en otros modelos existentes en los que no sea posible o útil aplicar normalización directamente, presentando así una alternativa prometedora (por ejemplo para *RNNs* o *RL*).

En general, encontré muy entretenido el artículo, con la propuesta bien explicada y los experimentos elegidos acorde a los objetivos del estudio. Algo que me gustaría haber visto es otras aplicaciones de esta estrategia, dado que en este trabajo sólo se probó con el área de *visual question answering*. Creo que este

modelo tiene bastante potencial en otros tipos de arquitecturas, por ejemplo las recurrentes, donde podría ser interesante ver si ofrece mejoras en áreas como *NLP*.