A Deep Feature Retrieved Network for Bitemporal Remote Sensing Image Change Detection

Shizhen Chang[1], Michael Kopp[1] and Pedram Ghamisi[1,2]

[1]Institute of Advanced Research in Artificial Intelligence (IARAI), 1030 Vienna, Austria
[2]Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Machine Learning Group, 09599 Freiberg, Germany

Abstract

The task of bitemporal change detection aims to identify the surface changes of specific scenes at two different points in time. In recent years, we have increasingly witnessed the success of deep learning in a variety of applications in remote sensing, including change detection and monitoring. In this paper, a novel deep feature retrieval neural network architecture for change detection is proposed that uses a trainable associative memory component to exploit potential similarities and connections of the deep features between image pairs. A key ingredient in our novel architecture is the use of a continuous modern Hopfield network component. The proposed method beats the current state-of-the-art on the well-known LEVIR-CD data set. The codes of this work will soon be available online (https://github.com/ShizhenChang).

Keywords

Remote sensing, change detection, modern Hopfield network, deep learning, Siamese network, convolutional neural network.

1. Introduction

With the rapid development of technologies for Earth observation, an ever-growing amount of very high resolution (VHR) remote sensing data has become available for geographical analysis and image processing [1]. VHR images can provide detailed information about land surfaces, and images collected at different time epochs from the scene are able to record changes regularly. Therefore, as one of the most important remote sensing tasks, change detection has been widely applied in many areas of land-use and land-cover analysis, such as environmental monitoring, urban growth, deforestation assessment, shifting cultivation evaluation, and so on.

A variety of deep neural networks, such as the convolutional neural networks (CNNs) [2], autoencoders (AEs) [3], recurrent neural networks [4], generative adversarial network (GAN) [5], and deep belief network (DBNs) [6], have been successfully utilized for remote sensing change detection over the last few years. Among them, CNN-based methods can take full use of the spatial information of VHR remote sensing images, thus, can better extract high-level deep features and abstract semantic contents to learn discriminative differences between the periods.

Strategies that have been applied to extract deep features of the inputs, can be broadly divided into two categories: early fusion [7, 8] and late fusion [9] networks. The early-fusion networks first concatenate multitemporal images into a unified data cube, and then, the parameters are hierarchically fine-tuned. The late-fusion networks usually learn single-temporal features individually and share the parameters by using a Siamese network. Compared to early-fusion networks, late-fusion methods can better utilize the features of the inputs and return clearer contours of the change objects. However, the features of shallower layers may not be sufficiently learned and utilized due to the gradient vanishing problem. Therefore, learning information from both shallow and deep layers are very important to effectively detect changes using deep-learning-based approaches.

In order to accurately extract features, deeper and more complex CNN-based networks have been designed, that include architecture components such as Long Short-Term Memory (LSTM) [10] and attention mechanisms (self-attention [11], spatial attention [12], and channel attention [8]). The successful combination of CNNs and other networks has shown that discriminative features within the image pairs can be better extracted and the detection accuracy can be greatly improved. However, limited by the architecture of CNNs, as the high-level features are only related to the shallower layers through larger receptive fields, the global and temporal information between the image pairs are still not sufficiently utilized.

To address this issue, we design a Hopfield pooling block to interactively retrieve the high-level concepts of changes. This idea is inspired by the successful application of the modern Hopfield network for continuous pattern retrieval [13]. Our assumption is that the semantic information between the image pairs in deeper layers

CDCEO 2022: 2nd Workshop on Complex Data Challenges in Earth Observation, July 25, 2022, Vienna, Austria
shizhen.chang@iarai.ac.at (S. Chang); michael.kopp@iarai.ac.at (M. Kopp); pedram.ghamisi@iarai.ac.at (P. Ghamisi)
0000-0002-9785-7937 (S. Chang); 0000-0002-1385-1109 (M. Kopp); 0000-0003-1203-741X (P. Ghamisi)

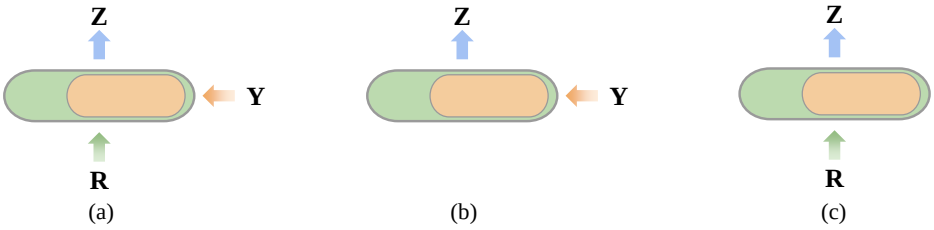

Figure 1: A brief illustration of three types of Hopfield layers for deep learning [13], where both the stored patterns **Y** and the query patterns **R** can be obtained from the previous layers or the input or can be learned. The output **Z** are the retrieved patterns for the queries, each being a linear combination of stored patterns lying in the convex hull of the simplex spanned by the stored patterns. (a) This *Hopfield* layer associates two sets **R** and **Y** to propagate sets of vectors. (b) Layer *Hopfield Pooling* layer performs a pooling operation to the set **Y** via learned queries. (c) The *Hopfield* layer learns a new set of stored patterns based on the input **R**.

can be represented using a common matrix, i.e. a query, that can be learned during the training process. We use this query to retrieve related semantic features between given images. These retrieved features reflect a common spatio-temporal context and are used by subsequent layers in our network. Concretely, we incorporate a Hopfield network block into a Siamese fully convolutional network (FCN) resulting in the design of our proposed deep feature retrieved network (FrNet) for bitemporal remote sensing change detection. It should be noted that different from previous change detection models, both semantic and temporal information can be fully considered; and it is our first attempt of using modern Hopfield networks in the remote sensing community.

The rest of this article is organized as follows. Section II briefly reviews continuous modern Hopfield networks. Section III describes the proposed method. Experiments are conducted and discussed in Section IV.

2. Continuous Modern Hopfield Network

Binary modern Hopfield networks are associative memories on binary data that can retrieve data of exponentially many stored patterns [14, 15], this being the key distinguishing feature to their classical binary counterparts [16, 17]. These binary modern Hopfield networks have been generalized to continuous modern Hopfield networks that, crucially, are differentiable and can thus be embedded in deep learning architectures trained by gradient descent [13, 18]. Moreover, continuous modern Hopfield networks retain the key ability to store exponentially many patterns and they can furthermore retrieve patterns in only one update step.

Given a matrix X of shape $d \times N$ formed of column vectors $\{x_1, ..., x_N\} \in \mathbb{R}^d$, a query pattern ξ, also a column vector, seeks to retrieve the best pattern in the convex hull of the simplex spanned by the $\{x_1, ..., x_N\}$, such the following energy function is minimized:

$$E = -\beta^{-1} \log(\sum_{i=1}^{N} \exp(\beta x_i^\top \xi)) + \beta^{-1} \log N + \frac{1}{2}\xi^\top \xi + \frac{1}{2}M^2,$$

where M is the largest norm of the $\{x_1, ..., x_N\}$ in $\mathbb{R}^d$. As shown in [13, 18], ξ^{new} is defined by the following update rule:

$$\xi^{new} = f(\xi; X, \beta) = X\text{softmax}(\beta X^\top \xi), \tag{1}$$

which will converge globally, almost always, to a local minima of the energy function in essentially one update step. Moreover, equation (1) is closely related to the well known transformer attention mechanism, showing that retrieval in modern Hopfield networks and transformer attention coincide [13, 18].

With changable structures in deeper networks (as shown in Fig. 1), continuous modern Hopfield networks have greater application prospects in deep learning. It has been successfully applied to solving large scale multi-instance learning tasks [19], to few- and zero-shot chemical reaction template prediction [20], to creating new reinforcement learning algorithms [21, 22], to improving contrastive learning of joint image- and text embedding representations [23] and to tabular data [24].

Inspired by continuous modern Hopfield networks, we design a Siamese Hopfield pooling layer and attempt to capture deep feature differences for remote sensing bitemporal change detection.

3. Deep Feature Retrieved Network for Change Detection

3.1. Overview

As shown in Fig. 2, the proposed deep feature retrieved network (FrNet) is a Siamese network that contains three

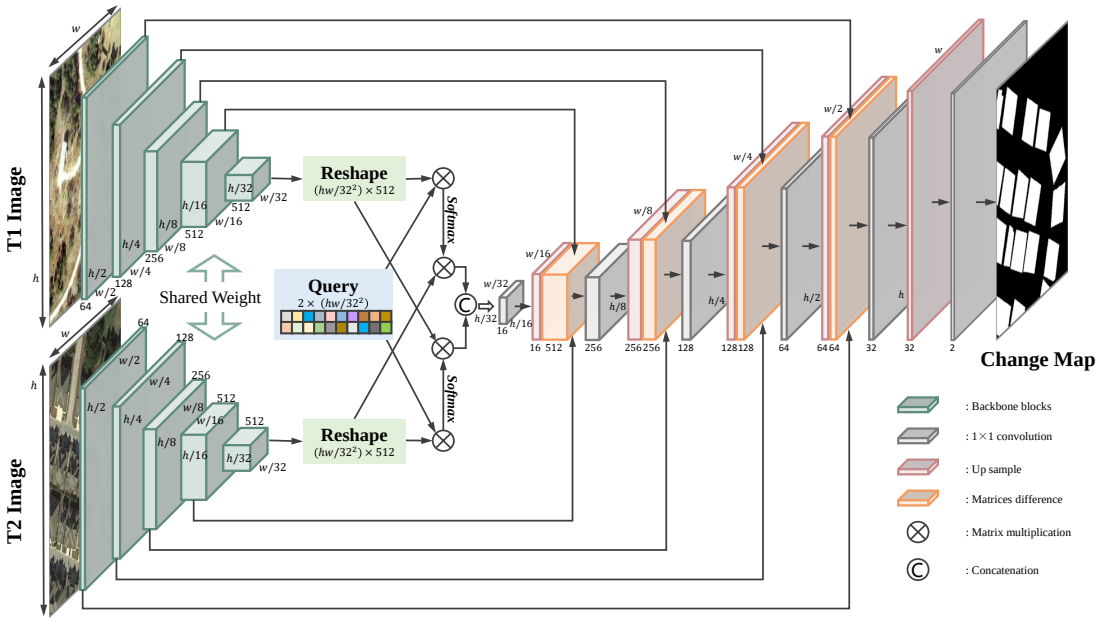

Figure 2: Flowchart of the proposed FrNet.

parts: a feature extractor, a Hopfield pooling block, and a decoder. Bitemporal change detection can be viewed as a segmentation task for image pairs that record the same geographic information at different times. Since the shapes and sizes of changed objects vary a lot, deeper layers of CNN-based approaches (e.g., U-Net and U-Net++) can effectively extract semantic features and retain details with a larger receptive field. To extract useful information from bitemporal images, a Siamese network with consistent architectures and shared weights are utilized as the feature extractor in our implementation (shown with green blocks in Fig. 2). The VGG-16 [25] with ImageNet pretrained parameters is chosen as the backbone network. Then the spatial dimensions of deep features are flattened and input into the Hopfield pooling block. The deep features of two periods are pooled and retrieved. After that, we feed the concatenation of the bitemporal retrieved features and the feature differences from shallower layers into the decoder and obtain the change map. The decoding modules are shown in the right part in Fig. 2.

3.2. Hopfield Pooling Block

The Hopfield layer is proven to be capable of retrieving key features of the input through one update. For the proposed bitemporal change detection task, the question is: "how can we obtain the most typical information that is related to changed objects from the bitemporal deep features?". We design a Hopfield pooling block to pool the features of various channels into fewer channels, and at the same time, attempt to interactively retrieve semantic information during the period of changes using the Hopfield update rule.

Let us assume two temporal VHR images are denoted by $X_i \in \mathbb{R}^{3\times h\times w}$, where $i = \{1, 2\}$ represents the i-th time period and h and w are the height and width of the images, respectively. Features obtained by the backbone are denoted as $F_i \in \mathbb{R}^{c\times\tilde{h}\times\tilde{w}}$, where c, $\tilde{h}$, and $\tilde{w}$ represent the number of channels, height, and width of the feature, respectively. For the proposed VGG-16 feature extractor, the channel size of F_i is 512, and the height and width of the features are 1/32 of the original image.

In the Hopfield pooling block, the features are first reshaped into $\mathbb{R}^{\tilde{h}\tilde{w}\times c}$ of row-wise vectors. Then, for the time 1 image, we introduce a trainable weight matrix $W_Q \in \mathbb{R}^{c_Q\times\tilde{h}\tilde{w}}$ to retrieve the related deep features of F_1 related to the 2nd period. The output can be written as:

$$Z_1 = \text{softmax}(\beta W_Q F_1^\top) F_2. \tag{2}$$

The number of rows c_Q in W_Q is set to 2 in this paper which represents the change/unchange semantic information we retrieved.

Similarly, the common weight matrix W_Q is utilized to

retrieve F_2 related to the 1st period:

$$Z_2 = \text{softmax}(\beta W_Q F_2^\top) F_1. \tag{3}$$

It should be noted that the retrieved output Z_1 and Z_2 have the same size and contain both global and temporal information of the image pairs.

We concatenate the retrieved outputs together: $Z = [Z_1; Z_2]$, restore their spatial dimensions, and feed them into a 1×1 2D convolutional layer with 16 filters to generate a new feature map. After bilinear interpolation, the features through the Hopfield pooling block is finally derived:

$$H = U(g(W * Z + b)), \tag{4}$$

where W and b represent the weight matrix and bias vector of the convolutional layers, $*$ denotes the 2D convolutional operation, $g(\cdot)$ denotes the batch normalization with ReLU activation, and $U(\cdot)$ denotes bilinear interpolation with an upsampling rate of 2.

4. Experiments

4.1. Data Set

In the experimental part, the LEVIR-CD data set [26] is utilized to compare the change detection methods. The LEVIR-CD data set is composed of 637 VHR (0.5m/pixel) Google Earth (GE) image pairs with the size of 1024×1024 pixels. These image pairs have been captured in different periods of 5 to 14 years and cover a total of 31,333 individual buildings for the task of building growth assessment. With the ratio of 7:1:2, these image pairs are split into the training set, validation set, and testing set. Following the initial settings, we crop each image into 16 non-overlapped small patches with the size of 256×256 pixels. Thus, there are a total of 7120 image pairs for training, 1024 for validation, and 2024 for testing.

4.2. Comparative method and Evaluation Metrics

To verify the effectiveness of the proposed FrNet method, four representative deep-learning-based change detection networks are taken into consideration. The FC-EF [7] is an early fusion method based on U-Net that concatenates the bitemporal image pairs as the input. And its extended versions, the FC-Siam-diff and FC-Siam-conc [7], use Siamese networks with shared weights to extract multi-level features and use feature difference and concatenation, respectively, to fuse bitemporal information. The bitemporal image transformer (BIT) network [12] designs a context-information-based enhancer to extract related concepts in the token-based space-time, and projects the context-rich tokens back to original features for prediction. To validate the effectiveness of the proposed FrNet, we also set a base model that consists of the CNN backbone (VGG-16) and the decoder for comparison.

Table 1
Quantitative Analysis of Different Networks on the LEVIR-CD Data Set. The Best Values are shown in Bold

Methods	Pre (%)	Rec (%)	F1 (%)	OA (%)
FC-EF	61.86	96.05	75.25	96.78
FC-Siam-conc	67.87	**97.53**	80.04	97.52
FC-Siam-diff	71.37	95.42	81.66	97.82
BIT	80.82	92.86	86.42	98.51
Base Model	85.24	92.26	88.61	98.79
FrNet	**86.32**	92.10	**89.12**	**98.85**

For the evaluation part, the precision (Pre), recall (Rec), F1 score, and overall accuracy (OA) are employed to quantitatively evaluate the performance of the studied methods. These metrics are calculated as follows:

$$Pre = \frac{TP}{TP + FP} \tag{5}$$

$$Rec = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = \frac{2Pre \cdot Rec}{Pre + Rec} \tag{7}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

where TP (True Positive) represents the number of pixels of real changes that are correctly detected, FP (False Positive) represents the number of pixels of unchanged objects that are falsely detected as changed objects, TN (True Negative) denotes the number of pixels of unchanged objects that are correctly regarded as nonchange, and FN (False Negative) denotes the number of changed pixels that are not detected as changed objects.

4.3. Experimental Results and Analysis

In our experiments, the proposed FrNet is implemented with the Pytorch platform using a single NVIDIA A100 GPU (with 40-GB RAM). During the training stage, the Adam optimizer with a weight decay of $1e-5$ was employed. The batch size is set to 32, and the learning rate is initially set to $1e-4$ and will linearly reduce to 0 over 50,000 iterations. The β of the Hopfield layer is set to $1/\sqrt{c_Q}$.

The quantitative results for the precision, recall, F1 score, and OA of all models are summarized in Table 1. It can be found that FC-EF obtains the lowest F1 score (75.25%) and OA (96.78%) among all the models. The FC-Siam-conc and FC-Siam-diff perform slightly better than FC-EF, which indicates the Siamese network and feature difference/concatenation have benefits for the preservation of useful information. The F1 score and OA of the

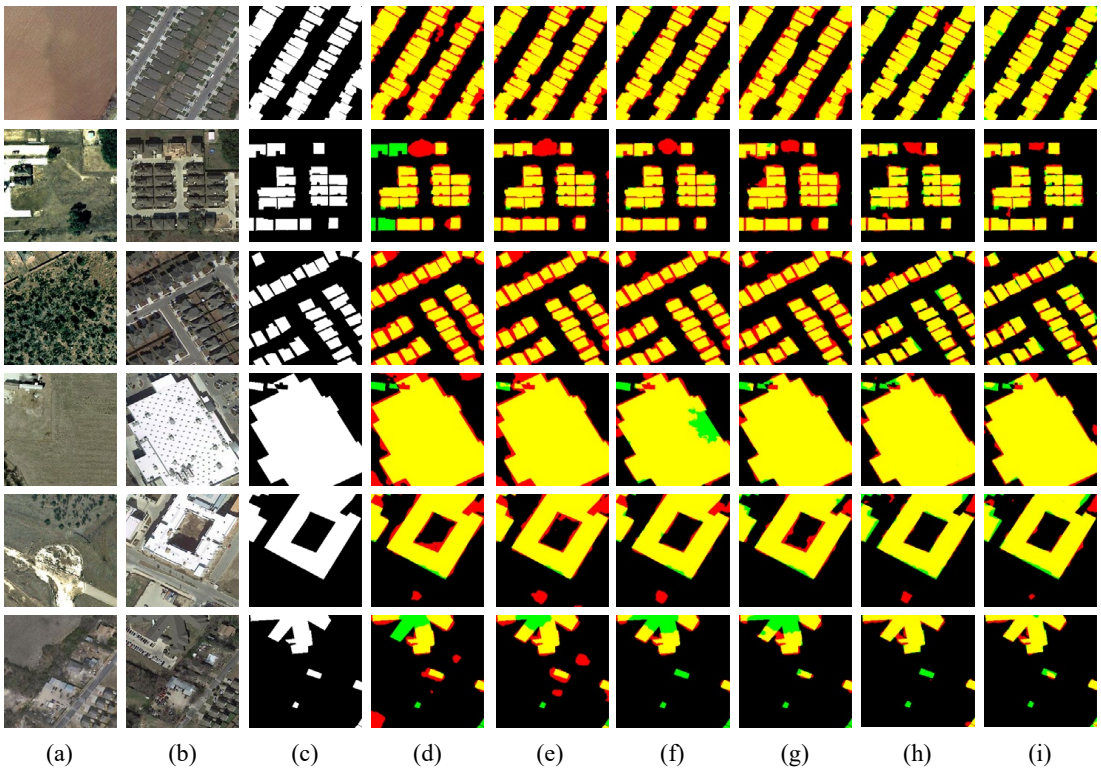

Figure 3: Visualization results of different methods using the LEVIR-CD data set. (a) T1 Image; (b) T2 Image; (c) Ground-truth; (d) FC-EF; (e) FC-Siam-conc; (f) FC-Siam-diff; (g) BIT; (h) Base Model; (i) FrNet. Yellow, black, red, and green represent TP, TN, FP, and FN, respectively.

BIT model are 83.22% and 98.06%, respectively, better than other FC-based models. This demonstrates that the tokens in spase-time can effectively capture the temporal changes and enhance the context information. The proposed FrNet achieves the highest F1 and OA among all the studied methods and has better performance than our base model. The improvements prove that the Hopfield layer helps retrieve the deep features and the shared query matrix can learn important information as part of the inputs for the decoder.

Fig. 3 illustrates change detection maps obtained by different methods, where TPs, TNs, FPs, and FNs are represented in yellow, black, red, and green, respectively. We can observe that FrNet achieves the best results among all the models. Firstly, FrNet can better distinguish small-sized changed buildings that have relatively regular shapes by reducing false alarms compared with other methods (e.g., the 1st, 2nd, and 3rd rows of Fig. 3). When the shapes of buildings are complex, our model can also preserve the boundary of the objects (e.g., the 4th, 5th, and 6th rows of Fig. 3).

5. Conclusion

Inspired by the successful application of continuous modern Hopfield for pattern retrieval, we propose a deep feature retrieved network (FrNet) for bitemporal change detection. Our Hopfield pooling block introduces a trainable weight matrix that aims to retrieve the global change of interests for high-level features and capture the discriminative representations of one period related to the other. To valuate the effectiveness of the proposed model, experiments are conducted on the LEVIR-CD data set. Our empirical evidence confirms the superiority of the proposed FrNet in comparison with other state-of-the-arts methods.

Acknowledgments

The authors would like to thank the contributors of the LEVIR-CD data set for making it publicly available, and the authors of the FC-EF, FC-Siam-conc, FC-Siam-diff, and the BIT methods for releasing their codes.

References

[1] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, et al., Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art, IEEE Geoscience and Remote Sensing Magazine 7 (2019) 6–39.

[2] Z. Li, F. Lu, H. Zhang, L. Tu, J. Li, X. Huang, C. Robinson, N. Malkin, N. Jojic, P. Ghamisi, et al., The outcome of the 2021 IEEE GRSS data fusion contest—track MSD: Multitemporal semantic change detection, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15 (2022) 1643–1655.

[3] Y. Wu, J. Li, Y. Yuan, A. Qin, Q.-G. Miao, M.-G. Gong, Commonality autoencoder: Learning common features for change detection from heterogeneous images, IEEE Transactions on Neural Networks and Learning Systems (2021).

[4] B. Bai, W. Fu, T. Lu, S. Li, Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection, IEEE Transactions on Geoscience and Remote Sensing (2021).

[5] X. Li, Z. Du, Y. Huang, Z. Tan, A deep translation (GAN) based change detection network for optical and SAR remote sensing images, ISPRS Journal of Photogrammetry and Remote Sensing 179 (2021) 14–34.

[6] F. Samadi, G. Akbarizadeh, H. Kaabi, Change detection in SAR images using deep belief network: a new training approach based on morphological images, IET Image Processing 13 (2019) 2255–2264.

[7] R. C. Daudt, B. Le Saux, A. Boulch, Fully convolutional siamese networks for change detection, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 4063–4067.

[8] X. Peng, R. Zhong, Z. Li, Q. Li, Optical remote sensing image change detection based on attention mechanism and image difference, IEEE Transactions on Geoscience and Remote Sensing 59 (2020) 7296–7307.

[9] B. Hou, Q. Liu, H. Wang, Y. Wang, From W-Net to CDGAN: Bitemporal change detection via deep learning techniques, IEEE Transactions on Geoscience and Remote Sensing 58 (2019) 1790–1802.

[10] H. Chen, C. Wu, B. Du, L. Zhang, L. Wang, Change detection in multisource VHR images via deep siamese convolutional multiple-layers recurrent neural network, IEEE Transactions on Geoscience and Remote Sensing 58 (2019) 2848–2864.

[11] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, H. Li, Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14 (2020) 1194–1206.

[12] H. Chen, Z. Qi, Z. Shi, Remote sensing image change detection with transformers, IEEE Transactions on Geoscience and Remote Sensing (2021).

[13] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, T. Adler, L. Gruber, M. Holzleitner, M. Pavlović, G. K. Sandve, et al., Hopfield networks is all you need, arXiv preprint arXiv:2008.02217 (2020).

[14] M. Demircigil, J. Heusel, M. Löwe, S. Upgang, F. Vermet, On a model of associative memory with huge storage capacity, Journal of Statistical Physics 168 (2017) 288–299. URL: https://doi.org/10.1007%2Fs10955-017-1806-y. doi:10.1007/s10955-017-1806-y.

[15] D. Krotov, J. J. Hopfield, Dense associative memory for pattern recognition, Advances in neural information processing systems 29 (2016).

[16] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proceedings of the national academy of sciences 79 (1982) 2554–2558.

[17] J. J. Hopfield, Neurons with graded response have collective computational properties like those of two-state neurons., Proceedings of the National Academy of Sciences 81 (1984) 3088–3092. URL: https://www.pnas.org/doi/pdf/10.1073/pnas.81.10.3088. doi:10.1073/pnas.81.10.3088.

[18] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, L. Gruber, M. Holzleitner, T. Adler, D. Kreil, M. K. Kopp, G. Klambauer, J. Brandstetter, S. Hochreiter, Hopfield networks is all you need, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=tL89RnzIiCd.

[19] M. Widrich, B. Schäfl, H. Ramsauer, M. Pavlović, L. Gruber, M. Holzleitner, J. Brandstetter, G. K. Sandve, V. Greiff, S. Hochreiter, G. Klambauer, Modern hopfield networks and attention for immune repertoire classification (2020). URL: https://arxiv.org/abs/2007.13505. doi:10.48550/ARXIV.2007.13505.

[20] P. Seidl, P. Renz, N. Dyubankova, P. Neves, J. Verhoeven, M. Segler, J. K. Wegner, S. Hochreiter, G. Klambauer, Modern hopfield networks for few- and zero-shot reaction template prediction, 2021. URL: https://arxiv.org/abs/2104.03279. doi:10.48550/ARXIV.2104.03279.

[21] F. Paischer, T. Adler, V. Patil, A. Bitto-Nemling, M. Holzleitner, S. Lehner, H. Eghbal-zadeh, S. Hochreiter, History compression via language models in reinforcement learning, 2022. URL: https://arxiv.org/abs/2205.12258. doi:10.48550/ARXIV.

2205.12258.

[22] M. Widrich, M. Hofmarcher, V. P. Patil, A. Bitto-Nemling, S. Hochreiter, Modern hopfield networks for return decomposition for delayed rewards, in: Deep RL Workshop NeurIPS 2021, 2021. URL: https://openreview.net/forum?id=t0PQSDcqAiy.

[23] A. Fürst, E. Rumetshofer, J. Lehner, V. Tran, F. Tang, H. Ramsauer, D. Kreil, M. Kopp, G. Klambauer, A. Bitto-Nemling, S. Hochreiter, Cloob: Modern hopfield networks with infoloob outperform clip, 2021. URL: https://arxiv.org/abs/2110.11316. doi:10.48550/ARXIV.2110.11316.

[24] B. Schäfl, L. Gruber, A. Bitto-Nemling, S. Hochreiter, Hopular: Modern hopfield networks for tabular data, 2022. URL: https://openreview.net/forum?id=3zJVXU311-Q.

[25] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[26] H. Chen, Z. Shi, A spatial-temporal attention-based method and a new dataset for remote sensing image change detection, Remote Sensing 12 (2020) 1662.