Neuron level Interpretation of Deep NLP models

Introduction

The proliferation of deep neural networks in various domains has seen an increased need for interpretability of these methods. A plethora of research has been carried out to analyze and understand components of the deep neural network models. Preliminary work done along these lines and papers that surveyed such were focused on a more high-level representation analysis. However, a recent branch of work has concentrated on interpretability at a more granular level, analyzing neurons and groups of neurons in these large models. In this survey, analysis is done on fine-grained neuron analysis including i) methods developed to discover and understand neurons in a network, ii) their limitations and evaluation, iii) major findings including cross architectural comparison that such analyses unravel and iv) direct applications of neuron analysis such as model behavior control and domain adaptation.

Neuron

State-of-the-art neural networks, such as RNNs or Transformer models consist of various components such as blocks, layers, attention heads, gates/cells, etc. We

use the term neuron to refer to the output of a single dimension from any of the aforementioned neural network components. For example, in the BERT base model, the output of a layer block has 768 neurons and the output of an attention head has 64 neurons. In the literature, neurons have also been referred to as features, experts, and units.

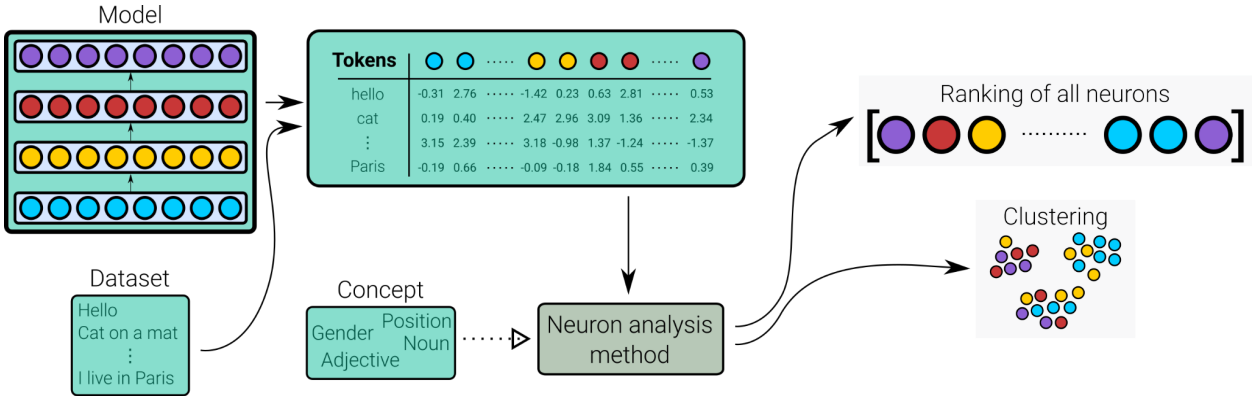
Concept

A concept is a property that we would like to interpret the model against. It can be a shallow phenomenon such as the presence of a particular lexical item (word, phrase, symbol emoticon, etc.) or its position in a sentence, length of a sentence, etc., or abstract concepts like the presence of a noun (morphology), the occurrence of an event (semantic) or preposition phrase (syntactic) in a sentence.

The objective of neuron analysis

Given a neural network model, neuron analysis aims to identify neurons that: i) learn a specific concept, or ii) are important to the model. An important facet is that some methods require a concept C as an additional input, in which case the importance is computed against the concept, while other methods just require the model M, where importance is measured with respect to the overall model and not a specific concept. For example, considering gender as a concept, the goal of neuron analysis would be to extract neurons from a model that encodes gender information.

Each neuron ni 2 N is represented as a vector of activation values over some dataset D. In NLP specifically, every element of the vector corresponds to some word in D. In order to study sentence-level concepts, an aggregation of neuron activations over words in a sentence is used. Alternatively, in the case of pre-trained models, the [CLS] token representation of a fine-tuned model is used for analysis. Figure 1 summarizes the pipeline of neuron analysis.



Overview of Neuron Analysis

Neuron Analysis methods:

Visualization

A simple way to discover the role of neurons is through visualizing their activations and manually identifying patterns over a set of sentences. However given a large number of neurons in a neural network model, it is cumbersome to visualize all of them. A number of clues have been used to shortlist the important neurons for visualization such as select saturated neurons: high/low variance neurons, or ignoring the dead neurons when using the ReLU activation function. While visualization has been effectively used for neuron analysis, it has a few limitations, for example, i) it is qualitative in nature, ii) it is hard to visualize polysemous neurons, and iii), not all neurons are visually interpretable.

Corpus-based methods

Corpus-based methods discover the role of a neuron by i) searching or generating sentences in the data that maximize its activation, or ii) based on various statistics computed over sentences in a corpus. We categorize corpus-based methods into the three types that we discuss below:

Corpus Rank

Given a neuron, the Corpus Rank method sorts sentences based on the activation values. The top sentences are then visualized to identify a concept the neuron is representing. For example, In a survey, it is mentioned that words that maximally activate a neuron by analyzing its top-k context from the corpus. The Corpus Rank method makes the process less cumbersome by reducing the search space to only the top sentences, as opposed to the Visualization method where the entire corpus or a random subset is analyzed by a human.

Corpus Generation

Corpus rank methods are efficacious in identifying neurons focusing on lexical concepts. However, their space of analysis is limited to the underlying corpus. It is possible that a neuron represents a diverse concept not featured in the corpus. The Corpus Generation method addresses this problem by generating novel sentences that maximize a neuron's activations. These sentences unravel hidden information about a neuron, facilitating the annotator to better describe its role. Corpus generation has widely been explored in Computer Vision e.g. Someone used gradient ascent to generate synthetic input images that maximize the activations of a neuron. However, in NLP, a similar approach using gradient ascent can not be directly applied because of the non-continuous input. This problem was worked

around using Gumble Softmax. In a qualitative evaluation, they claimed their method surpass corpus search in interpreting neurons.

Mask-based Corpus Selection

The corpus rank and corpus generation methods require humans in the loop to analyze the concepts that are represented by the neurons. Now we discuss a set of methods that automatically select a concept given a neuron. In a paper few researchers have proposed a Masked-based Corpus Selection method to determine important neurons with respect to a concept. The idea is to find the overlap between the presence of a concept in a set of sentences and the high activation values of a neuron. More specifically, by creating a binary mask of a neuron based on a threshold on its activation values, for every sentence in the corpus. And a binary mask for every concept based on its presence or absence in a sentence. Then they use an intersection-over-union (IoU) to compute the overlap between a given neuron mask vector and a given concept mask vector and rank all the neurons for a given concept by their IoU score. Different from them, another researcher used the values of neuron activations as the prediction score and compute the average precision score per neuron and per concept.

Neuron Probing

The visualization and corpus-based methods involve analyzing neurons from a data perspective, relying on sentences from corpora or automatically generating them. The Neuron Probing method trains a post-hoc classifier to identify neurons with respect to a pre-defined concept. Specifically, given supervised data for a concept, say Noun, it extracts activations of all the noun words in context. A classifier is then trained using these activations as features. A linear classifier is typically used. But some researchers have also tried a Random Forest classifier to derive neuron ranking.

Random Forest classifier defines the importance of each input feature by computing purity at various nodes in the trees. A ranking similar to the linear classifier can therefore be created using this importance score. Training both
linear and random forest classifiers involve various hyper-parameters such as regularization, number, and depth of trees, etc. which have a direct effect on the ranking produced. These settings are largely unexplored.

The limitation of the probing classifier is its dependence on supervised data for the concept of interest. Creating gold standard data requires careful and expensive annotations, which may not always be feasible. Therefore, most of the work using

probing classifiers is focused on pre-defined linguistic concepts which may not truly reflect all the concepts a model has learned.

Unsupervised methods

Unsupervised methods aim to unveil the knowledge patterns in the model without probing against any pre-defined concept. This enables unsupervised methods to uncover novel concepts other than the pre-defined human-engineered features. In contrast to the probing classifiers, another goal in the unsupervised method is to analyze what knowledge is captured within the most salient neurons of the model. A heterogeneous set of unsupervised methods have been proposed in the literature to analyze neurons. We discuss them as follows.

Ablation

The key idea in ablation is to identify the importance of a neuron or a group of neurons w.r.t the model or certain concept. This is typically done by masking the neurons from the network and observing their performance. Ablated neurons individually in an LSTM model and identified the most salient neurons for the network. Given the inherent redundancy in large pre-trained models, the ranking obtained using single-neuron ablation may not be very meaningful. However, trying all permutations of neurons is an intractable problem.

Few researchers have grouped redundant and similar neurons using correlation clustering. Perhaps performing a multi-neuron ablation on top of their method could be a potential solution to handle this problem.

Distribution based Probing

Distribution probing banks on the assumption that neuron activations
follow a certain distribution with respect to specific concepts. Individual probes for a neuron or a group of them are extracted from such a distribution. For instance, someone proposed that neuron activations exhibit Gaussian distributions with respect to concepts like past and present verbs. The first fit a multivariate gaussian using all neurons across a dataset to extract interesting linguistic concepts such as tense and number. A pitfall to their approach is that activations following a gaussian prior do not always hold in practice. Therefore the analysis is limited to the neurons that satisfy the criteria. Moreover, the interpretation is limited to single neurons and interpreting a group of neurons requires an expensive greedy search.

Matrix Factorization

Another method to group information is via Matrix Factorization (MF). The idea is to decompose a large matrix into a product of smaller matrices of factors, where each factor represents a group of elements performing similar function. Given activations of neurons for a sentence, matrix factorization can be effectively applied to identify neuron groups learning a concept, as shown on vision models. We could not find any research work using MF on the NLP models.

Compared to the previously discussed unsupervised methods, MF has an innate benefit of analyzing groups of neurons. However, the number of groups (factors) to decompose the activations matrix into is still a hyperparameter. Another con is that resulting analysis is per sentence. Generalizing the findings to a large set is a non-trivial frontier that requires further exploration.

Clustering Methods

Clustering is another effective way to analyze group of neurons in an unsupervised fashion. The intuition is that if a group of neurons target specific concepts, then their activations would naturally form a cluster. Meyes used UMAP to project neuron activations to a low dimension space and then performed K-means clustering to group neurons. One researcher aimed at identifying redundant neurons in the network. They first calculate correlation between neuron activation pairs and then used hierarchical clustering to group them. The neurons with highly correlated behavior are clustered together and are considered redundant in the network. Similar to the MF method, the number of clusters is a hyperparameter that needs to be pre-defined or selected empirically.

Multi-model Search

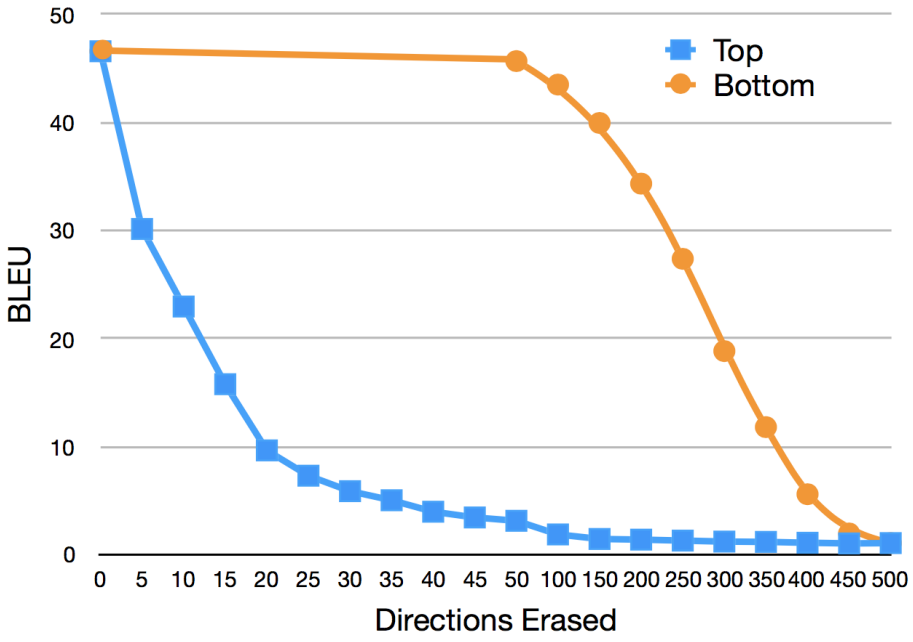
Multi-model search is based on the intuition that salient information is shared across the models trained towards a task i.e. if a concept is important for a task then all models learned to optimize the task should learn it. The search involves identifying neurons that behave similarly across the models. Someone used Pearson correlation to calculate a similarity score of each neuron of a model with respect to the neurons of other models. They aggregated the correlations for each neuron using several methods with the aim of highlighting different properties and findings. More specifically, they used Max Correlation to capture concepts that emerge strongly in multiple models, Min Correlation to select neurons that are correlated with many models though they are not among the top correlated neurons, Regression Ranking to find individual neurons whose information is distributed among multiple neurons of other models, and SVCCA to capture information that may be distributed in fewer dimensions than the whole representation.

Evaluation

The neuron analysis methods discussed so far provide a set of salient neurons or a ranking of neurons with respect to a model or a concept. Here, we survey methods to evaluate the correctness of the selected neurons or their ranking.

Ablation

While ablation has been used to discover salient neurons for the model, it has also been used to evaluate the efficacy of the rankings. The idea is to ablate neurons in the model in the order of their importance. For neuron ranking to be correct, removing the top neurons should result in larger drop in performance compared to removing same amount of random or bottom neurons from that list. Few of the researchers have used ablation to demonstrate correctness of their neuron ranking obtained via neuron-probing classifier method. Similarly, someone ablated the most salient neurons obtained using multi-model search in a neural machine translation model and showed that it leads to a much bigger drop as opposed to removing random neurons .



Ablating top vs. bottom SVCCA directions on a machine translation model

Classification Performance

Given salient neurons with respect to a concept, a simple method to evaluate their correctness is to train a classifier using them as features and predict the concept of

interest. The performance of the classifier relative to a classifier trained using random neurons and least important neurons is used as metric to gauge the efficacy of the selected salient neurons. Someone trained classifiers with the selected neurons to show the efficacy of their rankings w.r.t core-linguistic tasks of predicting morphology, syntax and semantics.

Cluster Comparison

Clustering-based methods result in groups of neurons that learn similar information. An accurate evaluation of such methods requires comparing them against ground-truth clusters. However, such gold-annotated clusters don't exist. A number of researchers, for instance, One researcher have used techniques such as B-cubed and Normalized Pointwise mutual information scores to compare the clusters to Named Entities and Universal Dependency datasets. Another way is to use the manually defined linguistic categories, such as part-of-speech tags, as a gold standard to evaluate their efficacy.

Qualitative Evaluation

The neuron analysis methods, particularly Visualization and Corpus-based methods have also been effectively used to provide a qualitative evaluation of the selected set of neurons. For instance, a Neuron Probing classifier identifies a set of salient neurons w.r.t a concept. The activations of these neurons are often visualized to verify the correctness of the selection. The corpus-based methods can be effective in reducing the search space of a large number of sentences to visualize. They enable ranking the sentences based on the activation values of a neuron, limiting the need for visualization for relevant sentences.

Applications

In this section, we cover various applications of neuron-level interpretation: i) Controlling model's behavior, ii) Model distillation and efficiency, and iii) Domain adaptation.

Controlling model behavior

Knowing the role of a neuron with respect to a concept enables controlling the model's behavior. Someone identified Switch Neurons in neuron machine translation models that activate positively for present-tense verbs and negatively for past tense verbs. By manipulating the values of the switch neuron at test time, researchers were able to successfully change output translations from present to past tense. The authors additionally found neurons that capture gender and number agreement concepts and manipulated them to control the system's output. Another effort along

this line was to manipulate the neurons responsible for a concept in the GPT model and generate sentences around specific topics of interest.

Model Distillation and Efficiency

Deep NLP models are trained using hundreds of millions of parameters, limiting their applicability in computationally constrained environments. Identifying salient neurons and sub-networks can be useful for model distillation and efficiency. Someone devised an efficient feature-based transfer learning procedure, stemmed from their redundancy analysis. By exploiting layer and neuron-specific redundancy in the transformer models, they were able to reduce the feature set size to less than 100 neurons for several tasks while maintaining more than 97% of the performance.
The procedure achieved a speedup of up to 6.2x in computation time for sequence labeling tasks as opposed to using all the features.

Domain Adaptation

Identifying the salient neurons with respect to a domain can be effectively used for domain adaptation and generalization. Researchers proposed a domain adaptation method using neuron pruning to target the problem of catastrophic forgetting of the general domain when fine-tuning a model for a target domain. They introduced a three-step adaptation process: i) rank the most important neurons based on their importance, ii) prune the unimportant neurons from the network and retrain with the student-teacher framework, iii) expand the network to its original size and fine-tune towards in-domain freezing the salient neurons and adjusting only the unimportant neurons. Using this approach helps to avoid catastrophic forgetting of the general domain while also obtaining an optimal in-domain model.

Compositional Explanations

Knowing the association of a neuron with a concept enables an explanation of the model's output. Someone identified neurons learning compositional concepts in vision and NLP models. Using a composition of logical operators, they provided an explanation of the model's prediction. The neuron activates for contradiction when the premise contains the word man. Such explanations provide a way to generate adversarial examples that change the model's predictions.