ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Balancing of Neural Contributions for Multi-modal Hidden State Association

Christian Emmerich, R. Felix Reinhart and Jochen J. Steil

Research Institute for Cognition and Robotics, Bielefeld University
Universitätsstr. 25, 33615 Bielefed - Germany
{cemmeric, freinhar, jsteil}@cor-lab.uni-bielefeld.de

Abstract. We generalize the formulation of associative reservoir computing networks to multiple input modalities and demonstrate applications in image and audio processing scenarios. Robust association with reservoir networks requires to cope with potential error amplification of output feedback dynamics and to handle differently sized input and output modalities. We propose a dendritic neuron model in combination with a modified reservoir regularization technique to address both issues.

1 Introduction

In the recent years, the idea of combining a nonlinear and high-dimensional random projection with a linear read-out layer has become popular under the notion of Reservoir Computing (RC, [1]) and Extreme Learning Machines (ELM, [2]). Standard RC networks and ELMs implement input-output mappings, where the input-driven representation in the hidden layer is utilized for a simple, linear read-out mapping. Associative reservoir computing networks [3, 2] extend this feed-forward network configuration to bidirectional information retrieval systems. In this paper, we generalize these previous ideas to a multi-modal formulation under the notion of Hidden State Association (HSA), where the hidden state holds a shared representation of multiple input modalities. The scheme can be applied in case of attractor- as well as transient-based computation, e.g. association of static patterns or sequence transduction.

Association in these networks, however, requires feedback connections from output neurons to the hidden layer which can potentially lead to error amplification of the retrieval dynamics [4]. Several techniques have been proposed to cope with output feedback dynamics in RC, e.g. [1, 5, 4]. In the context of bidirectional association, it is moreover crucial to control or to balance the contributions of inputs to the hidden state. In particular for very heterogeneous input modalities with different dimensions or diverse energy spectra, it is important that the network can still be driven by either input or output neurons to implement an effective bidirectional mapping. For these reasons, an explicit and also parameterized weighting of each modality is desirable.

In this paper, we complement the reservoir regularization approach introduced in [4] with an additional concept to balance contributions of several driving input modalities. The balancing is integrated in the regularization process such that both can be accomplished in one step. We apply the proposed method to image and audio processing scenarios.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

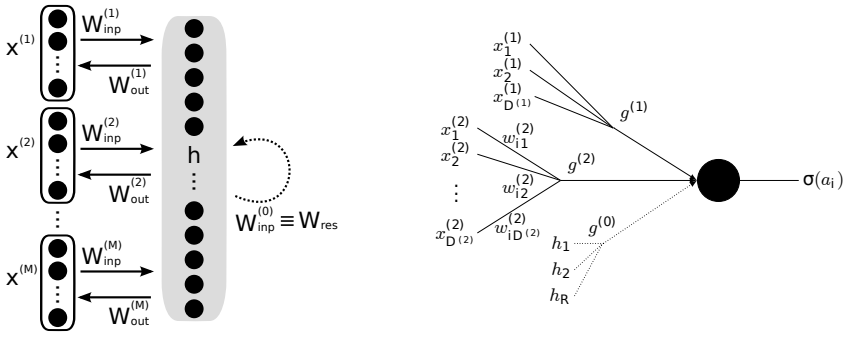

Fig. 1: Left: HSA network with multiple input modalities $\mathbf{x}^{(m)}$. Right: Formal neuron model with dendritic tree structure and weighting factors $g^{(m)}$.

2 Hidden State Association (HSA)

We consider network architectures as depicted in Fig. 1 (left) comprising a hidden layer of nonlinear neurons $\mathbf{h} = \sigma(\mathbf{a}) \in \mathbb{R}^R$, which is driven by M input modalities $\mathbf{x}^{(m)} \in \mathbb{R}^{D^{(m)}}$ with random connection strength $\mathbf{W}_{inp}^{(m)}$. Thereby, the nonlinear activation functions $\sigma(\cdot)$ are applied component-wise. Throughout this contribution we use $\sigma = \tanh$. The network activities $\mathbf{a}$ and reconstructed inputs $\hat{\mathbf{x}}^{(m)}$ are computed according to

$$\mathbf{a}(k) = \mathbf{W}_{inp}^{(0)}\mathbf{h}(k-1) + \sum_{m=1}^{M} \mathbf{W}_{inp}^{(m)}\mathbf{x}^{(m)}(k) \tag{1}$$

$$\hat{\mathbf{x}}^{(m)}(k) = \mathbf{W}_{out}^{(m)}\mathbf{h}(k) \qquad \forall m = 1, \ldots, M. \tag{2}$$

Note that this formulation also covers recurrent hidden layers, i.e. reservoirs, by optionally setting $\mathbf{W}_{inp}^{(0)} \neq \mathbf{0}$ and $\mathbf{x}^{(0)}(k) = \mathbf{h}(k-1)$. During network exploitation, the network is driven by given modalities $\mathbf{x}^{(m)}$ while the modality to be retrieved is iterated in a closed loop by feeding estimated values $\hat{\mathbf{x}}^{(n)}$ back into the hidden layer. In case of static pattern association, these feedback dynamics are iterated until convergence of the hidden state for each input sample.

Learning is restricted to the read-out connections $\mathbf{W}_{out}^{(m)}$. First, the network is teacher-forced by the training samples and its hidden states are collected in the matrix $\mathbf{H}$. Then, the read-out weights are computed by linear regression

$$(\mathbf{W}_{out}^{(m)})^T = \left(\mathbf{H}^T\mathbf{H} + \alpha^{(m)}\mathbb{1}\right)^{-1}\mathbf{H}^T\mathbf{X}^{(m)} \qquad \forall m = 1, \ldots, M, \tag{3}$$

where $\mathbf{X}^{(m)}$ are the training samples $\mathbf{x}^{(m)}$ collected in a matrix and $\alpha^{(m)}$ is a regularization parameter for each modality, respectively.

3 Weight Regularization with Activity Distribution

We control the contribution of each modality to the hidden state activity $\mathbf{a}$ by adopting the reservoir regularization approach from [4]. The idea is to re-

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

compute the weights $\mathbf{W}_{inp}^{(m)}$ such that they have a smaller norm but still implement the same input-to-state mapping. This reduces the overall gain of the feedback loops during network exploitation and gains stability of the system [4]. In addition, each modality shall contribute with a prescribed strength to the hidden state activity. We define the target activity to split into a weighted sum

$$\mathbf{a}(k) = \sum_{m=0}^{M} g^{(m)}\ \mathbf{a}(k) \qquad \text{with} \qquad \sum_{m=0}^{M} g^{(m)} = 1.$$

The regularized network weights per modality are then given by

$$(\hat{\mathbf{W}}_{inp}^{(m)})^T = g^{(m)} \left((\mathbf{X}^{(m)})^T \mathbf{X}^{(m)} + \beta^{(m)} \mathbb{1} \right)^{-1} (\mathbf{X}^{(m)})^T \mathbf{A} \qquad \forall m = 0, \ldots, M.$$

The contribution of each modality m to the network's activation can now be controlled by $g^{(m)}$, e.g. $g^{(m)} = 1/(M+1)$ for a uniform balancing. Regularization of the corresponding input weights is parameterized by $\beta^{(m)}$. This weighted contribution of input modalities models the dendritic tree structure of neurons (cf. Fig. 1 (right)).

4 Balancing Contributions for Robust Association

In the following, the scalability of the proposed approach is demonstrated in two dimensionality reducing scenarios. The idea is to associate high dimensional 'raw' data with a low-dimensional representation, thereby representing the reduction and the reconstruction mapping in a single network. Balancing contributions of the two input modalities, namely the original data $\mathbf{x}^{(1)}$ and its low-dimensional representation $\mathbf{x}^{(2)}$, is particularly important in order to compensate the strong discrepancy in size of the input modalities. In our experiments we choose a uniform weighting $g^{(1)} = g^{(2)} = 0.5$, where $\mathbf{W}_{inp}^{(0)} \equiv \mathbf{0}$.

Learning to embed and reconstruct handwritten digits
First, HSA networks are trained to embed handwritten digits into a plane and to reconstruct digits from their low-dimensional representation. This contributes to an upcoming thread of ideas that considers neighborhood-preserving embedding functions [6, 7]. This approach is appealing because embedding new data points after learning of the function does not require to rerun the embedding algorithm with the additional data points. Additionally, the reconstruction of data in the original space by "navigating" through the embedding space is enabled.

We consider images of handwritten digits from the MNIST data set [8]. We embed the data into the two-dimensional plane by t-Stochastic Neighbor Embedding (t-SNE) [6] in order to generate training data pairs $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ with a two-dimensional embedding vector $\mathbf{x}^{(2)}$ for each image $\mathbf{x}^{(1)} \in \mathbb{R}^{784}$.

We train 100 independently initialized HSA networks with $R = 400$ hidden neurons to learn the embedding projection and the inverse mapping. The input weights $\mathbf{W}_{inp}^{(m)}$ are initialized in $[-1/D^{(m)}, 1/D^{(m)}]$, where $D^{(m)}$ denotes the dimension of the respective input modality. For training, a randomly chosen subset of the data comprising 75% of the 2000 embedded images are chosen.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

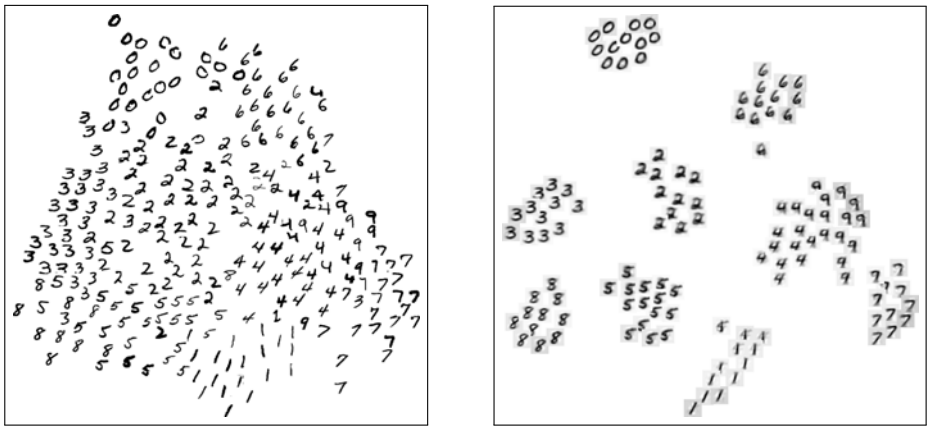

Fig. 2: Left: Embedding and generation of digits by the trained associative network: Input images $\mathbf{x}^{(1)}$ are plotted at estimated positions $\hat{\mathbf{x}}^{(2)}$. Right: Reconstructed images $\hat{\mathbf{x}}^{(1)}$ are plotted at embedding positions $\mathbf{x}^{(1)}$ of t-SNE.

The remaining 500 samples are used to evaluate the generalization abilities of the learned mapping. Read-out learning is conducted with $\alpha^{(1)} = \alpha^{(2)} = 0.1$ and the input weights are regularized by $\beta^{(1)} = 0.1$ and $\beta^{(2)} = 0.01$, which all are found by manual tuning.

We first consider the embedding projection from the image space to the plane. The estimated projections for a subset of the input images are shown in Fig. 2 (left). The embedding shows a decent and smooth structure with similar inputs at similar positions in the embedding space. To access the network performance quantitatively, we compute the dimension-normalized mean square error

$$NMSE^{(m)} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{D^{(m)}} ||\mathbf{x}_k^{(m)} - \hat{\mathbf{x}}_k^{(m)}||^2, \tag{4}$$

where K is the number of samples. In this scenario $\mathbf{x}_k^{(m)}$ with $m = 2$ are the embeddings according to t-SNE and $\hat{\mathbf{x}}_k^{(m)}$ the approximated embedding by the HSA network. The error statistics for training and test sets of the HSA embeddings with respect to the embedding generated by t-SNE are displayed in Tab. 1 indicating robust learning of the output-feedback-driven network dynamics.

The inverse mapping implemented by the network maps two-dimensional coordinates $\mathbf{x}^{(2)}$ to images $\hat{\mathbf{x}}^{(1)}$. Fig. 2 (right) shows such reconstructions of images for a subset of the embedded data points. The characteristics of all ten digits are reproduced well and variations of the digit shape and orientation are also associated with neighboring positions in the embedding space. The average reconstruction errors in Tab. 1 show a similar approximation capability of the networks for the backward data reconstruction in comparison to the forward embedding on the test sets.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Scenario	Set	Embedding	Reconstruction
MNIST	Train	$0.012 \pm 3.6 \cdot 10^{-4}$	$0.044 \pm 7.1 \cdot 10^{-4}$
	Test	$0.047 \pm 4.0 \cdot 10^{-3}$	$0.043 \pm 2.5 \cdot 10^{-4}$
MFCC	Train	$0.061 \pm 1.9 \cdot 10^{-3}$	$5.84 \cdot 10^{-4} \pm 9 \cdot 10^{-7}$
	Test	$0.070 \pm 2.8 \cdot 10^{-3}$	$6.81 \cdot 10^{-4} \pm 1 \cdot 10^{-6}$

Tab. 1: Normalized mean square errors (4) and standard deviations.

Learning to compress and decompress speech signals
In a second scenario, we apply HSA to encode and decode speech signals to and from a compact representation, implementing a continuous sequence transduction. For speech recognition, most state-of-the-art methods utilize the Mel Frequency Cepstral Coefficients (MFCCs). However, for speech generation and uncompressing compressed speech signals, it is desirable to also reconstruct a full power spectrum from its MFCC representation. The HSA approach enables this bi-directional sequence transduction in one network.

We process the German utterance "Ich möchte heute von München nach Frankfurt fahren." spoken by 58 female and male speakers at 16 kHz [9]. We calculate the normalized power spectrum and the corresponding 13 MFCCs each 10 ms with a window size of 25 ms, i.e. $\mathbf{x}^{(1)} \in \mathbb{R}^{256}$ and $\mathbf{x}^{(2)} \in \mathbb{R}^{13}$. We use networks with 500 hidden neurons. The input weights are regularized with $\beta^{(1)} = \beta^{(2)} = 0.1$ and the output weights with $\alpha^{(1)} = 0.1$ and $\alpha^{(2)} = 0.01$. Again, these parameters are tuned manually. All other learning and network parameters are set as above. We train 10 independently initialized networks in a 5-fold repeated random cross-validation. The inter-subject generalization ability of the networks is evaluated by collecting 44 ($\approx 75\%$) randomly chosen speakers for training and 14 for testing in each fold.

Fig. 3 shows the qualitative feature extraction abilities of the approach, where the extracted MFFCs $\hat{\mathbf{x}}^{(2)}$ (top right), estimated on the basis of the original power spectrogram $\mathbf{x}^{(1)}$ (top center) of an exemplary speech frame in a test utterance (top left), are compared to the target features $\mathbf{x}^{(2)}$ (bottom right). Both the characteristics of the 12 coefficients and the value of the log-energy of the frame is well approximated.

The quality of the inverse mapping, i.e. to reconstruct a full power spectrum from the MFCCs, is also revealed by Fig. 3. Given the low-dimensional MFCC representation $\mathbf{x}^{(2)}$ of a frame (bottom right), the network is able to estimate a power spectrogram $\hat{\mathbf{x}}^{(1)}$ (bottom center) qualitatively covering the characteristics of the original spectrogram (top center). Mapping the reconstructed spectra back to the time domain reconstructs a full utterance (bottom left)) very similar to the original speech signal. In this way, entire speech signals can be reconstructed from a very low-dimensional representation.

The entire evaluation results are collected in Tab. 1 and show competitive dimension-normalized mean square errors for both sequence transduction directions which underlines the balancing effect of the proposed method.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

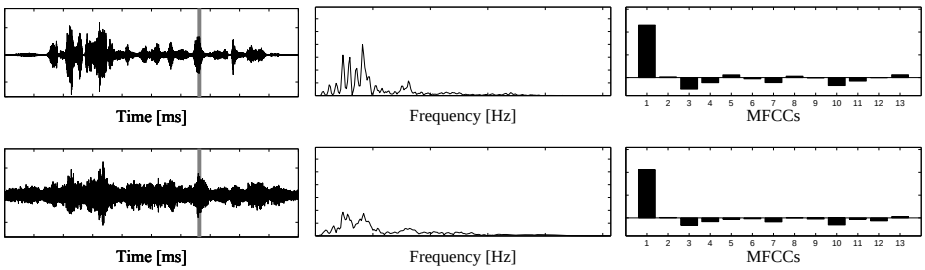

Fig. 3: Top left: original utterance with a highlighted frame. Top center: power spectrum of that frame. Top right: extracted MFFCs by the HSA network. Bottom right: target MFCCs of the highlighted frame. Bottom center: reconstruction of the frame's power spectrum by the network. Bottom left: generated speech signal based on the reconstructed power spectrograms.

5 Conclusion

The applications show that association of unequally sized modalities is possible with HSA and the balancing of contributions. We conclude from the small error variances in Tab. 1 that bidirectional mappings can be robustly learned in HSA networks. That is, error amplification of the feedback dynamics is prevented by applying regularization of the hidden layer in combination with the balanced contributions of the input modalities to the network activity. Balancing of contributions to the hidden states assures that the regularized model is still responsive to inputs or outputs, i.e. can be driven by the respective modality, which is of major importance for bidirectional association.

References

[1] H. Jaeger, M. Lukosevicius, D. Popovici, and U. Siewert. Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Networks*, 20:335–352, 2007.

[2] Reinhart and Steil. Neural learning and dynamical selection of redundant solutions for inverse kinematic control. In *Proc. Int. Conf. on Humanoid Robots*, pages 564–569, 2011.

[3] Antonelo, Schrauwen, and Campenhout. Generative modeling of autonomous robots and their environments using reservoir computing. *Neural Processing Letters*, 26:233–249, 2007.

[4] R. Felix Reinhart and Jochen J. Steil. Reservoir regularization stabilizes learning of Echo State Networks with output feedback. In *Proc. ESANN*, pages 59–64, 2011.

[5] F. wyffels, B. Schrauwen, and D. Stroobandt. Stable Output Feedback in Reservoir Computing Using Ridge Regression. In *Proc. ICANN*, pages 808–817. 2008.

[6] Laurens van der Maaten. Learning a parametric embedding by preserving local structure. In *Proc. AISTATS*, pages 384–391, 2009.

[7] Kerstin Bunte, Michael Biehl, and Barbara Hammer. Supervised dimension reduction mappings. In *Proc. ESANN*, pages 281–286, 2011.

[8] Yann Lecun and Corinna Cortes. MNIST, 1998. `http://yann.lecun.com/exdb/mnist/`.

[9] G.A. Fink and T. Plötz. Integrating speaker identification and learning with adaptive speech recognition. In *Proc. ODYS*, pages 185–192, 2004.