

Towards interpretable AI with quantum annealing feature selection

Extended Abstract — QTML 2026

Francesco Aldo Venturelli^{1,2}, Emanuele Costa², Sikha O K¹, Bruno Juliá-Díaz^{3,4},
Miguel A. González Ballester^{1,2,5}, Alba Cervera-Lierta²

¹BCN Medtech, Universitat Pompeu Fabra, Barcelona, Spain

²Barcelona Supercomputing Center (BSC)

³Departament de Física Quàntica i Astrofísica, Facultat de Física, Universitat de Barcelona

⁴Institut de Ciències del Cosmos, Universitat de Barcelona (ICCUB)

⁵ICREA, Barcelona, Spain

Abstract

Deep Learning (DL) models have significantly improved problem-solving across science and society, yet remain inherently difficult to interpret due to their black-box structure. In image classification, gradient-based activation mapping techniques - such as GradCAM [1] and GradCAM++ [2] - represent the current state of the art for post-hoc explainability. These methods, however, aggregate all feature maps (FMs) into a single saliency map, obscuring the role of individual learned representations. A more informative direction lies in selecting the most representative and non-redundant FM subset. However, this is an NP-hard combinatorial optimization problem, whose solution space grows exponentially with the number of features.

In this work, we introduce a quantum annealing-based approach to select the most representative and non-redundant FMs of a Convolutional Neural Network (CNN) trained on image classification. We formulate the problem as a Quadratic Unconstrained Binary Optimization (QUBO), whose objective function balances two competing terms: the maximization of each FM's gradient importance - inherited from the GradCAM framework - and the minimization of pairwise cosine similarity, which enforces geometric diversity and suppresses redundant representations. Since the resulting Ising Hamiltonian is directly implementable on quantum annealers, we study complexity and scalability via classical simulations, analyzing the minimum energy gap $\Delta_{\min}$ of the annealing schedule.

We demonstrate that the selected FMs achieve competitive performance with GradCAM and GradCAM++ in terms of average confidence drop, while improving class disentanglement, quantified via class-class correlation maps computed through the Bhattacharyya coefficient. Our results suggest that quantum annealing is a promising tool for explainable AI, with complexity properties favorable for near-term quantum hardware, opening new directions at the interface of quantum optimization, statistical physics, and deep learning. A detailed description of our method and further results can be found in [3].

1 Method

1.1 QUBO formulation

Given a pre-trained CNN, we extract the set of FMs from the last convolutional layer, $\{\mathbf{f}\} \in \mathbb{R}^{H_f \times W_f \times N_f}$, where N_f is the number of FMs and $H_f \times W_f$ the number of components of each FM. For each FM a , we compute the global average pooling of the gradient of the model output z :

$$\alpha_a = \frac{1}{H_f W_f} \sum_{ij}^{H_f, W_f} \frac{\partial z}{\partial f_{ij}^{(a)}}. \quad (1)$$

We retain only the *filtered* FMs with $\alpha_a > 0$, obtaining $d \leq N_f$ representations. The pairwise cosine similarity

$$J_{pq} = \left| \langle \tilde{\mathbf{f}}^{(p)}; \tilde{\mathbf{f}}^{(q)} \rangle \right| = \frac{\left| \sum_{ij} \tilde{f}_{ij}^{(p)} \tilde{f}_{ij}^{(q)} \right|}{\|\tilde{\mathbf{f}}^{(p)}\| \|\tilde{\mathbf{f}}^{(q)}\|} \quad (2)$$

measures geometric redundancy. The QUBO Hamiltonian reads

$$\hat{H}_{\text{QUBO}} = (1 - \beta) \frac{1}{2} \sum_{pq} J_{pq} \hat{n}_p \hat{n}_q + \beta \sum_p h_p \hat{n}_p, \quad (3)$$

where $\hat{n}_p = (1 - \sigma_z^{(p)})/2$, $h_p = \tilde{\alpha}_p / \max_p \tilde{\alpha}_p \in [0, 1]$, and $\beta \in [0, 1]$ controls the importance–diversity trade-off.

1.2 Quantum annealing protocol and complexity

The system is initialized in the uniform superposition $|\psi\rangle_{\text{init}} = 2^{-d/2} \sum_{z \in \{0,1\}^d} |z\rangle$ and evolved under $\hat{H}(s) = (1-s)\hat{H}_D + s\hat{H}_{\text{QUBO}}$. The final state is sampled with $n_{\text{shots}} = O(d^2)$ measurements. For $N_f = 512$ (full ResNet-18), exact simulation is intractable and we use Simulated Annealing (SA) via the D-Wave Ocean library [4].

The efficiency of the protocol is governed by the minimum energy gap $\Delta_{\min} \equiv \min_s (E_1(s) - E_0(s))$, where $E_i(s)$ are instantaneous eigenvalues of $\hat{H}(s)$. The adiabatic evolution should last at least $\tau \propto \Delta_{\min}^{-2}$, condition provided by the adiabatic theorem [5, 6]. The coupling distribution J_{pq} follows a Gaussian, placing our Ising model in the Sherrington-Kirkpatrick universality class [7, 8].

2 Results

2.1 Class disentanglement and Average Drop %

We validate on STL-10 [9] (10 classes, 5000 images) using a modified ResNet-18 [10] with 94% test accuracy. Fig. 1 shows the class-class correlation maps computed via the Bhattacharyya coefficient [11] for QA ($N_f = 16$) and SA ($N_f = 512$). For $N_f = 16$, only 14 out of 45 class pairs exhibit non-zero overlap (≥ 0.10), all below 0.50. For $N_f = 512$, the overlap is nearly uniform and below 0.30, indicating good disentanglement at scale.

The overall Average Drop % [2] for QA-*f*GradCAM ($N_f = 16$, $\beta = 0.7$) is $9.70 \pm 6.13\%$, comparable to GradCAM ($9.25 \pm 5.94\%$) and GradCAM++ ($6.95 \pm 5.41\%$). For the full ResNet-18 ($N_f = 512$), SA with $\beta = 0.99$ reaches $6.28 \pm 4.25\%$, competitive with GradCAM ($5.27 \pm 4.58\%$) and GradCAM++ ($6.73 \pm 4.79\%$), while SA with $\beta = 0.7$ yields a higher drop of $18.4 \pm 8.69\%$, confirming that β is a critical meta-parameter. Fig. 2 shows how increasing β makes the gradient importance term dominant, selecting more FMs and systematically reducing the Average Drop toward the GradCAM baseline values.

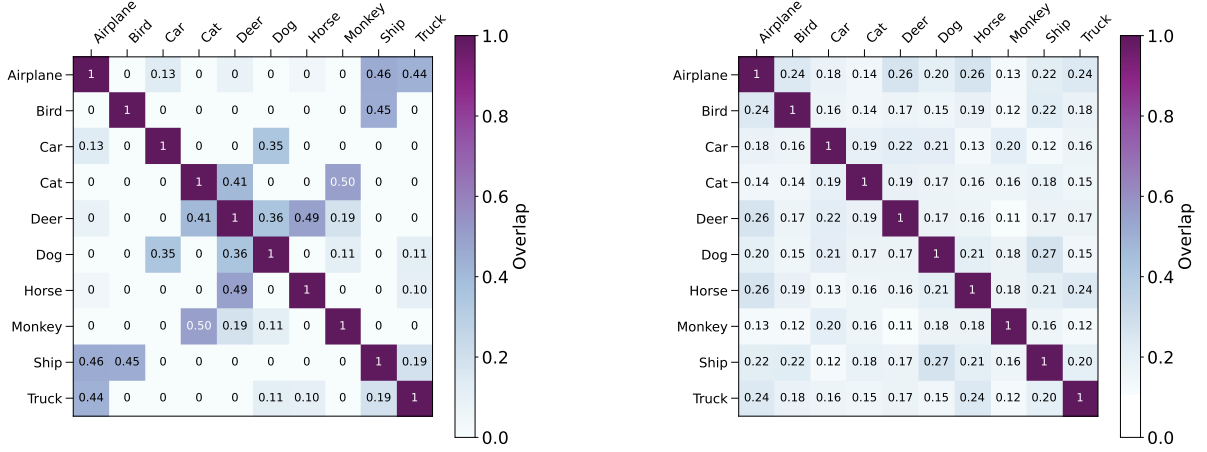


Figure 1: Class-class correlation maps via Bhattacharyya coefficient. **Left:** QA, $N_f = 16$, $\beta = 0.7$, $\tau = 50$. **Right:** SA, $N_f = 512$, $\beta = 0.99$. A fully diagonal matrix indicates complete class disentanglement.

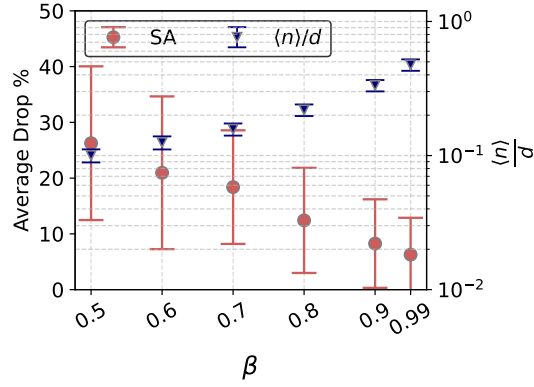


Figure 2: Average Drop % (left axis) and ratio $\langle n \rangle / d$ representing the fraction of selected features (right axis) as a function of β for SA. Increasing β makes the gradient term dominant: more FMs are selected and the Average Drop decreases.

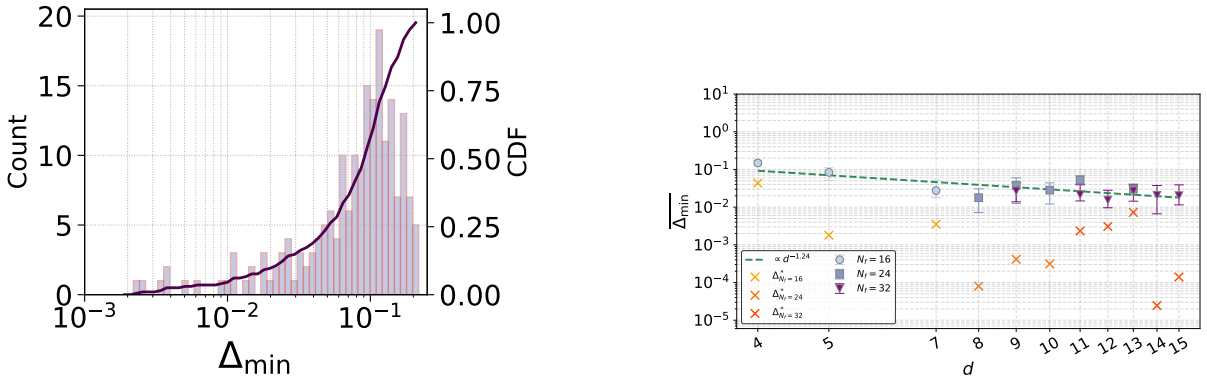


Figure 3: **Left:** Distribution of $\Delta_{\min}$ across 200 test images ($N_f = 16$). The gap is concentrated near 10^{-1} energy units; the probability of $\Delta_{\min} < 10^{-3}$ is essentially zero. **Right:** Average $\Delta_{\min}$ vs. number of filtered FMs d , showing $\Delta_{\min} \propto d^{-1}$.

2.2 Energy gap, fidelity, and complexity

The $\Delta_{\min}$ distribution (Fig. 3, left) confirms robust convergence to the $\hat{H}_{\text{QUBO}}$ ground state. The scaling $\overline{\Delta}_{\min} \propto d^{-1}$ (Fig. 3, right) implies $\tau \propto d^2$, a favorable polynomial complexity. A similar behaviour is provided by the hardest instances $\min_J(\Delta_{\min})$ with a scaling in agreement with a linear one.

References

- [1] Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* **128**, 336–359 (2019). URL <http://dx.doi.org/10.1007/s11263-019-01228-7>.
- [2] Chattopadhyay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–847 (IEEE, 2018). URL <http://dx.doi.org/10.1109/WACV.2018.00097>.
- [3] Venturelli, F. A. *et al.* Towards interpretable ai with quantum annealing feature selection. *arXiv preprint arXiv:2604.25649* (2026).
- [4] Systems, D.-W. D-wave quantum computing documentation. <https://www.dwavesys.com> (2021).
- [5] Rajak, A., Suzuki, S., Dutta, A. & Chakrabarti, B. K. Quantum annealing: an overview. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **381** (2022). URL <http://dx.doi.org/10.1098/rsta.2021.0417>.
- [6] Žnidarič, M. Scaling of the running time of the quantum adiabatic algorithm for propositional satisfiability. *Physical Review A* **71** (2005). URL <http://dx.doi.org/10.1103/PhysRevA.71.062305>.
- [7] K. Chakrabarti, B. & Das, A. *Transverse Ising Model, Glass and Quantum Annealing*, 1–36 (Springer Berlin Heidelberg, 2005). URL http://dx.doi.org/10.1007/11526216_1.
- [8] Panchenko, D. *The Sherrington-Kirkpatrick Model* (Springer New York, 2013). URL <http://dx.doi.org/10.1007/978-1-4614-6289-7>.
- [9] Coates, A., Ng, A. & Lee, H. An analysis of single-layer networks in unsupervised feature learning. In Gordon, G., Dunson, D. & Dudík, M. (eds.) *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, vol. 15 of *Proceedings of Machine Learning Research*, 215–223 (PMLR, Fort Lauderdale, FL, USA, 2011). URL <https://proceedings.mlr.press/v15/coates11a.html>.
- [10] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (IEEE, 2016). URL <http://dx.doi.org/10.1109/CVPR.2016.90>.
- [11] Patra, B. K., Launonen, R., Ollikainen, V. & Nandi, S. A new similarity measure using bhattacharyya coefficient for collaborative filtering in sparse data. *Knowledge-Based Systems* **82**, 163–177 (2015). URL <http://dx.doi.org/10.1016/j.knosys.2015.03.001>.