Structural Characteristics of Knowledge Graphs Determine the Quality of Knowledge Graph Embeddings Across Model and Hyperparameter Choices

Jeffrey Sardina

Submitted to the University of Dublin,
in partial fulfilment of the requirements for the degree of
Master of Science in Computer Science in the Data Science Strand

Supervised by Dr. Declan O'Sullivan

2021

Abstract

The realm of biomedicine is producing information at a rate far beyond the capacity of clinicians, researchers, and machine learning experts to analyze in full. Recently, developments in Knowledge Graphs (KGs) have facilitated the representation of all this information in an easily-queryable format. With increasing academic and clinical interest in Knowledge Graph Embeddings (KGEs), various KGE models have been developed and refined to allow machine learning to be run on these entire datasets and predict new, previously unseen information about the domain. However, the need to validate hyperparameters for every new dataset, especially considering the time and expertise needed for validation and model training, have limited the use of KGEs in biology to those who have expertise in machine learning and knowledge engineering. The main result of this dissertation is the development of a framework by which the effect of hyperparameters on model performance for a given dataset can be modelled as a function of KG structure. The results of this further provide an understanding of the relative performance of specific cross-dataset hyperparameter sets on KGE models for different biomedical datasets. While the hyperparameters identified have mixed success for their KGE models, a clear effect of graph structure on hyperparameter fitness is found, leading to the conclusion that more research into cross-dataset hyperparameter prediction and re-use has promise for increasing the accessibility and usability of KGE models.

Rialaíonn Airíonna Struchtúir na nGraf Eolais Cáilíocht Leabuithe Graif Eolais Fiú i gComhthéacs Samhlacha agus Hipearpharaiméadar Éagsúla

Jeffrey Sardina

Curtha isteach d'Ollscoil Átha Cliath,
mar chuid de chomhlíonadh céim
Máistreachta san Eolaíocht Ríomhaireachta i Roinn na hEolaíochta Sonraí

Faoi Stiúrthóireacht an Dochtúir Declan O'Sullivan

2021

Achoimre

Tá taighde bithleighis ag cur amach eolais nua i bhfad níos tapúla ná mar is féidir le cliniceoirí, le taighdeoirí, agus le saineolaithe meaisínfhoghlama anailísíocht iomlán a dhéanamh air. Ach le déanaí, tá dul chun cinn i réimse na nGraf Eolais (GE) tar éis deacrachtaí stórála, inrochtaineachta, agus samhlaithe a éascú. Anois agus suim acadúil agus cliniciúil i Leabuithe Graif Eolais (LGE) ag dul i méid, bhí roinnt samhlacha LGE forbartha agus beachtaithe chun cumas meaisínfhoghlama coibhneasta a chur ar fáil don eolas seo. Mar sin, tá deis ann na samhlacha seo a úsáid chun eolas nua i réimse an bhithleighis a réamhinsint. Tá gá le hipearpharaiméadair a fháil agus a fhíorú do ghach tacar sonraí, áfach, agus is riachtanach an-mhéid ama agus saineolais a chur isteach chun samhail LGE a ullmhú agus a thraenáil. Is mar sin nach raibh fáil ar LGEanna sa réimse bithleighis ach ag saineolaithe meaisínfhoghlama agus ag saineolaithe innealtóireachta eolais. Is é toradh an tráchtais seo ná creatlach nua a fhorbairt chun tionchar hipearpharaiméadar ar cháilíocht samhlacha LGE a thuiscint trí radharc struchtúir na GEanna. Thairis sin, cuireann an taighde seo le tuiscint ar shraitheanna hipearpharaiméadar a oibríonn le roinnt tacar sonraí ar leith i réimse an bhithleighis. Fiú nach bhfuil éirí mór ag na hipearpharaiméadair atá faighte anseo ar shamhlacha LGE, léirítear go bhfuil tionchar mór ag struchtúr GE ar céard iad na hipearpharaiméadair is fearr. Mar sin, is í conclúid eile an taighde seo ná go mbeadh rath ar thaighde eile dírithe ar réamhinsint hipearpharaiméadar a oibríonn thar GEanna ar leith, agus go gcuirfeadh an taighde sin le húsáideacht agus le hinrochtaineacht samhlacha LGE.