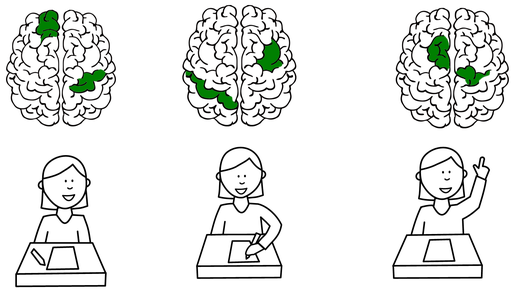
Input-Cell Attention Reduces Vanishing Saliency of Recurrent Neural Networks

Aya Abdelsalam Ismail[1], Mohamed Gunady[1], Luiz Pessoa[2], Héctor Corrada Bravo[*1] and Soheil Feizi[*1]

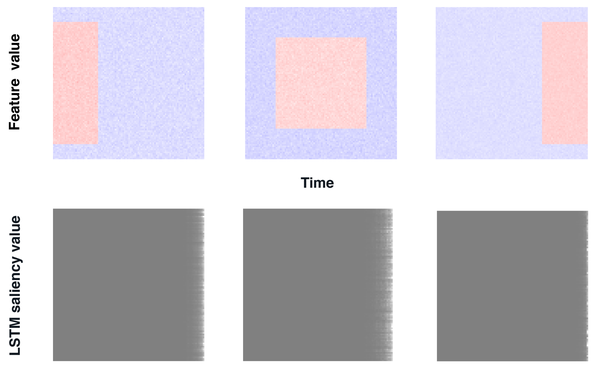
Saliency for Time Series



To characterize brain region activity pertinent to the task being performed by the subject, a saliency method should capture changes in feature importance (corresponding to brain regions) over time.

Saliency in RNNs Vanishes over Time

In RNNs, specifically LSTMs, saliency vanishes over time, biasing detection of salient features only to last few time steps.



Top row: samples from synthetic datasets, where red represents important features and blue is Gaussian noise. **Bottom row:** saliency maps produced by LSTM for each case, importance is reported only in the last few time steps.

For more details please check the original paper.

LSTM

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i)$$
$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f)$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o)$$
$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_{x\tilde{c}}\mathbf{x}_t + \mathbf{W}_{h\tilde{c}}\mathbf{h}_{t-1} + \mathbf{b}_{\tilde{c}})$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$$

Input-cell At.

$$\mathbf{i}_t = \sigma\left(\boxed{\mathbf{W}_{Mi}\mathbf{M}_t} + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i\right)$$
$$\mathbf{f}_t = \sigma\left(\boxed{\mathbf{W}_{Mf}\mathbf{M}_t} + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f\right)$$
$$\mathbf{o}_t = \sigma\left(\boxed{\mathbf{W}_{Mo}\mathbf{M}_t} + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o\right)$$
$$\tilde{\mathbf{c}}_t = \tanh\left(\boxed{\mathbf{W}_{M\tilde{c}}\mathbf{M}_t} + \mathbf{W}_{h\tilde{c}}\mathbf{h}_{t-1} + \mathbf{b}_{\tilde{c}}\right)$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$$

[1] Department of Computer Science, University of Maryland
[2] Department of Psychology, University of Maryland

Why does LSTM Saliency Vanish?

- The relevance (R) of each input feature at a given time step (x_{t_i}) to the output of network (S) is given by:

$$R_T(x_t) = \left|\frac{\partial S(x_T)}{\partial x_t}\right| = \left|\frac{\partial S}{\partial h_T}\left(\prod_{i=T}^{t+1}\frac{\partial h_i}{\partial h_{i-1}}\right)\frac{\partial h_t}{\partial x_t}\right|$$

- $\frac{\partial h_t}{\partial h_{t-1}}$ is the only term affected by the number of time steps.
- Solving the partial derivative:

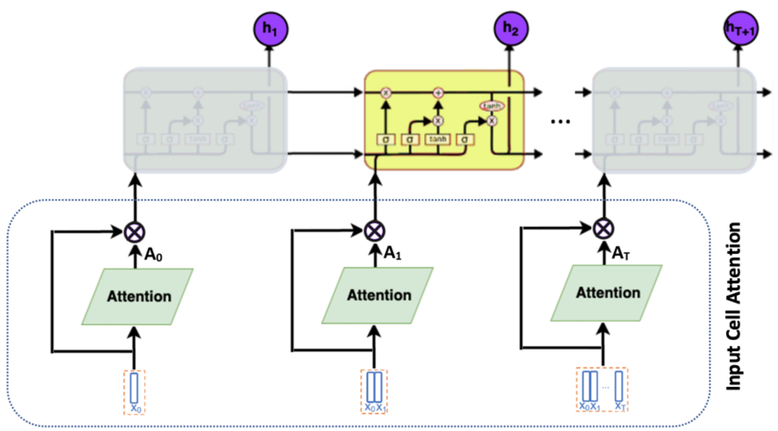
$$\frac{\partial h_t}{\partial h_{t-1}} = \tanh(c_t)\left(\boxed{W_{ho}}(o_t \odot (1-o_t))\right) + o_t \odot \left(1-\tanh^2(c_t)\right)\Big[c_{t-1}\left(\boxed{W_{hf}}(f_t \odot (1-f_t))\right) + \tilde{c}_t\left(\boxed{W_{hi}}(i_t \odot (1-i_t))\right) + i_t\left(\boxed{W_{h\tilde{c}}}(1-\tilde{c}_t \odot \tilde{c}_t)\right) + f_t\Big]$$

- As t decreases, those terms multiplied by the weight matrix will eventually vanish, $\frac{\partial h_t}{\partial h_{t-1}}$ will be reduced to :

$$\frac{\partial h_t}{\partial h_{t-1}} \approx o_t \odot \left(1-\tanh^2(c_t)\right)\Big[f_t\Big]$$

- The amount of information preserved depends on "forget gate" (f_t).
- The relevance of input at time t eventually disappears as t decreases.

Input-Cell Attention Reduces Vanishing Saliency



Input-cell attention uses a fixed-size matrix embedding, each row of the matrix attending to different inputs from current or previous time steps.

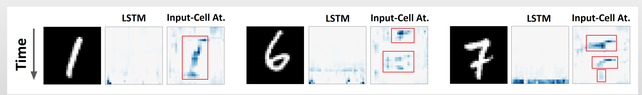
Experiments

Synthetic Datasets

The effect of changing the location of importance features in time on weighted Jaccard similarity and Euclidean distance for different architectures.

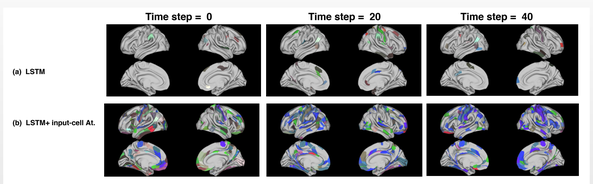
MNIST as a Time Series

To validate the resulting saliency maps in cases where important features have more structured distributions of different shapes, we treat MNIST dataset as a time series.



Human Connectome Project fMRI Data

RNNs are used to classify time series based on the task performed by the subject while scanned by an fMRI machine. The saliency map produced by each model is shown bellow:



Studying features identified as salient by different models: