Evaluating Neural Network Explanation Methods using Hybrid Documents and Morphosyntactic Agreement

Nina Poerner, Benjamin Roth, Hinrich Schütze
poerner@cis.uni-muenchen.de

The predictions of Deep Neural Networks (DNNs) are hard to understand from parameters/activations alone. This is problematic for

- developers, who want to improve their models
- non-experts who want explanations for decisions taken with respect to their data

Therefore, several approximative **post-hoc explanation methods** have been proposed to explain predictions of DNNs.

Contribution

We propose **two novel, highly scalable experimental paradigms** to evaluate explanation methods for NLP classification tasks.

We evaluate six families of explanation methods on five architectures, on small context and large context tasks and three corpora. This is the **most comprehensive** evaluation for NLP to date.

Results in a nutshell

LRP and DeepLIFT are the most consistent methods

Substring LIME (LIMSSE) works best on the hybrid document task (small context) but fails the morphosyntactic agreement task (large context)

Input gradient is competitive on CNN, but not on RNNs

Evaluated explanation methods

Definition: an explanation method is a function $\phi(t, k, \mathbf{x})$ that returns real-valued scores for positions t in text $\mathbf{x}$ for class k. For instance, in sentiment classification we would expect $\phi(1, \text{positive}, [\text{“great”}, \text{“bar”}]) > \phi(2, \text{positive}, [\text{“great”}, \text{“bar”}])$.

Input gradient

$\phi_{\text{grad}_1^{\text{L2}}}(t, k, \mathbf{x}) = ||\frac{\partial o(k, \mathbf{E})}{\partial \mathbf{e}_t}||_2$ [Bansal et al., 2016]

$\phi_{\text{grad}_1^{\text{dot}}}(t, k, \mathbf{x}) = \mathbf{e}_t \cdot \frac{\partial o(k, \mathbf{E})}{\partial \mathbf{e}_t}$ [Denil et al., 2015]

$\phi_{\text{grad}_f^{\text{L2}}}(t, k, \mathbf{x}) = ||\frac{1}{M} \sum_{m=1}^{M} \frac{\partial o(k, \frac{m}{M}\mathbf{E})}{\partial \mathbf{e}_t}||_2$

$\phi_{\text{grad}_f^{\text{dot}}}(t, k, \mathbf{x}) = \mathbf{e}_t \cdot \frac{1}{M} \sum_{m=1}^{M} \frac{\partial o(k, \frac{m}{M}\mathbf{E})}{\partial \mathbf{e}_t}$

[Sundararajan et al., 2017]

where k is a class, $\mathbf{e}_t$ is an embedding vector, $\mathbf{E}$ are concatenated embedding vectors, $o(k, \mathbf{E})$ is one of $p(k|\mathbf{E})$ or $s(k, \mathbf{E})$ (pre-softmax class scores).

Layer-wise relevance propagation (LRP)

$R(i) = \sum_j R(j) \frac{a_i w_{i,j}}{a'_j + \text{esign}(a'_j)}$ [Bach et al., 2015]

where j are neurons downstream from i, a' and a are activations before and after a nonlinearity, w is a weight and $\text{esign}(a')$ a small signed constant.

Modification for LSTM / GRU: treat sigmoid gates as weights rather than neurons [Arras et al., 2017].

DeepLIFT [Ancona et al., 2018], c.f., [Shrikumar et al., 2017]

Like LRP, but:

$R(i) = \sum_j R(j) \frac{(a_i - \bar{a}_i) w_{i,j}}{a'_j - \bar{a}'_j + \text{esign}(a'_j - \bar{a}'_j)}$

where $\bar{a}$ are activations during the forward pass of a baseline input.

LIMSSE

LIME (Local Interpretable Model-agnostic Explanations) probes DNN with random inputs drawn from $\mathbf{x}$ and fits a linear model to observed behavior [Ribeiro et al., 2016]. The original BOW sampling is inappropriate for word-order sensitive CNN/RNN. Therefore, LIMSSE draws random substrings:

$\text{limsse}(t, k, \mathbf{x}) = \text{argmin}_{v_{k,t}} \sum_n L(n, k)$

$\text{limsse}^{\text{ms}}$: $L(n, k) = (o(k, \mathbf{z}_n) - \mathbf{b}_n \cdot \mathbf{v}_k)^2$

$\text{limsse}^{\text{bb}}$: $L(n, k) = \log(\sigma(\mathbf{b}_n \cdot \mathbf{v}_k)) \mathbb{I}[f(\mathbf{z}_n) = k] + \log(1 - \sigma(\mathbf{b}_n \cdot \mathbf{v}_k)) \mathbb{I}[f(\mathbf{z}_n) \neq k]$

where $\mathbf{z}_n$ is a substring of $\mathbf{x}$, $\mathbf{b}_n \in \{0, 1\}^T$ indicates presence or absence of tokens in $\mathbf{z}_n$, $o(k, \mathbf{z}_n)$ is one of $p(k|\mathbf{z}_n)$, $s(k, \mathbf{z}_n)$.

Perturbation methods

Omit or occlude all N-grams that contain x_t and average over the change in output.

$\phi_{\text{occ}_N}(t, k, \mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} [s(k, \mathbf{E}) - s(k, [\mathbf{e}_1 \dots \mathbf{e}_{t-N-1+n} \mathbf{0}_1 \dots \mathbf{0}_N \mathbf{e}_{t+n} \dots \mathbf{e}_T])]$ [Li et al., 2016]

$\phi_{\text{omit}_N}(t, k, \mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} [s(k, \mathbf{E}) - s(k, [\mathbf{e}_1 \dots \mathbf{e}_{t-N-1+n} \mathbf{e}_{t+n} \dots \mathbf{e}_T])]$ [Kádár et al., 2017]

Cell decomposition [Murdoch and Szlam, 2017]

$\phi_{\text{decomp}}(t, k, \mathbf{x}) = \text{nl}(t, k) - \text{nl}(t - 1, k)$

LSTM: $\text{nl}(t, k) = \mathbf{w}_k \cdot \mathbf{o}_T \odot \tanh([\prod_{j=t+1}^{T} \vec{f}_j] \odot \vec{c}_t)$

GRU: $\text{nl}(t, k) = \mathbf{w}_k \cdot ([\prod_{j=t+1}^{T} \vec{z}_j] \odot \vec{h}_t)$

Hybrid document paradigm

- Train DNN on some document classification task.
- Sentence-tokenize documents in test set and shuffle.
- Reconcatenate sentences, ten at a time, into **hybrid documents**. Remember for every token the class of its document of origin.
- Let DNN classify a hybrid document: $f(\mathbf{x})$.
- Let explanation method find the most relevant token for this classification. If this token stems from a document with gold label $f(\mathbf{x})$, count sample as a hit point.
- Calculcate the ratio of hit points to the number of samples (pointing game accuracy) [Zhang et al., 2016]

Experiment

- Data: 20 newsgroups [Lang, 1995] for topic classification, subset of yelp dataset challenge for binary sentiment classification
- DNNs: GloVe [Pennington et al., 2014] pre-tained word embeddings, followed by bidirectional 1-layer LSTM [Hochreiter and Schmidhuber, 1997], GRU [Cho et al., 2014], Quasi-LSTM, Quasi-GRU [Bradbury et al., 2017] or CNN with max-pooling [Collobert et al., 2011], followed by softmax
- Also: Comparison with human relevance benchmark [Mohseni and Ragan, 2018] on subset of 20 newsgroups corpus

Morphosyntactic agreement paradigm

- Given: POS-tagged, parsed corpus
- Train DNN to predict a morphological feature of an agreeing word based on context (here: number of verb based on left context)
- For every test set sample, find most relevant token for the prediction: $\text{argmax}_t[\phi(t, f(\mathbf{x}), \mathbf{x})]$
- If $f(\mathbf{x})$ is correct, check if most relevant token is the head (here: the subject). If so, count a $\text{hit}_{\text{target}}$ point.
- Regardless of whether $f(\mathbf{x})$ is correct, check if most relevant token has the predicted feature (here: another noun with predicted number). If so, count a $\text{hit}_{\text{feature}}$ point.
- **Example:** *The man with the telescope [is...]* → maximal relevance on *telescope* gives $\text{hit}_{\text{feature}}$ point, maximal relevance on *man* gives $\text{hit}_{\text{target}}$ and $\text{hit}_{\text{feature}}$ point.
- Calculate pointing game accuracy.

Experiment

- Data: Automatically annotated English wikipedia [Linzen et al., 2016]
- DNNs: Randomly initialized word embeddings, followed by unidirectional 1-layer LSTM, GRU, Quasi-LSTM or Quasi-GRU, followed by softmax

	hybrid document experiment										man. groundtruth					morphosyntactic agreement experiment												
	yelp					20 newsgroups					20 newsgroups					$\text{hit}_{\text{target}}$ $f(\mathbf{X}) = y(\mathbf{X})$								hit_{feat} $f(\mathbf{X}) \neq y(\mathbf{X})$							
ϕ	GRU	QGRU	LSTM	QLSTM	CNN	GRU	QGRU	LSTM	QLSTM	CNN	GRU	QGRU	LSTM	QLSTM	CNN	GRU	QGRU	LSTM	QLSTM	GRU	QGRU	LSTM	QLSTM	GRU	QGRU	LSTM	QLSTM	
$\text{grad}_{1s}^{\text{L2}}$	.61	.68	.67	.70	.68	.45	.47	.25	.33	.79	.26	.31	.07	.18	.74	.48	.23	.63	.19	.52	.27	.73	.22	.09	.11	.19	.19	
$\text{grad}_{1p}^{\text{L2}}$	.57	.67	.67	.70	.74	.40	.43	.26	.34	.70	.18	.35	.07	.13	.66	.48	.22	.63	.18	.53	.26	.73	.21	.09	.09	.18	.11	
$\text{grad}_{\int s}^{\text{L2}}$	.71	.66	.69	.71	.70	.58	.32	.26	.21	.82	.23	.15	.11	.08	.76	.69	.67	.68	.51	.73	.70	.75	.55	.19	.22	.20	.20	
$\text{grad}_{\int p}^{\text{L2}}$	.71	.70	.72	.71	.77	.56	.34	.30	.23	.81	.13	.08	.14	.01	.78	.68	.77	.50	.70	.74	.82	.54	.78	.19	.21	.19	.30	
$\text{grad}_{1s}^{\text{dot}}$	.88	.85	.81	.77	.86	.79	.76	.59	.72	.89	.80	.70	.14	.47	.79	.81	.62	.73	.56	.85	.66	.81	.59	.42	.34	.46	.36	
$\text{grad}_{1p}^{\text{dot}}$	.92	.88	.84	.79	.95	.78	.72	.59	.72	.81	.71	.59	.20	.44	.69	.79	.58	.74	.54	.83	.61	.81	.56	.41	.33	.46	.35	
$\text{grad}_{\int s}^{\text{dot}}$	.84	.90	.85	.87	.87	.81	.68	.60	.68	.89	.82	.64	.21	.26	.80	.90	.87	.78	.84	.94	.92	.83	.89	.54	.51	.46	.52	
$\text{grad}_{\int p}^{\text{dot}}$	.86	.89	.84	.89	**.96**	.80	.69	.62	.73	.89	.80	.53	.40	.54	.78	.87	.85	.68	.84	.93	.92	.74	.93	.53	.48	.42	.51	
omit_1	.79	.82	.85	.87	.61	.78	.75	.54	.76	.82	.80	.48	.33	.48	.65	.81	.81	.79	.80	.86	.87	.86	.84	.43	.45	.44	.45	
omit_3	.89	.80	.89	.88	.59	.79	.71	.72	.81	.76	.77	.37	.36	.49	.61	.74	.77	.73	.73	.82	.84	.82	.79	.41	.45	.42	.46	
omit_7	.92	.88	.91	.91	.70	.79	.77	.77	.84	.84	.77	.49	.44	.55	.65	.76	.80	.66	.74	.85	.88	.78	.80	.40	.48	.43	.47	
occ_1	.80	.71	.74	.84	.61	.78	.73	.60	.77	.82	.77	.49	.19	.10	.65	**.91**	.85	**.86**	.86	.94	.88	.89	.88	.50	.44	.46	.47	
occ_3	.92	.61	.93	.85	.59	.78	.63	.74	.74	.76	.74	.37	.32	.35	.61	.74	.73	.71	.72	.78	.76	.76	.76	.43	.37	.41	.43	
occ_7	.92	.77	.93	.90	.70	.78	.62	.74	.77	.84	.74	.35	.43	.39	.65	.64	.65	.63	.65	.73	.73	.72	.73	.36	.35	.39	.43	
decomp	.79	.88	.92	.88	-	.75	.79	.77	.80	-	.54	.36	.72	.51	-	.84	.87	**.86**	.90	.90	.93	**.92**	**.96**	.52	.58	**.57**	**.63**	
lrp	.92	.87	.91	.84	.86	.82	.83	.79	.85	.89	**.85**	.72	.74	.81	.79	.90	**.90**	**.86**	**.91**	**.95**	**.95**	.91	.95	.58	**.60**	.52	**.63**	
deeplift	.91	.89	**.94**	.85	.87	.82	.83	.78	.84	.89	.84	.72	.70	.81	.80	**.91**	**.90**	.85	**.91**	**.95**	**.95**	.90	.95	**.59**	.59	.52	**.63**	
$\text{limsse}^{\text{bb}}$	.81	.82	.83	.84	.78	.78	.81	.78	.80	.84	.52	.53	.53	.54	.57	.43	.41	.44	.42	.54	.51	.56	.52	.39	.43	.42	.41	
$\text{limsse}_s^{\text{ms}}$	**.94**	**.94**	.93	**.93**	.91	**.85**	**.87**	**.83**	**.86**	.89	**.85**	**.84**	**.76**	**.84**	**.82**	.62	.62	.67	.63	.75	.74	.82	.75	.52	.53	.55	.53	
$\text{limsse}_p^{\text{ms}}$	.87	.88	.85	.86	.94	**.85**	.86	**.83**	**.86**	**.90**	.81	.80	.74	.76	.76	.62	.62	.67	.63	.75	.74	.82	.75	.51	.53	.55	.53	
random	.69	.67	.70	.69	.66	.20	.19	.22	.22	.21	.09	.09	.06	.06	.08	.27	.27	.27	.27	.33	.33	.33	.33	.12	.13	.12	.12	
last	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.66	.67	.66	.67	.76	.77	.76	.77	.21	.27	.25	.26	
N	$7551 \leq N \leq 7554$					$3022 \leq N \leq 3230$					$137 \leq N \leq 150$					$N \approx 1400000$								$N \approx 20000$				

Discussion

- Modified LRP and DeepLIFT are the most consistent methods across tasks and architectures
- Magnitude-sensitive LIMSSE wins the hybrid document task, but fails on the morphosyntactic agreement task → failure to capture dependencies that span large contexts
- Gradient L2-norm is not competitive, due to its inability to distinguish evidence for and against k
- Gradient-embedding dot product is competitive on CNN, with decent results on GRU. It fails on (Q)LSTM. Hypothesis: LSTM memory vectors can become indefinitely big and may saturate the final tanh nonlinearity. GRU hidden vectors are constantly kept in $[-1, 1]$.
- Gradient integration leads to small improvements, but does not remedy the situation on LSTM.
- Input perturbation mostly not competitive.
- Cell decomposition works well on LSTM, but not consistently on other architectures.

decomp	initially a pagan culture , detailed information about the return of the christian religion to the islands during the *norse-era* [is ...]
deeplift	initially a pagan culture , detailed information about the return of the christian religion to the islands during the *norse-era* [is ...]
$\text{limsse}_p^{\text{ms}}$	initially a pagan culture , detailed information about the return of the christian religion to the islands during the *norse-era* [is ...]

Morphosyntactic agreement experiment. Verb context classified singular by LSTM. Underlined: subject. Green: evidence for singular.

$\text{grad}_{1p}^{\text{L2}}$	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The Armenians just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . [...]
$\text{limsse}_s^{\text{ms}}$	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The Armenians just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . [...]

Hybrid newsgroup post classified talk.politics.mideast by QGRU. Underlined: talk.politics.mideast fragment. Green: evidence for talk.politics.mideast.

lrp	From : *kolstad* @ cae.wisc.edu (Joel Kolstad) Subject : Re : Can Radio Freq . Be Used To Measure Distance ? [...] What is the difference between vertical and horizontal ? Gravity ? Does n't gravity pull down the photons and cause a *doppler* shift or something ? (Just kidding !)

Manually annotated sci.electronics post classified by CNN. The explanation method highlights overfitting (“Kolstad” has 11 out of 13 appearances in sci.electronics documents), but this is not rewarded by the manual ground truth.

occ1	[...] FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . ; +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 231 755-2386
lrp	[...] FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . ; +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 231 755-2386

Hybrid newsgroup post classified comp.windows.x by LSTM. The explanation methods highlight overfitting (the address only appears in comp.windows.x posts).

References

[Ancona et al., 2018] Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, Vancouver, Canada.

[Arras et al., 2017] Arras, L., Montavon, G., Müller, K.-R., and Samek, W. (2017). Explaining recurrent neural network predictions in sentiment analysis. In *Eighth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark.

[Bach et al., 2015] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

[Bansal et al., 2016] Bansal, T., Belanger, D., and McCallum, A. (2016). Ask the GRU: Multi-task learning for deep text recommendations. In *ACM Conference on Recommender Systems*, pages 107–114, Boston, USA.

[Bradbury et al., 2017] Bradbury, J., Merity, S., Xiong, C., and Socher, R. (2017). Quasi-recurrent neural networks. In *International Conference on Learning Representations*, Toulon, France.

[Cho et al., 2014] Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar.

[Collobert et al., 2011] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

[Denil et al., 2015] Denil, M., Demiraj, A., and de Freitas, N. (2015). Extraction of salient sentences from labelled documents. In *International Conference on Learning Representations*, San Diego, USA.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Kádár et al., 2017] Kádár, A., Chrupała, G., and Alishahi, A. (2017). Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

[Lang, 1995] Lang, K. (1995). Newsweeder: Learning to filter netnews. In *International Conference on Machine Learning*, pages 331–339, Tahoe City, USA.

[Li et al., 2016] Li, J., Monroe, W., and Jurafsky, D. (2016). Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

[Linzen et al., 2016] Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

[Mohseni and Ragan, 2018] Mohseni, S. and Ragan, E. D. (2018). A human-grounded evaluation benchmark for local explanations of machine learning. *CoRR*, abs/1801.05075.

[Murdoch and Szlam, 2017] Murdoch, W. J. and Szlam, A. (2017). Automatic rule extraction from long short term memory networks. In *International Conference on Learning Representations*, Toulon, France.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco, California.

[Shrikumar et al., 2017] Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153, Sydney, Australia.

[Sundararajan et al., 2017] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, Sydney, Australia.

[Zhang et al., 2016] Zhang, J., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S. (2016). Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*, pages 543–559, Amsterdam, Netherlands.