Think Architecture First: Benchmarking Deep Learning Interpretability in Time Series Predictions

3 April 2020

Presenter: Sanchit Sinha

https://qdata.github.io/deep2Read/

Motivation

- Saliency methods used a lot to explain and rank the most important features of a data
- Conundrum: Do we focus on better saliency methods or better model architecture or somewhere in between
- Understanding usefulness of saliency methods on less studied time-series data
- Considering a slightly different metric - precision and recall of the features
- No ground-truth to compare to, come up with a way to compare.
  - It is easy to **SEE** in visual problems
  - Easy to **INTERPRET** in NLP tasks
  - No way to tell in raw feature tasks
- Comparison between various saliency techniques (7) over multiple models (4) and datasets (11) = 308

Background

- Saliency methods:
  - Gradient (GRAD) $\frac{\partial S_c(X)}{\partial x_{t_i}}$
  - Input*Gradient $x_{t_i} \times \frac{\partial S_c(X)}{\partial x_{t_i}}$
  - Integrated Gradient (IG) $(x_{t_i} - \overline{x}_{t_i}) \times \int_{\alpha=0}^{1} \frac{\partial S_c\left(\overline{X} + \alpha\left(X - \overline{X}\right)\right)}{\partial x_{t_i}} d\alpha$
  - SmoothGrad (SG) $\frac{1}{n} \sum_{1}^{n} \frac{\partial S_c(X + \mathcal{N}(0, \sigma^2))}{\partial x_{t_i}}$
  - DeepLift - Only x-x’
  - GradientSHAP - Adding Gaussian Noise to SmoothGrad
  - DeepSHAP - DeepLift + Shapley
  - Random

Related Work

- The saliency methods mentioned before
- Ismail, A. A., Gunady, M., Pessoa, L., Corrada Bravo, H., and Feizi, S. Input-cell attention reduces vanishing saliency of recurrent neural networks: **On generating time series data**
- **Transformers:** Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Atten- tion is all you need.
- **Temporal Convolutional Network:** (Oord et al., 2016; Lea et al., 2017; Bai et al., 2018)

Synthetic Data Generation

- Gaussian noise with 0 mean and unit variance
- Informative features highlighted by adding constant mu (=1) to positive class and subtracting mu from negative class
- Dataset size is 1000 (training) and 300 (testing)
- Embedding Size = 100, timesteps = 100
- Types of data generated:
  - Earlier Boxes
  - Middle Boxes
  - Latter Boxes
  - Moving Box -
  - Right Triangle - information grows over time
  - Equilateral Triangle - important for the majority of the time points
  - Unequal Events
  - Rare - anomalies in time series variables
  - Sine - info varies widely over time.

Figure for data generation

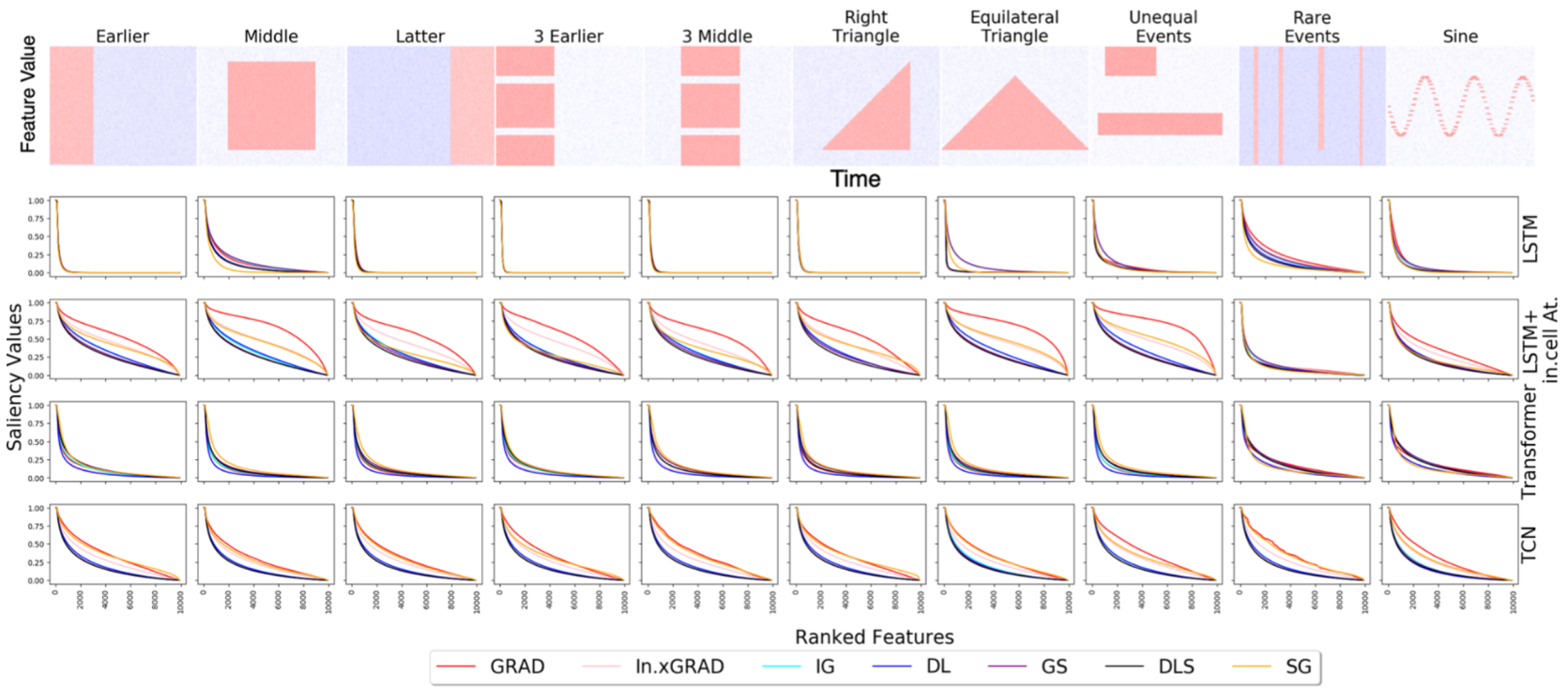

Figure 1. First row shows samples from our Synthetic Datasets, where red represents informative features and blue is Gaussian noise $\mathcal{N}(0, \sigma^2)$. Rows 2-5 shows the distribution of saliency values of ranked features produced by different saliency methods. Neural nets were trained to classify positive and negative examples for each dataset. The distribution of saliency is highly dependent on the neural architecture rather then the saliency method itself.

Proposed Solution

- Sort relevance $R(X)$, so that $R_e\left(x_{t_i}\right)$ is the e^{th} element in ordered set $\{R_e\left(x_{t_i}\right)\}_{e=1}^{T\times N}$.
- Find top k relevance in the order set such that $\frac{\sum_{e=1}^{k} R_e\left(x_{t_i}\right)}{\sum_{i=1,t=1}^{N,T} R(x_{t_i})} \approx d$ (where d is percentage of distortion over all saliency).
- Adversarially attack x_{t_i}, where $R\left(x_{t_i}\right) \in \{R_e\left(x_{t_i}\right)\}_{e=1}^{k}$; we used an L_∞ projected gradient descent (Madry et al., 2017) attack.
- We calculate the drop in model accuracy after the attack, this is repeated at different levels of degradation $d = [0, 10, 20, \ldots, 100]$.

Proposed Solution - 2

- The above process an result in:
  - A steep drop in accuracy, meaning that the attacked feature is necessary for a correct prediction **(important feature)**
  - Unchanged accuracy:
    - The saliency method incorrectly identified the feature as important **(wrong method)**
    - Removal is not sufficient for the model to behave incorrectly **(ambiguous)**
- So solution which the authors have proposed is to use Precision and Recall

Metrics

Saliency \ Actual	Informative	Noise
High	True Positive (TP)	False Positive (FP)
Low	False Negative (FN)	True Negative (TN)

Table 1. Confusion Matrix, for precision and recall calculation.

- Precision:

$$\frac{\sum R\left(x_{t_i}\right)\left\{x_{t_i} : x_{t_i} \in TP\right\}}{\sum R\left(x_{t_i}\right)\left\{x_{t_i} : x_{t_i} \in TP\right\} + \sum R\left(x_{t_i}\right)\left\{x_{t_i} : x_{t_i} \in FP\right\}}$$

- Recall:

$$\frac{\sum R\left(x_{t_i}\right)\left\{x_{t_i} : x_{t_i} \in TP\right\}}{\sum R\left(x_{t_i}\right)\left\{x_{t_i} : x_{t_i} \in TP\right\} + \sum R\left(x_{t_i}\right)\left\{x_{t_i} : x_{t_i} \in FN\right\}}$$

Experimental Results

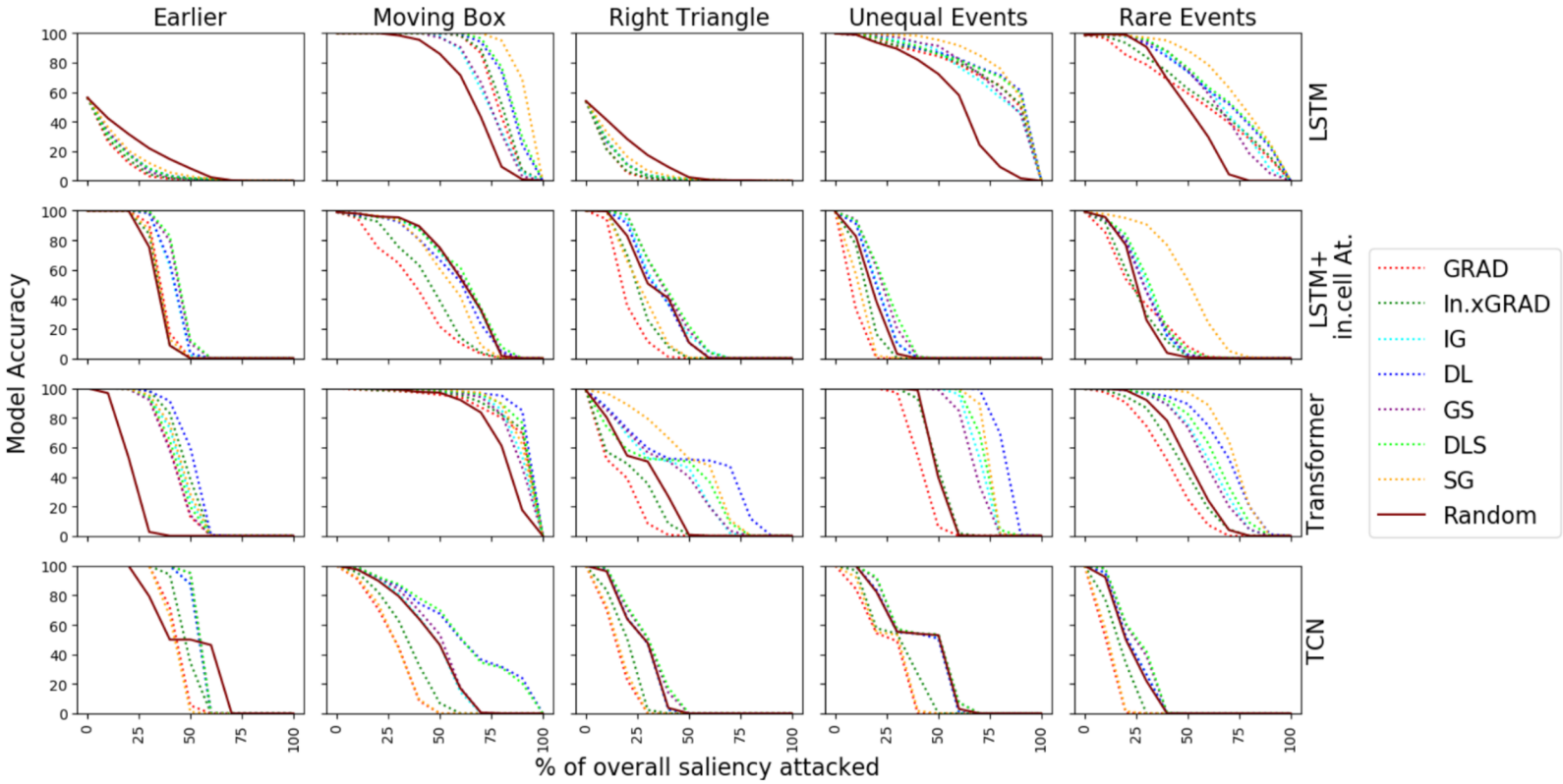

Figure 2. The effect of adversarially perturbing features identified as salient by different interpretability methods against a random baseline. The rate of accuracy drop is not consistent across datasets and architectures. The variance between the saliency method curves is relatively small (except for Transformers). In many cases there isn't a clear distinction between random baseline and other saliency methods.

Experimental Results

Table 2. Precision and recall for different feature groups. **Bold** numbers show the highest precision or recall for a *(neural architecture, dataset)* pair. **Green** numbers show the highest precision or recall for a dataset. All numbers are represented as factors of 10^{-5}.

	Precision												Recall											
	Earlier Box				Right Triangle				Rare Events				Earlier Box				Right Triangle				Rare Events			
Methods	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN
GRAD	0	104	44	82	-	25	-	15	159	60	96	-	0	85	24	42	-	47	-	25	43	**299**	26	-
In.×GRAD	0	80	**78**	80	-	**53**	-	14	154	60	113	0	0	**101**	35	50	-	**63**	-	**31**	34	298	23	0
IG	0	76	30	**91**	0	39	9	10	100	69	87	0	0	89	36	51	0	52	46	21	**48**	279	27	0
DL	0	82	32	**91**	0	47	**32**	10	**171**	**96**	**140**	0	0	90	37	51	0	50	**61**	21	40	275	26	0
GS	0	100	11	87	0	45	21	**29**	91	42	77	0	0	91	24	**53**	0	53	43	30	44	255	27	0
DLS	0	**109**	14	88	0	47	17	25	169	87	98	0	0	92	17	51	0	52	59	26	43	262	29	0
SG	0	76	26	31	0	18	12	10	159	79	95	-	0	81	**52**	**29**	0	35	50	21	35	271	**29**	-
Random	0	63	1	20	0	35	2	5	71	22	39	0	0	84	15	37	0	55	40	26	43	273	26	0

(a) Features that result in 30% drop in accuracy when attacked. '-' indicates that the smallest attack caused accuracy drop $> 30\%$

	Precision												Recall											
	Earlier Box				Right Triangle				Rare Events				Earlier Box				Right Triangle				Rare Events			
Methods	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN
GRAD	0	85	24	37	0	61	55	32	44	249	26	25	0	**104**	44	53	0	**99**	97	**57**	**121**	**145**	96	83
In.×GRAD	0	**101**	26	**44**	0	**70**	63	**38**	35	244	23	20	0	80	**45**	**52**	0	88	**104**	55	119	140	**113**	87
IG	0	89	23	40	0	61	65	30	**59**	248	31	27	0	76	13	43	0	71	33	46	59	84	46	79
DL	0	90	22	40	0	59	**71**	30	42	248	32	27	0	82	12	43	0	84	56	46	107	123	64	79
GS	0	89	24	42	0	61	47	36	50	229	28	27	0	71	11	40	0	78	34	54	49	54	39	81
DLS	0	91	17	39	0	60	66	34	49	233	**35**	27	0	79	14	38	0	80	49	49	112	113	60	78
SG	0	81	**35**	29	0	53	50	29	33	**262**	34	**28**	0	76	11	31	0	79	12	42	73	50	41	**96**
Random	0	84	25	37	0	62	56	33	44	249	28	24	0	63	7	20	0	60	12	23	50	30	28	32

(b) Highest ranked features that represent 30% of overall saliency.

	Precision												Recall											
	Earlier Box				Right Triangle				Rare Events				Earlier Box				Right Triangle				Rare Events			
Methods	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN
GRAD	0	161	182	138	332	168	256	160	228	330	199	136	0	84	43	50	0	65	56	47	39	**202**	25	21
In.×GRAD	0	208	212	192	332	209	**277**	211	202	328	199	135	0	**124**	53	77	0	**87**	69	**70**	26	180	18	15
IG	0	200	236	**238**	332	215	273	228	247	308	239	243	0	83	68	88	0	66	68	60	36	89	29	**34**
DL	0	200	246	**238**	332	211	285	229	256	328	236	242	0	82	64	88	0	61	64	60	38	178	23	**34**
GS	0	209	219	253	**333**	204	246	238	229	306	223	239	0	87	68	**98**	0	62	54	63	33	85	30	33
DLS	0	**215**	223	246	**333**	**218**	269	**241**	265	**329**	224	**248**	0	88	55	91	0	62	62	65	**41**	177	26	**34**
SG	0	132	**259**	149	**333**	120	258	153	**267**	308	**247**	139	0	62	**95**	59	0	34	**71**	50	26	89	**35**	20
Random	**97**	99	97	100	102	105	105	99	100	106	97	95	**78**	79	78	80	**62**	64	64	60	31	33	30	30

(c) Top 30% of ranked features

Experimental Results

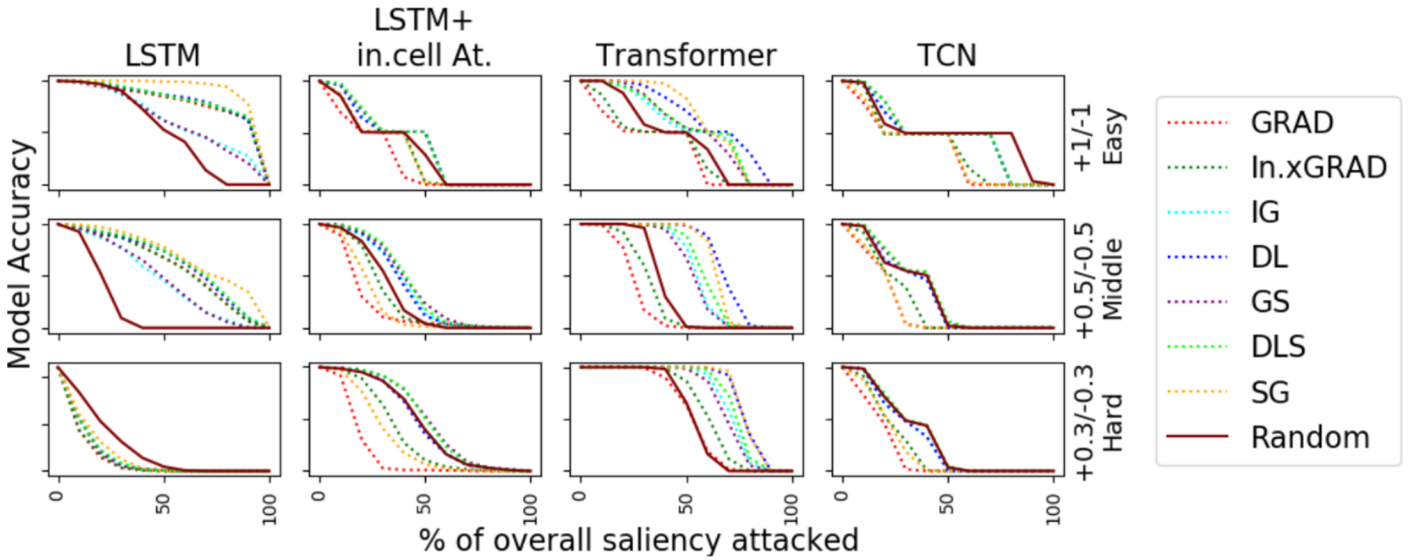

Figure 3. The effect of adversarially perturbing features identified as salient by different attribution methods when changing the difficulty of classification task. First row, shows the result for our standard informative features for positive class are generated as $\mathcal{N}(0,\, \sigma^2) + 1$ and negative $\mathcal{N}(0,\, \sigma^2) - 1$. Second row, positive class $\mathcal{N}(0,\, \sigma^2) + 0.5$ and negative $\mathcal{N}(0,\, \sigma^2) - 0.5$. Third row, positive class $\mathcal{N}(0,\, \sigma^2) + 0.3$ and negative $\mathcal{N}(0,\, \sigma^2) - 0.3$. The variance between the saliency method curves increases as the difficulty of the task increases (except for LSTM).

Experimental Results

Table 3. Precision and recall of features that result in 30% drop in accuracy when attacked, for the same dataset but different classification difficulty. All numbers are represented as factors of 10^{-5}.

	Precision												**Recall**											
	Middle Box +1/-1				Middle Box +0.5/-0.5				Middle Box +0.3/-0.3				Middle Box +1/-1				Middle Box +0.5/-0.5				Middle Box +0.3/-0.3			
Methods	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN	LSTM	LSTM+ in.cell At.	Trans.	TCN
GRAD	0	192	37	141	**1**	90	53	39	-	136	91	0	0	49	16	37	1	23	43	11	-	33	116	0
In.×GRAD	0	181	34	134	**1**	95	53	39	-	134	98	0	0	46	**25**	35	2	52	**55**	11	-	62	**155**	0
IG	0	**195**	30	135	0	**120**	53	37	0	133	94	0	0	42	5	60	0	79	31	10	0	81	77	0
DL	0	192	**66**	130	**1**	119	48	37	0	131	88	0	0	45	19	61	**3**	**85**	47	10	0	90	107	0
GS	0	184	42	136	0	115	51	**102**	0	**141**	100	**22**	0	**86**	8	59	0	74	22	**51**	0	119	72	**9**
DLS	0	191	65	**142**	**1**	118	61	101	0	135	104	**22**	0	**86**	11	**64**	2	81	35	49	0	**127**	110	8
SG	0	191	53	126	0	93	**70**	39	0	133	**105**	0	0	48	20	23	2	21	42	9	0	56	105	0
Random	0	192	41	138	0	104	52	42	0	134	91	**22**	0	39	2	12	0	41	6	4	0	77	21	3