Investigating the interpretability of hidden layers in deep text mining

Stephan Raaijmakers
TNO, Data Science Department, The Hague, The Netherlands

Maya Sappelli
TNO, Data Science Department, The Hague, The Netherlands

Wessel Kraaij
TNO, Data Science Department, The Hague, The Netherlands
LIACS, Leiden University

ABSTRACT

In this short paper, we address the interpretability of hidden layer representations in deep text mining: deep neural networks applied to text mining tasks. Following earlier work predating deep learning methods, we exploit the internal neural network activation (latent) space as a source for performing k-nearest neighbor search, looking for representative, explanatory training data examples with similar neural layer activations as test inputs. We deploy an additional semantic document similarity metric for establishing document similarity between the textual representations of these nearest neighbors and the test inputs. We argue that the statistical analysis of the output of this measure provides insight to engineers training the networks, and that nearest neighbor search in latent space combined with semantic document similarity measures offers a mechanism for presenting explanatory, intelligible examples to users.

CCS CONCEPTS

• **Computing methodologies → Neural networks**;

ACM Reference format:
Stephan Raaijmakers, Maya Sappelli, and Wessel Kraaij. 2017. Investigating the interpretability of hidden layers in deep text mining. In *Proceedings of Semantics2017, Amsterdam, Netherlands, September 11–14, 2017,* 4 pages.
https://doi.org/10.1145/3132218.3132240

1 INTRODUCTION

In the current wave of *human-aware* and deep learning-dominated AI [4], explainability of analyses becomes an important factor for the acceptance of AI by humans. Yet, with deep learning models becoming increasingly complex cognitive architectures, explainability becomes challenging, and the concept itself is elusive and often not clearly defined (see e.g. [8] for a critical review). A common thread running through the various current approaches to making AI explainable, i.e. transparent to humans, is the desire to monitor the information a trained network stores internally and deploys for analyzing new input. Explainable AI leads to a deeper insight in the nature of cognition, may lead to more accurate systems by addressing their weaknesses, and may become a necessity in the light of initiatives such as the upcoming EU General Data Protection Regulation (GDPR, [3]), where citizens obtain the "right to explain" for algorithms working on their data. Since the information encoded in machine learning models typically is of a low-level nature, special steps need to be undertaken, such as making connections with human-interpretable representations, preferably with overt semantics. One of the tenets of deep learning is that hidden layers close to the output layer encode semantically more abstract information compared to hidden layers close to the input layer. In deep learning image processing, for instance, visualizations of higher layers indeed reveal the encoding of higher-order visual information, such as edges, or human-interpretable concepts like facial parts (cf. [16]). Such inspectability of the internal data representations of a deep neural network allows for a certain amount of transparency or *explainability* of the network to users. Engineers, for instance, may optimize a deep learner by taking a peek at its internal data representations ([14]). For text mining, transparency of network-internal representations is not trivial. Usually, input texts are encoded into vectors with neural word embeddings (produced by e.g. word2vec [9]). These vectors encode a shallow form of semantics based on distributional information: two words are semantically 'similar' if they occur roughly in the same context. Neural word embeddings like word2vec are learned with shallow neural networks. It is hard to imagine what document vector representations encode once they get processed by several layers in a deep learner, and whether the depth of the network impacts the abstractness of the information encoded at the various hidden layers. The purpose of this paper is to assess empirically if and how we can make the information encoded in deep neural networks for text analysis available for human inspection. Specifically, we are interested in *case-based explainable text mining*: explanations for analyses of test data based on similarities with training data. This paper addresses the following research questions:

(1) Can we make the internal activations of deep learning networks for textual analysis human-interpretable, i.e. translate them back to textual representations?
(2) How can we leverage the information encoded in a neural network for explanatory purposes, for both engineers and end-users?

We demonstrate that by applying nearest neighbor methods, we can indeed associate internal network activations with human-interpretable text: the text of nearest neighbors in the training data. We argue that a statistical comparison of the semantic distance of these nearest neighbors with input cases provides important insights to engineers pertaining to training/test data matches. Further, we demonstrate the explanatory use of these nearest neighbors for end-users. We stress from the onset that our results are initial, and

Semantics2017, September 11–14, 2017, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5296-3/17/09…$15.00
https://doi.org/10.1145/3132218.3132240

need to be manually evaluated by humans; yet, our results display many interesting observations, and give rise to new research questions. As for related work, in [8], an overview of various approaches to explainable AI is presented, with critical remarks on the proper delineation of the topic. The topic of explainability is diverse, and -given our current lack of understanding what good explanations actually are- it is not surprising that there is an abundance of different approaches. Attention-based models like [10] attempt to model the selective focus of deep learners on certain 'important' aspects of input data over time. Related approaches like [17] identify latent factors in input data for explainable recommendation engines. Our work builds upon the early work of [1], antedating the rise of deep learning methods, which presents the idea of using internal activation layers-based nearest neighor search for case-based explanation of machine learning methods. Our approach explicitly allows for measuring the dynamics of nearest neighbors in the network: nearest neighbors may fluctuate across layers, or become stable and migrate from one layer to another.

2 SEMANTIC DOCUMENT DISTANCE

In [7], a variant of the Earth Movers Distance, especially tailored for measuring textual similarity, was proposed. This distance measure, the Word Movers Distance (WMD), takes into account the word2vec representations of separate words in different documents, and is proportional to the effort it takes to transform the word2vec representations from one document into those of the other document.[1] WMD is defined as follows.

$$\begin{aligned} & WMD(x_i, x_j) = \min_{T \geq 0} \sum_{i,j=1}^{n} T_{ij} \mid\mid x_i - x_j \mid\mid_2 \\ & \text{subject to } \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i^a \text{ and } \sum_{i=1}^{n} \mathbf{T}_{ij} = d_j^b \end{aligned} \tag{1}$$

Given $X \in \mathbb{R}^{d \times n}$, a word2vec embedding for a vocabulary of n words, with d the dimension of the word2vec embedding, a vector $x_i \in \mathbb{R}^d$ is the word2vec representation in X of the i-th word. Further, d_i^a is the normalized frequency of the i-th word in the bag of word representation d^a of document a. $\mathbf{T} \in \mathbb{R}^{n \times n}$ is a transport matrix that determines how much of this frequency mass needs to be transported between documents a and b in order to minimize the objective function D: $\mathbf{T}_{ij}$ describes the transport of frequency mass from d_i^a to d_i^b. As an example, [7] discusses the following pair of sentences: *Obama speaks to the media in Illinois* and *The President greets the press in Chicago.*. The semantically close distance between these two sentences cannot be described by a simple *bag-of-words* model alone (describing exact matches). The use of a semantic distance measure (word2vec-based) alleviates this problem, and addresses the similarity for the word pairs (Obama, President); (speaks, greets); (media, press); and (Illinois, Chicago). We will refer to the vector space of activations of a neural network, consisting of the activations of a pre-specified layer in response to input from lower layers, as *latent space*. In the latent space for a given training dataset, we first fit a nearest neighbor tree that determines per activation vector (for a given training instance) the nearest activation space neighbors. Subsequently, for all training instances x, we determine the nearest neighbor set in the latent space, across different depths. We reason back from these nearest neighbors to the underlying textual representation in the training data. In order to evaluate the semantic relation between the input case and its nearest neigbors at the different layer depths, we apply WMD to the original input instance text and the text of the nearest neighbor, and analyze statistically the progression of distance scores across layers. The k-NN algorithm is outlined in Algorithm 1.

```
Input  : Training data ($X_{train}$, $y_{train}$) and test data ($X_{test}$, $y_{test}$),
         with $X$ data vectors and $y$ the labels; parameter $k$ for
         number of nearest neighbors; layer depth $d$; deep learner $D$;
         $f_{\alpha,M}(x, d)$ obtaining the activation for input $x$ at network
         layer $d$, using the model $M$ of trained deep learner $D$; a word
         embedding $E$.
Output : Per layer, per test case $x$, the textual representation $N$ of the
         nearest neighbor of $x$ in the training data ($W$), and the
         WMD distance between $x$ and $N$.
begin
    $A \longleftarrow \emptyset$;
    $W \longleftarrow \emptyset$;
    $L \longleftarrow \emptyset$;
    Train $D$ on ($X_{train}$, $y_{train}$), deriving model $M$.
    for $x \in X_{train}$ do
        $A \longleftarrow A \cup \{f_{\alpha,M}(x, d)\}$
        $L \longleftarrow L \cup \{< f_{\alpha,M}(x, d), x >\}$
    end
    Fit a 1-NN tree to $A$, yielding $f_{NN} : A \mapsto A$.
    for $x \in X_{test}$ do
        $N = L(f_{NN}(f_{\alpha,M}(x, d)))$
        $W \longleftarrow W \cup \{< x, d, N, WMD(x, N) >\}$
    end
end
```

Algorithm 1: Latent space k-NN.

3 EXPERIMENTS

We apply a deep neural network to a number of text mining datasets, with the purpose of 'explaining' analyses of test data inputs with representative, semantically related and formally similar training data examples. Our deep neural network is a generic deep multilayer perceptron with 5 hidden activation layers. Its structure is depicted in Figure 1. We made no effort to optimize its architecture for the datasets at hand. For every dataset, the network was run for 100 iterations. It was trained on 80% of the data (with 10-fold cross-validation), and, once trained, tested on the remaining 20%. We used a fixed batch[2] size of 32. The various activation layers deploy the ReLU (rectified linear unit) activation function. The network was implemented in Python with the Keras[2] deep learning library. Our data consists of the 6 datasets listed in Table 1. We represent every text as the summed word2vec vectors of its words, which produces a vector of dimension 300[3]. Once the model has been trained, the latent-space nearest neighbor search is applied

[1] word2vec expresses semantic relations between words based on distributional similarity.

[2] The number of training examples in a single forward/backward pass of the network during training.

[3] Our word2vec model is based on pre-trained, 300-dimensional Google News vectors (`GoogleNews-vectors-negative300.bin`, see [5]).

to produce textual training data correlates for the test data. These textual nearest neighbors are subjected to the semantic distance measure WMD, and the distances of the nearest neighbors produced per layer are analyzed statistically, in order to monitor the progression of semantic distances across the network.

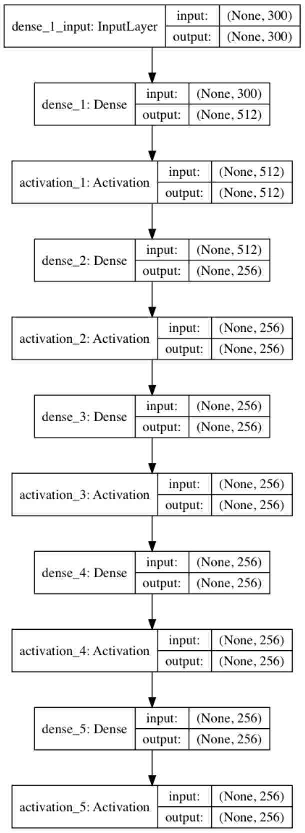

Figure 1: Our deep MLP.

Dataset	Description	Labels	Instances
Yelp[6]	Polarity (sentiment) labeled restaurant reviews	2	1,000
Amazon[6]	Polarity (sentiment) labeled product reviews	2	1,000
IMDB[6]	Polarity (sentiment) labeled movie reviews	2	1,000
Subj[11]	Subjectivity labeled movie reviews (sentences)	2	10,000
Pol[12]	Polarity (sentiment) labeled movie reviews (sentences)	2	10,662
Intent[13]	Email snippets (sentences) labeled for the presence of intent (speech acts)	2	3,656

Table 1: Datasets.

For analyzing the process of nearest neighbor formation in our network, we applied a one-tailed Wilcoxon rank sum test. This test evaluates the null hypothesis that two populations have the same distribution with the same median. In our case, we compare the semantic distances of the textual representations (nearest training neighbors) of two consecutive hidden layers with the text of the input (test data) cases. For all input test cases, we activate hidden layers l and $l+1$, compute the nearest neighbors in latent space, and gather the WMD distances of these neighbors to the input as two populations. The Wilcoxon test measures the difference between these populations. The result of the analysis is given by Table 2. In Pol, Subj and Intent, semantic distances increases monotonously up to the highest layer, with statistical significance. In Yelp, this effect occurs for the first 4 layers; it is virtually absent in IMDB (only occurring for one pair of layers). In Amazon, we observe it for two pairs of layers. If, for any pair of layers, the null hypothesis of the Wilcoxon test is rejected, we typically have a case where nearest neighbors for a certain layer differ significantly (according to the WMD distance) with the input case[4]. In the context of WMD, this indicates 'semantic drift': a larger WMD distance between two documents means they are semantically more remote. Nearest neighbors in this case may be semantically related but not superficially similar to the input cases. On the other hand, if the null hypothesis of the Wilcoxon test is confirmed, distances remain fairly constant across two layers l and $l + 1$. This indicates persistence of nearest neighbors: the network settles strongly on a nearest neighbor that 'survives' from l up to $l + 1$. It suggests the training data fits quite tightly with the test data, and, if observed in the initial activation layer (directly following the input layer), nearest neighbors interpretations of the first layer can be quite literal. The dynamics of neighbor distance in the network may reveal aspects of the learning process. For instance, we observed the following case in the Amazon data (notice the typo in 'satisifed'):

Input: `I'm very disappointed with my decision.`
Layer 1: `I am very pleased with my purchase.`
Layer 2: `Very satisifed with that.`
Layer 3: `very disappointed.`
Layer 4: `very disappointed.`
Layer 5: `very disappointed.`

which shows that, after an initial derailment at the first three layers, the network manages to restore a plausible nearest neighbor in the higher layers. The exact study of the dynamics of this phenomenon is beyond the scope of this paper. A simple heuristic that searches for the nearest neighbor in the set of neighbors produced by all layers[5] with minimal distance compared to the input case produces interesting results, however (see Table 3 for semantically similar, yet orthographically diverse examples).

4 CONCLUSIONS

In this paper, we proposed a nearest neighbor mechanism in the activation space of deep neural networks for text mining. The mechanism deploys a semantic document similarity metric, and identifies statistical patterns of neighbor similarity across hidden layers. The neighbor search can be operationalized for users by displaying the best matching neighbors as explanatory examples. The use of

[4] A negative z-value indicates for a one-tailed test that the second distribution has a higher median than the first distribution.

[5] In our current setup: a set of 5 nearest neighbors, since we have 5 hidden layers, each producing one nearest neighbor.

Dataset	layers 1, 2	layers 2, 3	layers 3, 4	layers 4, 5
Pol	$p < .001$ z=-9.28	$p < .001$ z=-6.98	$p < .001$ z=-4.30	$p < .004$ z=-2.65
Subj	$p < .001$ z=-7.12	$p < .001$ z=-5.27	$p < .001$ z=-4.50	$p < .04$ z=-1.78
Intent	$p < .001$ z=-3.90	$p < .02$ z=-2.06	$p < .005$ z=-2.62	$p < .02$ z=-2.3
Amazon	$p < .25$ z=-0.70	$p < .03$ -z=2.0	$p < .02$ z=-2.13	$p < .16$ z=-0.99
IMDB	$p < .12$ z=-1.19	$p < .04$ z=-1.86	$p < .49$ z=-0.05	$p < .41$ z=-0.25
Yelp	$p < .002$ z=-2.82	$p < .01$ z=-2.33	$p < .03$ z=-1.89	$p < .27$ z=-0.63

Table 2: One-tailed Wilcoxon rank sum results for paired activation layers.

Dataset	**Input**	**Nearest neighbor**
Yelp	Ordered burger rare came in we'll done	Damn good steak
	The ambience is wonderful and there is music playing	The decor is nice, and the piano music soundtrack is pleasant
Amazon	After receiving and using the product for just 2 days it broke	I purchased this and within 2 days it was no longer working!
	I tried talking real loud but shouting on the telephone gets old and I was still told it wasn't great	People couldnt hear me talk and I had to pull out the earphone and talk on the phone
IMDB	It's like a bad two hour TV movie.	I wouldn't say they're worth 2 hours of your time, though
	a captivating new film	a compelling film
Subj	a good-natured ensemble comedy that tries hard to make the most of a bumper cast , but never quite gets off the ground	attal pushes too hard to make this a comedy or serious drama . he seems to want both , but succeeds in making neither
	just about the best straight-up , old-school horror film of the last 15 years	a haunting film about one of the great escapes of all time
Pol	entertaining enough , but nothing new	interesting , but not compelling
	a whale of a good time for both children and parents seeking christian-themed fun	boomers and their kids will have a barrie good time
Intent	If you do not wish to receive further communications like this, please click here to unsubscribe	If you would no longer like to receive such offers, please click on the link and and we will remove you
	Please review and give me you comments	Please give your thoughts on this

Table 3: Illustrative input/output examples for all 6 datasets, thresholded on WMD-scores ($\theta < 1.0$).

rank sum tests on pairs of deep layers provides a facility for inspecting eventual persistence of information across layers, and may contribute to estimating uncertainty (or confidence) of a network in its internal semantic representations. Our results are currently presented anecdotically, and are in need of human evaluation and comparison to a baseline. Nonetheless, many striking cases emerge from our results. Our future work will consist of the human evaluation of the use of rank sum tests for network optimization purposes, and the explanatory merits of the examples. Additional topics to be addressed in our future work include the analysis of the semantic relations between explanatory nearest neighbors and the input cases, the replicability of our observations for more elaborate, fine-tuned neural network architectures, and the use of semantic distance measures based on lexical semantics (such as WordNet). Our current results and code are available from GitHub ([15]).

ACKNOWLEDGMENTS

This research was sponsored by the TNO Programme Big Data Analytics for Defense and the TNO Early Research Programme for Applied AI.

REFERENCES

[1] Rich Caruana, Hooshang Kangarloo, John D. N. Dionisio, Usha Sinha, and David B. Johnson. 1999. Case-based explanation of non-case-based learning methods. In *AMIA 1999, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 6-10, 1999.*

[2] François Chollet et al. 2015. Keras. https://github.com/fchollet/keras. (2015).

[3] EU. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* L119/59 (May 2016).

[4] Eugene C. Freuder. 2017. Explaining Ourselves: Human-Aware Constraint Reasoning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.* 4858–4862.

[5] Google. 2016. word2vec. https://code.google.com/archive/p/word2vec/. (2016).

[6] Dimitrios Kotzias, Misha Denil, Nando de Freitas, and Padhraic Smyth. 2015. From Group to Individual Labels Using Deep Features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15).* ACM, New York, NY, USA, 597–606. https://doi.org/10.1145/2783258.2783380

[7] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. 2015. From Word Embeddings To Document Distances. In *Proceedings ICML.* 957–966.

[8] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *CoRR* abs/1606.03490 (2016). http://arxiv.org/abs/1606.03490

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems.* 3111–3119.

[10] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent Models of Visual Attention. *CoRR* abs/1406.6247 (2014). http://arxiv.org/abs/1406.6247

[11] Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity. In *Proceedings of ACL.* 271–278.

[12] Bo Pang and Lillian Lee. 2005. Seeing Stars: Exploiting Class Relationships For Sentiment Categorization With Respect To Rating Scales. In *Proceedings of ACL.* 115–124.

[13] ParakweetLabs. 2014. EmailIntentDataSet. https://github.com/ParakweetLabs/EmailIntentDataSet. (2014). Accessed: May 2017.

[14] Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the Black Box of Deep Neural Networks via Information. *CoRR* abs/1703.00810 (2017). http://arxiv.org/abs/1703.00810

[15] TNO. 2017. Deeptext, explainable deep text mining. https://github.com/stephanraaijmakers/deeptext. (2017).

[16] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding Neural Networks Through Deep Visualization. In *Deep Learning Workshop, International Conference on Machine Learning (ICML).*

[17] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit Factor Models for Explainable Recommendation Based on Phrase-level Sentiment Analysis. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval (SIGIR '14).* ACM, New York, NY, USA, 83–92.