IB Attribution and Beyond:

A Comprehensive Review and Expansion of Information Bottlenecks to Transformer Models

Hooman Ramezani

University of Toronto

A report submitted for MAT1510

Instructor: Vardan Papyan

December 20, 2023

Information Bottlenecks for Attribution

Explainability in deep learning is pivotal for models to be trusted by the public. The Information Bottleneck (IB) principle offers a promising avenue for disentangling the complexities of neural computation.

The Information Bottleneck (IB) approach quantifies how much a neural network compresses input data while retaining important information for output predictions Tishby et al. [2000]. In deep learning, attribution is the process of quantifying which parts of the input data a model uses to make decisions. Its essential in order to demistify model reasoning to highlight important inputs. The Information Bottleneck (IB) is well-suited for attribution because it focuses on isolating the most crucial features of input data by evaluating their contribution to information gain.

Schulz et al. [2020] introduces an attribution method grounded in the information bottleneck principle. This approach involves injecting noise into intermediate network features, thereby quantifying information flow in terms of bits. The proposed Information Bottleneck Attribution (IBA) method outperforms existing baselines across various metrics on popular architectures like VGG-16 and ResNet-50.

The IBA method estimates the informational value of image regions, ensuring that regions with negligible information do not contribute to the model's decision-making process. The authors employ two variants: Per-Sample and Readout Bottleneck, each tailored to single data points or entire datasets, respectively. The Per-Sample Bottleneck is emphasized for its flexibility and superior performance, despite the convenience of the Readout Bottleneck's single-pass attribution map generation.

$$\max I[Y;Z] - \beta I[X,Z], \tag{1}$$

$$Z = \lambda(X)R + (1-\lambda(X))\epsilon, \tag{2}$$

$$I[R,Z] = \mathbb{E}_R[D_{KL}[P(Z|R) \parallel P(Z)]] - D_{KL}[P(Z) \parallel Q(Z)], \tag{3}$$

The objective is to maximize the information about the output Y that the bottleneck Z retains while minimizing the information about the input X, formalized in Equation (1). The bottleneck Z itself is a weighted combination of the original feature representation R and noise ϵ, Equation (2) Alemi et al. [2017]. This introduces a constraint on the flow of information, ensuring that the resulting bottleneck captures only the most salient features for making predictions. Subsequently, Equation (3) provides an estimation of the mutual information, which assesses the significance of the features that pass through the constriction.

The papers quantitative analysis presents a comparison of information attribution methods using a sensitivity-n test and bounding box method. Sensitivity-n, Ancona et al. [2017], masks the network's input randomly and then measures how strongly the amount of attribution correlates with the drop in accuracy. The bounding box method calculates the ratio of top-scored pixels correctly identified within an object's bounding box. The per-sample bottleneck performs best on both metrics, indicating its efficiency in identifying relevant features and preserving model performance.

Figure 1 demonstrates a successful reproduction of the results of the IBA paper on a pretrained ResNet-50 model, with IBA injected into layer 3. In conclusion information-theoretic foundation provides a quantifiable guarantee that regions with zero-valued attribution are non-essential for accurate classification. IBA is expected to enhance trust in neural networks for sensitive applications.

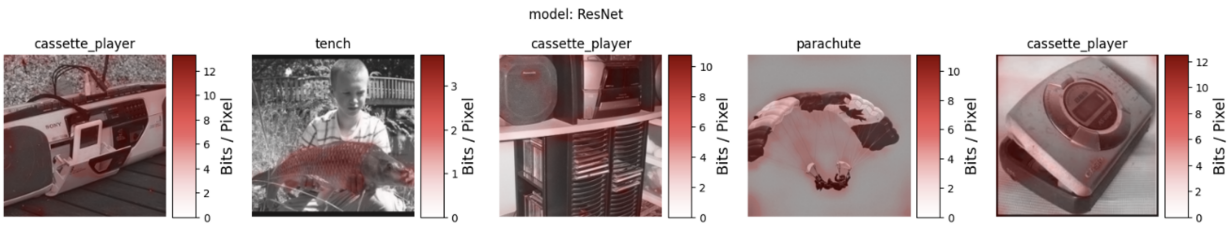

Figure 1: Heatmap of Information Contribution Generated by IBA with ResNet-50

Expanding IBA To BERT Transformers

Motivating the Problem The increasing complexity of Transformer-based models, especially in natural language processing (NLP) tasks, has highlighted a critical gap in model interpretability. While these models achieve state-of-the-art results, understanding the causal relationship between input features and predictions remains elusive. We want to adapt the IBA to a novel framework involving BERT-based transformers. This paves the way for more transparent and understandable language models backed by information theory.

Method Moving on to the methodology, we delve into the instance level, focusing specifically on token-level features and how they interact across different layers of the transformer model. By analyzing these interactions, we can understand how various types of information are encoded at different stages of processing. Our goal is to pinpoint the most information-rich layers within the model. Why is this important? Because not all layers contribute equally to the model's understanding of the input data. Some layers capture more nuanced meanings or complex relationships between tokens.

Layer Analysis Figure 2 demonstrates layer-wise analysis of a BERT-large Transformer finetuned on sentiment analysis for IMDb movie reviews. Results suggests the placement of these bottlenecks is critical, with certain layers being more conducive to informative feature extraction than others. Analysis of attribution at each layer demonstrates how information is sequentially consolidated. By layer 7 information is distilled to one or two key features from approximately twenty candidates. Further analysis of layers 9 through 11 suggests that by then the model successfully concentrates the input into the classification token [CLS], indicating its preparedness for the final classification task.

Degradation Testing Degradation testing was utilized to asses the impact of feature removal on the model's predictive performance. The approach involved systematically occluding input tokens based on the attribution scores and observing the effect on the model's output. Features are degraded in order of most to least important, and the consequent drop in the model's confidence or accuracy is monitored. This decline in performance serves as an indicator of the attribution method's ability to correctly identify relevant input features.

Results, demonstrated in Figure 3, indicate the model's performance deteriorates significantly after the removal of the highest attributed features. We see a smooth decline of accuracy proportionate, suggesting that the attribution method is effectively capturing the influential elements of the input data.

Interpretability Insights In conclusion, the extension of IBA to Transformers successfully was able to capture significant offers a significant advancement in the interpretability of transformer-based models. It not only provides a more effective method for feature attribution compared to existing techniques but also offers insights into the internal information processing of these models. This contributes to a deeper understanding of how deep learning models operate, particularly in the realm of natural language processing.

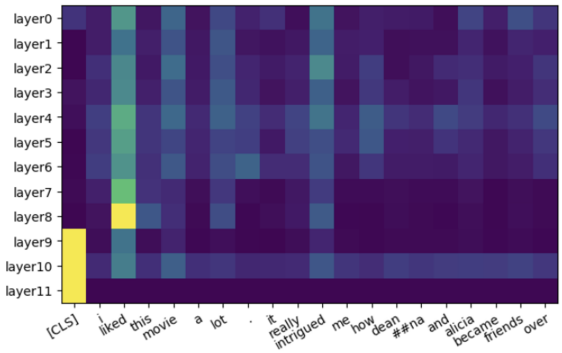

Figure 2: Finetuned BERT model, Information Bottleneck Aattribution on Layers 1-11

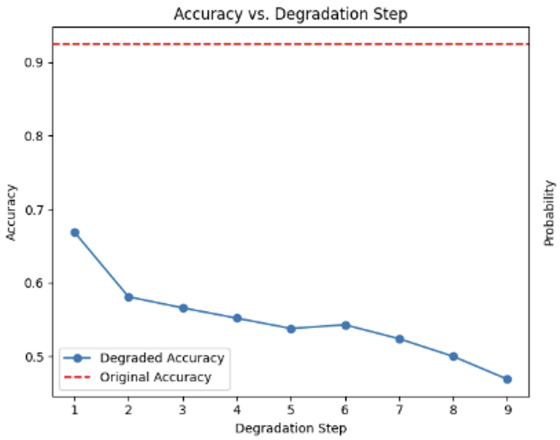

Figure 3: Degradation Testing for n-highest Scored Tokens

References

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. *ICLR*, 2020.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *5th International Conference on Learning Representations (ICLR 2017)*, 2017.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. A unified view of gradient-based attribution methods for deep neural networks. In *NIPS 2017-Workshop on Interpreting, Explaining and Visualizing Deep Learning*. ETH Zurich, 2017.