DODRIO: Exploring Transformer Models with Interactive Visualization

Zijie J. Wang **Robert Turko** **Duen Horng (Polo) Chau**
College of Computing, Georgia Tech
{jayw, rturko3, polo}@gatech.edu

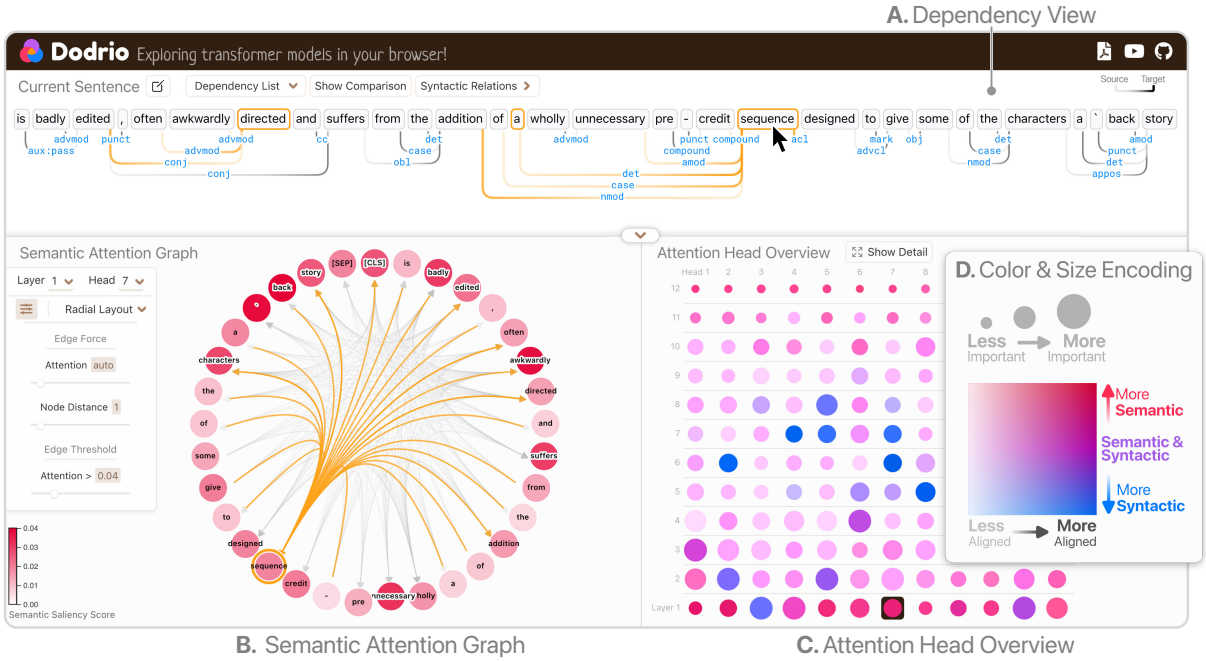

Figure 1: The DODRIO user interface showing user exploration of connections between attention weights from a fine-tuned BERT model and syntactic dependencies as well as semantic saliency scores on the SST2 dataset. **(A) Dependency View** enables users to hover over a word from the input sentence to highlight its associated dependency directed links as **orange** arcs (**lighter** is *source*; **darker** is *target*). **(B) Semantic Attention Graph** highlights the word's related tokens and their attentions; nodes are tokens (darker means more salient); a directed edge encodes attention weight between two tokens. **(C) Attention Head Overview** shows all attention heads in a multi-layer and multi-head model as a grid of circles, each head is **(D) colored** based on its linguistic knowledge in the model (more **red**→more semantic-aligned, more **blue**→more syntactic-aligned; darker→more aligned), and **sized** based on its importance score in the model (larger→more important).

Abstract

Why do large pre-trained transformer-based models perform so well across a wide variety of NLP tasks? Recent research suggests the key may lie in multi-headed attention mechanism's ability to learn and represent linguistic information. Understanding how these models represent both syntactic and semantic knowledge is vital to investigate why they succeed and fail, what they have learned, and how they can improve. We present DODRIO, an open-source interactive visualization tool to help NLP researchers and practitioners analyze attention mechanisms in transformer-based models with linguistic knowledge. DODRIO tightly integrates an overview that summarizes the roles of different attention heads, and detailed views that help users compare attention weights with the syntactic structure and semantic information in the input text. To facilitate the visual comparison of attention weights and linguistic knowledge, DODRIO applies different graph visualization techniques to represent attention weights scalable to longer input text. Case studies highlight how DODRIO provides insights into understanding the attention mechanism in transformer-based models. DODRIO is available at https://poloclub.github.io/dodrio/.

1 Introduction

The rise of transformer-based models have brought dramatic performance improvements across many NLP tasks (Wang et al., 2019). In particular,

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 132–141, August 1st - August 6th, 2021.

BERT (Devlin et al., 2019) has demonstrated that transformer-based models pre-trained on large-scale corpora can be effectively fine-tuned for a wide variety of downstream tasks, such as sentiment analysis, question answering, and text summarization. However, how these language models generalize text representations learned from an unsupervised training process to downstream sentence understanding tasks remains unclear. There is a growing research body in interpreting transformer-based models, as understanding what these models have learned and why they succeed and fail is vital for NLP researchers to develop better models, and critical for decision makers to trust these models.

The current approach on interpreting transformer-based models focuses on probing and attention weight analysis (Hewitt and Liang, 2019). There is an active discussion on whether attention weights are explanations (Jain and Wallace, 2019), but more recent work has shown that they do provide insights on what the models have learned (Atanasova et al., 2020). In particular, research has shown that transformer-based models have learned to represent semantic knowledge and lexical structure in text (Rogers et al., 2020). Furthermore, interaction visualization systems have shown great potential in explaining complex deep learning models (Hohman et al., 2018; Wang et al., 2020). Some visualization tools have been developed for transformer-based models (Vig, 2019; Hoover et al., 2020; DeRose et al., 2021). However, these systems usually focus on visualizing and analyzing attention weights, instead of visually connecting them to linguistic knowledge that is crucial to investigate why transformer-based models work so well across different tasks (Rogers et al., 2020).

To address this research challenge, we present **DODRIO** (Figure 1), an interactive visualization tool to help NLP researchers and practitioners analyze and compare attention mechanisms with linguistic knowledge. For a demo video of DODRIO, visit https://youtu.be/qB-T9j7UTgE. In this work, our primary contributions are:

1. **DODRIO, a novel interactive visualization system** that helps users better understand the attention mechanisms in transformer-based models by linking attention weights to semantic and syntactic knowledge.
2. **Novel interactive visualization design** of DODRIO, which integrates overview + detail, linking + brushing, and graph visualizations that simultaneously summarizes a complex multi-layer and multi-head transformer model, and provides linguistic context for users to interpret attention weights at different levels of abstraction.
3. **An open-source**[1] and **web-based implementation** that broadens the public's access to modern deep learning techniques. We also provide thorough documentations to encourage users to extend DODRIO to their own models and datasets.

2 Background

Attention heads are comprised of weights incurred from words when calculating the next representation of the current word (Clark et al., 2019), which are known as attention weights. Easily interpretable, using attention to understand model predictions across domains is a very popular research area (Xu et al., 2015; Rocktäschel et al., 2016). In NLP, there has been a growing body of research on attention used as a tool for interpretability across many language tasks (Wiegreffe and Pinter, 2019; Vashishth et al., 2019; Kobayashi et al., 2020).

Existing visualization systems and techniques do not visually connect attention mechanisms to linguistic knowledge (Tenney et al., 2020; DeRose et al., 2021), we propose novel visualization approaches that foster exploration across semantically and syntactically significant attention heads in complex model architectures. For example, for every attention head in the 144 heads of BERT, the entry $A_{i,j}$ in the attention map A, represents the attention weight from token i to token j. With 144 $\times$ number of tokens $\times$ number of tokens attention weights in BERT for each input instance, it is challenging to systematically analyze these attention weights without abstraction and linguistic context. DODRIO aims to address this challenge by applying novel interactive visualization techniques.

3 Interface

3.1 Attention Head Overview

As a user explores the attention weights, the Attention Head Overview (Figure 1C) serves as a guide to effectively navigate the remaining views of the interface. With visual linking and brushing (McDonald, 1988), we unify attention head selection with the state of the remainder of the interface. This view of a grid of attention heads

[1] https://github.com/poloclub/dodrio

guides the user to inspect semantically and syntactically important heads. Attention heads are encoded as circles where color encodes the head's linguistic alignment (more red→more semantic-aligned, more blue→more syntactic-aligned; darker→more aligned), and sized represents its importance score in the model (larger→more important) (Figure 1B).

We calculate the **semantic score** m by computing the cosine similarity between the sum of attentions received for each token at a given head, and the sentiment score of each token. If the sentiment score is not available in a dataset, we use the saliency score for each token instead. The saliency score of a token measures how important that token contributes to the final model prediction (Barredo Arrieta et al., 2020), and it is shown to correlate with word semantics (Atanasova et al., 2020).

Following Clark et al. (2019)'s framework, we use the source token's most-attended token as its predicted dependency target. For each existing dependency relationship, we compute each head's average accuracy across all instances. Finally, we calculate the head's **syntactic score** n by taking the maximum of its average accuracy across all existing dependency relationships (ground truth or generated by a parser).

There are multiple metrics to measure the importance of a given attention head. By default, we calculate the **importance score** c of an attention head by the average of its maximum attention for all instances in the dataset (Voita et al., 2019). DODRIO also supports using the sum of absolute gradients of attention weights in an attention head as its importance score c (Clark et al., 2019).

After computing these three scores, we create a linear color scale and a linear size scale to encode them in the Attention Head Overview (Figure 1C, D). We use the Hue-chroma-luminance (HCL) color space to represent colors in DODRIO. The HCL color space is designed to better align with human perception of colors, so that interpolations in this space is smoother and more consistent (Zeileis et al., 2009). We use the hue value (H) in the HCL color space to encode $m - n$ with range [-1, 0, 1] as [**blue**, **purple**, **red**]; the luminance value (L) to encode $\max(m, n)$ (range [0, 1]); and the size of circles to encode c (range [0, 1]). With our color and size encoding, the Attention Head Overview (Figure 1C and Figure 2) provides an accurate and efficient summarization of attention heads.

In the Attention Head Overview, users can also click a button to show the expanded Attention Head Overview (Figure 2) that additionally provides a preview of the attention pattern in each attention head through the *Radial Layout* visualization. Hovering over one attention head displays its linguistic and importance information.

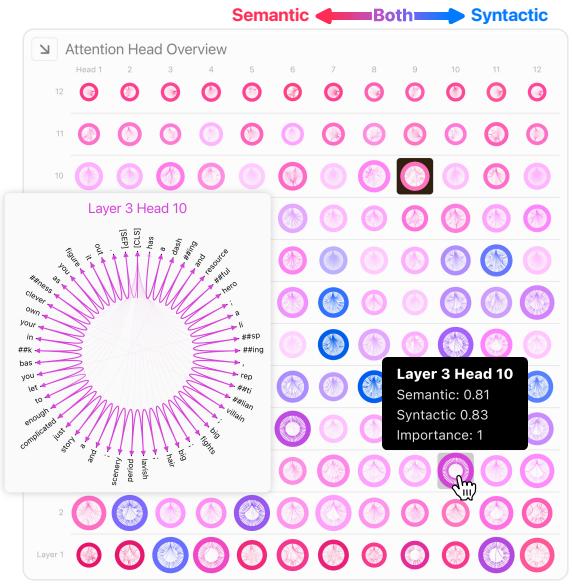

Figure 2: The expanded Attention Head Overview provides a preview of all attention heads for the input sentence. Attention heads are represented as a grid of rings (right) where their attention weights are shown in the middle. Each ring's color and size encode the attention head's linguistic knowledge alignment and importance score (**red**→semantic; **purple**→semantic and syntactic; **blue**→syntactic; larger→more important). Users can click an attention head to inspect its attention weights in detail in a radial layout window (left).

3.2 Syntactic Dependencies

Word relations in a sentence are important features to understand the lexical makeup of a sentence, which can help users further deduce model decisions in the context of sentence structure. In DODRIO, a user can explore an attention head with input sentence's dependency relationships.

Dependency View (Figure 1A). We visualize true dependency relations, if available, or relations tagged by the CoreNLP pipeline (Manning et al., 2014) linked with the Semantic Attention Graph for users to investigate syntax-sensitive behavior at different attention heads. The user can further explore the dependency representation in a hierarchical structure by filtering dependency relations.

Comparison View (Figure 3). Understanding raw attention weights are best interpreted relative

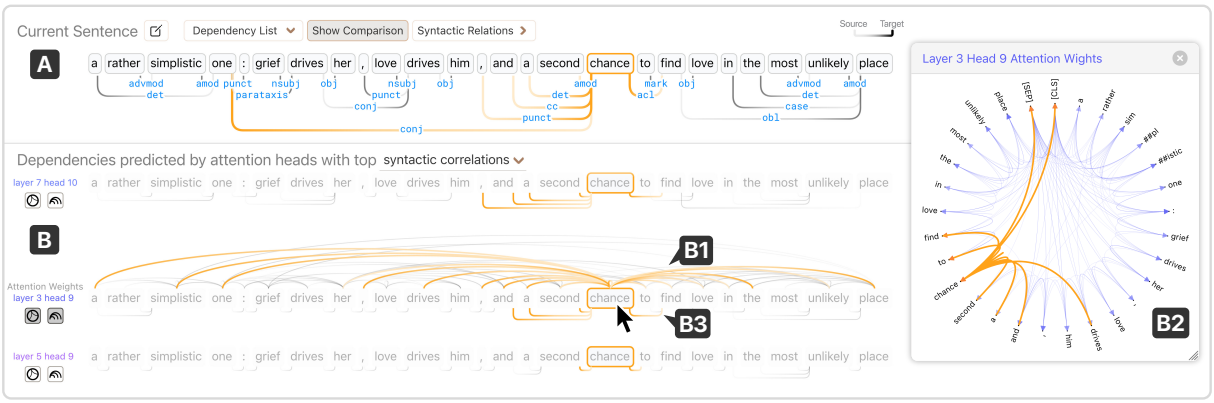

Figure 3: The Comparison View allows users to compare multiple attention heads and explore the connection between attention weights and the syntactic structure of the input sentence. (**A**) The top *rectangular arc diagram* visualizes dependencies generated by a parser (lighter is source; darker is target). (**B**) Each attention head is represented as a row of tokens where (**B1**) the top *curved arc diagram* and (**B2**) the *radial layout window* display the selected head's attention weights on demand. (**B3**) The *rectangular arc diagram* below the tokens shows the dependencies predicted using attentions. Hovering over one token highlights all associated attentions and dependency links.

to the attention weights at other attention heads in the model. The Comparison View enables users to examine the dependencies predicted by attention heads (Figure 3-B3). A user can select additional attention representations under each attention head label within this view to supplement their analysis of attention with respect to the grammatical structure of the sentences. By viewing the attention edges drawn above the tokens, which encode attention weight magnitude with opacity *in the Arc Layout* (Figure 3-B1), a user can maintain word-order context in the sentence, while the attention representation utilizing a *Radial Layout* (Figure 3-B2) of attention edges allows for a clearer interpretation the attention distribution. The edge linking with interaction between this view and the Dependency View further reinforces the syntax-sensitive behavior present in attention heads

3.3 Semantic Attention Graph

The attention map at each head can be interpreted as an adjacency matrix, which can be visualized using different graph visualization techniques (Figure 4). Users can primarily use this interactive graph view to inspect semantically significant attention heads, as defined the Attention Head Overview. Since the node color encodes the saliency score, linked to word's semantics (Li et al., 2016), the behavior of the attention mechanism in the model can be evaluated from a semantic perspective.

Similarly to representations in the Comparison View, the Semantic Attention Graph representations can be customized with interaction to allow

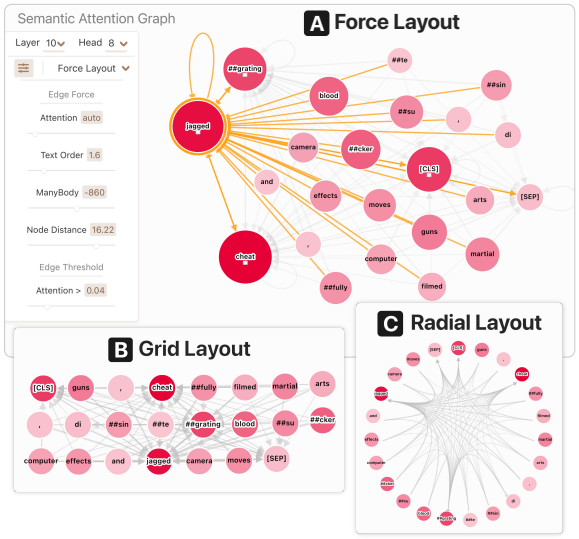

Figure 4: The Semantic Attention Graph employs three graph visualization techniques to show the attention weights. (**A**) The *force layout* allows users to flexibly change token positions; (**B**) the *grid layout* enhances the readability of input sentence; (**C**) the *radial layout* compactly highlights attention patterns.

for detailed attention inspection for selected tokens (Figure 4A), preserve token-order context in the *Grid Layout* (Figure 4B), or allow for clear attention analysis in the *Radial Layout* (Figure 4C). Adjusting graph parameters in the side panel of this view encourages the user to customize the graph representation to ease attention analysis (eg. adjusting the *edge threshold* parameter will only show attention weights with a greater magnitude) (Figure 4-A left). We utilize linking to allow the user to interpret tokens in the context of their attention

weights and dependence relations simultaneously as both nodes and edges are highlighted when a user hovers over a node in either the Semantic Attention Graph or the Dependency View.

3.4 Instance Selection View

For a robust understanding of the attention mechanisms in Transformers, it is important to explore the behavior of attention across interesting components of a sentence (eg. coreferences, word sense, etc.) present in various instances in a dataset.

The **Embedding View** (Figure S1-A) uses UMAP (McInnes et al., 2018) to project text instance's model representation computed by concatenating the last four hidden state layers of BERT to a 2D space and visualizes it with a scatter plot.

The **Table View** (Figure S1-B) allows for instance selection while providing the user with instance's true and predicted labels. Users can hover over a dot in the Embedding View to view the sentence text, and click a dot or a row in the Table View to change DODRIO's input sentence.

4 Case Study

4.1 Understanding Sentiment in BERT

How does a Transformer handle conflicting sentiment in opinionated phrases when resolving coreferences? In DODRIO, we can explore the attention mechanism within a text instance from a movie review dataset, SST2 (Socher et al., 2013), such as "*A coming-of-age film that avoids the cartoonish clichés and sneering humor of the genre as it provides a fresh view of an old type*." Using this sentence, we can explore the concept of *sentiment consistency* as proposed by (Ding and Liu, 2010) in the context of coreference resolution.

When interpreting the sentence above, it is clear to us that "it" refers to the "film" because the first half of the sentence expresses positive sentiment towards the "film" and negative towards the "genre," while the second half of the sentence represents a positive opinion on the "film." We can deduce that "it" refers to the "film" as sentiment is expressed in a consistent manner as discussed by (Ding and Liu, 2010). By exploring the Attention Head Overview of DODRIO (Figure S3), we can select an attention head that conveys semantically significant information as indicated by the 2D color scale (eg. layer 1, head 7). As we begin to analyze the Semantic Attention Graph (Figure S3-left), we can hover over the node representing "it" to visualize the attention behavior. "It" attends highly to "film," which validates the coreference resolution policy that we discussed above (Figure S3-right). Users are encouraged to explore other attention heads as well to compare the behavior of the attention mechanism across various linguistic features.

4.2 Penn Treebank Analysis

Understanding attention across natural language tasks is pivotal for a systematic understanding of the attention mechanism as it relates to interpretability (Vashishth et al., 2019). If we visualize BERT on a text corpus with annotated syntactic sentence structure, like Penn Treebank (Marcus et al., 1993), can attention accurately predict syntactic heads, and what patterns will we observe?

To investigate these ideas, we navigate to the Dependency View within DODRIO. Beginning in the Dependency View, we observe edges of human annotated dependency relations connecting each token to its syntactic head, rather than part of speech (POS) tagging and dependency parsing annotations by the CoreNLP pipeline (Manning et al., 2014) when human annotations are not provided. To identify whether some attention heads more accurately attend to the syntactic heads of each token, we will enter the Comparison View (Figure 3) by clicking the *Show Comparison* button in the toolbar.

As we see in Figure 3-B3, DODRIO highlights correct syntactic head predictions by attention with a gradient edge, which is linked with the true dependencies in the Dependency View. After exploring various instances, we begin to understand patterns of certain attention heads. For example, we observe that attention head 9 in layer 3 attends to nominals (group of nouns and adjectives: `obj`, `nmod`, `obl`, etc.) across unique instances (Figure S2). This behavior highlights the syntax-aware attention that exists in BERT as discussed by (Clark et al., 2019). Visualizing consistent behavior by attention heads in Transformers outlines how the attention mechanism lends itself to model interpretability.

4.3 Exploring DistilBERT

The computational barrier to achieve state-of-the-art performance on natural language tasks with large pre-trained Transformers like BERT (Devlin et al., 2019) was lowered when DistilBERT (Sanh et al., 2019), a smaller version of BERT, was presented. DistilBERT is 40% smaller and retains up to 97% performance compared to BERT with half as many self-attention layers. With DODRIO,

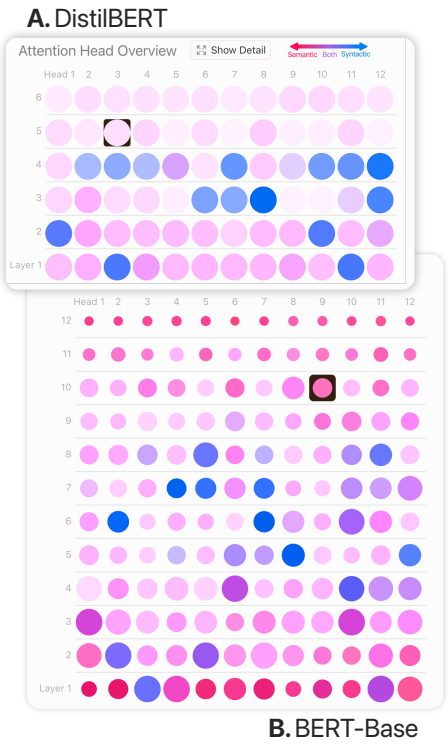

Figure 5: The Attention Head Overview showing attention head roles for two transformer-based models. (**A**) All heads in DistilBERT are important and heads in early layers tend to have stronger linguistic alignment. (**B**) Attention heads in earlier layers tend to be more important and more semantic-aligned in BERT-Base.

we can analyze attention mechanisms at various attention heads in DistilBERT to understand how attention compares to its larger version, BERT.

Using the Attention Head Overview from DODRIO to visualize DistilBERT (Figure 5), we immediately notice that all radial attention head representations have the same diameter, unlike in the case of BERT. Upon further inspection, we see that all attention heads have a confidence score that is very close to one via the tooltip present when hovering over an attention head, which indicates that every attention head has highly attended to tokens on average. As we continue to explore the attention heads, we recognize a similar pattern of syntactic and semantic attention heads, but in the later layers the attention head rings have a much higher luminance in DistilBERT than they did in BERT. According to the 2D color scale (Figure 1D), this represents a lower overall score meaning that these attention heads neither attend to primarily text semantics of grammatical structure. It might imply that DistilBERT has learned some other linguistic knowledge beyond simple word semantics and syntactice dependencies. We can then conduct quantitative experiment to test this hypothesis formed by using DODRIO.

5 Discussion

DODRIO aims to help NLP researchers and practitioners to explore attention mechanisms in transformer-based models with linguistic knowledge. With overview + detail, linking + brushing, graph visualization techniques, DODRIO enables the users to investigate attention weights with different levels of abstraction in a context with both semantic and syntactic information. Through use cases, we demonstrate that DODRIO not only helps users validate existing research results regarding the connections between attention weights with linguistic information, but also inspires the users to form hypothesis regarding the behavior and roles of attention heads across different models.

We acknowledge that there is an active discussion on whether attention weights can help people interpret transformer-based models (Jain and Wallace, 2019) and whether the attentions can be directly linked to the corresponding tokens in interpretation tasks (Brunner et al., 2020). Our work joins the growing research body in NLP interpretability and human-centered NLP, highlighting novel visualization designs that can be generalized to other interactive NLP systems. Despite the increasing popularity of applying Human-computer Interaction techniques to help people from various fields interact with complex NLP systems, little work have been done to evaluate how effective these tools are (Wang et al., 2021). To fill this research gap, we plan to run a user study to evaluate the usability and usefulness of DODRIO.

6 Conclusion

We present DODRIO, an interactive visualization system that fosters the exploration of the attention mechanism in transformer-based models with linguistic knowledge. Through analysis from the model to the attention head level, users can explore how attention differs across a complex, state-of-the-art architecture over any instance within a dataset. Our tool runs in modern web browsers and is open-sourced, broadening the public's access to modern AI techniques. We hope our work will inspire further research in understanding attention mechanisms and development of visualization tools that help people interact with complex NLP models.

7 Broader Impact

We designed DODRIO with good intentions — to help researchers and practitioners more easily explore attention weights in transformer-based models and investigate why their models succeed and fail. However, bad actors could exploit this knowledge of whether and how the models may perform under different situations for malevolent purposes, such as manipulating the model prediction by injecting arbitrary keywords (Kurita et al., 2020). The potential vulnerability warrants further study.

Acknowledgments

We thank Haekyu Park, Rahul Duggal, and Nilaksh Das for their constructive feedback. This work was supported in part by NSF grants IIS-1563816, CNS-1704701, DARPA GARD, gifts from Intel, NVIDIA, Bosch, Google, Amazon. Use, duplication, or disclosure is subject to the restrictions as stated in Agreement number HR00112030001 between the Government and the Performer.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

J. F. DeRose, J. Wang, and M. Berger. 2021. Attention flows: Analyzing and comparing attention mechanisms in language models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1160–1170.

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. Association for Computational Linguistics.

Xiaowen Ding and Bing Liu. 2010. Resolving object and attribute coreference in opinion mining. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 268–276, Beijing, China. Coling 2010 Organizing Committee.

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.

Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–196, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pretrained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2793–2806, Online. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. 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.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330.

John Alan McDonald. 1988. Orion i: Interactive graphics. *Dynamic Graphics Statistics*, page 179.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online. Association for Computational Linguistics.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across NLP tasks.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting Humans in the Natural Language Processing Loop: A Survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics.

Zijie J. Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Chau. 2020. CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.

Achim Zeileis, Kurt Hornik, and Paul Murrell. 2009. Escaping rgbland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, 53(9):3259–3270.

8 Appendix

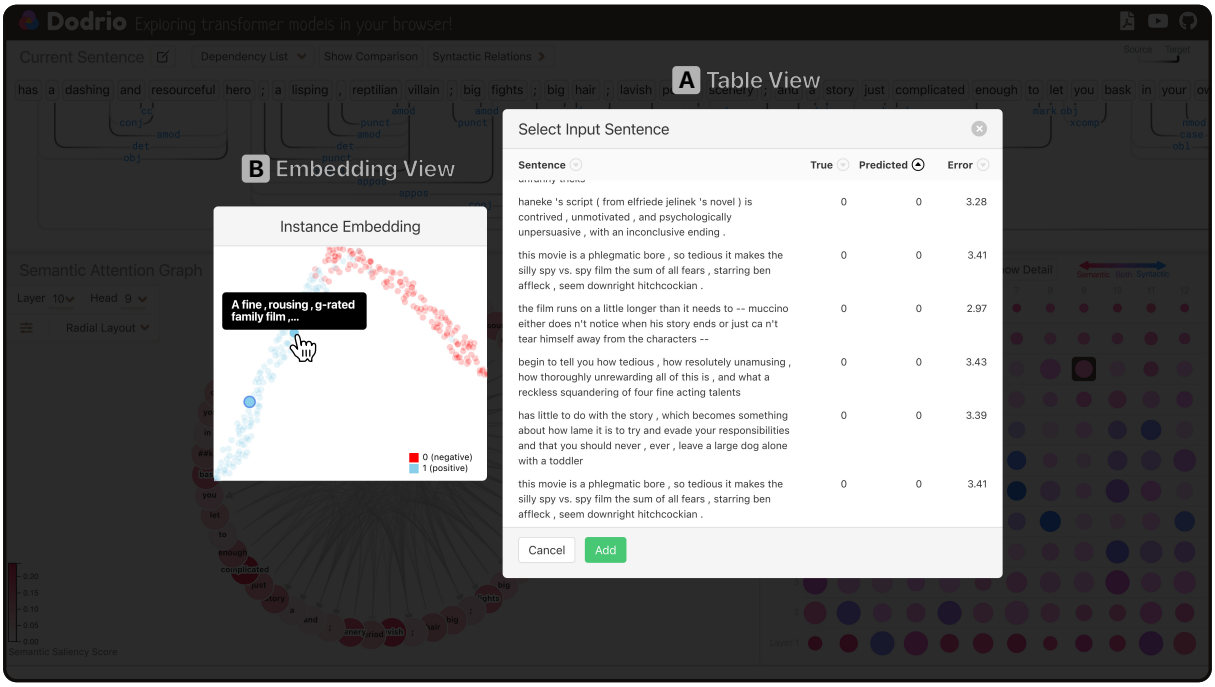

Figure S1: The Instance Selection View within DODRIO encourages users to explore sentences with interesting linguistic features to understand how various attention heads throughout a model attend to them. (**A**) **Table View** presents all text instances in a tabular format with other dataset and task-specific information as well with sortable columns for efficient instance browsing. (**B**) **Embedding View** motivates users to inspect text clustered by dataset label to explore semantically interesting phrases. These views are linked, so that clicking an instance in either view will update the state of the other view, while setting the instance will update the global state of the entire interface.

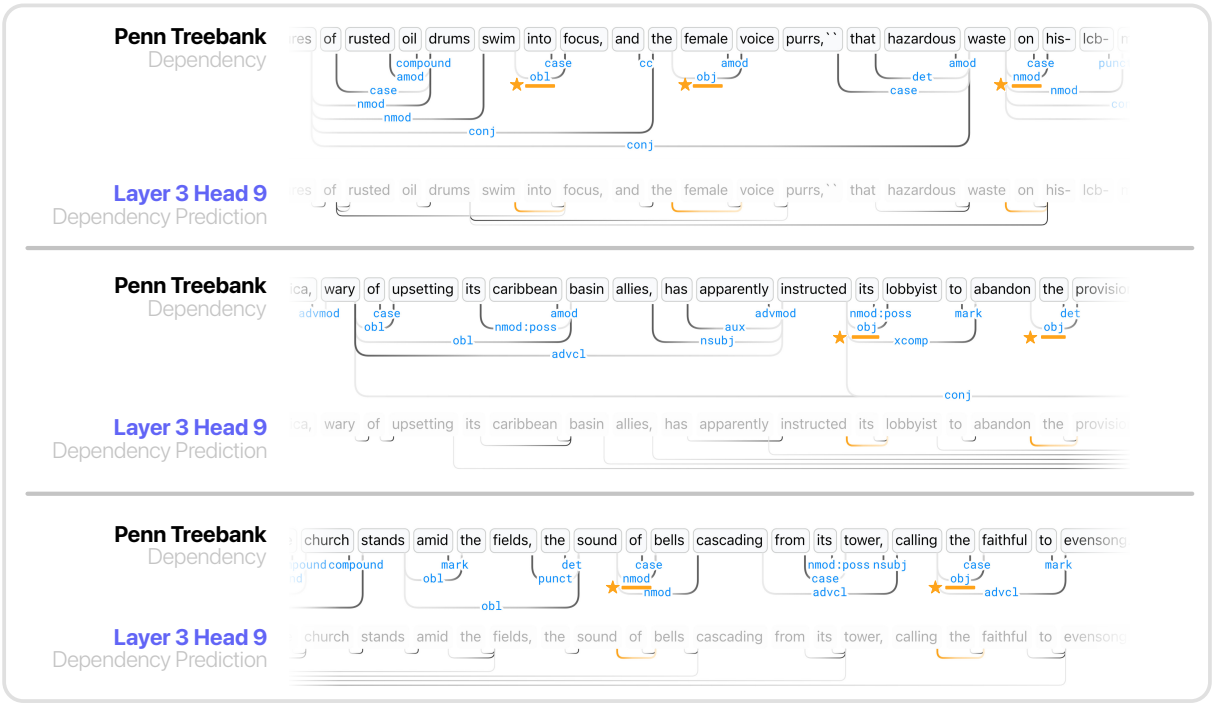

Figure S2: The Comparison View visualizes syntactic relationships on the Penn Treeback dataset. It highlights attention head (Layer 3 Head 9) that can accurately predict the nominal relationships (group of nouns and adjectives: `obj`, `nmod`, `obl`, etc.) across multiple unique instances.

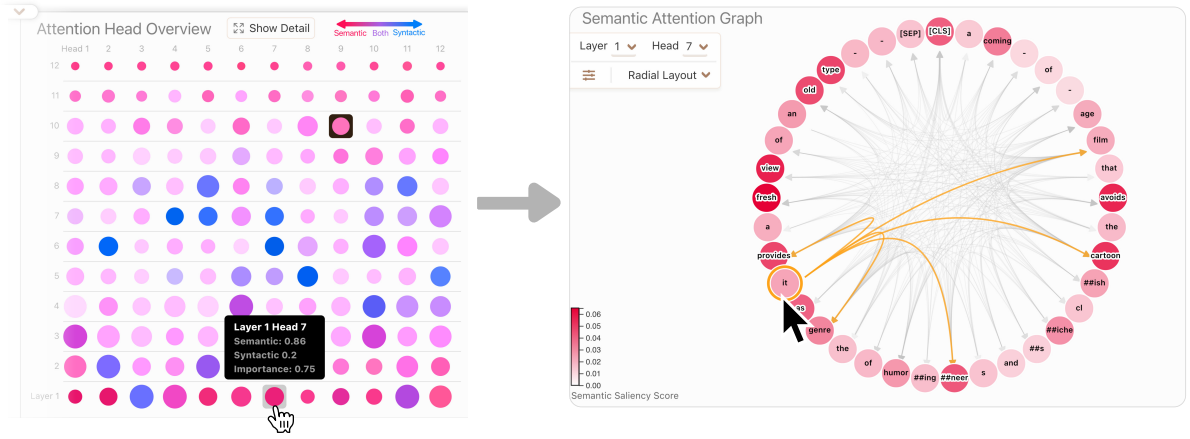

Figure S3: The **Attention Head Overview** (left) helps users identify interesting attention heads (e.g., more semantic-aligned and important heads), and then the **Semantic Attention Graph** (right) quickly visualizes the attention weight pattern of the selected head on the current input sentence, allowing users to rapidly validate their hypothesis regarding attention head's linguistic knowledge.