Latent Concept-based Explanation of NLP Models[*]

Xuemin Yu
Faculty of Computer Science,
Dalhousie University, Canada
xuemin.yu@dal.ca

Fahim Dalvi
Qatar Computing Research
Institute, HBKU, Qatar
faimaduddin@hbku.edu.qa

Nadir Durrani
Qatar Computing Research
Institute, HBKU, Qatar
ndurrani@hbku.edu.qa

Marzia Nouri
nouri.marzia.1999
@gmail.com

Hassan Sajjad
Faculty of Computer Science,
Dalhousie University, Canada
hsajjad@dal.ca

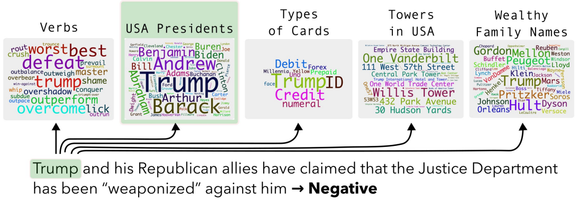

Verbs
USA Presidents
Types of Cards
Towers in USA
Wealthy Family Names
Trump and his Republican allies have claimed that the Justice Department has been "weaponized" against him → Negative

Figure 1: An example of various facets of word "trump"

Extended Abstract

The opaqueness of deep neural network (DNN) models is a major challenge to ensuring a safe and trustworthy AI system. Extensive and diverse research works have attempted to interpret and explain these models. One major line of work strives to understand and explain a neural network model's prediction using input words' attribution to prediction (Sundararajan et al., 2017; Denil et al., 2014).

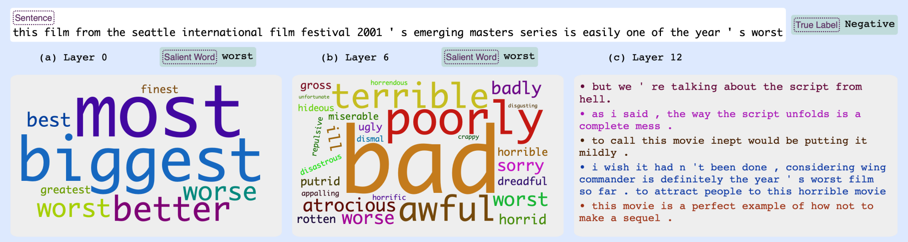
However, the explanation based solely on input words is less informative due to the discrete nature of words and the lack of contextual verbosity. A word consists of multifaceted aspects such as semantic, morphological, and syntactic roles in a sentence. Consider the word "trump" in Figure 1. It has several facets such as a verb, a verb with specific semantics, and a named entity representing a certain aspect such as tower names, family names, etc. We argue that given various contexts of a word in the training data, the model learns these diverse facets during training. Given a test instance, depending on the context a word appears, the model uses a particular facet of the input words in making the prediction. The explanation based on salient words alone does not reflect the facets of the word the model has used in the prediction and results in a less informed explanation.

Dalvi et al. (2022) showed that the latent space of DNNs represents the multifaceted aspects of words learned during training. The clustering of training data contextualized representations provides access to these multifaceted concepts, hereafter referred to as *latent concepts*. Given an input word in context at test time, we hypothesize that the alignment of its contextualized representation to a latent concept represents the facet of the word being used by the model for that particular input. We further hypothesize that this latent concept serves as a correct and enriched explanation of the input word. To this end, we propose the LAtent COncept ATTribution (LACOAT) method that generates an explanation of a model's prediction using the latent concepts. LACOAT discovers latent concepts of every layer of the model by clustering contextualized representations of words in the training corpus. Given a test instance, it identifies the most salient input representations of every layer with respect to the prediction and dynamically maps them to the latent concepts of the training data. The shortlisted latent concepts serve as an explanation of the prediction. Lastly, LACOAT integrates a plausibility module that generates a human-friendly explanation of the latent concept-based explanation.

Methodology & Results Consider a sentiment classification dataset and a sentiment classification model as an example. LACOAT works as follows: ConceptDiscoverer takes the training dataset and the model as input and outputs latent concepts of the model. At test time, given an input sentence, PredictionAttributor identifies the most salient input representations with respect to the prediction. ConceptMapper maps these salient input representations to the training data latent concepts and provides them as an explanation of the prediction. PlausiFyer takes the test sentence and its concept-based explanation and generates a human-friendly and insightful explanation of the prediction.

We perform a qualitative evaluation of LACOAT to

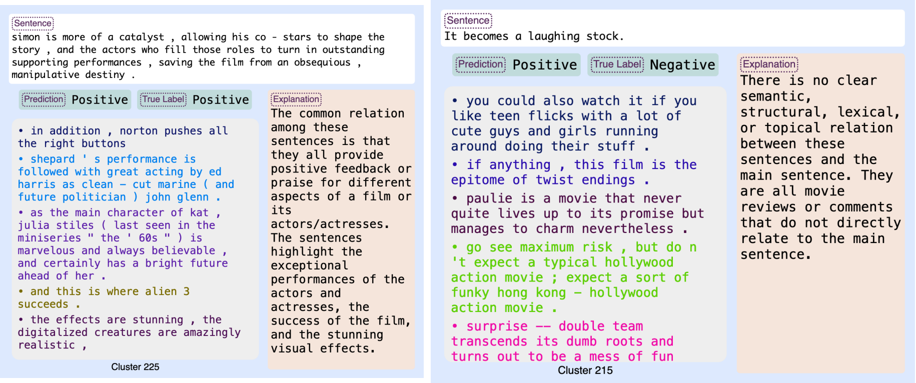
[*] The full paper link: https://arxiv.org/pdf/2404.12545



this film from the seattle international film festival 2001 ' s emerging masters series is easily one of the year ' s worst

- but we ' re talking about the script from hell.
- as i said , the way the script unfolds is a complete mess .
- to call this movie inept would be putting it mildly .
- i wish it had n 't been done , considering wing commander is definitely the year ' s worst film so far . to attract people to this horrible movie
- this movie is a perfect example of how not to make a sequel .

Figure 2: Sentiment task: Latent concepts of the most attributed words in Layers 0, 6 and 12



simon is more of a catalyst , allowing his co - stars to shape the story , and the actors who fill those roles to turn in outstanding supporting performances , saving the film from an obsequious , manipulative destiny .

- in addition , norton pushes all the right buttons
- shepard ' s performance is followed with great acting by ed harris as clean – cut marine (and future politician) john glenn .
- as the main character of kat , julia stiles (last seen in the miniseries " the ' 60s ") is marvelous and always believable , and certainly has a bright future ahead of her .
- and this is where alien 3 succeeds .
- the effects are stunning , the digitalized creatures are amazingly realistic ,

The common relation among these sentences is that they all provide positive feedback or praise for different aspects of a film or its actors/actresses. The sentences highlight the exceptional performances of the actors and actresses, the success of the film, and the stunning visual effects.

(a) Sentiment: A positive labeled test instance correctly predicted by the model.

It becomes a laughing stock.

- you could also watch it if you like teen flicks with a lot of cute guys and girls running around doing their stuff .
- if anything , this film is the epitome of twist endings .
- paulie is a movie that never quite lives up to its promise but manages to charm nevertheless .
- go see maximum risk , but do n 't expect a typical hollywood action movie ; expect a sort of funky hong kong – hollywood action movie .
- surprise –– double team transcends its dumb roots and turns out to be a mess of fun

There is no clear semantic, structural, lexical, or topical relation between these sentences and the main sentence. They are all movie reviews or comments that do not directly relate to the main sentence.

(b) Sentiment: A negatively labeled test instance that is incorrectly predicted as positive.

Figure 3: A few examples of LACOAT explanations for BERT using Sentiment tasks

evaluate the usefulness of the latent concept-based explanation and the generated human-friendly explanation with the sentiment task.

Evolution of Concepts LACOAT generates the explanation for each layer with respect to a prediction, which shows the evolution of concepts in making the prediction. Figure 2 shows layers 0, 6 and 12's latent concept of the most attributed input token for RoBERTa fine-tuned on the sentiment task. We found that the initial layer latent concepts do not always align with the sentiment of the input instance and may represent a general language concept. Figure 2(a) shows the concept of comparative and superlative adjectives of both positive and negative sentiments. In the middle layers, the latent concepts evolved into concepts that align better with the sentiment of the input sentence. The latent concept of Figure 2(b) shows a mix of adjectives and adverbs of negative sentiment. In the sentiment task, the most attributed word in the last layer is [CLS] which resulted in latent concepts consisting of [CLS] representations of the most related sentences to the input. We randomly pick five [CLS] instances from the latent concept and show their corresponding sentences in the figure (see Figure 2(c)). We found that the last layer's latent concepts are best aligned with the input instance and its prediction and are the most informative explanation of the prediction. Then, we deepen our analysis of the explanations generated using the last layer only.

Analyzing Last Layer Explanations Figure 3 presents various examples of LACOAT for Sentiment tasks using BERT.

Correct prediction with correct gold label Figure 3a presents a case of correct prediction with latent-concept explanation and human-friendly explanation. In the case of sentence-level latent concepts (Figure 3a), it is harder to interpret compared to latent concepts consisting of words. However, PlausiFyer still highlights additional information about the relation between the latent concept and the input sentence. For example, it captures that the reason of positive sentiment in 3a is due to praising different aspects of a film and its actors and actresses.

Wrong prediction with correct gold label Figure 3b shows the predicted label is wrong. The input sentence has a negative sentiment but the model predicted it as positive. The instances of latent concepts show sentences with mixed sentiments such as “manages to charm” is positive, and “never quite lives up to its promise” is negative. This provides the domain expert an evidence of a possible wrong prediction. The PlausiFyer's *explanation* is even more helpful as it clearly states that “there is no clear ... relation between these sentences ...".

References

Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. 2022. Discovering latent concepts learned in BERT. In *International Conference on Learning Representations*.

Misha Denil, Alban Demiraj, and Nando de Freitas. 2014. Extraction of salient sentences from labelled documents. *CoRR*, abs/1412.6815.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. *arXiv preprint*. ArXiv:1703.01365 [cs].