2. LITERATURE SURVEY

In the literature, some approaches to calculate the feature importances have been proposed with the intent of interpretability. These approaches mentioned below subsections alter certain inputs or neurons while observing the effects on subsequent neurons in the network.

2.1. Backpropagation-based Methods

DeepLIFT [5] is an algorithm that rates the significance of inputs for a specific outcome. It is distinctive in two ways. First, it defines the important question in terms of variations from "reference" state, where the "reference" is selected based on the current issue. Using a difference variable from reference enables DeepLIFT to transmit an importance signal even when the gradient is zero and prevents artifacts brought on by gradient discontinuities, in contrast to other gradient-based algorithms. Second, DeepLIFT may identify dependencies that other techniques have overlooked by optionally taking into account separately the impacts of positive and negative contributions at nonlinearities. DeepLIFT scores can be effectively produced in a single backward pass once a prediction has been made since they are calculated using a backpropagation-like approach, making it efficient.

2.2. Forward Propagation-based Methods

[6] displayed the change in the activations of subsequent layers by obscuring various portions of an input picture. Using "in-silico mutagenesis" [7], virtual mutations were introduced at specific locations in a genomic sequence, and their effects on the result were measured.

[8] is based on an instance-specific method by [9] so called the prediction difference analysis. The proposed methodology by [8] is a method that similar to [6], but

the difference is both removing information from the image and evaluating the effect of this. For explaining classification decisions made by deep neural networks, the method is used to produce a saliency map for each $(instance, node)$ pair that highlights the parts (features) of the input that constitute most evidence for or against the activation of the given (internal or output) node.

[10, 11] used a technique for plotting the weights of a linear classifier or their p-values (as determined by permutation testing) [12,13] to visualize feature importances. These are independent of the input picture, and reading these weights in general may be deceptive, according to [14] and [15].

2.3. Layer-wise Relevance Propagation-based Methods

[16] suggested Layerwise Relevance Propagation (LRP) as a method for distributing significance ratings. [17, 18] demonstrated that the LRP rules for rectified linear units (ReLU) networks were equal within a scaling factor to an element-wise product between the saliency maps of [19] and the input ($gradient * input$) in the absence of adjustments to address numerical stability. The saturation issue or the thresholding phenomenon are still unaddressed, despite the fact that $gradient * input$ is often superior than gradients alone since it makes use of the input's sign and intensity.

2.4. Deconvolutional Network-based Methods

[19] suggested computing an image's saliency map utilizing the gradient of the output with reference to pixels of the input picture, while doing image classification tasks. Except for how they handled the nonlinearity at ReLU, the authors demonstrated that this was similar to deconvolutional networks [6]. When backpropagating importance using gradients, if the input to the ReLU during the forward pass is negative, the gradient flowing into the ReLU during the backward pass is zeroed out. The importance signal entering a ReLU during the backward pass of deconvolutional networks, in contrast, is zeroed out if and only if it is negative, regardless of the sign of

the input into the ReLU during the forward pass. The importance signal at a ReLU is zeroed out if either the input to the ReLU during the forward pass or the importance signal during the backward pass is negative, according to guided backpropagation [20].

With the exception that gradients that become negative during the backward run are eliminated at ReLUs, guided backpropagation may be compared to calculating gradients. Both guided backpropagation and deconvolutional networks may be unable to identify inputs that have a negative impact on the output due to the zeroing out of negative gradients.

2.5. Gradient-based Localization Methods

[19] showed that, first, the numerical optimization of the input picture may be used to produce intelligible visualisations of CNN classification models [21]. The final fully-connected classification layer's optimal neuron should be maximized to display the class of interest since, unlike [21], the net is trained in a supervised way (to identify the neuron in charge of each class in the unsupervised instance, [22] needed to consult a different collection of annotated picture data).

Second, using a single backpropagation run through a classification Convolutional Neural Network (CNN), [19] developed a technique for calculating the spatial support of a particular class in an image (image-specific class saliency map). Weakly supervised object localization may be done using these saliency maps.

Grad-CAM [23] creates a coarse-grained feature-importance map by classifying the final convolutional layer's feature maps according to the gradients of each class with respect to each feature map, and then using the weighted activations of the feature maps to determine which inputs are most crucial. The authors suggested conducting an elementwise product between the scores acquired from Grad-CAM and the scores received from guided backpropagation, known as Guided Grad-CAM, to get more finely grained feature significance. Yet since negative gradients are zeroed out during backpropaga-

tion, this approach inherits the drawbacks of guided backpropagation. Moreover, it is unique to convolutional neural networks.

[24] integrated the gradients as the inputs are scaled up from a beginning value (such as all zeros) to their present value, instead of calculating the gradients simply at the input's current value. Nevertheless, numerically deriving high-quality integrals adds processing complexity. This fixes the saturation and thresholding issues. Moreover, this strategy may still provide false findings.

2.6. Latent Tree-based Methods

[25] use a general shift-reduce parser, whose training depends on ground-truth parsing trees, to construct trees and combine semantics. Combining latent tree learning with Tree-structured Recurrent Neural Networks (Tree-RNN) has been shown to be a successful strategy for sentence embedding since it simultaneously optimizes the sentence compositions and a task-specific target. For instance, [26] train a shift-reduce parser using reinforcement learning (RL) without using any ground truth.

[27] replace the shift-reduce parser with a Cocke–Younger–Kasami chart parser (CYK) [28–30] and completely differentiate it using the softmax annealing method. Nevertheless, since the chart parser needs $O(n^3)$ time and space complexity, their approach has problems with both time and space. According to the easy-first parsing technique proposed by [31], each pair of neighboring nodes is scored using a query vector, and the best pair is greedily combined into one parent node at each stage. They allow end-to-end training by computing parent embedding in a hard categorical gating approach using the Straight-Through Gumbel-Softmax estimator (STG) [32]. Using various datasets, [33] compare the aforementioned models and show that [31] perform the best.

2.7. Attention-based Methods

Inter-attention [34,35], which needs a pair of sentences to attend with each other, and intra-attention [36, 37], which just requires the phrase, may be categorized as attention-based approaches. The latter is more flexible than the former. [38] use graphical models rather than recursive trees to include structural distributions into attention networks. Notice that current latent tree-based models treat all input words identically as leaf nodes and neglect the fact that various words contribute to sentence semantics to differing degrees, despite the fact that this is the basic driver of the attention process.