Enhancing Solar Magnetogram Retrieval with Deep Semantic Hashing and Hierarchical Graph Indexing

Rafał Grycuk
Czestochowa University of Technology
Faculty of Computer Science and Artificial Intelligence
Czestochowa, Poland *rafal.grycuk@pcz.pl*

Rafał Scherer
Czestochowa University of Technology
Faculty of Computer Science and Artificial Intelligence
Czestochowa, Poland
AGH University of Krakow
Faculty of Computer Science
and Center of Excellence in Artificial Intelligence
Krakow, Poland *rafal.scherer@pcz.pl*

Abstract

We present a method for content-based retrieval of solar magnetograms using semantic hashing. Based on HMI data from SDO and implemented with SunPy and PyTorch, the approach encodes magnetic regions as fixed-length vectors, avoiding full-disk image processing. A fully connected autoencoder compresses 400-dimensional descriptors into 50-dimensional semantic hashes. Experiments show that our method outperforms state-of-the-art techniques in precision and is also applicable to solar image classification.

Keywords: fast image hash, solar activity analysis, solar image description,

1. Introduction

The Solar Dynamics Observatory (SDO), part of NASA's Living With a Star (LWS) program, continuously observes the Sun at high spatial and temporal resolutions across multiple wavelengths [1]. Among its instruments, the Helioseismic and Magnetic Imager (HMI) measures solar surface oscillations and magnetic fields, providing data such as dopplergrams, continuum filtergrams, and magnetograms. This work addresses efficient retrieval of HMI magnetograms with similar magnetic structures. Given SDO's high data volume, manual analysis is infeasible. Existing image retrieval techniques are not well-suited for solar data. To overcome this, we apply semantic hashing, which maps high-dimensional inputs to compact binary codes while preserving similarity [3], [8], [11]. Section 2 details our hashing method. Experimental validation is provided in Section3, followed by conclusions and future work in Section 4.

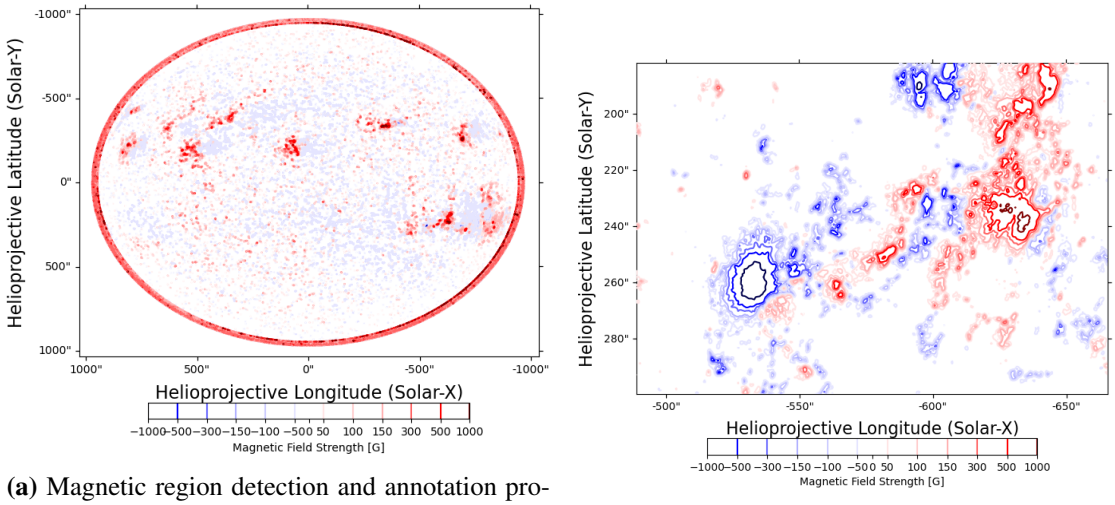
2. Proposed Method for Solar Magnetogram Hashing

SDO instruments enable multi-wavelength solar imaging (e.g., via AIA) and the creation of detailed magnetograms depicting the solar magnetic field. In active regions, field strengths can exceed typical values by over three orders of magnitude. Given their full-disk coverage, magnetograms support diverse solar physics analyses. We utilize magnetograms for solar image representation and hashing, positing that magnetic data improves retrieval precision by suppressing irrelevant noise. Unlike intensity-based images (often affected by transient phenomena like flares) magnetograms provide a more stable and physically grounded view of solar activity.

This makes them well-suited for image retrieval and classification tasks. This section presents the key steps of the proposed hashing approach.

2.1. Magnetic Region Detection

The first step involves preprocessing the magnetogram to enhance magnetic region visibility—termed magnetic region detection (MRD) (Fig. 1a). Magnetogram data were acquired using the SunPy library [10], [12], enabling estimation of field strength across the solar surface. As shown in Fig.1a, magnetic fields intensify near active regions. We map field strength to image intensity (Fig.1b), revealing magnetic regions (MRs) that influence solar dynamics. MRs are closely linked to CMEs and flares, especially where opposite polarities–red (north) and blue (south)–interact. Identifying and tracking MRs is thus essential for flare prediction and forms a critical input for the next algorithmic stages.



(a) Magnetic region detection and annotation process. Magnetic regions can be clearly visible.

(b) A magnification of magnetic regions.

Fig. 1. Magnetic region detection and their magnification.

2.2. Calculation of Magnetic Field Grid-based Descriptor

This section describes the Magnetic Field Grid-based Descriptor. To reduce the computational load of full-disk image analysis, we represent the Magnetic Region Intensity Image (MRI, see Section 2.1) using a regular $N \times M$ grid. Each row yields a magnetic field intensity histogram by aggregating values across its cells, producing M histograms. These are concatenated into a single descriptor vector (MFGDV) for a compact representation.

All the steps described above are presented in Fig. 2 and in Alg. 2 respectively. Based on a series of experiments and simulations, we determined that setting both parameters N and M to 5 produces the most effective results for our solution. Every histogram has fixed values between [-1000;1000] Therefore, we have 16 values for each cell of the grid, thus finally $MFGDV$ has 400 values.

2.3. Hash Generation

This section outlines the hash generation process, which encodes the Magnetic Field Grid-Based Descriptor Vector (MFGDV) into a compact binary form. The goal is to produce a representative hash of solar magnetic structures at a given time, enabling efficient retrieval (see Section 2.4). We use a fully connected autoencoder (AE) to compress MFGDV into a lower-dimensional

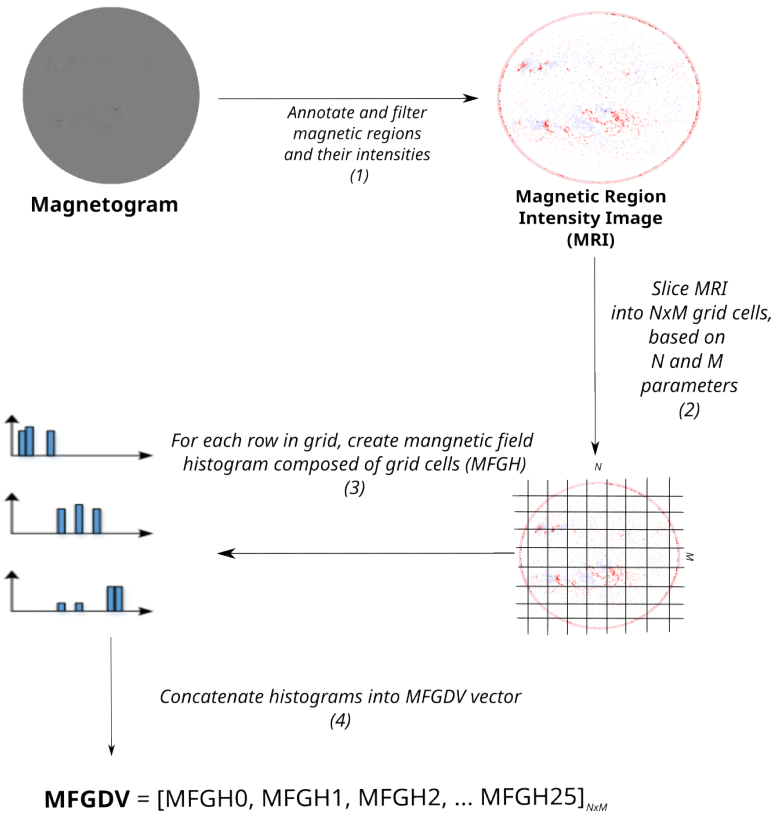


Fig. 2. Algorithm steps for calculating the Magnetic Field Grid-Based Descriptor Vector.

latent space. Due to their ability to preserve semantic content in an unsupervised manner, AEs are well-suited for generating compact, content-aware hashes [4],[7]. The autoencoder architecture is shown in Table 1. It provides effective dimensionality reduction while preserving key magnetic region features. Only the encoder is used for hash generation; the decoder serves training purposes to minimize reconstruction error. Training for 40 epochs yielded a good balance between generalization and overfitting, producing stable and meaningful hashes.

2.4. Retrieval

For retrieval, we use the *Hierarchical Navigable Small World* (HNSW) algorithm [9] to perform approximate nearest neighbor (ANN) search in the image embedding space. HNSW constructs a multilayer proximity graph $G = (V, E)$, where each node represents an image embedding vector $\mathbf{x}_i \in \mathbb{R}^d$. The graph is organized into a hierarchy of layers $L_0, L_1, \ldots, L_{\max}$, with L_0 containing all data points and each layer L_l ($l > 0$) being a progressively sparser subset. The probability of assigning a node to level l follows $P(l_{\max} = l) \sim e^{-\lambda l}$, ensuring an average logarithmic hierarchy depth. During indexing, each vector $\mathbf{x}_i$ is inserted using greedy search from the top layer down to L_0, connecting to M nearest neighbors while preserving the small-world property. At query time, a query vector $\mathbf{q} \in \mathbb{R}^d$ initiates a top-down search, refining a candidate list C by minimizing a distance function, typically $D(\mathbf{q}, \mathbf{x}_i) = \|\mathbf{q} - \mathbf{x}_i\|_2$, or, for normalized embeddings, cosine similarity $\cos(\mathbf{q}, \mathbf{x}_i) = (\mathbf{q} \cdot \mathbf{x}_i)/(\|\mathbf{q}\| \, \|\mathbf{x}_i\|)$. The final nearest neighbors are selected in layer L_0, with optional reranking. The expected time complexity of HNSW queries is $\mathcal{O}(\log N)$, making it suitable for high-dimensional image retrieval. We incorporate HNSW into our system to ensure scalable and accurate retrieval of magnetogram-based image hashes.

Table 1. Tabular representation of the fully-connected autoencoder (input = 400, latent = 50).

Layer (type)	Output	Filters (in, out)	Params
$Input\,(InputLayer)$	[1, 400]		0
$Linear\_1\,(Linear)$	[1, 200]	400, 200	80,200
$ReLU\_1$	[1, 200]		0
$Linear\_2\,(Linear)$	[1, 100]	200, 100	20,100
$ReLU\_2$	[1, 100]		0
$Linear\_3\,(Linear)$	[1, 50]	100, 50	5,050
$ReLU\_3$	[1, 50]		0
$Encoded\,(latent\text{-}space)$	[1, 50]		
$Linear\_4\,(Linear)$	[1, 100]	50, 100	5,100
$ReLU\_4$	[1, 100]		0
$Linear\_5\,(Linear)$	[1, 200]	100, 200	20,200
$ReLU\_5$	[1, 200]		0
$Linear\_6\,(Linear)$	[1, 400]	200, 400	80,400
$ReLU\_6$	[1, 400]		0
$Decoded\,(Tanh)$	[1, 400]		

3. Experimental Results

This section presents simulation results and our evaluation approach using unlabeled solar images. Due to the lack of ground-truth labels, we applied unsupervised learning to encode descriptors, leveraging the Sun's rotation as a form of natural supervision to define sets of similar images (SI). Our key assumption is that images taken within short temporal intervals depict the same active regions with minor positional shifts. Images were sampled every 6 minutes, with high visual similarity expected across consecutive frames. Based on experiments, we defined a temporal similarity window of 48 hours. For instance, an image from 2012-06-15 00:00:00 has its SI set drawn from the 24 hours before and after that timestamp. This assumption enables practical performance evaluation of our hashing method. Each experiment followed these steps: (1) submit a query image and retrieve results using the hash; (2) compare retrieved image timestamps; (3) classify images as similar if they fall within the 48-hour window.

Once the set of similar images (SI) is identified, standard performance metrics such as precision and recall can be calculated, as discussed in works such as [5],[13]. These metrics are based on the following sets: SI - set of similar images; RI - set of retrieved images for query; $PRI(TP)$ - set of positive retrieved images (true positive); $FPRI(FP)$ - false positive retrieved images (false positive); $PNRI(FN)$ - positive, not retrieved images; $FNRI(TN)$ - false, not retrieved images (TN). We compared our method with state-of-the-art approaches [3]. Our method achieved an average precision of ≈ 0.930, outperforming Banda et al. (0.848), Angryk et al. (0.850) [3],[2], and Grycuk et al. [6].

Most solar images that were temporally and structurally close to the query were correctly retrieved. In contrast, relevant but not retrieved images (PNRI) generally had larger temporal offsets. However, their occurrence was notably lower than in previous studies. Elevated PNRI values are primarily due to solar dynamics, especially rotation which can shift or obscure active regions even within a 48-hour window.

4. Conclusions

We proposed a semantic hashing method for retrieving structurally similar solar magnetograms based on HMI data from SDO, using SunPy and PyTorch. Magnetic regions were encoded as fixed-length vectors, enabling efficient comparison in a 50-dimensional space. A fully connected

autoencoder reduced the 400-dimensional MFGDV to a compact semantic hash. Our method achieved the highest precision among the state-of-the-art techniques. Beyond retrieval, it is also suitable for classification tasks. Using magnetograms, less affected by transient noise than AIA images, the method yields robust and noise-resistant representations.

References

[1] Banda, J.M., Angryk, R.A.: Selection of image parameters as the first step towards creating a cbir system for the solar dynamics observatory. In: 2010 International Conference on Digital Image Computing: Techniques and Applications. pp. 528–534. IEEE (2010)

[2] Banda, J.M., Angryk, R.A.: Large-scale region-based multimedia retrieval for solar images. In: International Conference on Artificial Intelligence and Soft Computing. pp. 649–661. Springer (2014)

[3] Banda, J.M., Angryk, R.A.: Regional content-based image retrieval for solar images: Traditional versus modern methods. Astronomy and computing 13, pp. 108–116 (2015)

[4] Brunner, C., Kő, A., Fodor, S.: An autoencoder-enhanced stacking neural network model for increasing the performance of intrusion detection. Journal of Artificial Intelligence and Soft Computing Research 12(2), pp. 149–163 (2022)

[5] Buckland, M., Gey, F.: The relationship between recall and precision. Journal of the American society for information science 45(1), pp. 12 (1994)

[6] Grycuk, R., Scherer, R.: Grid-based concise hash for solar images. In: International Conference on Computational Science. pp. 242–254. Springer (2021)

[7] Krizhevsky, A., Hinton, G.E.: Using very deep autoencoders for content-based image retrieval. In: ESANN. vol. 1, p. 2 (2011)

[8] Kucuk, A., Banda, J.M., Angryk, R.A.: A large-scale solar dynamics observatory image dataset for computer vision applications. Scientific data 4, pp. 170096 (2017)

[9] Malkov, Y.A., Yashunin, D.A.: Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE transactions on pattern analysis and machine intelligence 42(4), pp. 824–836 (2018)

[10] Mumford, S., Freij, N., et al.: Sunpy: A python package for solar physics. Journal of Open Source Software 5(46), pp. 1832 (2020), https://doi.org/10.21105/joss.01832

[11] Salakhutdinov, R., Hinton, G.: Semantic hashing. International Journal of Approximate Reasoning 50(7), pp. 969–978 (2009)

[12] The SunPy Community et al.: The sunpy project: Open source development and status of the version 1.0 core package. The Astrophysical Journal 890, pp. 1–12 (2020), https://iopscience.iop.org/article/10.3847/1538-4357/ab4f7a

[13] Ting, K.M.: Precision and recall. In: Encyclopedia of machine learning, pp. 781–781. Springer (2011)