What Part of the Neural Network Does This? Understanding LSTMs by Measuring and Dissecting Neurons

Ji Xin, Jimmy Lin, and **Yaoliang Yu**
David R. Cheriton School of Computer Science
University of Waterloo
{ji.xin,jimmylin,yaoliang.yu}@uwaterloo.ca

Abstract

Memory neurons of long short-term memory (LSTM) networks encode and process information in powerful yet mysterious ways. While there has been work to analyze their behavior in carrying low-level information such as linguistic properties, how they directly contribute to label prediction remains unclear. We find inspiration from biologists and study the affinity between individual neurons and labels, propose a novel metric to quantify the sensitivity of neurons to each label, and conduct experiments to show the validity of our proposed metric. We discover that some neurons are trained to specialize on a subset of labels, and while dropping an arbitrary neuron has little effect on the overall accuracy of the model, dropping label-specialized neurons predictably and significantly degrades prediction accuracy on the associated label. We further examine the consistency of neuron-label affinity across different models. These observations provide insight into the inner mechanisms of LSTMs.

1 Introduction

In recent years, the application of deep learning to natural language processing (NLP) has been a success. Many consider the employment of distributed representations to be one of the reasons for deep learning's success (LeCun et al., 2015; Young et al., 2018). However, how these distributed representations encode information in deep neural networks, especially long short-term memory (LSTM) networks that are prevalent in NLP, still remains unclear (Feng et al., 2018). One of the potential ways to understand how neural networks function is to analyze the behavior of individual neurons that carry the distributed representation. While there have been a number of works that analyze low-level information stored in individual LSTM neurons, such as linguistic properties (Bau et al., 2019; Qian et al., 2016), syntax of source code (Karpathy et al., 2015), and sentiment (Radford et al., 2017), how each neuron contributes directly to the final classification layer remains unclear.

We find inspiration to analyze individual neurons of LSTMs from how biologists analyze neurons of roundworms (White et al., 1986). Biological neural systems consist of a huge number of neurons, and can react to the environment in complicated ways. Biologists start with analyzing basic components of reactions, what stimuli trigger them, and which neurons are excited during the process. To verify the relationship between stimuli, neurons, and reactions, biologists further dissect neurons which are correlated with specific basic reactions, and see if the reaction still occurs for the same stimuli.

We adopt the same methodology to study LSTMs, using a representative task in NLP: named-entity recognition (NER) (Ratinov and Roth, 2009; Lample et al., 2016). Even though the output of a neural network may be complicated, we focus on basic components of the output: whether a label is predicted or not. We feed into the neural model various input instances, and analyze the relationship between the value of each LSTM neuron and the predicted label. We quantify the sensitivity of neurons to each label, and study how label-specific information is distributed among all neurons. We discover that each individual neuron is specialized to carry information for a subset of labels, and the information of each label is only carried by a subset of all neurons. We further conduct experiments to gradually drop out individual neurons. This significantly lowers the accuracy of labels that the neuron is specialized on, while having little effect on the overall performance of the model. We also study the corre-

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5823–5830, Hong Kong, China, November 3–7, 2019.

lation between labels, and discover some patterns that are shared among different models.

Our contributions are as follows: (1) To the best of our knowledge, we are the first to have taken this neuron-label affinity focused approach to understanding the inner workings of LSTMs. (2) We propose a novel metric to quantify such affinity, and conduct experiments to verify the validity and consistency of this metric.

2 Related Work

Recently, work has been done to analyze continuous representations in NLP. Shi et al. (2016) and Qian et al. (2016) analyze linguistic properties carried by representation vectors using external supervision. Bau et al. (2019) and Dalvi et al. (2019) further analyze linguistic information in individual neurons from neural machine translation representations in an unsupervised manner. For LSTMs of language models, Karpathy et al. (2015) identify individual neurons that trigger for specific information, such as bracket and sequence length, and Radford et al. (2017) discover neurons that encode sentiment information.

In computer vision, Zhou et al. (2018) analyze the relationship between individual units of a CNN and label prediction. To the best of our knowledge, however, in the field of NLP, there has been little work on analyzing the affinity between labels and neurons of recurrent networks. This paper aims to address this problem.

3 Model and Experiments

3.1 Model Selection

Named-entity recognition is a sequence labeling task. The input of the model is a sequence of words $x^{(t)}, t = 1, 2, \cdots$. Each input word has a corresponding label $z^{(t)} \in \mathcal{L}$, where $\mathcal{L}$ is the set of all labels $\{l_j\}, j = 1, 2, \cdots, m$. The label indicates whether the word is an entity or not, and if yes, which kind of entity it is.

A typical modern NER model consists of a bi-directional LSTM and a conditional random field (CRF) on top of the LSTM (Collobert et al., 2011; Huang et al., 2015). Sometimes there is also a convolutional neural network (CNN) (Ma and Hovy, 2016; Chiu and Nichols, 2016). However, the goal of this paper is not to achieve state-of-the-art performance on this task, but rather we are trying to understand the mechanisms of LSTMs. Therefore, we choose a relatively simple model (see Figure 1) for the experiments: a single layer uni-directional LSTM with a fully-connected layer on top of it.

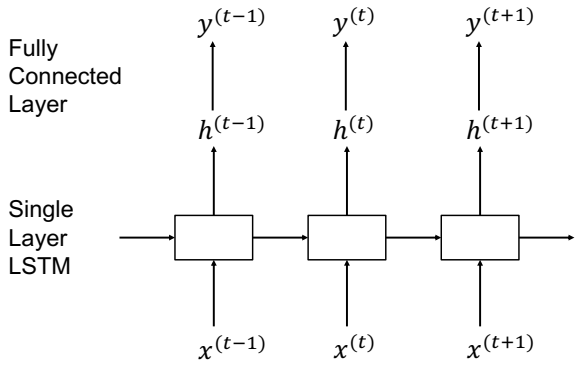

Figure 1: The model we use for this task: a single layer uni-directional LSTM with a fully-connect layer on top of the LSTM.

We denote $\mathbf{h}^{(t)} \in \mathbb{R}^n$ as the LSTM's hidden state at timestep t, and $\mathbf{h}_i^{(t)}$ its i-th entry. $\mathbf{W} \in \mathbb{R}^{n \times m}$ is the weight matrix of the fully-connected layer. The output of the entire model at timestep t is therefore the vector

$$\mathbf{y}^{(t)} = \text{Softmax}(\mathbf{W}^\top \mathbf{h}^{(t)}), \quad (1)$$

where $\mathbf{y}^{(t)} \in \mathbb{R}^m$ and each entry is the predicted probability of a label in $\mathcal{L}$:

$$\Pr(z^{(t)} = l_j) = \mathbf{y}_j^{(t)} \quad (2)$$

$$= \frac{\exp\left(\mathbf{W}_{:,j}^\top \mathbf{h}^{(t)}\right)}{\sum_{j'} \exp\left(\mathbf{W}_{:,j'}^\top \mathbf{h}^{(t)}\right)}, \quad (3)$$

where $\mathbf{W}_{:,j}^\top$ is the transpose of the j-th column vector of the matrix $\mathbf{W}$. The final prediction $\tilde{z}^{(t)}$ is chosen as the label with greatest probability.

3.2 Experiment Setup

The model is trained on the CoNLL2003 (Sang and De Meulder, 2003) training dataset. Development and test sets of CoNLL2003 will be used in experiments in Section 4. In this dataset there are nine labels in total, under the BIO tagging schema. See the first row of Figure 3 for the complete set of labels.

Code for this paper is adopted from the toolkit by Yang and Zhang (2018). We set the hidden size of the LSTM to 50, since a larger hidden size does not significantly improve the results. Other hyper-parameters, such as learning rate, batch size, and drop out rate, are kept unchanged. The model is trained for 10 epochs, and we pick the checkpoint

cell	B-PER	I-PER	B-LOC	I-LOC	B-ORG	I-ORG	B-MISC	I-MISC	O
7	0.10	-0.01	0.02	0.01	0.08	0.01	-0.08	0.00	0.10
10	0.10	0.18	0.06	0.02	0.07	0.13	1.05	-0.01	-0.01
13	0.72	1.81	0.09	0.00	0.00	0.05	-0.01	0.00	-0.03
17	0.03	0.08	1.17	0.13	0.08	0.15	-0.01	1.29	0.09
19	0.07	0.25	0.27	-0.19	0.85	-0.09	0.50	0.08	2.40
24	-0.04	0.00	0.03	0.06	-0.04	-0.03	0.02	0.17	-0.03
27	-0.23	0.06	0.21	-0.23	-0.11	0.11	1.29	0.05	1.35
34	-0.04	0.26	0.24	0.02	0.04	-0.05	1.39	0.12	0.13
37	1.23	0.11	-0.02	-0.04	0.74	-0.08	0.01	0.04	0.56
48	0.17	0.30	0.00	-0.05	-0.02	0.14	1.09	0.25	0.57

Figure 2: Sensitivity of the top ten neurons for all labels; red for positive values, blue for negative ones, and deeper colors stand for larger absolute values.

based on the best F1 score on the development set. The chosen checkpoint has an F1 score of 86.4. For comparison, the F1 score obtained by the same toolkit using a bi-directional LSTM and a CRF is 89.5 (Yang et al., 2018).

4 Analyzing Neuron-Label Affinity

In this section, we first identify important neurons by quantifying the sensitivity of a neuron to a label, and then verify the quantification by neuron ablation experiments.

4.1 Identifying Important Neurons

A neuron of an LSTM corresponds to an entry (dimension) of $\mathbf{h}^{(t)}$. For a certain label l_j, we try to identify neurons that are important for its prediction in the following way.

We define the *contribution* of the i-th neuron to the j-th label at timestep t as

$$u_{i,j}^{(t)} = \mathbf{W}_{i,j}\mathbf{h}_i^{(t)}. \quad (4)$$

Note that contribution is defined with the number after multiplied by $\mathbf{W}$ in the fully-connected layer. Therefore the contribution value itself is what matters, not its absolute value.

The *sensitivity* of the i-th neuron to the j-th label is further defined as

$$s_{i,j} = \mathbb{E}(u_{i,j}^{(t)}|\tilde{z}^{(t)} = l_j) - \mathbb{E}(u_{i,j}^{(t)}|\tilde{z}^{(t)} \neq l_j), \quad (5)$$

where $\mathbb{E}$ stands for taking average over t. This is the difference of the mean contribution over l_j entity words versus other words. The higher $s_{i,j}$ is, the more sensitive the i-th neuron is for predicting the label l_j.

We compute $s_{i,j}$ for all i and j pairs, and the average is done over the entire development set. A part of the results is shown in Figure 2, and the full results are shown in the appendix.

From the figure we can see that information is distributed across different neurons in a highly non-uniform way:

B-PER	I-PER	B-LOC	I-LOC	B-ORG	I-ORG	B-MISC	I-MISC	O
11	13	45	28	1	28	34	28	19
44	9	29	14	19	44	27	17	42
37	44	17	3	37	16	48	30	32
9	14	2	30	28	38	10	16	39
13	31	39	31	9	36	29	29	23
42	11	32	44	47	3	45	42	27
29	30	38	45	44	12	28	45	22
46	36	21	25	32	43	30	26	46
14	16	9	41	6	33	19	14	38
2	40	23	2	21	47	21	1	20

Figure 3: Top ten neurons in terms of importance rankings for all labels. The green and yellow labels are the ones used in the experiments described in Section 4.2.

- Each neuron has a different sensitivity to different labels. Some neurons are only sensitive to one label, e.g., neuron #10 for B-MISC; some are sensitive to multiple labels, e.g., neuron #17 for B-LOC and I-MISC; some are even not sensitive to any, e.g., neuron #7.

- For each label, there are multiple neurons that are sensitive to it, as well as multiple neurons that are not.

From these, we can come to the conclusion that the prediction of each label is based on information that is distributed among multiple, but not all, neurons. Furthermore, different types of information are distributed differently.

For each label, we further rank all neurons based on their sensitivity, and obtain an *importance ranking* for the label. The top ten neurons for each label are shown in Figure 3, and the full results are shown in the appendix.

4.2 Verifying the Importance of Neurons

We try to verify whether the *sensitivity* we define in the previous subsection is a valid and consistent indicator of a neuron's importance for a label.

The way to do this is to perform model evaluation on the test set,[1] while incrementally ablating neurons[2] from the model, in a certain order. If the sensitivity of neurons we obtain is valid and consistent, when we ablate neurons in the order of *importance ranking* of the label l_j, the performance on the test set should drop fastest for predicting l_j, and slower for other labels.

We choose two pairs of labels: (B-PER, B-MISC) and (B-LOC, I-ORG). In each pair, we conduct neuron ablation according to each label's

[1] Recall that we obtain the value of sensitivity only from the development set.

[2] Ablating a neuron here means setting $\mathbf{h}_i^{(t)}$ to 0.

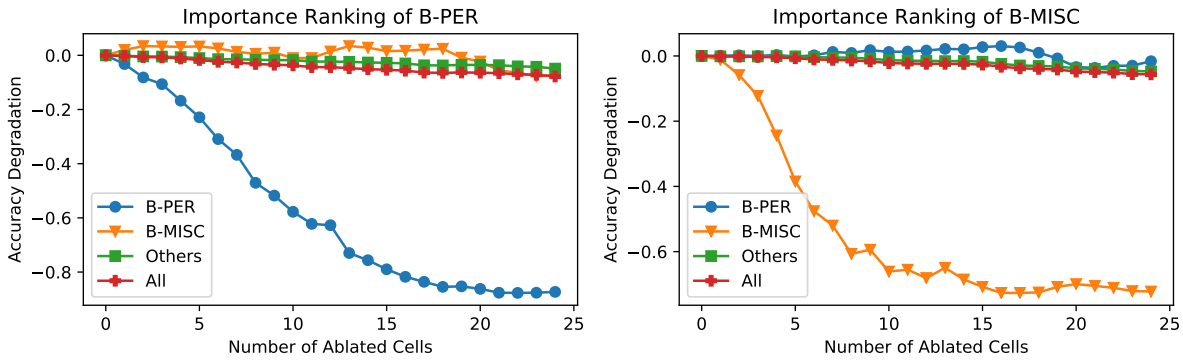

Figure 4: Ablation according to importance ranking of B-PER and B-MISC (yellow columns in Figure 3).

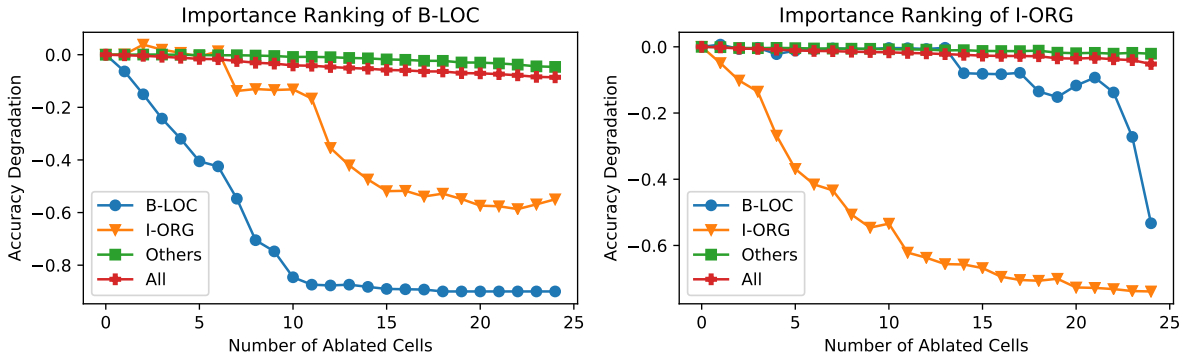

Figure 5: Ablation according to importance ranking of B-LOC and I-ORG (green columns in Figure 3).

importance ranking, and compare the model's performance for predicting each label. The results are shown in Figures 4 and 5. We only show the first half of the importance ranking, since the latter half not only is less important, but also has more overlap between different labels.

From the figures we can see that when ablating neurons according to a certain label's importance ranking, the accuracy of the label drops much faster than the other labels. The overall performance, however, remains more or less unaffected. This shows that while a single neuron can be important for a subset of labels, the overall performance is more robust to neuron ablation. This further verifies our observations from the previous subsection: information is distributed among multiple neurons in various ways. A neuron may have encoded important information for a certain label, but it is unlikely that all important information is concentrated in one neuron.

It is worth noting that a neuron can be important for multiple labels. Therefore, when ablating neurons according to one label's importance ranking, the performance for other labels may also degrade. This can be seen in the left plot of Figure 5. Neuron #38 appears in both the top ten lists of B-LOC and I-ORG (shaded boxes in Figure 3), and when it is ablated (the seventh ablated neuron), not only the performance of B-LOC, but also that of I-ORG, is compromised. The fourth ablated neuron in the right plot of Figure 5 has a similar behavior, but it is less significant, probably because this neuron is ranked seventh for B-LOC and is therefore less important than it is for I-ORG. This phenomenon is less significant in Figure 4, since the top neurons from importance rankings of B-PER and B-MISC have fewer overlaps.

4.3 Correlation Between Labels

Even though the distribution of information in neurons may seem arbitrary, we want to see if multiple, independently-trained models share any common traits.

In addition to the model we have used in previous sections, we train three more models with the same model architecture and hyperparameters but different random seeds. We compute neuron-label sensitivity for all four models using both develop-

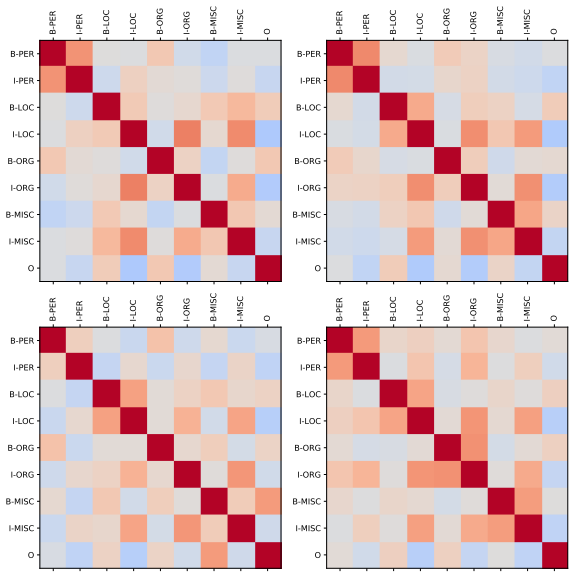

Figure 6: Correlation between all labels for four different models.

ment set and test set. For each of the models, we compute the correlation between all labels among different neurons. The sensitivity matrix for each model has 50 rows (neurons) and 9 columns (labels); see Figure 7 in the appendix for example. Correlation is computed among all rows in the matrix. The results are shown in Figure 6.

While there are differences among the four correlation plots, they share the following patterns:

- Label pairs of the form B-x and I-x (where x is PER/LOC/ORG/MISC) are generally positively correlated. We can observe some dark-red 2×2 blocks on the diagonal. Although for each trained model, it might be different neurons (i.e., neuron #) that encode information about B-x, these neurons typically also carry information about I-x.
- The label triples I-LOC, I-ORG, and I-MISC are also positively correlated.
- Label pairs of the form B-x and I-y (where x and y are different entities) are generally negatively correlated, e.g., I-PER with any of B-LOC, B-ORG, and B-MISC.
- The label O is negatively correlated with all I-x labels.

Although it remains unclear what information the neurons exactly encode, we speculate that there are at least two kinds of information, based on the observed patterns:

- Coarse-grain types of the current word. For example, whether the word is related to PER, or LOC/ORG/MISC, or O.
- Entity boundary location. If the previous prediction is O, it means the current word should be either another O or the left boundary of an entity, and thus the model should only predict O or B-x, but never I-x. Hence, I-x is negatively correlated with B-y and O.

5 Conclusion

In this paper, we try to understand the mechanisms of LSTMs by measuring and dissecting LSTM neurons. We discover that the prediction of each label is based on label-specific information, which is distributed among different groups of neurons. We propose a method to quantify and rank the importance of each neuron for each label, and further conduct ablation experiments to verify the validity and consistency of such importance rankings. Results show that the importance of a neuron is very different for different labels.

Future work. We consider the following three directions as future work. (1) While we now know how label-specific information is distributed among neurons and how important each neuron is to different labels, we can only make speculations about what the information is. It would be meaningful to study how the neurons are trained to encode label-specific information and what exactly the information is. (2) From the sensitivity figures, some neurons do not seem important to any label. It would be interesting to see what will happen if they are removed, and furthermore, how this can help model compression and hidden size selection. (3) Transformers (Vaswani et al., 2017; Devlin et al., 2019) have been increasingly popular in NLP, and it would be important to extend our work to understanding these architecture.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and enabled by computational resources provided by Compute Ontario and Compute Canada.

References

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*.

Jason P. C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James R. Glass. 2019. What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 6309–6317.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016. Analyzing linguistic knowledge in sequential model of sentence. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 826–835, Austin, Texas.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

John G White, Eileen Southgate, J. Nichol Thomson, and Sydney Brenner. 1986. The structure of the nervous system of the nematode Caenorhabditis elegans. *Philosophical Transactions of the Royal Society B*, 314(1165):1–340.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889, Santa Fe, New Mexico.

Jie Yang and Yue Zhang. 2018. NCRF++: an open-source neural sequence labeling toolkit. In *Proceedings of ACL 2018, System Demonstrations*, pages 74–79, Melbourne, Australia.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Revisiting the importance of individual units in CNNs via ablation. *arXiv preprint arXiv:1806.02891*.

cell	B-PER	I-PER	B-LOC	I-LOC	B-ORG	I-ORG	B-MISC	I-MISC	O
0	0.00	0.27	0.01	0.02	-0.05	0.45	0.00	0.26	0.03
1	-0.04	0.41	0.04	0.01	1.40	0.21	-0.05	0.36	0.09
2	0.37	0.03	0.92	0.62	-0.08	0.12	0.19	0.17	0.43
3	-0.13	0.37	-0.01	0.83	-0.03	0.64	-0.04	0.11	0.52
4	0.02	-0.01	0.01	0.02	0.03	0.35	0.05	0.02	0.10
5	0.22	0.26	0.07	0.38	0.11	-0.03	-0.02	0.27	0.21
6	0.14	0.08	0.06	0.00	0.45	0.04	-0.02	-0.02	0.51
7	0.10	-0.01	0.02	0.01	0.08	0.01	-0.08	0.00	0.10
8	-0.04	0.17	0.00	0.03	0.25	0.00	0.00	0.03	0.05
9	0.89	1.50	0.63	0.05	0.61	0.01	0.06	-0.01	0.00
10	0.10	0.18	0.06	0.02	0.07	0.13	1.05	-0.01	-0.01
11	1.93	0.94	-0.01	0.00	0.40	0.03	0.10	0.25	0.01
12	0.06	0.10	0.15	0.19	0.40	0.49	-0.05	-0.10	0.23
13	0.72	1.81	0.09	0.00	0.00	0.05	-0.01	0.00	-0.03
14	0.39	0.95	-0.06	1.30	-0.04	0.39	-0.04	0.36	0.04
15	0.29	-0.02	-0.01	0.20	0.36	0.19	0.01	0.01	-0.13
16	0.27	0.58	0.00	0.61	-0.07	0.75	0.02	0.84	0.01
17	0.03	0.08	1.17	0.13	0.08	0.15	-0.01	1.29	0.09
18	0.00	0.03	0.55	0.32	-0.01	0.15	0.11	-0.22	0.10
19	0.07	0.25	0.27	-0.19	0.85	-0.09	0.50	0.08	2.40
20	0.28	0.36	0.22	0.04	-0.10	-0.03	0.00	0.00	0.59
21	0.03	0.03	0.65	0.43	0.41	0.25	0.44	0.14	0.13
22	0.29	-0.04	0.00	0.01	0.13	0.00	0.25	0.00	0.83
23	0.12	0.13	0.59	-0.02	0.16	0.16	-0.03	-0.01	1.36
24	-0.04	0.00	0.03	0.06	-0.04	-0.03	0.02	0.17	-0.03
25	0.00	0.04	0.05	0.71	0.07	0.32	0.03	0.21	0.01
26	-0.01	0.17	0.03	0.23	0.00	0.27	0.04	0.37	0.05
27	-0.23	0.06	0.21	-0.23	-0.11	0.11	1.29	0.05	1.35
28	0.13	0.20	0.52	1.71	0.62	1.48	0.68	1.73	-0.04
29	0.65	0.19	1.34	0.49	0.02	0.11	0.81	0.56	0.02
30	0.19	0.75	0.25	0.80	0.03	0.40	0.52	1.05	0.03
31	0.04	0.94	0.11	0.80	0.15	0.13	0.39	0.27	-0.22
32	0.22	0.46	0.77	0.08	0.48	0.09	0.02	-0.13	1.88
33	0.20	-0.02	-0.02	0.00	0.24	0.47	-0.03	-0.03	-0.05
34	-0.04	0.26	0.24	0.02	0.04	-0.05	1.39	0.12	0.13
35	0.03	0.26	0.02	0.00	0.16	0.20	-0.01	-0.07	0.03
36	0.19	0.72	-0.02	0.50	0.04	0.69	0.05	0.11	0.02
37	1.23	0.11	-0.02	-0.04	0.74	-0.08	0.01	0.04	0.56
38	0.07	0.01	0.67	0.26	0.25	0.73	0.20	0.00	0.60
39	0.29	0.22	0.90	0.23	0.38	0.13	-0.15	0.01	1.74
40	0.01	0.47	0.00	0.01	0.01	0.00	0.21	0.01	-0.02
41	-0.02	0.17	-0.12	0.71	0.17	-0.02	0.39	-0.07	0.03
42	0.69	0.36	0.11	0.13	0.28	-0.26	0.08	0.51	1.93
43	0.07	0.14	0.38	-0.06	0.40	0.49	0.05	0.14	0.29
44	1.37	1.12	0.42	0.74	0.53	0.89	-0.05	-0.04	-0.02
45	0.00	0.06	1.50	0.72	0.00	0.36	0.73	0.49	0.12
46	0.65	0.41	0.00	0.00	0.09	0.05	0.01	0.00	0.69
47	0.18	0.00	0.00	0.02	0.54	0.47	0.01	0.00	-0.10
48	0.17	0.30	0.00	-0.05	-0.02	0.14	1.09	0.25	0.57
49	0.07	0.45	0.01	0.01	0.03	0.01	-0.03	-0.02	-0.10

Figure 7: Sensitivity of all neurons. This is the full result of Figure 2.

B-PER	I-PER	B-LOC	I-LOC	B-ORG	I-ORG	B-MISC	I-MISC	O
11	13	45	28	1	28	34	28	19
44	9	29	14	19	44	27	17	42
37	44	17	3	37	16	48	30	32
9	14	2	30	28	38	10	16	39
13	31	39	31	9	36	29	29	23
42	11	32	44	47	3	45	42	27
29	30	38	45	44	12	28	45	22
46	36	21	25	32	43	30	26	46
14	16	9	41	6	33	19	14	38
2	40	23	2	21	47	21	1	20
39	32	18	16	43	0	41	5	48
22	49	28	36	12	30	31	31	37
15	46	44	29	11	14	22	0	3
20	1	43	21	39	45	40	48	6
16	3	19	5	15	4	38	11	2
5	20	30	18	42	25	2	25	43
32	42	34	38	8	26	18	2	12
33	48	20	39	38	21	11	24	5
30	0	27	26	33	1	42	21	34
36	35	12	15	41	35	9	43	21
47	5	31	12	23	15	4	34	45
48	34	42	42	35	23	43	3	7
6	19	13	17	31	18	36	36	18
28	39	5	32	22	17	26	19	4
23	28	6	24	5	48	25	27	1
10	29	10	9	46	39	16	37	17
7	10	25	20	17	31	32	8	8
38	41	1	8	7	10	24	4	26
19	26	26	47	25	2	37	40	14
43	8	24	10	10	29	46	15	41
49	43	7	4	34	27	47	39	30
12	23	35	34	36	32	15	20	35
31	37	4	0	49	46	20	7	0
21	12	49	7	4	13	0	47	36
17	17	0	22	30	6	8	13	29
35	6	8	49	29	11	17	38	11
4	45	48	40	40	49	13	46	16
40	27	46	1	26	9	35	22	25
18	25	47	46	45	7	6	10	9
0	2	22	6	13	8	5	9	10
25	21	16	33	18	40	49	23	44
45	18	40	13	48	22	23	49	40
26	38	15	11	3	41	33	6	24
41	24	11	35	24	5	3	33	13
24	47	3	23	14	24	14	44	28
8	4	37	37	0	20	1	41	33
1	7	36	48	16	34	12	35	47
34	15	33	43	2	37	44	12	49
3	33	14	19	20	19	7	32	15
27	22	41	27	27	42	39	18	31

Figure 8: Importance rankings for all labels. This is the full result of Figure 3.