EMNIST Classification using a Hopfield Network

Tamir Shoresh, Liora Cohen, Dan Goren

# Abstract

Understanding handwriting is an essential feat for many computer vision tasks such as reading checks and understanding doctor prescriptions, and has many benefits such as: better data storage by digitizing handwritten texts, faster information retrieval and Improved accessibility for the public. This study proposes the Hopfield network for the task of classifying handwritten letters from the EMNIST letter dataset. The Hopfield network is an associative memory network that resembles the transformer attention mechanism with an additional focal depth parameter called the inverse temperature (Ramsauer et al., 2020). The performance of this model yielded promising results, with a test accuracy which topped at 91.33%, outperforming some SOTA benchmarks, including the OPIUM classifier (Cohen et al., 2017) and [OptConv+Log+Perc](https://paperswithcode.com/paper/efficient-neural-vision-systems-based-on) (Ghadekar et al., 2018). The main takeaway from this study is the use of a Hopfield layer as an attention layer, treating the inverse temperature as a focal depth, thus creating a tunable attention mechanism.

**Keywords**: Hopfield network, artificial intelligence, computer vision, character recognition, image classification, EMNIST.

# Introduction

Associative memory, or content addressable memory, is an approach in which state variables are assigned to the conceptual items (memories), as well as to the connections that obviously exist between them (Kohonen, 2012). As opposed to addressable memory, in which content is retrieved by directly searching a specific memory address, associative memory is optimized for performing searches through data. Using the stored patterns as keys and the input pattern as a query, the associative memory will retrieve the stored pattern which most closely resembles or relates to the query.

Hopfield networks are a type of associative memory first suggested by John Hopfield in 1982 (Hopfield, 1982). In this paper, the simple Hopfield network is presented, a binary associative memory that retrieves patterns by minimizing an energy function. In Ramsauer et al., 2020, the modern Hopfield network is presented, with an adaptation to the energy function that increases the network’s memory capacity. Using this new and improved energy function, Ramsauer et al. also proposes a strategy for creating continuous associative memory, as opposed to the existing binary memory, making it applicable to many fields, such as image classification. Ramsauer et al. Proceeds to show that a continuous-valued Hopfield layer can be seen as a general case of attention, adding on to it the inverse temperature , a hyper-parameter that controls the focal length of the attention.

Due to their continuous nature, Hopfield networks can be integrated as a single layer into deep learning architectures, equipping their layers with associative memories. In this paper, a new Hopfield based network architecture is proposed for the EMNIST letter classification task. The hypothesis of this paper is that the Hopfield layer will serve as a tunable attention mechanism, remembering certain features that contribute to the classification, and using to test different focal depths.

# Dataset

The NIST Special Database 19 (Cohen et al., 2017) contains handwritten digits and characters collected from over 500 writers. The dataset contains binary scans of the handwriting sample collection forms, and individually segmented and labelled characters which were extracted from the forms. The characters include numerical digits and both uppercase and lowercase letters. From NIST, several smaller datasets can be derived, such as MNIST, which contains hand-written numbers, and EMNIST which contains hand-written letters. Both these datasets have become popular benchmarks learning, classification and computer vision systems.

The EMNIST dataset consists of 145,600 images, each 28X28 pixels. The dataset is made up of 26 balanced classes, one for every letter in the English alphabet. Notice the dataset contains lower-case letters as well as upper-case letters, meaning corresponding lower-case and upper-case letters belong to the same class.

# Prior Work

**Simple Hopfield Networks**

The simple Hopfield network was first presented by John Hopfield in 1982 (Hopfield, 1982). In his paper, the simple Hopfield network is described as a fully connected network of binary neurons, i.e., . Storing patterns in this network is done by calculating the sum of outer products of these patterns. This sum encodes within it the stored patterns, and corresponds to the weight matrix :

Equation

When presented with a new state pattern , also called a query, the corresponding stored pattern can be retrieved either in a synchronous or an asynchronous manner. When using the synchronous update strategy, a stored pattern can be retrieved by repeatedly applying the following update rule:

Equation

Where is a bias vector, which can be interpreted as an activation threshold for every neuron. The asynchronous update rule performs this update only for one component of at a time, then randomly uniformed selects the next component for update, until finally updating all components. Convergence is said to be reached when . The asynchronous version of the update rule minimizes the energy function E:

Equation

The energy function is designed such that the stored patterns are the local minima of the function, meaning that converging to any local minima will retrieve one of said patterns Mathematically, this corresponds to the equation:

Equation

In [8], it is proven that the storage capacity for retrieval of patterns with a small error ( for each component) is:

Equation

This means the memory capacity of a simple Hopfield network grows linearly with respect to its size.

**Modern Hopfield Networks**

The storage capacity is a crucial characteristic of any memory system. In a paper by Ramsauer et al., 2020, a new and improved energy function is introduced to the simple Hopfield network. This new energy function allows a higher storage capacity. The resulting network is referred to as a modern Hopfield Network.

The new energy function is defined as:

Equation

Where is referred to as the interaction function.

As shown by Demircigil et al., 2017, using an exponential interaction function of the shape results in a memory capacity that grows exponentially with respect to its size. The Energy function can be described as:

Equation

where is the data matrix of stored patterns. Allowing the same small error as in the simple Hopfield network, the storage capacity of the modern Hopfield network can be estimated by:

Equation

The main effect of a wisely chosen energy function is the decorrelation of the stored patterns, allowing the network to retain more patterns with the same low error probability. Essentially, the interaction function pulls apart close patterns, allowing the network to distinguish between more strongly correlated patterns.

**Continuous-Valued Patterns and States**

Paper [7] proves that the modern Hopfield energy function can be generalized for the case of continuous-valued patterns. The corresponding energy function is defined as:

Equation

Constructed from N continuous stored patterns by the matrix , where is the largest norm of all stored patterns and is the inverse temperature which will be discussed further on.

This energy function allows deriving an update rule for a state pattern by the Concave-Convex-Procedure (CCCP) [6]. The update rule for a state pattern therefore reads:

Equation

Applying the Concave-Convex-Procedure to obtain the update rule guarantees the monotonical decrease of the energy function. The most important properties of our new energy function are:

1. Global convergence to a local minimum
2. Exponential storage capacity
3. Convergence after one update step.

The new continuous energy function allows extending our example to continuous patterns. When sufficiently different from each other, these patterns can be retrieved perfectly. However, similar stored patterns may cause the energy function to converge to a metastable state, which is a local minimum created by the superposition of many close local minima. The learning dynamics can be controlled by the inversed temperature . High values of correspond to a low temperature and mean that the attraction basins of the individual patterns remain separated, making it unlikely that metastable states appear. For low values of the formation of metastable states becomes more likely.

**Hopfield as a General Case of Attention**

It can be shown that the continuous-valued Hopfield is a general case of transformer attention from a paper by Vaswani et al., 2017. By mapping the stored patterns matrix and the state pattern matrix to a key matrix and a query matrix , Ramsauer et al., 2020 shows that the update rule can be formulated as:

Equation

Where is the inverse temperature and is a weight matrix referred to as the projection matrix. For a value of , this is exactly transformer self-attention. In this case, can be seen as a scaling parameter controlling the focus of the attention mechanism. Large values of correspond to a more focused attention, able to distinguish patterns very well by creating many separated local minima in the energy function. Small values of correspond to a less focused attention, able to generalize and cluster similar patterns by creating metastable states in the energy function, each containing several similar patterns.

# Methods

The hypothesis of this study is that a Hopfield layer can be used as a tunable attention layer in respect to . This hypothesis is tested on the EMNIST letter classification task.

The main challenge regarding this dataset is the classification of lower-case and upper-case letters to the same class. The EMNIST letter dataset contains 26 classes, one for each letter of the alphabet, meaning lower-case and upper-case letters belong to the same class despite of the obvious difference in their shape.

This study proposes the use of a Hopfield layer, using a large valued to better separate lower-case and upper-case letters, while also taking advantage of the Hopfield layer’s inherent property of equipping the layers with associative memories.

The network architecture proposed for solving this task resemble the transformer encoder presented by Vaswani et al., 2017. The network consists of an input layer of 784 neurons, corresponding to the number of pixels in the EMNIST images. Next is a Hopfield layer with batch normalization, followed by a feed-forward layer with batch normalization and leaky ReLU activation function, ending in a fully connected output layer with 26 neurons. Architecture schematics can be seen in Figure 1. The loss function chosen for this task is cross-entropy.

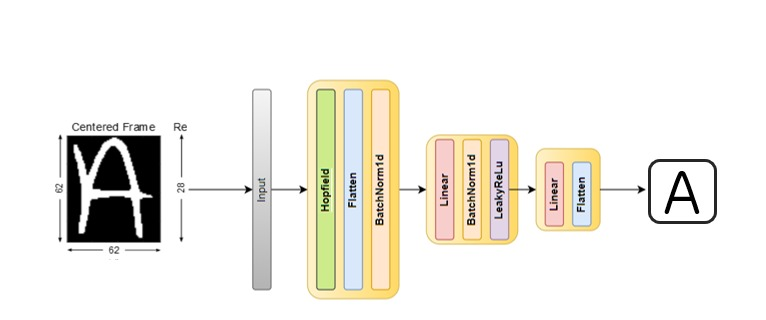


Figure – Model architecture

To reduce the chance of overfitting over many epochs, augmentations were introduced to each image, including a random (uniform) rotation of up to and random (uniform) horizontal and vertical shifts of up to of the image axis.

With the model at hand, some of the architecture parameters were optimized. After testing different sizes of the feed-forward layer, ranging from 60 to 4000, the optimal size was found to be 3000 neurons. The number of parallel Hopfield layers was also tested, ranging from 1 to 198, showing optimal results with one head alone. Finally, two optimizers were tested, AdamW and RMSprop, with RMSprop showing 1-2% higher accuracy across all experiments. In addition, after testing different amounts of epochs, the training process started to diverge after roughly 360 epochs. The final number of epochs was chosen to be 350. The cross-entropy loss function over 400 epochs is shown in Figure 2.

A picture containing histogram

Description automatically generated

Figure – Cross Entropy Loss over 400 epochs

With the architecture set in place, two hyper-parameters of the model were optimized. The learning rate was tested with values of , with showing more promising results. The second hyper-parameter to be optimized is the inverse temperature . The values chosen were , with corresponding to the transformer attention mechanism. As expected, higher values of helped the algorithm distinguish better between patterns, with being the optimal value.

As a final experiment, a dropout layer was added after the feed-forward layer, with dropout probabilities ranging . In all these experiments, the dropout layer had a negative impact on the training process, making the loss diverge at a much sooner epoch. For our results, the dropout layer was removed.

# Results

As explained in the Methods, the optimization was done in 3 parts, each optimizing over a grid of potential hyper-parameters. A held-out validation set was used to evaluate the performance of the model at each iteration. The optimal values for the model are shown in Table 1.

Table – Hyperparameter optimized.

|  |  |  |
| --- | --- | --- |
|  | Parameter | Optimized Value |
| 1 | Inverse Temperature | 3.6 |
| 2 | Size of feed-forward Layer | 3000 |
| 3 | # Of Parallel Hopfield Layers | 1 |
| 4 | Optimizer | RMSprop |
| 5 | # Of epochs | 350 |
| 6 | Learning Rate η |  |

The accuracy of the training and validation sets over all 350 epochs is shown in Figure 3. The results show a higher accuracy on the validation set than on the training set. A possible cause of this phenomenon are the augmentations introduced to the training data. It is very likely that some of the images were slightly corrupted by the augmentation making it harder for the model to distinguish between letters. The validation set on the other hand had no augmentations.

A picture containing shape

Description automatically generated

Figure – Accuracy over 350 epochs

After training the model the test set is evaluated, showing an accuracy of 91.33% and mean cross-entropy loss of 0.2714. The confusion matrix of the true and predicted labels is shown in Figure 4.

Chart, scatter chart

Description automatically generated

Figure – Confusion Matrix

# Discussion

In this study a new architecture is proposed, incorporating a Hopfield layer as an essential part of a letter classification model.

One of the main goals of this study was to test the effects of the inverse temperature , treating it as a focal depth for the attention layer. The hypothesis proposed was that a larger value of would help distinguish better between letters. This hypothesis was proven true using , which is 100 times larger that the that corresponds to the transformer attention mechanism. Additionally, the optimization proves that cannot be enlarged indefinitely, given that led to a slightly lower accuracy than .

Looking closely at the confusion matrix in Figure 4, it can be seen that most errors made by the model are reasonable and could have also been made by the human eye, errors such as mixing between “L” and “I”, “G” and “Q”, and some others. It is important to keep in mind that lower-case and upper-case characters of the same letter belong to the same class, meaning these errors might have been between any of the above.

In the future, this phenomenon can be treated by creating stand-alone binary classifiers for specific letter pairs that show higher errors. These binary classifiers may add additional evidence that will enrich the decision process. Another suggested solution is to incorporate a convolutional neural network before the Hopfield layer, thus using tunable attention on a feature vector of the image instead of the original image. Other options exist for improving the performance of the proposed classifier, such as introducing other types of augmentations to the dataset or even enlarging the dataset, but those are beyond the scope of this study.

# References

Cohen, G., Afshar, S., Tapson, J., & Van Schaik, A. (2017). EMNIST: Extending MNIST to handwritten letters. *2017 International Joint Conference on Neural Networks (IJCNN)*, 2921–2926.

Demircigil, M., Heusel, J., Löwe, M., Upgang, S., & Vermet, F. (2017). On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, *168*, 288–299.

Ghadekar, P., Ingole, S., & Sonone, D. (2018). Handwritten digit and letter recognition using hybrid dwt-dct with knn and svm classifier. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1–6.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558.

Kohonen, T. (2012). *Self-organization and associative memory* (Vol. 8). Springer Science & Business Media.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., & others. (2020). Hopfield networks is all you need. *ArXiv Preprint ArXiv:2008.02217*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \Lukasz, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.