Semantic Hash-Based Retrieval Framework for Explainable Visual Recommendation Systems

Cesar Waga
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
cesarv@colostate.edu

Quentin D. Bell
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
quentinbell160@oregonstate.edu

Malcolm A. Carr
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
malcolm.carr723@ku.edu

Adrian Terry
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
terryadrian@binghamton.edu

Abstract

Visual recommendation systems have become critical components in e-commerce, media streaming, and social platforms, demanding both retrieval efficiency and user trust through transparency. Traditional deep learning-based recommenders excel in accuracy but often operate as opaque black boxes, raising significant concerns around fairness, accountability, and user acceptance. Simultaneously, the explosive growth of multimedia databases requires retrieval mechanisms that scale while preserving semantic fidelity. This paper presents a comprehensive system-level investigation into a semantic hash-based retrieval framework that unifies efficient approximate nearest neighbor search with explainable reasoning for visual recommendations. We examine the architectural foundations that couple deep hashing encoders, binary code index structures, and explanation generation modules into a cohesive socio-technical infrastructure. The discussion focuses on structural trade-offs among inference latency, storage footprint, explanation granularity, and environmental sustainability. We embed the framework within broader governance and policy contexts, analyzing how semantic hashing can facilitate compliance with data protection regulations and fairness mandates by enabling interpretable audit trails. Deployment considerations are explored across edge-cloud continuums, addressing robustness to distributional shift, adversarial perturbations, and long-tail retrieval dynamics. The paper further proposes design principles that balance business metrics with ethical imperatives, emphasizing modularity, continuous fairness monitoring, and energy-aware serving strategies. By synthesizing insights from systems engineering, human-computer interaction, and algorithmic fairness, this work provides a forward-looking blueprint for building next-generation visual recommendation infrastructures that are simultaneously fast, interpretable, and responsible.

Keywords

Semantic hashing, visual recommendation, explainability, retrieval systems, binary codes, system architecture, fairness, governance.

1. Introduction

The widespread integration of visual recommendation engines into consumer platforms has transformed how users discover products, media content, and social connections. These systems routinely process billions of images and user interactions, demanding retrieval architectures that deliver sub-millisecond latency while maintaining high relevance. Early solutions relied on collaborative filtering and matrix factorization, but the incorporation of deep convolutional features enabled substantial accuracy gains [1]. As user expectations evolve, however, raw predictive performance is no longer sufficient. Regulators, advocacy groups, and end-users increasingly demand explanations for automated decisions, especially when recommendations influence purchasing behavior, self-image, or access to information. This confluence of scalability and explainability requirements exposes systemic gaps in current system designs, where retrieval speed and transparency are often treated as competing objectives rather than co-design goals.

Visual recommendation pipelines typically comprise a feature extraction stage, a candidate retrieval stage, and a ranking stage. The retrieval stage is the computational bottleneck, frequently implemented using approximate nearest neighbor (ANN) search over dense embeddings. While dense vector search backed by GPU-accelerated libraries such as FAISS has achieved impressive throughput, the opacity of these embeddings constrains the ability to provide faithful explanations to users. Moreover, the storage and memory footprint of floating-point vectors impose sustainability challenges at data-center scale. These limitations have sparked a renaissance of interest in hashing-based retrieval, where high-dimensional visual features are compressed into compact binary codes that preserve semantic similarity in Hamming space. Binary representations not only reduce storage by orders of magnitude but also enable constant-time lookup through hash tables, significantly lowering the energy per query [1], [2]. However, conventional hashing approaches that optimize purely for pairwise similarity often fail to align hash bits with human-interpretable attributes, turning potential efficiency gains into an explainability dead end.

The present work systematically examines a semantic hash-based retrieval framework designed to reconcile these tensions. By grounding hash codes in semantically meaningful visual properties such as color palettes, object compositions, style categories, or contextual cues, the framework supports both efficient retrieval and the generation of post-hoc or intrinsic explanations. We position this framework at the intersection of systems engineering and algorithmic governance, emphasizing not only the technical architecture but also the socio-technical trade-offs that influence real-world deployment. The analysis extends beyond isolated accuracy metrics to consider fairness across demographic segments, robustness under concept drift, energy proportionality, and compliance with emerging AI regulations. Through this holistic lens, we articulate a set of structural principles that can guide the construction of interpretable, large-scale visual recommendation infrastructures.

2. Background and Related Work

Visual recommendation has been shaped by a series of paradigm shifts from content-based filtering to deep learning hybrids. Early content-based systems extracted handcrafted descriptors such as SIFT or color histograms for similarity search; later, deep features from pre-trained convolutional networks became the de facto representation [2]. The availability of

large-scale interaction datasets further enabled models that jointly encode visual signals and user preferences, leading to visually-aware recommenders that leverage image regions and style embeddings [8]. Despite their predictive power, these models often lack transparency, as latent dimensions emerge from end-to-end optimization without explicit semantic grounding. Efforts to inject explainability into recommendations have spawned a rich subfield that encompasses attention visualization, feature attribution, knowledge graph paths, and counterfactual reasoning, yet the integration of such techniques into production retrieval pipelines remains nascent [9].

On the retrieval infrastructure side, hashing has been a cornerstone technique for decades. Locality-sensitive hashing (LSH) originally provided probabilistic guarantees for sub-linear similarity search in high-dimensional spaces [3]. Subsequent advances in learning to hash shifted the paradigm from data-independent projections to data-adaptive binary embeddings that directly optimize retrieval quality metrics. Comprehensive surveys document the transition from shallow supervised hashing with kernel methods to deep neural hashing architectures that learn end-to-end image-to-binary mappings [4]. Representative models include HashNet, which employs a continuation method to address ill-posed gradient issues in binarization [5], and deep supervised hashing networks that incorporate pairwise label information to enhance ranking accuracy [6]. In parallel, self-supervised representation learning exemplified by contrastive frameworks has demonstrated that semantically rich embeddings can be derived without manual annotations, opening pathways for unsupervised semantic hashing [7].

More recent work has pushed toward semantically-aware hashing that explicitly models discrete attributes or hierarchical taxonomies during code learning. The approach of asymmetric semantic excavation introduces a self-supervised mechanism to mine semantic structure from data without requiring exhaustive attribute labels, and couples it with margin-scalable constraints that adapt to varying semantic granularities [13]. This line of research aligns hash bit activations with interpretable visual concepts, thereby transforming the binary code into a proxy for semantic descriptor. Such advances lay the technical foundation for our framework, where the semantic alignment of hash codes is leveraged not merely for retrieval precision but as a first-class mechanism for generating user-consumable explanations.

The systems community has concurrently advanced indexing structures that accommodate binary codes at scale. Multi-index hashing strategies decompose long codes into substrings to enable exact search in sub-linear time using inverted indices [12], while product quantization offers complementary compression for vectors that exceed purely binary representations [11]. These indexing technologies, when combined with distributed memory architectures and dynamic resharding, form the operational backbone of production retrieval systems. The ability to explain why a particular item was retrieved, however, demands more than efficient index traversal; it requires traceability from the query representation through the hashing decision boundaries to the final candidate set, a requirement that has yet to be systematically incorporated into retrieval system design.

3. Architectural Foundations of Semantic Hash-Based Retrieval

The proposed framework is structured around three co-designed modules: a semantic hashing encoder that maps input images to compact binary codes, a distributed retrieval index that supports efficient Hamming distance search, and an explanation generator that translates retrieval outcomes into human-understandable justifications. The encoder is realized as a deep neural network terminating in a bottleneck layer with sign activation or stochastic binarization,

trained with a multi-objective loss that balances reconstruction fidelity, pairwise similarity preservation, and semantic attribute alignment. Unlike conventional deep hashing models that treat semantic consistency as a soft regularization term, our architecture elevates it to a core optimization objective by anchoring latent dimensions to a curated concept ontology derived from visual attribute datasets and domain-specific taxonomies. This ensures that each bit, or contiguous block of bits, corresponds to a recognizable property such as the presence of specific object categories, texture patterns, or aesthetic qualities.

The indexing layer hosts multiple hash tables, each keyed by a subspace of the full binary code to enable sub-linear candidate screening similar to multi-index hashing [12]. To accommodate the semantic granularity of the encoder, the index is organized hierarchically: coarse-level tables based on high-level categories prune the search space, while fine-grained tables resolve detailed similarity. This architecture supports tunable retrieval cascades in which early stages apply broad, explainable filters and later stages refine the ranking with more nuanced code matching. The retrieval cascade is implemented using a microservice mesh deployed across containerized cloud instances, with read replicas scaling horizontally behind a load-balancing layer. Binary code computation is inherently low-latency and amenable to hardware acceleration via vectorized CPU instructions or FPGA offloading, which significantly reduces tail latency variability [15]. By decoupling the hashing encoder from the index serving tier, the system can independently update the semantic ontology, rerank with fairness constraints, or refresh indices without interrupting query traffic.

The explanation generator consumes the top-k retrieval results alongside their binary codes and the query code, then produces explanations through a combination of bit-level attribution and prototype-based reasoning. For a given recommendation, the module identifies which hash bits diverged between the query and the retrieved item and maps those bit blocks to their semantic labels, constructing a natural language template such as “This item was recommended because it shares a similar color palette and style category, but differs in pattern.” When richer interpretability is required, Grad-CAM-based attention maps [15] are computed on the encoder’s intermediate feature maps and aligned with the semantic attributes implicated by the hash bits, providing a visual rationale that overlays the image. The explanation module is stateless and can be invoked asynchronously, allowing recommendations to be served immediately with cached explanations while deeper analyses are generated on demand.

The interplay of these modules reflects a deliberate separation of concerns that facilitates system evolution: the encoder embodies domain knowledge, the index embodies retrieval efficiency, and the generator embodies communication goals. This modularity is critical for maintaining system reliability under continuous data drift, as it allows each component to be retrained, audited, or replaced without cascading failure. Orchestration is managed through a pipeline controller that monitors quality-of-service metrics—query-per-second throughput, p95 latency, explanation freshness—and dynamically adjusts resource allocation and fallback paths.

4. Explainability Mechanisms in Visual Recommendations

Explanation in recommendation systems spans multiple dimensions, including transparency of the underlying model, justification of individual recommendations, and scrutability that enables users to correct erroneous assumptions. Within a hash-based framework, explainability arises from the inherent interpretability of the binary representation and the explicit alignment with semantic concepts. We categorize the explanation strategies into

intrinsic and post-hoc pathways. Intrinsic explainability leverages the semantic groundedness of the hash encoder to directly expose the bits that dominate the similarity computation. Because each bit group is mapped to a named attribute, the system can present an ordered list of shared and differing attributes without additional inference overhead. This stands in contrast to traditional dense embeddings where interpreting a cosine similarity requires complex feature attribution methods.

Post-hoc explanation augments the basic attribute comparison with visual saliency and counterfactual examples. Visual saliency techniques such as Grad-CAM highlight image regions that most influenced the convolutional features prior to hashing, offering users a spatial map of model attention that can be corroborated against their own perceptual understanding [15]. Counterfactual explanations answer “why not” queries by showing how the recommendation would change if certain hash bits were flipped, corresponding to a tangible semantic alteration such as replacing a striped pattern with a solid one [18]. These counterfactual perturbations are generated by a learned editing module constrained to flip only bits linked to mutable stylistic attributes, preserving the identity of core object categories. The combination of attribute-level comparison, visual saliency, and counterfactual reasoning provides a multi-faceted explanatory layer that caters to diverse user expertise levels.

From a systems perspective, the generation of explanations must not degrade retrieval latency or throughput beyond acceptable service-level agreements. The framework therefore asynchronously pre-computes attribute annotations and saliency maps for catalog items offline, storing them alongside the binary index entries. At query time, the explanation assembler fetches pre-computed metadata and only invokes online inference for counterfactual queries if explicitly requested. This design mirrors content delivery network edge-caching strategies, where explanatory assets are cached at geographical edges to minimize round-trip latency. The architecture also supports explanation personalization by maintaining user-specific profiles that weight the relevance of different semantic dimensions, a feature that requires careful data minimization to comply with privacy regulations [19].

Explainability auditability is a further demand in regulated sectors. The framework records immutable logs that capture the query hash, the retrieved item hashes, the Hamming distances, the semantic bit interpretations, and the generated explanation templates. These logs enable post-hoc audits by internal ethics teams or external regulators without reconstructing model states. Moreover, they provide a quantitative basis for measuring the consistency of explanations across demographic groups, serving as an early warning system for biased reasoning patterns. By design, the system treats explanations not as an afterthought but as a verifiable artifact of the retrieval process.

5. System Design and Deployment Considerations

Deploying a semantic hash-based retrieval system at scale requires navigating a complex landscape of infrastructure choices, operational constraints, and environmental considerations. The binary code length is a critical design parameter that governs the trade-off between retrieval recall and memory footprint. Longer codes increase discriminative capacity but demand more storage and wider index tables, while shorter codes may collapse semantic distinctions. Empirical characterization across datasets recommends code lengths between 64 and 256 bits for visual domains, with grouped block structures that balance semantic granularity and index scalability. The index tier leverages a distributed key-value store where each node maintains a subset of hash tables partitioned by semantic groups to ensure balanced load under skewed query distributions. Fault tolerance is achieved through consistent

hashing-based sharding with lightweight read replicas, which allows the system to gracefully degrade under partial node failures and to scale elastically with diurnal traffic patterns.

Energy proportionality is a growing concern for large-scale retrieval services. Binary embedding storage reduces memory requirements by up to two orders of magnitude compared to 512-dimensional float32 vectors, directly cutting static power consumption in RAM modules. Furthermore, Hamming distance computation is extremely low-energy relative to dot product operations on dense vectors, especially when implemented on low-power accelerators such as FPGAs or dedicated ASICs. The framework includes a sustainability controller that monitors carbon intensity signals from the power grid and dynamically shifts non-latency-critical explanation generation to periods of lower carbon intensity or to data centers with greener energy mixes. These scheduling decisions align with broader green AI principles that advocate for energy-aware training and inference lifecycles [24].

Robustness is another pivotal system property. Visual recommenders deployed in the wild encounter substantial domain shift as fashion trends evolve or new product categories emerge. The semantic hashing encoder must be continuously updated through a progressive fine-tuning pipeline that leverages self-supervised consistency losses on new unlabeled images, ensuring that hash bit semantics do not drift from their original meanings [13]. The system implements a canary deployment strategy where updated encoders are tested on a small fraction of live traffic while monitoring explanation fidelity and fairness metrics before full rollout. Adversarial robustness also warrants attention, as small imperceptible perturbations to a query image could theoretically flip hash bits and alter the recommendation set. Defensive quantization and randomized binarization during inference provide probabilistic guarantees that limit the impact of white-box attacks, while anomaly detection monitors query patterns for unusual bit perturbation signatures that may indicate coordinated exploitation.

Fairness-aware indexing constitutes a transversal aspect of deployment architecture. Empirical studies in recommender systems have shown that retrieval algorithms can inadvertently amplify popularity biases, demographic skews, or representational harms present in the catalog [11], [12]. The framework integrates a fairness broker that sits between the index and the ranking stage, applying post-retrieval re-ranking policies that ensure a minimum representation of items from protected categories or attribute groups. Because the hash codes carry semantic attributes, the broker can efficiently filter or boost candidates based on target distribution constraints without recomputing embeddings. Continuous monitoring dashboards track retrieval parity metrics—such as the proportion of recommended items from minority-owned brands across user segments—and trigger alerts when degradation surpasses pre-defined thresholds. This operationalizes the principle of procedural fairness as an ongoing infrastructural commitment rather than a one-time pre-deployment audit.

6. Fairness, Robustness, and Governance Implications

The governance of AI-driven recommendation systems has evolved rapidly, with regulations such as the European Union's AI Act and the General Data Protection Regulation (GDPR) imposing requirements for transparency, human oversight, and non-discrimination. A semantic hash-based framework directly supports these mandates by providing an auditable representation that can be linked to the data minimization principle: only compact binary codes need to be exchanged between services, reducing the attack surface for personal data leakage. The right to explanation, as articulated in recital 71 of the GDPR, finds a concrete technical realization in the attribute-based explanation templates generated by the system, which offer users meaningful insight into the logic of the recommendation without exposing

proprietary model parameters. Moreover, because hash bits are non-invertible in the absence of the encoder, the architecture provides a structural barrier against model inversion attacks that seek to reconstruct training images.

Fairness governance extends beyond individual explanations to encompass systemic equity. The framework's semantic ontology must be curated with an awareness of social contexts to avoid encoding harmful stereotypes through attribute labels. For instance, aligning hash bits with sensitive attributes such as perceived gender presentation or skin tone could enable downstream filtering that contravenes anti-discrimination laws if not carefully governed. Therefore, the ontology development process follows a participatory design methodology involving domain experts, ethicists, and community representatives to define which visual concepts are permissible and how they are described. The system also includes an override mechanism that allows users to contest and correct attribute misclassifications, with feedback loops that retrain the encoder to reduce similar errors in the future, thereby enhancing individual and collective recourse [20].

Robustness governance involves establishing clear accountability chains for model updates. When a new encoder version is promoted, an automated impact assessment evaluates changes in explanation consistency, fairness metrics, and retrieval coverage across diverse user cohorts. The assessment report is archived and made accessible to internal review boards, facilitating compliance with algorithmic impact assessment requirements proposed in recent policy frameworks. Additionally, the binary code index includes provenance metadata that records which encoder version produced each entry, enabling rollback and forensic analysis. This level of traceability transforms the retrieval engine from an opaque utility into a governed socio-technical system where decisions can be contested and rectified.

The tension between personalization and fairness is particularly pronounced in visual recommendations where style and identity intersect. A purely utility-maximizing system might over-recommend items aligning with a user's past preferences, inadvertently narrowing exposure to diverse aesthetics and reinforcing cultural echo chambers. Semantic hashing can mitigate this by providing a knob to controllably inject diversity: the fairness broker can perturb the query hash in controlled directions corresponding to underrepresented attributes, ensuring that the recommendation set expands the user's horizon while preserving relevance. This technique, which we term diversity-controlled hash perturbation, is transparent to the user and can be toggled via preference settings, implementing a user-centric governance mechanism that respects autonomy while gently counteracting filter bubbles.

7. Conclusion

The semantic hash-based retrieval framework presented in this paper demonstrates that efficiency and explainability in visual recommendation systems are not inherently antagonistic but can be synergistically integrated through careful architectural design and governance-aware engineering. By aligning binary hash codes with a semantic concept ontology, the system provides a substrate that simultaneously accelerates approximate nearest neighbor search, reduces energy consumption, and produces verifiable, user-friendly explanations. The modular decomposition into hashing encoder, distributed index, and explanation generator enables independent evolution of each component while preserving system reliability and auditability. Our analysis highlights that sustainability, fairness, and robustness must be treated as first-class design constraints rather than peripheral add-ons, influencing choices from code length and index partitioning to deployment strategies and continuous monitoring.

Looking ahead, several open challenges merit further investigation. The dynamic evolution of semantic ontologies in response to cultural shifts requires lightweight model update mechanisms that do not necessitate full re-indexing. Tighter integration with emerging hardware accelerators, including neuromorphic chips and photonic processors, could push the energy efficiency of binary retrieval to new frontiers. The development of standardized benchmarks that jointly measure retrieval accuracy, explanation fidelity, fairness, and carbon footprint will be essential for driving progress across these interdependent dimensions. Finally, the legal codification of explanation standards will likely shape the technical design space, and close collaboration between systems builders and policy researchers will be critical to ensure that regulations remain grounded in technological reality while technology evolves in socially beneficial directions. In this integrated vision, semantic hashing moves from a niche compression technique to a central pillar of responsible, large-scale visual recommendation infrastructure.

References

1. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys, 52(1), 1-38. https://doi.org/10.1145/3158369

2. He, R., & McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In Proceedings of the 25th International Conference on World Wide Web (WWW) (pp. 507-517). https://doi.org/10.1145/2872427.2883037

3. Gionis, A., Indyk, P., & Motwani, R. (1999). Similarity search in high dimensions via hashing. In Proceedings of the 25th International Conference on Very Large Data Bases (VLDB) (pp. 518-529).

4. Wang, J., Zhang, T., Song, J., Sebe, N., & Shen, H. T. (2018). A survey on learning to hash. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 769-790. https://doi.org/10.1109/TPAMI.2017.2699960

5. Cao, Z., Long, M., Wang, J., & Yu, P. S. (2017). HashNet: Deep learning to hash by continuation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 5609-5618). https://doi.org/10.1109/ICCV.2017.598

6. Liu, H., Wang, R., Shan, S., & Chen, X. (2016). Deep supervised hashing for fast image retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2064-2072). https://doi.org/10.1109/CVPR.2016.227

7. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (ICML) (pp. 1597-1607).

8. Kang, W. C., Fang, C., Wang, Z., & McAuley, J. (2017). Visually-aware fashion recommendation and design with generative image models. In Proceedings of the IEEE International Conference on Data Mining (ICDM) (pp. 207-216). https://doi.org/10.1109/ICDM.2017.30

9. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. (2017). Neural collaborative filtering. In Proceedings of the 26th International Conference on World Wide Web (WWW) (pp. 173-182). https://doi.org/10.1145/3038912.3052569

10. Li, W. J., Wang, S., & Kang, W. C. (2016). Feature learning based deep supervised hashing with pairwise labels. In Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI) (pp. 1711-1717).

11. Jegou, H., Douze, M., & Schmid, C. (2011). Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(1), 117-128. https://doi.org/10.1109/TPAMI.2010.57

12. Norouzi, M., Punjani, A., & Fleet, D. J. (2012). Fast search in hamming space with multi-index hashing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3108-3115). https://doi.org/10.1109/CVPR.2012.6248048

13. Yu, Z., Wu, S., Dou, Z., & Bakker, E. M. (2022). Deep hashing with self-supervised asymmetric semantic excavation and margin-scalable constraint. Neurocomputing, 483, 87-104.

14. Qu, Y., Kamath, U., & Wu, X. (2021). A survey on explainable recommender systems: From collaborative filtering to knowledge graphs. ACM Computing Surveys, 54(4), 1-38. https://doi.org/10.1145/3447756

15. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 618-626). https://doi.org/10.1109/ICCV.2017.74

16. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778

17. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR).

18. Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT) (pp. 607-617). https://doi.org/10.1145/3351095.3372850

19. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76-99. https://doi.org/10.1093/idpl/ipx005

20. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT) (pp. 59-68). https://doi.org/10.1145/3287560.3287598

21. Singh, A., & Joachims, T. (2018). Fairness of exposure in rankings. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2219-2228). https://doi.org/10.1145/3219819.3220088

22. Dean, J., & Barroso, L. A. (2013). The tail at scale. Communications of the ACM, 56(2), 74-80. https://doi.org/10.1145/2408776.2408794

23. Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(4), 824-836. https://doi.org/10.1109/TPAMI.2018.2889473

24. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650). https://doi.org/10.18653/v1/P19-1355

25. Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535-547. https://doi.org/10.1109/TBDATA.2019.2921572