BEHZAD VAHEDI
Department of Geography
University of California, Santa Barbara
Email: behzad@ucsb.edu

Behzad Vahedi is a PhD student at the Department of Geography, University of California, Santa Barbara. He is generally interested in spatiotemporal question answering and his research is currently focused on question-based spatial computing. Particularly, he is interested in using a combination of natural language processing and spatial computing techniques to translate natural language spatial questions to a high-level language of core spatial operations. Prior to this, he has worked on spatial data modeling with a focus on spatial fields.

Vahedi has an engineering background and holds a B.Sc. in Geodesy and Geomatics Engineering, and a M.Sc. in GIS engineering.

Challenges in Spatial Data Science

Unprecedented performance of deep learning algorithms in solving general computer science tasks at the beginning of current decade has made such algorithms and techniques appealing for scientists in many different disciplines to the point that deep learning and artificial intelligence in general have been called "the new electricity.[1]" While this may sound bold at the moment, there is no denying that machine learning has had huge impact on data science in particular. Geography and GIScience are no exceptions and machine learning algorithms have extensively been used to solve spatial problems recently. In this position paper, I will focus on applications of machine learning in Spatial Data Science (SDS) and discuss two challenges that SDS faces because of the special nature of spatial data and analysis.

The first challenge is a familiar one, but in a new form: the challenge of representation. Data representation plays an important role in machine learning and representation learning models in particular. Bengio et al (2013) state "the success of machine learning algorithms generally depends on data representation, and this is because different representations can entangle and hide more or less the different explanatory factors of variation behind the data." In light of this, the vector versus raster or field versus object dilemma in GIScience will, in my opinion, become an even bigger issue in spatial data science. Creating optimal representations that are computationally sound on one hand, and are able to capture the nuances of spatial data is a must in SDS era. Ironically, vector representation is by far the most-commonly used in representation learning and data, from text data to image and audio, are often represented as high-dimensional vectors. As an example in GIScience,

[1] https://www.fast.ai/2016/10/11/fortune/

place representations has been studied well both before and after deep learning boom. Yan et al. (2017) proposed a place embedding based on word embedding to facilitate reasoning about place types, but what would happen if one were to do the same for places on a map or an aerial image. Can we think of place embeddings not based on the place-name, but based on the place as a concept? Could such embedding be used in different applications to reason about places? These are questions that lead me to think that we need a general approach to represent spatial features. The omnipresence of vector representation in existing machine learning methods may mean that we are bounded to vector representations for now, and this makes this challenge even more difficult: Are vector representations able to sufficiently represent different phenomena and data types used in spatial data?

The second challenge, which is an open problem in computer science, is to move from machine learning to machine reasoning (Bottou, 2014). Deep learning approaches often struggle to perform well on tasks with a compositional nature. They are also known to be sensitive to statistical priors and are essentially correlation engines that “will hone in on any statistical, potentially spurious pattern that allows them to model the observed data more accurately” (Hudson & Manning, 2018). Therefore they are not able to perform inference and are not well-suited for problem solving tasks. Machines reasoning, on the other hand, not only does provide (advanced) problem solving capabilities, but also needs less training data to achieve a comparable performance. Performing inferences is of high importance in spatial problem solving tasks such as spatial question answering where usually more than one (spatial) operation is needed to answer a given question. Models that can reason spatially could potentially learn geographic principles such as first law of geography from data and apply it in solving problems.

In conclusion, spatial data science faces the same computational and theoretical challenges as data science in general, as well as additional challenges that arise as a result of the geographic structure and complexity of spatial data. Here, I mentioned two of these particular challenges that if addressed, could open the door to a lot of opportunities in spatial data science.

References

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on pattern analysis and machine intelligence* 35(8): 1798–1828.

Bottou, L. (2014). From machine learning to machine reasoning. *Machine learning* 94(2): 133–149.

Hudson, D. A., & Manning, C. D. (2018). Compositional attention networks for machine reasoning. arXiv preprint arXiv:1803.03067.

Yan, B., Janowicz, K., Mai, G., & Gao, S. (2017). From ITDL to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (p. 35). ACM.