Interpreting Convolutional Neural Networks by Constraining the Selection of Feature Maps through Quantum Annealer

Francesco Aldo Venturelli[1,2]
Emanuele Costa[2], Sikha O K[1], Bruno Juliá Díaz[3], Miguel A. González Ballester[1,2,4], Alba Cervra-Lierta[2]

[1]*BCN Medtech, Universitat Pompeu Fabra, Barcelona, Spain*
[2]*Barcelona Supercomputing Center, Barcelona, Spain*
[3]*Universidad de Barcelona, Barcelona, Spain*
[4]*ICREA, Barcelona, Spain*

francescoaldo.venturelli@upf.edu

Abstract

Interpreting deep learning (DL) model predictions has become essential as learning algorithms are increasingly deployed in safety-critical applications. Understanding why models produce specific predictions is crucial for validating assumptions and designing more robust architectures[1]. We propose a novel interpretability framework for convolutional neural networks (CNNs) in image classification tasks by reformulating input-specific feature selection (FS) algorithm as a Quadratic Unconstrained Binary Optimization (QUBO) problem. The overall pipeline is represented in **Figure 1**. Since FS algorithms may scale exponentially with the number of variables[2], we employ quantum annealing (QA) to efficiently solve the resulting combinatorial problem. The proposed approach identifies feature maps that are most relevant to individual predictions, enabling faithful explanations that are compared against established methods such as Grad-CAM, Grad-CAM++, and Score-CAM, observing improved class disentanglement. Finally, we analyse the computational complexity of the QA process by studying the minimum gap of the global Hamiltonian and the corresponding success probability[3].

Figures

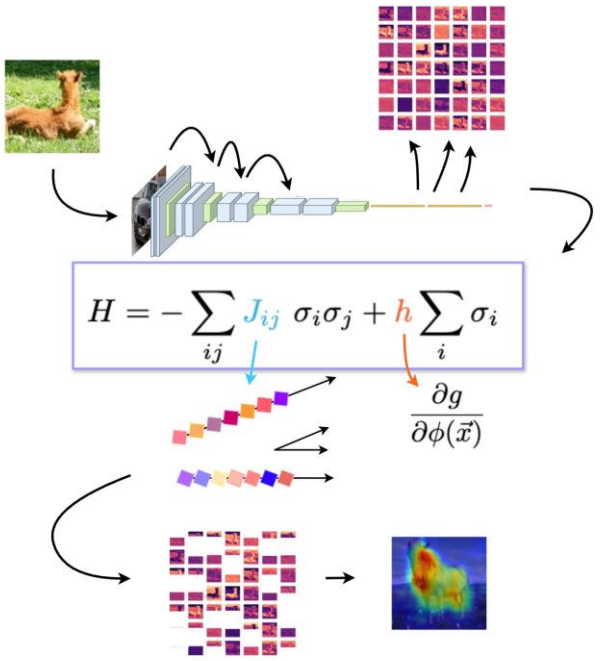

Figure 1: For a given image we extract a set of feature maps from the last block. We evaluate positive-gradient contribution and cosine similarity among distinct maps that are used to construct the QUBO Hamiltonian. Eventually, QA samples a solution bit-string containing the selected elements.

References

[1] Tehreem Qamar and Narmeen Zakaria Bawany. In: PeerJ Computer Science 9 (2023), e1629.
[2] Bolón-Canedo, Veronica, et al. Knowledge and Information Systems 56.2 (2018): 395-442.
[3] Rajak, Atanu, et al. Philosophical Transactions of the Royal Society A 381.2241 (2023): 20210417.