Fine-grained Interpretation and Causation Analysis in Deep NLP Models

Hassan Sajjad   Narine Kokhlikyan[*]   **Fahim Dalvi   Nadir Durrani**

{hsajjad,faimaduddin,ndurrani}@hbku.edu.qa
Qatar Computing Research Institute, HBKU Research Complex, Doha 5825, Qatar
narine@fb.com
[*]Facebook AI, 1 Facebook Way, Menlo Park, CA 94025, USA

1 Introduction

Deep neural networks have constantly pushed the state-of-the-art performance in natural language processing (NLP) and are now considered as the de facto modeling approach in solving most complex NLP tasks such as machine translation, summarization and question-answering. Despite the benefits and the usefulness of deep neural networks at-large, their opaqueness is a major cause of concern. Interpreting neural networks is considered important for increasing trust in AI systems, providing additional information to decision makers, and assisting ethical decision making (Lipton, 2016).

Work done on neural interpretation can be broadly divided into two types: i) intrinsic analysis, ii) causation analysis. The former class analyze representations learned within neural networks (Adi et al., 2016; Belinkov et al., 2017a,b; Dalvi et al., 2017; Conneau et al., 2018; Tenney et al., 2019; Durrani et al., 2019).The latter tries to understand the network with various behavioural studies (Gulordava et al., 2018; Linzen et al., 2016; Marvin and Linzen, 2018), or by studying importance of input features and neurons with respect to a prediction (Dhamdhere et al., 2018a; Lundberg and Lee, 2017; Tran et al., 2018). These works on interpretation have been presented in earlier tutorials, for example ACL 2020 (Belinkov et al., 2020b) and EMNLP 2020 (Wallace et al., 2020). The former focused on representation analysis and behavioral studies and the latter discusses work done on the causal interpertation such as assessing model's behavior using constructed examples. Both these great tutorials serve as a starting point for the new researchers in this area.

The representation analysis, also called as structural analysis, is useful to understand how various core linguistic properties are learned in the model (Belinkov et al., 2020a). However, the analysis suffers from a few limitations. It mainly focuses at interpreting full vector representations and does not study the role of fine-grained components in the representation i.e. neurons. Also the findings of representation analysis do not link with the cause of a prediction (Belinkov and Glass, 2019). While the behavioral analysis evaluates model predictions, it does not typically connect them with the influence of the input features and the internal components of a model (Vig et al., 2020).

In this tutorial, we present and discuss the research work on interpreting fine-grained components of a model from two perspectives, i) fine-grained interpretation, ii) causation analysis. The former introduces methods to analyze individual neurons and a group of neurons with respect to a language property or a task. The latter studies the role of neurons and input features in explaining decisions made by the model. We will cover important research questions such as i) how is knowledge distributed across the model components? ii) what knowledge learned within the model is used for specific predictions? iii) does the inhibition of specific knowledge in the model change predictions? iv) how do different modeling and optimization choices impact the underlying knowledge? The contents of this tutorial are covered our survey paper (Sajjad et al., 2021).

Recent work on interpreting neurons has shown that in-addition to gaining better understanding of the inner workings of neural networks, the neuron-level interpretation has applications in model distillation (Rethmeier et al., 2020), domain adaptation (Gu et al., 2021) or efficient feature selection (Dalvi et al., 2020) e.g., by removing unimportant neurons, facilitating architecture search, and mitigating model bias by identifying neurons responsible for sensitive attributes like gender, race or politeness (Bau et al., 2019; Vig et al., 2020). These recent works are not only enabling better understanding of these networks, but are also leading towards better, fairer and more environmental-friendly models,

which are all important goals for the Artificial Intelligence community at large.

2 Description

The tutorial is divided into two main parts: i) fine-grained interpretation, and ii) causation analysis. The first part of the tutorial covers methods that align neurons to human interpretable concepts or study the most salient neurons in the network. We cluster these methods into four groups i) Visualization Methods (Karpathy et al., 2015; Li et al., 2016a), ii) Corpus Selection (Kádár et al., 2017; Poerner et al., 2018; Na et al., 2019; Mu and Andreas, 2020), iii) Neuron Probing (Dalvi et al., 2019a; Lakretz et al., 2019; Valipour et al., 2019; Durrani et al., 2020, 2021) and iv) Unsupervised Methods (Bau et al., 2019; Torroba Hennigen et al., 2020; Wu et al., 2020; Michael et al., 2020). We also discuss evaluation methods that are used to measure the effectiveness of an interpretation method, such as accuracy, control tasks (Hewitt and Liang, 2019) and ablation studies (Li et al., 2016b; Lillian et al., 2018; Dalvi et al., 2019a; Lakretz et al., 2019). Moreover, we cover various applications of these methods that go beyond interpretation such as efficient transfer learning (Dalvi et al., 2020), controlling system's behavior (Bau et al., 2019; Suau et al., 2020), generating explanations (Mu and Andreas, 2020) and domain adaptation (Gu et al., 2021).

The second part, *Causation Analysis*, focuses on the methods that characterize the role of neurons and layers towards a specific prediction. More concretely, we discuss gradient and perturbation-based attribution algorithms such as Integrated Gradients (Sundararajan et al., 2017), Layer Conductance (Dhamdhere et al., 2018b), Saliency (Simonyan et al., 2014), SHapley Additive exPlanations(SHAP) (Lundberg and Lee, 2017) and showcase how they can help us to identify important neurons in different layers of a deep neural network. Besides that we also dive deep into more recent and advanced attribution algorithms that take feature or neuron interactions into account. More specifically, we look into Integrated Hessians (Janizek et al., 2020), Shapely Taylor index (Dhamdhere et al., 2020) and Archipelago (Tsang et al., 2020).

Lastly, we present various open source toolkits and libraries that provide implementation of notable techniques in the area. A few examples include: Captum (Kokhlikyan et al., 2020), Inter-

pretML[1], NeuroX (Dalvi et al., 2019b), Ecco[2] and Diagnnose (Jumelet and Hupkes, 2019). We give a walk-through of how some of these tools can be used for fine-grained interpretation and causation analysis. Throughout the tutorial, we critically evaluate the strengths and weakness of the presented methods, and discuss future directions.

3 Outline

I. **Introduction:** We introduce the topic and motivate it by providing the vision of model interpretability, and how it leads towards fair and ethical models and towards better generalization. We then describe various forms of interpretation and outline the scope of the tutorial (15 minutes).

II. **Fine-grained Interpretation:** We present and discuss the work on neuron-level interpretation. (90 minutes)

- Methods (30 minutes)
- Evaluation (15 minutes)
- Findings (30 minutes)
- Practical (15 minutes)

III. **Causation Analysis:** In causation analysis we present various methods on interpreting model predictions with respect to input features and individual neurons. (60 minutes)

- Methods (30 minutes)
- Evaluation (10 minutes)
- Practical (20 minutes)

IV. **Concept-based Interpretation of Prediction:** This part aims to bridge the gap between fine-grained interpretation and causation analysis. We discuss how fine-grained interpretation and causation analysis can be combined to establish concept-based interpretation of model predictions. (10 minutes)

V. **Discussion:** The last part discusses the overall challenges that the current work faces and suggests future directions. (10 minutes)

4 Prerequisites

We assume a basic knowledge of the deep learning and familiarity with the LSTM-based and

---

[1]https://interpret.ml/
[2]https://www.eccox.io/

transformer-based pre-trained models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). Additionally, some familiarity with natural language processing tasks such as, named entity tagging, natural language inference, etc. would be useful but not mandatory. We do not expect participants to have familiarity with the research on the interpretation and analysis of deep models. Familiarity with Python, Pytorch and Transformers library (Wolf et al., 2019) would be useful to understand the practical part.

5   Reading List

- In order to get an overview of the interpretation field, trainees may look at the following survey papers: Belinkov and Glass (2019) and Danilevsky et al. (2020).

- Fine-grained analysis and its Applications: Bau et al. (2019); Dalvi et al. (2019a); Mu and Andreas (2020); Suau et al. (2020) etc.

- Causation analysis: Lundberg and Lee (2017) provides an overview of various methods introduced in literature. For more details, see the following papers: Voita et al. (2020); Sundararajan et al. (2017); Dhamdhere et al. (2018b); Ribeiro et al. (2016); Janizek et al. (2020)

In addition to the above list, interested trainees may look at the papers mentioned in Section 2.

6   Instructor Information (Alphabetic order

Fahim Dalvi, Software Engineer, Qatar Computing Research Institute, Qatar
Email: faimaduddin@hbku.edu.qa
Website: https://fdalvi.github.io

Fahim Dalvi is an experienced Software Engineer with a demonstrated history of working in the research industry and is currently employed at the Qatar Computing Research Institute. Fahim's research is centered around the intersection of Natural Language Processing and Deep Learning, and he has worked on wide variety of problems in these fields including Machine Translation, Language Modelling and Explainability in Deep Neural Networks. He also spends his time converting research into practical applications, with a focus on scalable web applications. Fahim also spends some time every year mentoring and teaching Deep Learning at Fall and Summer schools.

Hassan Sajjad, Senior Research Scientist, Qatar Computing Research Institute, Qatar

Email: hsajjad@hbku.edu.qa
Website: https://hsajjad.github.io

Hassan Sajjad is a Senior Research Scientist at the Qatar Computing Research Institute (QCRI), HBKU. His research interests include the interpretation of deep neural models, machine translation, domain adaptation, and natural language processing involving low-resource and morphologically-rich languages. His research work has been published in several prestigious venues such as CL, CSL, ICLR, ACL, NAACL and EMNLP. His work in collaboration with MIT and Harvard on the interpretation of deep models has also been featured in several tech blogs including MIT News. Hassan co-organized BlackboxNLP 2020, and the WMT 2019/2020 machine translation robustness task. He served as an area chair for the analysis and interpretability, NLP Application, and machine translation tracks at various *CL conferences. In addition, Hassan has been regularly teaching courses on deep learning internationally at various spring and summer schools.

Narine Kokhlikyan, Research Scientist, Facebook AI

Email: narine@fb.com
Website: https://www.linkedin.com/in/narine-k-88916721/

Narine is a Research Scientist focusing on Model Interpretability as part of PyTorch team at Facebook. Her research interests include the understanding of Deep Neural Network internals and their predictions across different applications such as Natural Language Processing, Computer Vision and Recommender Systems. In the recent years she gave talks and presented tutorials on Model Interpretability at KDD 2020 and NeurIPS 2019. Before joining Facebook Narine worked on Natural Language Processing, Time Series Analysis and numerical optimizations.

Nadir Durrani, Senior Research Scientist, Qatar Computing Research Institute, Qatar

Email: ndurrani@hbku.edu.qa

Website: http://alt.qcri.org/~ndurrani/

Nadir Durrani is a Senior Research Scientist at the Arabic Language Technologies group at Qatar Computing Research Institute. His research interests include interpretation of neural networks, neural and statistical machine translation (with focus on reordering, domain adaptation, transliteration, dialectal translation, pivoting, closely related and morphologically rich languages), eye-tracking for MT evaluation, spoken language translation and speech synthesis. His recent work focuses on analyzing contextualized representations with the focus of linguistic interpretation, manipulation, feature selection and model distillation. His work on analyzing deep neural networks has been published at venues like Computational Linguistics, *ACL, AAAI and ICLR. Nadir has been involved in co-organizing workshops such as simultaneous machine translation and WMT 2019/2020 Machine translation robustness task. He regularly serves as program committee and has served as Area chair at ACL and AAAI this year.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *CoRR*, abs/1608.04207.

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*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do Neural Machine Translation Models Learn about Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver. Association for Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2020a. On the linguistic representational power of neural machine translation models. *Computational Linguistics*, 46(1):1–52.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020b. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, D. Anthony Bau, and James Glass. 2019a. 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, Oral presentation)*.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, and Stephan Vogel. 2017. Understanding and Improving Morphological Learning in the Neural Machine Translation Decoder. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*.

Fahim Dalvi, Avery Nortonsmith, D. Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, and James Glass. 2019b. Neurox: A toolkit for analyzing individual neurons in neural networks. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Fahim Dalvi, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. 2020. Analyzing redundancy in pretrained transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP-2020)*, Online. Association for Computational Linguistics.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

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), Minneapolis, Minnesota. Association for Computational Linguistics.

Kedar Dhamdhere, Ashish Agarwal, and Mukund Sundararajan. 2020. The shapley taylor interaction index.

Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. 2018a. How important is a neuron? CoRR, abs/1805.12233.

Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. 2018b. How important is a neuron?

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, and Preslav Nakov. 2019. One size does not fit all: Comparing NMT representations of different granularities. 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 1504–1516, Minneapolis, Minnesota. Association for Computational Linguistics.

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep nlp models? In Findings of the Association for Computational Linguistics: ACL 2021, Online. Association for Computational Linguistics.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4865–4880, Online. Association for Computational Linguistics.

Shuhao Gu, Yang Feng, and Wanying Xie. 2021. Pruning-then-expanding model for domain adaptation of neural machine translation.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. 2020. Explaining explanations: Axiomatic feature interactions for deep networks.

Jaap Jumelet and Dieuwke Hupkes. 2019. diagnnose: A neural net analysis library.

Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. Computational Linguistics, 43(4):761–780.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for PyTorch.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. 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 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016a. Visualizing and understanding neural models in NLP. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 681–691, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. Understanding neural networks through representation erasure. CoRR, abs/1612.08220.

Peter Lillian, Richard Meyes, and Tobias Meisen. 2018. Ablation of a robot's brain: Neural networks under a knife. CoRR, abs/1812.05687.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. Transactions of the Association for Computational Linguistics, 4:521–535.

Zachary C Lipton. 2016. The Mythos of Model Interpretability. In ICML Workshop on Human Interpretability in Machine Learning (WHI).

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 4765–4774. Curran Associates, Inc.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1192–1202,

Brussels, Belgium. Association for Computational Linguistics.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6792–6812, Online. Association for Computational Linguistics.

Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. *CoRR*, abs/2006.14032.

Seil Na, Yo Joong Choe, Dong-Hyun Lee, and Gunhee Kim. 2019. Discovery of natural language concepts in individual units of cnns. *CoRR*, abs/1902.07249.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics.

Nina Poerner, Benjamin Roth, and Hinrich Schütze. 2018. Interpretable textual neuron representations for NLP. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 325–327, Brussels, Belgium. Association for Computational Linguistics.

Nils Rethmeier, Vageesh Kumar Saxena, and Isabelle Augenstein. 2020. Tx-ray: Quantifying and explaining model-knowledge transfer in (un-)supervised NLP. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, page 197. AUAI Press.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2021. Neuron-level Interpretation of Deep NLP Models: A Survey. *CoRR*, abs/2108.13138.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps.

Xavier Suau, Luca Zappella, and Nicholas Apostoloff. 2020. Finding experts in transformer models. *CoRR*, abs/2005.07647.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The Importance of Being Recurrent for Modeling Hierarchical Structure. *arXiv preprint arXiv:1803.03585*.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. 2020. How does this interaction affect me? interpretable attribution for feature interactions.

Mehrdad Valipour, En-Shiun Annie Lee, Jaime R. Jamacaro, and Carolina Bessega. 2019. Unsupervised transfer learning via BERT neuron selection. *CoRR*, abs/1912.05308.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias. *arXiv preprint arXiv:2004.12265*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2020. Analyzing the source and target contributions to predictions in neural machine translation.

Eric Wallace, Matt Gardner, and Sameer Singh. 2020. Interpreting predictions of NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

John Wu, Hassan Belinkov, Yonatan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2020. Similarity Analysis of Contextual Word Representation Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Seattle. Association for Computational Linguistics.