XXIX Congresso de Iniciação Científica UNICAMP 2021

Estudo e Aplicação de Redes Neurais Profundas com Estados de Eco

Aluno: Gabriel Araujo Pinheiro
Orientador: Prof. Levy Boccato
Faculdade de Engenharia Elétrica e de Computação

Palavras-chave: redes neurais, aprendizado profundo, redes neurais com estados de eco, equalização de canais

1 Introdução

As redes neurais artificiais têm a capacidade de resolver problemas utilizando como inspiração elementos do sistema nervoso. Tipicamente, sua estrutura é composta por elementos de processamento de informação, denominados neurônios artificiais, que estão massivamente interligados [1]. Embora a operação de cada neurônio seja relativamente simples, uma rede neural é capaz de sintetizar complexos mapeamentos entrada-saída compondo vários neurônios em camadas.

Dentro desse contexto, é possível separar as variadas propostas de redes neurais em duas grandes categorias: as redes neurais *feedforward*, nas quais a informação flui através da rede em um único sentido, da entrada para a saída, e as redes recorrentes (RNNs, do inglês *recurrent neural networks*) [1,2], que contêm conexões que retroalimentam as ativações de neurônios de volta para a rede, de modo que o sistema passa a ter comportamento dinâmico (e, também, memória).

Na última década, o uso de arquiteturas profundas, compostas por cadeias mais longas de múltiplas camadas de processamento, se consolidou como uma opção poderosa para o tratamento de problemas do mundo real, como no processamento de sinais de áudio, de imagens e de linguagem natural [3].

Os modelos profundos possuem a habilidade de aprender representações para os dados em diferentes níveis de abstração. Ao longo das sucessivas camadas da rede é criada uma hierarquia de representações, na qual atributos de mais alto nível são construídos por meio da agregação dos atributos mais elementares identificados nas camadas iniciais [3].

Este tipo de arquitetura também foi transportado para o domínio das RNNs, e tem se mostrado interessante, pois possibilita um processamento hierárquico de dados temporais. Em outras palavras, é possível o surgimento de dinâmicas internas associadas a diferentes escalas de tempo no domínio de interesse, o que pode ser vantajoso para a modelagem de comportamentos da entrada e para a aproximação da resposta desejada.

Esta ideia foi trazida ao contexto das redes neurais com estados de eco (ESNs, do inglês *echo state networks*) [2,4] em 2017, dando origem à ESN profunda (ou DESN, de *deep echo state network*). O diferencial das DESNs reside na possibilidade de explorar, em certa medida, os benefícios de um modelo recorrente sem ter que realizar o processo de treinamento de toda a estrutura, evitando assim o custo computacional e os riscos inerentes à otimização [2,5,6]. Além disso, as DESNs podem explorar a informação temporal de uma forma mais efetiva devido às múltiplas camadas recorrentes.

Este trabalho teve como objetivo estudar as DESNs e realizar sua aplicação ao problema de equalização de canais de comunicação [2,7], no qual é fundamental que a estrutura temporal dos sinais envolvidos seja bem aproveitada e, também, que o modelo utilizado para cancelar as distorções do canal tenha a flexibilidade de criar mapeamentos não-lineares.

2 Redes Neurais com Estados de Eco

Uma DESN é caracterizada pela presença de uma hierarquia de camadas recorrentes (ou reservatórios dinâmicos) empilhadas, na qual a saída de uma camada atua como a entrada da próxima camada, conforme podemos observar na Figura 1.

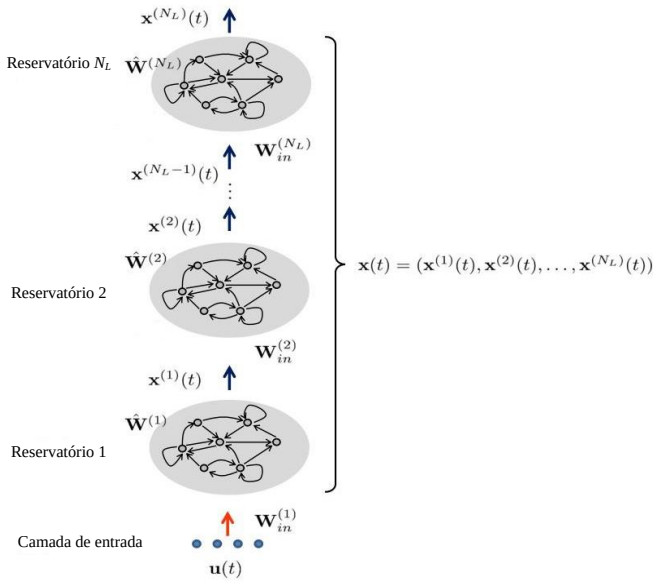

Figura 1 – Estrutura do Reservatório de uma DESN. Adaptada de [4].

Da Figura 1, pode-se observar que N_L indica o número de reservatórios, $\mathbf{u}(t)$ representa a entrada externa no tempo t, enquanto $\mathbf{x}^{(l)}(t)$ é o estado do l-ésimo reservatório no tempo t. A composição dos estados de todas as camadas recorrentes dá origem ao vetor $\mathbf{x}(t) = (\mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(N_L)}(t))$, que representa o estado global da rede e constitui a entrada da camada de saída (*readout*). Assim, a função de transição do estado da DESN, utilizando o modelo de vazamento (*leaky integrator*) nos reservatórios, pode ser expressa da seguinte maneira:

$$\mathbf{x}^{(l)}(t) = \left(1 - a^{(l)}\right)\mathbf{x}^{(l)}(t-1) + a^{(l)}\tanh\left(\mathbf{W}_{in}^{(l)}\mathbf{i}^{(l)}(t) + \widehat{\mathbf{W}}^{(l)}\mathbf{x}^{(l)}(t-1)\right)$$

Aqui, para cada camada $l = 1, \ldots, N_L$, $\mathbf{W}_{in}^{(l)}$ representa a matriz de pesos de entrada, $a^{(l)} \in [0,1]$ é a taxa de vazamento e $\widehat{\mathbf{W}}^{(l)}$ denota a matriz com os pesos das conexões recorrentes.

É pertinente ressaltar que enquanto a primeira camada deve reagir à entrada externa ($\mathbf{u}(t)$), as demais camadas são alimentadas pelo vetor de estados da camada anterior. Por isso, adota-se na expressão do vetor de estados a variável $\mathbf{i}^{(l)}(t)$, a qual indica a entrada correta da camada l:

$$\mathbf{i}^{(l)}(t) = \begin{cases} \mathbf{u}(t), \text{se } l = 1 \\ \mathbf{x}^{(l-1)}(t), \text{se } l > 1 \end{cases}$$

Em relação ao cálculo da saída da DESN, o procedimento é similar àquele explorado na ESN clássica, ou seja, todas as saídas das unidades de reservatório são combinadas linearmente. Levando em conta a organização hierárquica da arquitetura profunda, podemos escrever que a saída da rede é dada por:

$$y(t) = \mathbf{W}_{out}[\mathbf{x}^{(1)}(t)\ \mathbf{x}^{(2)}(t) \ldots\ \mathbf{x}^{(N_L)}(t)]^T$$

onde, assim como no caso clássico, $\mathbf{W}_{out}$ é a matriz de pesos de saída da DESN.

O projeto de uma DESN segue o mesmo esquema consagrado para uma ESN clássica: os parâmetros dos reservatórios são definidos de forma antecipada e aleatória, de modo que apenas os pesos da camada de saída são ajustados de maneira supervisionada para minimizar a função custo (e.g., erro quadrático médio), o que traz uma grande simplificação ao treinamento do modelo.

Para isto, é necessário assegurar que cada reservatório satisfaça a propriedade de estados de eco, o que nos leva à seguinte condição para a rede como um todo [8]:

$$\max_{1 \le l \le N_L} \rho\left(\left(1 - a^{(l)}\right)\mathbf{I} + a^{(l)}\widehat{\mathbf{W}}^{(l)}\right) = \max_{1 \le l \le N_L} \rho_l < 1,$$

onde $\rho(\cdot)$ denota o operador que calcula o raio espectral (i.e., o maior autovalor em módulo). Ou seja, a condição imposta é a de que todos os reservatórios apresentem um raio espectral menor do que um em relação à matriz $\left(1 - a^{(l)}\right)\mathbf{I} + a^{(l)}\widehat{\mathbf{W}}^{(l)}$.

A construção de uma DESN passa, portanto, pela especificação dos seguintes hiperparâmetros: (1) o número N_L de reservatórios; (2) o número de neurônios em cada reservatório; (3) o fator de esquecimento $a^{(l)}$ e (4) o raio espectral ρ_l para cada reservatório. Ademais, é preciso definir estratégias para a criação das matrizes de pesos de entrada $\left(\mathbf{W}_{in}^{(l)}\right)$ e das matrizes de pesos recorrentes $\left(\widehat{\mathbf{W}}^{(l)}\right)$. No primeiro caso, os elementos são tipicamente tomados a partir de uma distribuição uniforme em um intervalo $[-w_{i,\max}, w_{i,\max}]$, sendo a extensão deste intervalo outro hiperparâmetro do modelo. No segundo caso, uma opção indicada em [9] consiste em inicializar os elementos de $\widehat{\mathbf{W}}^{(l)}$ uniformemente no intervalo $[-1,1]$ e, então, re-escalar a matriz para que a condição referente à ESP seja satisfeita.

3 Equalização de canais

Sistemas de comunicação são projetados para viabilizar a transmissão de informações de interesse a partir de uma fonte para um receptor. Durante a transmissão, algumas distorções são introduzidas pelo canal utilizado (ar, fibra óptica), de maneira que o sinal que chega ao receptor precisa ser processado. Uma estratégia clássica para tentar recuperar a informação da fonte no receptor consiste em construir um dispositivo, denominado equalizador, cujo papel é o de desfazer a ação do canal, i.e., remover as distorções e entregar uma estimativa confiável do conteúdo transmitido. Este desafio caracteriza o problema de equalização de canais de comunicação [2,7].

Um modelo de canal classicamente utilizado, em especial no contexto de sistemas de comunicações digitais, é o de um sistema linear e invariante com o tempo, cuja resposta ao impulso tem duração finita (FIR, do inglês *finite impulse response*). Neste cenário, a distorção introduzida pelo canal é conhecida como interferência inter-simbólica, e a saída do canal é dada por uma combinação linear de amostras atrasadas (ou símbolos) da fonte:

$$s'(n) = h_0^* s(n) + h_1^* s(n-1) + \cdots + h_{D-1}^* s(n-D+1)$$

onde os parâmetros h_i são denominados de coeficientes do canal, $(\cdot)^*$ denota o complexo conjugado e D denota o comprimento da resposta ao impulso do canal. Adicionalmente, considera-se que o sinal transmitido também sofre perturbações aleatórias, as quais são representadas como um ruído aditivo, de maneira que o sinal que efetivamente chega ao receptor pode ser escrito como:

$$r(n) = s'(n) + \eta(n),$$

onde $\eta(n)$ denota o valor do ruído no instante n.

3.1 Equalização supervisionada

No cenário supervisionado, tem-se acesso a uma sequência de símbolos da fonte durante o projeto do equalizador, de modo que é possível utilizar uma medida do erro entre a estimativa gerada pelo equalizador e o símbolo verdadeiramente transmitido como critério de treinamento.

Tipicamente, os parâmetros do equalizador são ajustados tendo em vista a minimização do erro quadrático médio (MSE, do inglês *mean squared error*) determinado para T pares entrada-saída $\{\mathbf{r}(n); s(n-\alpha)\}$, onde $\mathbf{r}(n) = (r(n) \ldots r(n-K+1))$ é o vetor de entrada do equalizador com as últimas K amostras do sinal recebido e α denota o atraso de equalização.

3.2 Equalização cega

No cenário não-supervisionado (ou cego), não há mais um sinal de referência para guiar o treinamento do equalizador. Neste contexto, o uso de uma estrutura não-linear para cancelar o canal é possível graças a uma elegante estratégia baseada em predição [2]. A ideia consiste em treinar o modelo para prever o próprio sinal recebido $r(n)$ a partir de algumas amostras passadas, $r(n-1), \ldots, r(n-K)$. Então, o erro cometido pelo preditor pode ser genericamente escrito como:

$$\begin{aligned} e(n) &= r(n) - F\{r(n-1), \ldots, r(n-K)\} \\ &= h_0^* s(n) + \cdots + h_{D-1}^* s(n-D+1) \\ &\quad - F\{r(n-1), \ldots, r(n-K)\} + \eta(n), \end{aligned}$$

onde $F\{\cdot\}$ denota o mapeamento não-linear produzido pelo preditor.

Interessantemente, é possível mostrar que se o preditor consegue aproveitar toda a informação dos símbolos da fonte subjacentes às amostras do sinal recebido em sua entrada, toda a redundância entre seu conjunto de entradas e o valor futuro $r(n)$ é eliminada, de modo que o erro ótimo de predição se torna:

$$e^{\text{ótimo}}(n) = h_0^* s(n) + \eta(n)$$

Nesta condição, vemos que o erro de predição oferece uma estimativa do símbolo $s(n)$, a menos de um ruído aditivo e de um fator de escala proporcional ao primeiro coeficiente do canal (h_0). É pertinente ressaltar que o treinamento do preditor é supervisionado, mas como ele se baseia apenas em amostras do próprio sinal recebido, não fazendo uso de valores conhecidos da fonte, a equalização do canal é, de fato, não-supervisionada.

4 Metodologia

A fim de analisar o comportamento da DESN no problema de equalização, foram considerados três canais, cujas funções de transferência são: (1) $H(z) = 0{,}5 + z^{-1}$; (2) $H(z) = 1 + z^{-1}$ e (3) $H(z) = 0{,}5 + 0{,}71z^{-1} + 0{,}5z^{-2}$. Optamos por fixar o número de entradas da rede no valor mínimo ($K = 1$) para implicitamente verificar a capacidade de extração de informação e de memória dos modelos estudados. Além disso, adotamos um atraso de equalização nulo ($\alpha = 0$) nos experimentos, criando, nos três cenários, condições desafiadoras de equalização: no primeiro cenário, é necessário que o equalizador crie uma fronteira não-linear para separar os estados do canal, enquanto, nos dois casos seguintes, também é preciso que o modelo tenha realimentação para lidar com os estados coincidentes. A relação sinal-ruído adotada nos experimentos foi de 30 dB.

Utilizando uma busca em grade, analisamos diferentes combinações de valores para os quatro principais hiperparâmetros da DESN: o número N_L de reservatórios; o número N de neurônios em cada reservatório; o fator de esquecimento $a^{(l)}$ e o raio espectral ρ_l para cada reservatório[1]. Nesta etapa, exploramos uma validação cruzada do tipo *holdout*, usando 10.000 amostras tanto para treinamento quanto para validação. Ao final, a melhor configuração da DESN, determinada com base na média do MSE[2] de validação após 10 repetições independentes, foi aplicada ao conjunto de teste, que contém também 10.000 amostras, no qual determinamos o MSE e a taxa de erro (BER, do

[1] Para reduzir o custo desta busca, consideramos que todos os reservatórios têm os mesmos valores de N, $a^{(l)}$ e ρ_l.

[2] No cenário não-supervisionado, o MSE calculado é entre o erro de predição ótimo e o erro cometido pela rede.

inglês *bit error rate*). Nos três conjuntos de dados, as primeiras 500 amostras são utilizadas somente para inicializar a rede e remover efeitos transitórios, não interferindo no cálculo de desempenho.

A título de comparação, consideramos também uma ESN com um único reservatório contendo a mesma quantidade de neurônios que o total utilizado pela DESN ótima. Então, analogamente ao que foi feito para a DESN, buscamos os melhores valores para os demais hiperparâmetros da ESN (a e ρ) e, em seguida, aplicamos a rede ótima ao conjunto de teste.

5 Resultados

Primeiramente, o canal $H(z) = 0{,}5 + 0{,}71z^{-1} + 0{,}5z^{-2}$ foi escolhido como o cenário base para uma apresentação mais detalhada da sensibilidade paramétrica da DESN. As Figuras 2 e 3 exibem a progressão do desempenho médio de validação da DESN em função dos quatro parâmetros estudados, considerando o caso supervisionado e o não-supervisionado, respectivamente.

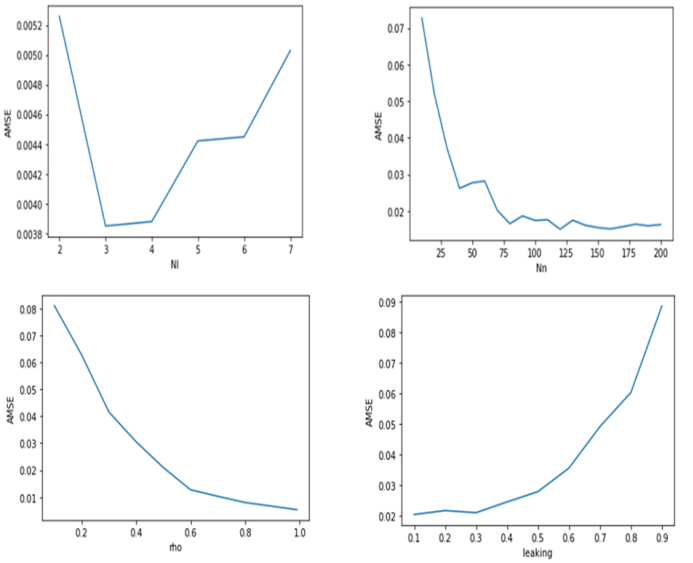

Figura 2 – AMSE de validação da DESN em função dos quatro hiperparâmetros – cenário supervisionado.

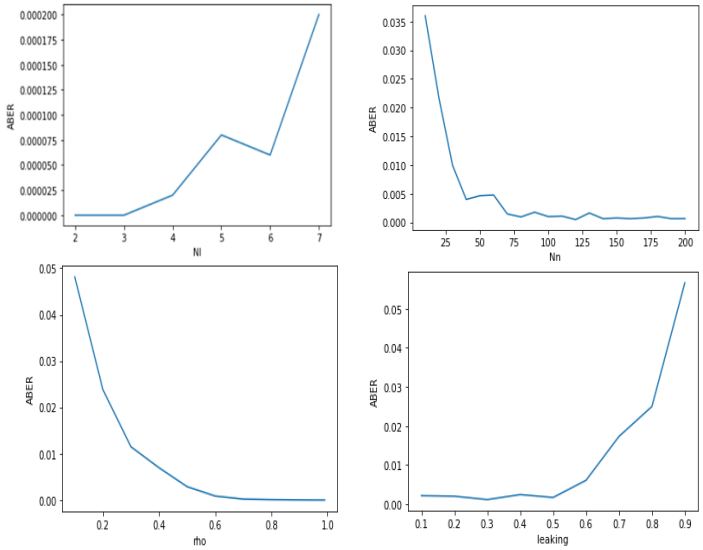

Figura 3 – AMSE de validação da DESN em função dos quatro hiperparâmetros – cenário não-supervisionado.

Observando as Figuras 2 e 3, é possível perceber que tendências relativamente claras em relação aos parâmetros N, ρ_l e $a^{(l)}$: para os dois primeiros, quanto maior seu valor, menor tende a ser o erro médio de validação; em contrapartida, quanto menor o valor do fator de vazamento, melhor é o desempenho da rede. Quanto ao número de reservatórios (N_L), notamos que nos dois cenários (supervisionado e cego), uma quantidade de camadas limitada a quatro se mostra mais adequada, pois a partir desse valor há uma degradação significativa no desempenho da rede.

Selecionando, então, os valores ótimos, aplicamos a DESN ao conjunto de teste, assim como uma ESN com um único reservatório contendo a mesma quantidade total de neurônios. Os valores do MSE e da BER obtidos pelos dois modelos foram os seguintes:

Tabela 1. MSE e BER de teste para o canal $H(z) = 0{,}5 + 0{,}71z^{-1} + 0{,}5z^{-2}$ no caso supervisionado

	MSE	BER
DESN	0.000206	0.0000
ESN ($N_L = 1$)	0.001935	0.0000

Tabela 2. MSE e BER de teste para o canal $H(z) = 0{,}5 + 0{,}71z^{-1} + 0{,}5z^{-2}$ no caso não-supervisionado

	MSE	BER
DESN	0.003988	0.0000
ESN ($N_L = 1$)	0.006459	0.0001

É possível constatar que a DESN obteve um desempenho melhor que a ESN clássica: inclusive, no caso supervisionado, a redução do MSE alcançada pela DESN foi bem expressiva. Isto indica que o uso de uma hierarquia de reservatórios, em vez de um único reservatório com mais neurônios, foi benéfica para lidar com as características deste canal.

Para os dois canais restantes, apresentamos na Tabela 2 os valores ótimos identificados para os hiperparâmetros da DESN, assim como o MSE e a BER obtidos para o conjunto de teste. Por sua vez, a Tabela 3 traz as mesmas informações, mas para a ESN ótima de um único reservatório.

Comparando as duas tabelas, vemos que no cenário supervisionado a DESN se saiu um pouco melhor que a ESN clássica no caso do canal $H(z) = 0{,}5 + z^{-1}$, mas um pouco pior para o canal $H(z) = 1 + z^{-1}$.

Já no cenário cego, a DESN alcançou um desempenho ligeiramente inferior à ESN clássica nos dois casos. Isto indica que provavelmente a busca por construir reservatórios que reajam à entrada com diferentes escalas de tempo não é tão

Tabela 2. Hiperparâmetros ótimos da DESN e desempenho no conjunto de teste.

Canal	critério matemático	N_l	N	ρ	a	MSE de teste	BER de teste
$[0.5, 1]$	supervisionado	4	100	0.95	0.2	0.000001	0.000000
	não-supervisionado	3	10	0.9	0.2	0.002081	0.000000
$[1, 1]$	supervisionado	3	100	0.95	0.9	0.002855	0.000380
	não-supervisionado	2	50	0.95	0.5	0.017954	0.000680

Tabela 3. Hiperparâmetros ótimos da ESN padrão (um reservatório) e desempenho no

Canal	critério matemático	N_l	N	ρ	a	MSE de teste	BER de teste
$[0.5, 1]$	supervisionado	1	400	0.99	0.1	0.000003	0.000000
	não-supervisionado	1	30	0.99	0.15	0.001961	0.000000
$[1, 1]$	supervisionado	1	300	0.99	0.76	0.001822	0.000320
	não-supervisionado	1	100	0.94	0.72	0.016878	0.000260

favorável no contexto da predição do sinal recebido ($r(n)$) a partir de sua única amostra passada ($r(n-1)$).

De qualquer forma, o desempenho alcançado pela DESN na equalização dos canais estudados pode ser considerado bastante satisfatório, tanto na abordagem supervisionada quanto na abordagem cega, incluindo os casos com estados coincidentes, o que posiciona as DESNs como alternativas promissoras para problemas de desconvolução e extração da informação.

5 Conclusão

Neste trabalho, foi realizado um estudo das redes neurais com estados de eco em sua versão profunda, na qual múltiplos reservatórios são empilhados com o intuito de criar representações dinâmicas com escalas de tempo distintas, e sua aplicação ao problema de equalização de canais de comunicação. O treinamento de uma DESN segue o mesmo procedimento típico de uma ESN, de modo que somente os parâmetros da camada de saída são efetivamente obtidos a partir da minimização da função de erro, enquanto os conjuntos de pesos dos reservatórios são gerados aleatoriamente e obedecendo à propriedade de estados de eco.

Os experimentos realizados em três cenários e para as abordagens supervisionada e cega mostram que há situações em que a profundidade da DESN é benéfica e leva a uma equalização mais efetiva que uma ESN convencional com um único reservatório.

Referências

[1] Haykin, S. (2008). Neural Networks and Learning Machines. 3ª ed. Pearson.

[2] Boccato, L. (2013). “Novas propostas e aplicações de redes neurais com estados de eco”. Tese de Doutorado, Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas.

[3] Géron, A. (2019), Hands-on machine learning with scikit-learn, keras and tensorflow, 2ª ed., O’Reilly Media.

[4] Lukoševičius, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. Computer Science Review, 3, 127-149.

[5] Gallicchio, C., & Micheli, A. (2017). Deep Echo State Network (DeepESN): A Brief Survey, arXiv preprint arXiv:1712.04323.

[6] Gallicchio, C., Micheli, A. & Pedrelli, L. (2017). Deep reservoir computing: A critical experimental analysis, Neurocomputing, 268, pp. 87-99.

[7] Haykin, S., (1996). Adaptive filter theory. 3ª ed., Prentice Hall.

[8] Gallicchio, C., Micheli, A. (2017). Echo state property of deep reservoir computing networks., Cognitive Computation, vol. 3, 337-350.

[9] Gallicchio, C., Micheli, A., Pedrelli, L. (2018). Design of deep echo state networks, Neural Networks, vol. 108,
pp. 33-47