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 12

Crítica

Vision-Language Navigation with Self-Supervised Auxiliary Reasoning Tasks

En este paper se presenta una arquitectura que permite mejorar el desempeño de los modelos actuales en tareas de *VLN* (navegación a partir de instrucciones en lenguaje natural). En la introducción se presenta el problema: mover a un agente en un cierto entorno 3D (aunque con navegación limitada a ciertos puntos del entorno que son fijados por un grafo), de forma que siga paso a paso las instrucciones que se le entregan en lenguaje natural. Se mencionan varios enfoques del estado del arte, los que logran buenos resultados al combinar las características visuales con las instrucciones textuales, pero algo que les falta es considerar la información acumulada a través de cada paso (no sólo el último paso), además de aprovechar mejor la información semántica del entorno. La propuesta es una red llamada *AuxRN*, que está compuesta por varias *RNNs* en su interior, cada una enfocada en algo específico (extraer características visuales de imágenes, generar representaciones de texto, combinar visión con instrucciones, entre otras). Lo más importante de esta red, es que además de predecir a qué punto el agente debe moverse a continuación, también se encarga de realizar 4 tareas auxiliares de razonamiento, las que al ser entrenadas permiten mejorar la robustez y eficiencia del modelo al atacar el problema de interés principal (mover al agente). Estas actividades auxiliares son: generar un texto que exprese los pasos realizados desde el inicio, estimar el progreso de la ruta, predecir el siguiente ángulo y alinear las instrucciones con la visión del estado actual del agente. Encuentro sumamente interesante este *approach*, ya que las personas también razonan de forma similar al navegar sobre un entorno, en el sentido de que van considerando los pasos realizados anteriormente y la relación con las instrucciones para ir ubicándose en el entorno (no se basan solamente en la visión del estado actual). En la sección de experimentos se muestra como se probó el modelo en distintas modalidades, llegando a transformarse en el nuevo estado del arte en esta tarea en particular. Además, se analiza el aporte de cada tarea auxiliar, concluyendo que cada una aporta significativamente y que ninguna es prescindible (de hecho al estar juntas la red aprende mejor, son complementarias). En lo personal me parece un trabajo muy útil, ya que busca un enfoque más general sobre el problema, lo que en mi opinión también agrega otra ventaja que es el poder utilizar lo aprendido por el modelo en las tareas auxiliares directamente. Creo que además de mejorar su rendimiento en la tarea de navegación, se podría aprovechar también que el modelo sabe manejar las tareas auxiliares (que pueden perfectamente ser otros problemas de interés relacionados al de navegación), logrando ser multi-funcional.