# Hidden Neuron Activation Analysis on Labeled Text Data

Avishek Das
Computer Science
Kansas State University
Manhattan, KS, USA
avishek@ksu.edu

Abhilekha Dalal
Computer Science
Kansas State University
Manhattan, KS, USA
adalal@ksu.edu

Pascal Hitzler
Computer Science
Kansas State University
Manhattan, KS, USA
hitzler@ksu.edu

## Abstract

Understanding the internal mechanisms of deep neural networks remains a central challenge in the field of Explainable Artificial Intelligence (XAI). With the rapid advancement of neural architectures in natural language processing (NLP), analyzing the role of hidden neurons in capturing and processing linguistic features has become increasingly important. This study investigates Hidden Neuron Activation Analysis on labeled text data to reveal how individual neurons contribute to a model's decision-making process. We propose a model-agnostic explainability framework for text classifiers that identifies concepts activating specific neurons involved in classification. An LSTM-based network is trained on the AG News topic classification dataset, comprising four distinct classes, and the final Dense layer with 64 neurons was analyzed. In addition, statistical analyses like the Mann-Whitney U Test is conducted to assess the robustness and reliability of the system. The statistical analysis shows that, concepts plays important role in the decision making process of neural network. Our findings enhance interpretability in NLP models and offer a foundation for optimizing neural architectures in text classification tasks.

## CCS Concepts

• **Computing methodologies** → **Information extraction**; **Causal reasoning and diagnostics**; **Neural networks**.

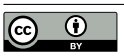
## Keywords

Explainable AI, Hidden Neuron Analysis, Dense Layer Analysis, Concept-based Explanation

**ACM Reference Format:**
Avishek Das, Abhilekha Dalal, and Pascal Hitzler. 2025. Hidden Neuron Activation Analysis on Labeled Text Data. In *Knowledge Capture Conference 2025 (K-CAP '25), December 10–12, 2025, Dayton, OH, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3731443.3771369

## 1 Introduction

Deep neural networks have revolutionized the field of NLP by achieving remarkable performance across a wide range of tasks including sentiment analysis, machine translation, and text classification. Despite their success, these deep learning solutions are often criticized for their "black box" nature, where the internal decision-making processes remain largely opaque. This lack of transparency has become a significant barrier to their adoption in sensitive domains such as healthcare, law, and finance. As a result, the field of Explainable Artificial Intelligence (XAI) has gained much spotlight in recent years, aiming to make deep learning models more interpretable and trustworthy [6, 33]. Traditional deep learning evaluations rely on statistical metrics, which often fail to explain specific model behaviors [7]. In contrast, concept-based XAI offers more intuitive, human-aligned explanations by mapping internal representations to discrete, interpretable units such as sentiment or linguistic roles [12, 23]. These models enable transparent reasoning by aligning neuron activations with semantically meaningful concepts. Notably, research has shown that even without supervision, hidden neurons can spontaneously encode human-interpretable concepts [4, 11]. In this study, we explore Hidden Neuron Analysis in text classification tasks to uncover the functional roles of individual neurons in trained NLP models. Our aim is to bridge the gap between performance and interpretability by identifying patterns or concepts associated with specific class-label activations.




*K-CAP '25, Dayton, OH, USA*

ACM ISBN 979-8-4007-1867-0/25/12
https://doi.org/10.1145/3731443.3771369

## 2 Related Works

There have been many works in the domain of XAI for image data, particularly in computer vision. Techniques such as saliency maps (e.g., Grad-CAM [27], LRP [21]), feature attribution(e.g., SHAP [17]), feature visualization [31], and concept attribution methods [9] have been widely used to identify which regions or visual features contribute most to a model's decision. In a work by Bau et al., they introduced network dissection to interpret convolutional neurons as detectors of human-understandable visual concepts such as textures and objects [5]. Similarly, concept-based approaches like TCAV [12] have provided intuitive explanations by quantifying the influence of user-defined concepts on model predictions. However, these methods have dependency on user-defined concepts and susceptibility to biased classifiers. Another work used concept induction method [13, 14] to generate explanations of an image classifier [8]. They used a Wikipedia-derived concept hierarchy [26] with approximately 2 million classes as background knowledge and ECII heuristic concept induction system [25]. Barua et al. used a somewhat similar framework of concept induction on image data but eliminated the use of Wikipedia background knowledge and instead used LLM prompting for generating concepts [3]. They found concepts from GPT-4 are more comprehensible to human compared to those generated by ECII.

XAI has also been applied to several studies on tabular data [24]. DeepLIFT, introduced by Shrikumar et al., explains neural network predictions on tabular data by assigning feature importance scores via Backpropagation [28]. Mollas et al. proposed a method for interpreting random forests using unsupervised learning and a "6-models clustering" algorithm to simplify decision paths [20].

Lundberg et al. developed Tree SHAP, an extension of SHAP for tree-based models, which computes exact Shapley values by leveraging the tree structure [16], though it can be computationally intensive. SHAP has also been applied to interpretable models like logistic regression [10, 29].

These methods have helped bridge the gap between model interpretability and human understanding in vision and tabular tasks and have motivated the application of similar ideas in the NLP domain. In the NLP domain, there are only a few works and most of them used SHAP or variations of it like, KernelSHAP, DeepSHAP, TreeSHAP [22]. Li et al. visualized hidden layer embeddings as heatmaps and t-SNE plots to analyze how a deep learning model handles Negation (e.g., "not good" vs. "not bad"), Intensification (e.g., "very good"), Clause composition (e.g., "I like it, but it's boring") [15]. The proposed methods offer initial insights but are limited in capturing the full complexity and nonlinearity of deep neural network behavior. Yeh et al. developed a concept based explanation method named ConceptSHAP, that does not require human-labeled concepts as they are discovered automatically from the model's internal activations using unsupervised learning [30]. However, the automatically discovered concepts are not always aligned with human-understandable semantics. Saliency maps methods like Grad-CAM and LRP were originally developed for image models, but have been adapted for analyzing DL-based text classifiers by attributing relevance to input tokens [2]. But these methods are only applicable to a CNN model with fixed pre-trained word embeddings.

In our study, we have developed a model-agnostic XAI pipeline for deep learning-based text classifiers that learns concepts from the given samples automatically based on the activation of the last dense layer of the model[1].

## 3 Methodology

This section covers the whole methodological workflow that we have conducted in our study. Fig. 1 shows the overall workflow and we will go through each steps of it. The workflow is inspired from the work by Dalal et al [8].

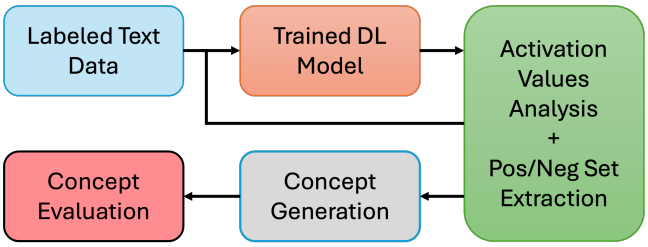



**Figure 1: Workflow for generating concepts from text data.**

**Labeled Text Dataset:** For the labeled text dataset we used the AG News Topic Classification dataset [32] consisting of 4 classes and a total of 127,600 data samples. Table 1 shows the distribution of the dataset. Some samples of the dataset are shown in Appendix A Table 5.

[1]Source code and result files are available online at https://github.com/avishek-018/text_hna

**Table 1: Distribution of AG News Topic Classification dataset.**

| Class | Training Samples | Testing Samples |
|---|---|---|
| World | 30,000 | 1,900 |
| Sports | 30,000 | 1,900 |
| Business | 30,000 | 1,900 |
| Sci/Tech | 30,000 | 1,900 |
| **Total** | **120,000** | **7,600** |

**Trained DL model:** We trained some popular and widely used deep learning models for the text classification like LSTM, BiLSTM, 1D CNN. The results on the test data are represented in Table 2 and we can see LSTM performed better. We kept the number of layers and hyperparameters same for all of the DL models and didn't emphasize on improving the performance of the DL models with hyperparameter tuning. This is because our target is to analyze the mechanism of the hidden layers' decision making process.

**Table 2: Performance of different DL models. W-Pr, W-Re and W-F1, Acc stands for Weighet Precision, Weighted Recall, Weighted F1-score and Accuracy respectively.**

| Model | W-Pr | W-Re | W-F1 | Acc |
|---|---|---|---|---|
| 1D CNN | 0.89 | 0.88 | 0.88 | 0.88 |
| BiLSTM | 0.90 | 0.89 | 0.89 | 0.89 |
| **LSTM** | 0.90 | 0.90 | 0.89 | 0.90 |

**Activation values analysis and Pos/Neg Set extraction:** After training, we passed the test set to the saved model and recorded the activation values of the last Dense layer having 64 neurons. The activation function used in the last Dense layer is Rectified Linear Unit (ReLU). Now for each neuron, we calculated the maximum activation value, take the 80% of it and we call this the pos_threshold and also take the 20% of the maximum activation value and call it the neg_threshold. For a specific neuron, out of all the test samples those who have the activation value greater than or equal the pos_threshold we tag them as *positive set* and those are below or equal the neg_threshold we tag them as *negative set*.

**Wordnet concept generation:** At this stage, we have positive and negative set for all the neurons. For each neuron, we extracted the top TF-IDF-weighted terms from the *positive set* and filtered out terms that were also prevalent in the *negative set*. The remaining salient terms were then mapped to their most common WordNet [19] noun synsets, which served as concepts. This process produced a compact set of interpretable concepts for each neuron, which we call the Target Label. At this stage, these labels are supposed to effect the model's decision making process.

**Concept evaluation:** For our evaluation, we excluded 11 neurons because they either showed no activation across the dataset or their extracted concepts did not produce any noun synsets in WordNet. As a result, the evaluation was conducted on the remaining 53 neurons and their associated candidate labels. To evaluate the candidate labels of the activated neurons, we generated 20 *target* and 20 *non target* data for each neuron using GPT 4 [1], where the *target* sentences must have the concepts and *non target* sentences must not have them. 50% of this new test dataset was used for stage 1 evaluation and remaining 50% for stage 2.

## 4 Results and Discussion

Table 3 has the results of the stage 1 evaluation as well as the labels. Here the Target % column shows the percentage of the target sentences that activate each neuron(having activation values greater than the 80% of the maximum activation). Same goes for Non-Target %. We define a label for a neuron to be confirmed if it activates for>=80% of its target set and 31 neurons are found confirmed after stage 1 evaluation. Also we can see, the Non-target % is too low which represents that the neurons are poorly activated by the non-target sentences.

**Table 3: Stage 1 evaluation. Here only the confirmed neurons are shown. Full version of the Table is shown in Appendix B Table 6.**

| Neuron No. | Labels | Target % | Non-target % |
|---|---|---|---|
| 1 | iraq,gaza_strip,baghdad | 100% | 10% |
| 3 | iraq,gaza_strip,baghdad | 90% | 10% |
| 4 | iraq,gaza_strip,baghdad | 100% | 10% |
| 10 | company,institution | 80% | 20% |
| .. | ... | ... | ... |
| 61 | software,code,internet | 80% | 20% |
| 62 | company,institution | 90% | 20% |
| 63 | gaza_strip,iraq,israeli | 80% | 10% |

For stage 2 evaluation, we used the remaining 50% of the new test data and did the same evaluation on the 31 neurons from stage 1. Now out of 31 neurons, 25 neurons showed consistency (i.e., have Target %>=80). Table 4 shows the percentage values of Target and Non-Target along with some statistical evaluation. Here we are showing only the values that have Target %>=80. We compute the z-score and p-value using the non-parametric Mann-Whitney U test [18] to quantify the statistical significance of the evaluation. The negative z-score indicates that the activation values for non-target samples are indeed lower than for the target samples. The corresponding null hypothesis is that activation values are not different between target and non-target. Only neuron 52 and 62 have p>0.05, thus we can not reject the null hypothesis for them. Out of the 25 neurons, 23 neurons reject the null hypothesis at p<0.05.

Taking a closer look at Table 4, we can see both strong alignments and outliers. For example, neuron 55, having concepts *software, code, internet*, was activated in 90% of target samples and only 20% of non-target samples, with a very significant p-value (p = 0.0002). This suggests that the neuron is reliably capturing technological topics, mostly related to Sci/Tech class. On the other hand, neuron 52 and 62 share the label *company, institution* but yield non-significant p-values (0.1405, and 0.0757, respectively). This represents these neurons are less consistent in encoding the concept.

Based on our hypothesis and verification process, we conclude that neurons are activated by the presence of specific concepts, which in turn influence the network's output.

## 5 Conclusion and Future Works

This study demonstrates that concept-based hidden neuron activation analysis is feasible and effective for text classification tasks. We show that individual neurons in an LSTM model are selectively activated by semantically meaningful concepts, and that this activation is statistically significant for the majority of neurons examined. To our knowledge, this is the first study to apply this framework to textual data—thereby validating the generality of the method beyond image classification. In future works, we will try to incorporate different concept generation methods such as, semantically generated concepts or concepts generated by LLMs. Our findings open the door to building interpretable text classifiers where neurons are semantically grounded in human-understandable concepts. Such systems could be extended to support debugging, bias detection, or even interactive explanations, where users can query which concepts drive specific decisions. This contributes toward the goal of making deep learning systems more transparent, trustworthy, and controllable.

## Acknowledgments

The authors acknowledge partial funding through the Kansas State University GRIPex: AI in the Disciplines program, as part of the "Mapping Seed-to-Plant Life Cycle to Predict Yields — Integrated Artificial Intelligence and Data Analytics Approach" project.

## References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in NLP. *arXiv preprint arXiv:1606.07298* (2016).

[3] Adrita Barua, Cara Widmer, and Pascal Hitzler. 2024. Concept induction using llms: a user experiment for assessment. In *International Conference on Neural-Symbolic Learning and Reasoning*. Springer, 132–148.

[4] Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2018. Identifying and controlling important neurons in neural machine translation. *arXiv preprint arXiv:1811.01157* (2018).

[5] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6541–6549.

[6] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30071–30078.

[7] Hyuk-Il Choi, Seok-Ki Jung, Seung-Hak Baek, Won Hee Lim, Sug-Joon Ahn, Il-Hyung Yang, and Tae-Woo Kim. 2019. Artificial intelligent model with neural network machine learning for the diagnosis of orthognathic surgery. *Journal of Craniofacial Surgery* 30, 7 (2019), 1986–1989.

[8] Abhilekha Dalal, Rushrukh Rayan, Adrita Barua, Eugene Y Vasserman, Md Kamruzzaman Sarker, and Pascal Hitzler. 2024. On the value of labeled data and symbolic methods for hidden neuron activation analysis. In *International Conference on Neural-Symbolic Learning and Reasoning*. Springer, 109–131.

[9] Mara Graziani, Vincent Andrearczyk, Stéphane Marchand-Maillet, and Henning Müller. 2020. Concept attribution: Explaining CNN decisions to physicians. *Computers in biology and medicine* 123 (2020), 103865.

[10] Kun Guo, Bo Zhu, Lei Zha, Yuan Shao, Zhiqin Liu, Naibing Gu, and Kongbo Chen. 2025. Interpretable prediction of stroke prognosis: SHAP for SVM and nomogram for logistic regression. *Frontiers in Neurology* 16 (2025), 1522868.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *European conference on computer vision*. Springer, 630–645.

[12] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*. PMLR, 2668–2677.

**Table 4: Stage 2-Statistical evaluation with Mann Whitney U test. Bold rows indicate neurons with p-value ≥ 0.05.**

| Neuron | Labels | Target % | Non-Target % | Z-score | P-value |
|---|---|---|---|---|---|
| 1 | iraq,gaza_strip,baghdad | 80% | 10% | -3.7796 | 0.0001 |
| 3 | iraq,gaza_strip,baghdad | 90% | 10% | -3.7796 | 0.0001 |
| 4 | iraq,gaza_strip,baghdad | 90% | 10% | -3.7796 | 0.0001 |
| 12 | game,activity,time_period | 100% | 20% | -3.7796 | 0.0002 |
| 13 | time_period,activity,league | 80% | 30% | -3.7796 | 0.0002 |
| 18 | software,code,internet | 90% | 30% | -3.7796 | 0.0002 |
| 24 | software,code,internet | 80% | 10% | -3.7796 | 0.0002 |
| 25 | time_period,activity,league | 80% | 10% | -3.7796 | 0.0001 |
| 33 | activity,league,association | 100% | 20% | -3.7796 | 0.0002 |
| 34 | gaza_strip,iraq,baghdad | 90% | 10% | -3.7796 | 0.0001 |
| 35 | iraq,gaza_strip,baghdad | 90% | 20% | -3.7796 | 0.0001 |
| 36 | gaza_strip,iraq,baghdad | 90% | 10% | -3.7796 | 0.0001 |
| 37 | space,attribute | 90% | 10% | -3.7041 | 0.0002 |
| 38 | company,institution | 80% | 40% | -2.0410 | 0.0452 |
| 40 | time_period,activity,league | 80% | 20% | -3.7796 | 0.0002 |
| 41 | activity,time_period,rest_day | 100% | 20% | -3.7796 | 0.0001 |
| 42 | iraq,baghdad,gaza_strip | 90% | 10% | -3.7796 | 0.0001 |
| 50 | space,attribute,internet | 80% | 10% | -3.7796 | 0.0001 |
| 51 | software,code | 100% | 20% | -3.7796 | 0.0002 |
| **52** | **company,institution** | **80%** | **30%** | **-1.5119** | **0.1405** |
| 55 | software,code,internet | 90% | 20% | -3.7796 | 0.0002 |
| 59 | league,association,activity | 100% | 20% | -3.7796 | 0.0001 |
| 60 | software,code,company | 80% | 20% | -3.6285 | 0.0003 |
| 61 | software,code,internet | 90% | 20% | -3.7796 | 0.0002 |
| **62** | **company,institution** | **80%** | **30%** | **-1.8142** | **0.0757** |


[13] Jens Lehmann and Pascal Hitzler. 2010. Concept learning in description logics using refinement operators. *Machine Learning* 78, 1 (2010), 203–250.

[14] Jens Lehmann and Johanna Völker. 2014. *Perspectives on ontology learning*. Vol. 18. IOS Press.

[15] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and understanding neural models in NLP. *arXiv preprint arXiv:1506.01066* (2015).

[16] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* 2, 1 (2020), 56–67.

[17] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[18] Patrick E McKnight and Julius Najab. 2010. Mann-whitney U test. *The Corsini encyclopedia of psychology* (2010), 1–1.

[19] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[20] Ioannis Mollas, Nick Bassiliades, and Grigorios Tsoumakas. 2022. Conclusive local interpretation rules for random forests. *Data Mining and Knowledge Discovery* 36, 4 (2022), 1521–1574.

[21] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), 193–209.

[22] Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-Based Explanation Methods: A Review for NLP Interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 4593–4603. https://aclanthology.org/2022.coling-1.406/

[23] Tuomas Oikarinen and Tsui-Wei Weng. 2022. CLIP-Dissect: Automatic description of neuron representations in deep vision networks. In *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*. https://openreview.net/forum?id=Hcx62bSwg9

[24] Maria Sahakyan, Zeyar Aung, and Talal Rahwan. 2021. Explainable artificial intelligence for tabular data: A survey. *IEEE access* 9 (2021), 135392–135422.

[25] Md Kamruzzaman Sarker and Pascal Hitzler. 2019. Efficient concept induction for description logics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3036–3043.

[26] Md Kamruzzaman Sarker, Joshua Schwartz, Pascal Hitzler, Lu Zhou, Srikanth Nadella, Brandon Minnery, Ion Juvina, Michael L Raymer, and William R Aue. 2020. Wikipedia knowledge graph for explainable AI. In *Iberoamerican Knowledge Graphs and Semantic Web Conference*. Springer, 72–87.

[27] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[28] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*. PMlR, 3145–3153.

[29] JM Imtinan Uddin, Hayden Hall, Jessica Long, Connor Kimbrell, Isabel Obrien, Sudave Mendiratta, Patrick Koo, Yingfeng Wang, and Hong Qin. 2024. Assessing the Impact of Homelessness on COVID-19 Hospitalization Rates in Patients with Underlying Medical Conditions Through Explainable AI. In *World Congress in Computer Science, Computer Engineering & Applied Computing*. Springer, 105–119.

[30] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems* 33 (2020), 20554–20565.

[31] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[32] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems* 28 (2015).

[33] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. 2018. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence* 41, 9 (2018), 2131–2145.

## A Dataset

Some data samples from the AG News topic classification are shown in Table 5. The relevant concepts are shown in bold in the sentences.

**Table 5: Data samples from AG News Topic Classification Dataset.**

| Text | Class |
|---|---|
| GAZA CITY: The Israeli army demolished 13 Palestinian houses during an incursion in the southern **Gaza strip** town of Rafah on Thursday, Palestinian security sources and witnesses said. | World |
| BAGHDAD (Reuters) - At least 110 people were killed across **Iraq** on Sunday in a sharp escalation of violence that saw gun battles, car bombs and bombardments rock the capital. | World |
| AP - Arsenal extended its unbeaten streak in the Premier **League** to 48 games Saturday, getting two goals from Thierry Henry in a 4-0 victory over Charlton and bouncing back from a Champions **League** tie. | Sports |
| AP - Three New York Giants have filed complaints with the NFL Players **Association** after being fined by coach Tom Coughlin for not being "early enough" to team meetings. | Sports |
| **Stocks** fell on Wednesday after investment bank Morgan Stanley (MWD.N: Quote, Profile, Research) said quarterly profit fell, casting doubt on corporate profit growth, while a brokerage downgrade on Cisco Systems Inc. | Business |
| US **stocks** fell as setbacks for drugmakers, including a study showing Pfizer Inc. #39;s Celebrex painkiller increased the risk of heart attacks, sent health-care shares tumbling. | Business |
| CHICAGO - Hewlett-Packard(HP) has moved its Active Counter Measures network security **software** into beta tests with a select group of European and North American customers in hopes of readying the product for a 2005 release, an HP executive said at the HP World conference here in Chicago Wednesday. | Sci/Tech |
| AUGUST 18, 2004 (IDG NEWS SERVICE) - A majority of US home **Internet** users now have broadband, according to a survey by NetRatings Inc. | Sci/Tech |

## B Stage 1 Evaluation

Full version of the stage 1 evaluation is shown in Table 6.

**Table 6: Stage 1 evaluation. Only confirmed neurons are shown (Target % >= 80).**

| Neuron No. | Labels | Target % | Non-Target % |
|---|---|---|---|
| 1 | iraq,gaza_strip,baghdad | 100% | 10% |
| 3 | iraq,gaza_strip,baghdad | 90% | 10% |
| 4 | iraq,gaza_strip,baghdad | 100% | 10% |
| 10 | company,institution | 80% | 20% |
| 12 | game,activity,time_period | 100% | 10% |
| 13 | time_period,activity,league | 90% | 30% |
| 18 | software,code,internet | 100% | 30% |
| 21 | oil,lipid,monetary_value | 80% | 20% |
| 23 | iraq,corporate_executive | 80% | 10% |
| 24 | software,code,internet | 80% | 10% |
| 25 | time_period,activity,league | 80% | 0% |
| 32 | time_period,league,association | 90% | 10% |
| 33 | activity,league,association | 90% | 10% |
| 34 | gaza_strip,iraq,baghdad | 80% | 20% |
| 35 | iraq,gaza_strip,baghdad | 90% | 0% |
| 36 | gaza_strip,iraq,baghdad | 80% | 0% |
| 37 | space,attribute | 100% | 20% |
| 38 | company,institution | 80% | 20% |
| 40 | time_period,activity,league | 80% | 40% |
| 41 | activity,time_period,rest_day | 100% | 20% |
| 42 | iraq,baghdad,gaza_strip | 100% | 10% |
| 48 | game,activity,league | 100% | 10% |
| 49 | oil,lipid,monetary_value | 90% | 20% |
| 50 | space,attribute,internet | 100% | 10% |
| 51 | software,code | 100% | 40% |
| 52 | company,institution | 80% | 30% |
| 55 | software,code,internet | 90% | 20% |
| 59 | league,association,activity | 100% | 20% |
| 60 | software,code,company | 100% | 10% |
| 61 | software,code,internet | 80% | 20% |
| 62 | company,institution | 90% | 20% |
| 63 | gaza_strip,iraq,israeli | 80% | 10% |