Neural Networks for Information Retrieval

Tom Kenter[1(✉)], Alexey Borisov[2], Christophe Van Gysel[3], Mostafa Dehghani[3], Maarten de Rijke[3], and Bhaskar Mitra[4]

[1] Booking.com, Amsterdam, Netherlands
tom.kenter@gmail.com
[2] Yandex, Moscow, Russia
alborisov@yandex-team.ru
[3] University of Amsterdam, Amsterdam, Netherlands
{cvangysel,dehghani,derijke}@uva.nl
[4] Microsoft, University College London, London, UK
bmitra@microsoft.com

Abstract. Machine learning plays a role in many aspects of modern IR systems, and deep learning is applied in all of them. The fast pace of modern-day research has given rise to many approaches to many IR problems. The amount of information available can be overwhelming both for junior students and for experienced researchers looking for new research topics and directions. The aim of this full-day tutorial is to give a clear overview of current tried-and-trusted neural methods in IR and how they benefit IR.

Prompted by the advances of deep learning in computer vision, neural networks (NNs) have resurfaced as a popular machine learning paradigm in many other directions of research, including IR. Recent years have seen NNs being applied to all key parts of the typical modern IR pipeline, such as click models, core ranking algorithms, dialogue systems, entity retrieval, knowledge graphs, language modeling, question answering, and text similarity. A key advantage that sets NNs apart from many learning strategies employed earlier, is their ability to work from raw input data. Where designing features used to be a crucial aspect and contribution of newly proposed IR approaches, the focus has shifted to designing network architectures instead. As a consequence, many different architectures and paradigms have been proposed, such as auto-encoders, recursive networks, recurrent networks, convolutional networks, various embedding methods, and deep reinforcement learning. The aim of this tutorial is to provide an overview of the main network architectures currently applied in IR and to show how they relate to previous work. The tutorial covers methods applied in industry and academia, with in-depth insights into the underlying theory, core IR tasks, applicability, key assets and handicaps, efficiency and scalability concerns, and tips & tricks. We expect the tutorial to be useful both for academic and industrial researchers and practitioners who either want to develop new neural models, use them in their own research in other areas or apply the models described here to improve actual IR systems.

G. Pasi et al. (Eds.): ECIR 2018, LNCS 10772, p. 837, 2018.
https://doi.org/10.1007/978-3-319-76941-7