Research Statement

Danish Pruthi
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
ddanish@cs.cmu.edu

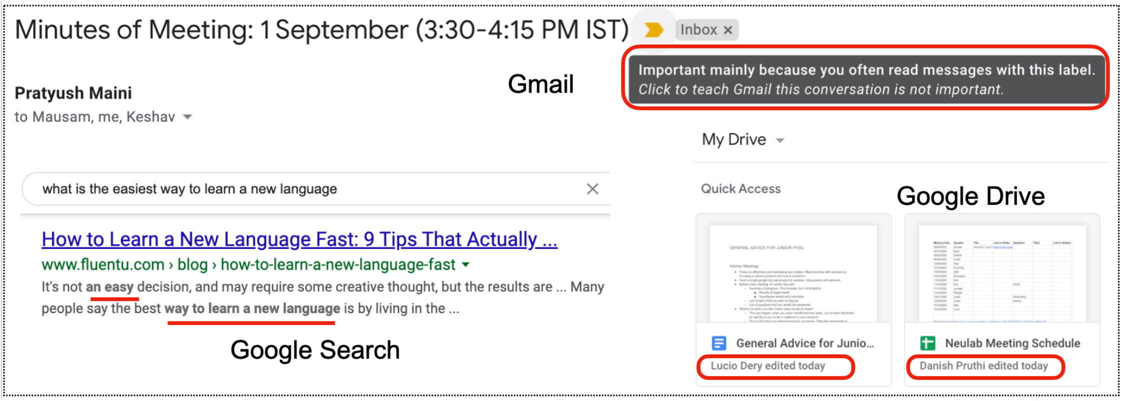
1 Introduction

In the last decade, we have witnessed remarkable progress in natural language understanding, largely due to the resurgence of deep neural networks. Supervised deep learning models boast impressive predictive abilities. However, they remain *uninterpretable* and are commonly treated as “black boxes”. Researchers, practicioners, policy makers, and journalists have begun to express concerns that we must back our predictions by interpretations for the purpose of assessing the quality and morality of these predictions. Amidst growing concerns for interpretability, I am interested in the following questions:

- How to supplement predictions with evidence that users can use to verify the validity of predictions?
- How to evaluate the quality of a given explanation?
- How to ensure (and measure the degree to which) an explanation is *faithful* to the model?

Below, I outline some of the proposed and past research that aims to address these questions.

For many applications, end users desire not only predictions but also supporting evidence so that they can readily verify the prediction. This ability to verify results engenders trust among users and increases adoption of machine learning systems [2, 4, 13]. Fortunately, for many problems, a localized portion of the input is sufficient to validate the predicted label. In a large image, a small patch of an image containing a hamster may be sufficient to render the hamster label applicable. Similarly, in a long medical record, a single sentence may suffice to identify a certain diagnosis. For the task of evidence extraction, we propose several new methods to combine scarce evidence annotations (strong semi-supervision) with abundant document-level labels (weak supervision). We find that our methods outperform baselines adapted from the interpretability literature to our task (refer to §2.1 for details). Our techinques could potentially enhance explanatory information provided by many Google services including Gmail, Ads, Search, etc. (Figure 1, 2).



Minutes of Meeting: 1 September (3:30-4:15 PM IST)

Pratyush Maini
to Mausam, me, Keshav

Important mainly because you often read messages with this label.
Click to teach Gmail this conversation is not important.

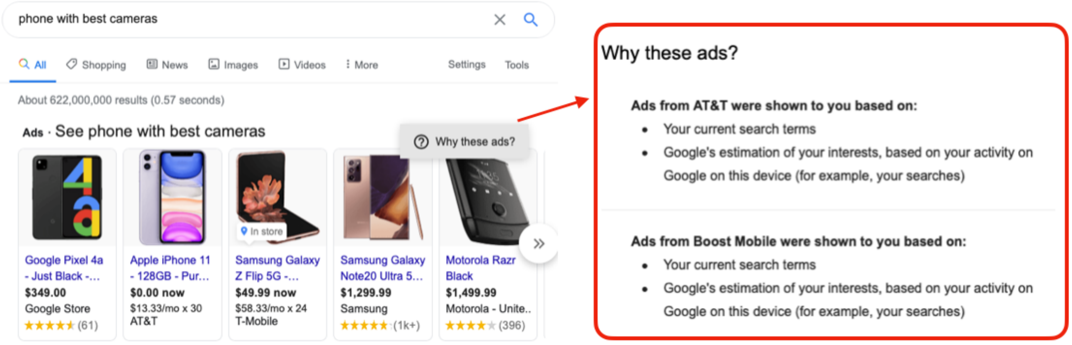
what is the easiest way to learn a new language

How to Learn a New Language Fast: 9 Tips That Actually ...
www.fluentu.com › blog › how-to-learn-a-new-language-fast
It's not **an easy** decision, and may require some creative thought, but the results are ... Many people say the best **way to learn a new language** is by living in the ...

General Advice for Junio...
Lucio Dery edited today

Neulab Meeting Schedule
Danish Pruthi edited today

Figure 1: Several Google services provide some form of explanatory information. Top: Gmail Magic tool suggests why an email was marked important. Bottom left: a search result highlights article tokens that are similar to the query tokens. Bottom right: the quick access tool offers brief reasons for the recommended documents in Google Drive.



About 622,000,000 results (0.57 seconds)

Ads · See phone with best cameras

Google Pixel 4a - Just Black -...	Apple iPhone 11 - 128GB - Pur...	Samsung Galaxy Z Flip 5G -...	Samsung Galaxy Note20 Ultra 5...	Motorola Razr Black
$349.00	$0.00 now	$49.99 now	$1,299.99	$1,499.99
Google Store	$13.33/mo x 30 AT&T	$58.33/mo x 24 T-Mobile	Samsung	Motorola - Unite..
★★★★★ (61)			★★★★★ (1k+)	★★★★☆ (396)

Why these ads?

Ads from AT&T were shown to you based on:

- Your current search terms
- Google's estimation of your interests, based on your activity on Google on this device (for example, your searches)

Ads from Boost Mobile were shown to you based on:

- Your current search terms
- Google's estimation of your interests, based on your activity on Google on this device (for example, your searches)

Figure 2: Why these ads? Partial explanations for the displayed advertisements.

Despite the large body of recent work on interpretability research (see xxx for a survey), there is little consensus on how to evaluate the quality of explanations. To make matters worse, across papers, the motives of intepretability are diverse and occasionally discordant [9]. In an ongoing research effort (in collaboration with Google researchers) to alleviate these concerns, we propose a novel framework to assess the value of explanations. Our framework draws upon the common use case of explanations—to *communicate information about how the decisions are made*, and thereby indicate how future decisions will be taken. We operationalize this usage via a teacher-student framework, where the teacher and the student could be humans or computer systems. As per the framework, the goal of the student is to build a model that approximates the teacher's model using the input, output, and explanations from the teacher. Explanations provided by the teacher are effective if they help students predict better. Furthermore, the framework ensures that unfaithful explanations will not improve student's predictive ability (see §2.2 for more details).

In a recent project, we also characterize the manipulability of popularly used attention-based explanations.

2 Ongoing and Past Research

2.1 Evidence Extraction

Despite the success of deep learning for countless prediction tasks, practitioners often desire that these models not only be accurate but also provide *interpretations* or *explanations* [1, 15]. Unfortunately, these terms lack precise meaning, and across papers, such explanations purport to address such a wide spectrum of desiderata that it seems unlikely any one method could address them all [9]. In both computer vision [13, 14] and natural language processing [8, 7], proposed explanation methods often take the form of highlighting salient features of the input. These so-called *local explanations* are intended to highlight sets of features that elucidate "the reasons behind predictions" However, this characterization of the problem remains under-specified. Further, due to confounding, many features may be predictive but do not constitute *evidence* [5].[1]

Instead, we focus on supplementing predictions with evidence that gives users the ability to quickly verify the correctness of machine predictions. Fortunately, for many problems, a localized portion of the input is sufficient to validate the predicted label. Thus motivated, we cast our problem as learning to extract evidence from both strong and weak supervision. The former takes the form of explicit, but scarce, manual annotations of evidence segments, whereas the latter is provided by only input documents and their class labels, which we assume are available in abundance. In the extreme case where evidence annotations are available for all examples, our task collapses to a standard multitask learning problem, and in the opposite extreme, where only weak supervision is available, we find ourselves back in the under-specified realm addressed by the interpretability literature.

We optimize the joint likelihood of class labels and evidence spans, given the input examples ($P(y, e|x)$)

[1] For instance, in the IMDb movie review dataset the token "horror" is predictive of negative sentiment because horror movies tend to receive poorer ratings than movies from other genres [5]. However, no expert would mark it to be the evidence justifying the negative review.

Movie Review
I don't know what movie the critics saw, but it wasn't this one. The popular consensus among newspaper critics was that **this movie is unfunny and dreadfully boring** . In my personal opinion, they couldn't be more wrong. If you were expecting Airplane! - like laughs and Agatha Christie - intense mystery, then yes, this movie would be a disappointment. However, if you're just looking for **an enjoyable movie and a good time** , this is **one** to see ...
Lean, mean, escapist thrillers are a tough product to come by. **Most are unnecessarily complicated** , and others have no sense of expediency–the thrill-ride effect gets lost in the **cumbersome** plot. Perhaps the ultimate escapist thriller was the fugitive, which featured none of the flash-bang effects of today's market but rather a bread-and-butter, **textbook example of what a clever script and good direction is all about.** ...

Table 1: Non cherry-picked evidence extractions from our approach. We condition our extraction model on both the **positive** and the **negative** label. Our approach is able to tailor the extractions as per the conditioned label.

We factorize our objective such that we first *classify* ($P(y|x)$), and then *extract* the evidence conditioned on the predicted label ($P(e|y,x)$). For classification, we use BERT. The extraction task (a sequence tagging problem) is modeled using a linear-chain CRF that takes representations and attention scores from BERT as emission features, allowing the two tasks to benefit from shared parameters. Further, the evidence extraction module is conditioned on the predicted label, enabling the CRF to output different evidence spans tailored to the predicted class label. This is illustrated in in Table 1.

We directly compare our methods against input attribution methods from the interpretability literature. Many approaches in this category first *extract, and then classify*Across two text sequence classification and evidence extraction tasks, we find our methods to outperform baselines. Encouragingly, we observe gains by using our approach with as few as 100 evidence annotations.

2.2 Explanations as Communication

One familiar purpose of explanations for humans is to communicate information about how decisions are made, which also indicate how future decisions will be made. We formalize this specific use case of explanations: assume a teacher t, a student s, and a series of inputs $x_1, \ldots, x_n$, and suppose the student's goal is to predict what the teacher will do in the future. The student's problem is thus a learning task, where the student needs to build a model $f_s(x)$ that approximates the teacher's model $f_t(x)$. Explanations are modeled as additional side-information $e_t(x)$ that can be provided by the teacher in addition to label $f_t(x)$ for input x. Explanations are effective if a student can approximate $f_t(x)$ better or faster with explanations than with teacher-labeled examples alone. The degree by which different explanations help the student model provides a quantitative assessment of the value of explanations. The above framework can be decomposed into two subproblems: i) for a teacher model t, how to obtain the explanatory information $e_t(x)$; and ii) how can the student model leverage this side-information $e_t(x)$ for training?

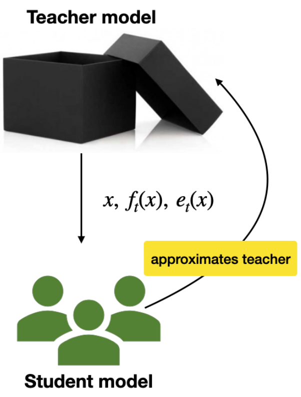


Figure 3: Explanation as communication framework.

While the framework is broadly applicable, we apply it for question answering (QA) tasks. We start with gold explanations collected from human experts. Each explanation contains necessary entities in the question and their relevant references in the passage. We propose two new techniques to incorporate the explanation during training. The first approach regularizes a (few) attention heads in a (few) layers to align with the symbolic mappings given by $e_t(x)$. In the second method, we jointly predict i) the participating entities in the question and the passage using a linear-chain CRF [6], and ii) the start and end index of the answer span. For this multi-task setup, we share the encoder representations so that the task of QA would benefit from predicting the relevant entities. We observe that regularizing attention in such a manner results in an improvement of 6.4 F1 points in the student QA model, and jointly predicting entities and answers improves the QA performance by 5.2 F1 points. Furthermore, we observe that student models perform better with increasing quantity and quality of explanations. These preliminary results indicate that our framework can be an effective paradigm to formalize the value of explanations. A more comprehensive assessment of this framework is an active ongoing effort.

2.3 Manipulating Explanations

Since their introduction as a means to cope with unaligned inputs and outputs in neural machine translation, attention mechanisms have emerged as popular and effective components in various neural network architectures. Attention works by aggregating a set of tokens via a weighted sum, where the *attention weights* are calculated as a function of both the input encodings and the state of the decoder. Because attention mechanisms allocate weight among the encoded tokens, these coefficients are sometimes thought of intuitively as indicating which tokens the model *focuses on* when making a particular prediction. Based on this loose intuition, attention weights are often purported to *explain* a model's predictions.

In this study [12], we elucidate two potential pitfalls why one should exercise caution in interpreting attention scores as indicative of models' inner workings or relative importance of input tokens. First, we demonstrate that attention scores are surprisingly easy to manipulate by designing a training scheme whereby the resulting models appear to assign little attention to any among a specified set of *impermissible* tokens while nevertheless continuing to rely upon those features for prediction (see Table **??** for an example). The ease with which attention can be manipulated without significantly affecting predictions suggests that even if a vanilla model's attention weights conferred some insight (still an open and ill-defined question), these insights would rely on knowing the precise objective on which models were trained. Second, practitioners often overlook the fact that the attention is not over words but over final layer representations, which themselves capture information from neighboring words. Our results present troublesome implications for proposed uses of attention in the context of fairness, accountability, and transparency. For example, malicious practitioners asked to justify *how their models work* by pointing to attention weights could mislead regulators with this scheme.

2.4 Interpretable Representations

Distributed word representations (or word embeddings) are ubiquitous components of neural architectures applied in natural language processing. However, word embeddings are dense representations that people find difficult to interpret. For instance, we are often clueless as to what a high value along a given dimension of a vector signifies when compared to a low value. Here, our notion of interpretibility — one that requires each dimension to capture a semantic concept — is pragmatically motivated. In many classification problems, a linear (pre-softmax) layer projects representations hitherto to (un-normalized) class probabilities. Larger values of weights sway the output class probabilities greatly. In order to explain a prediction, one necessarily has to understand the semantic concepts that each of the dimensions corresponding to these large weights represent. Hence, this notion of post-hoc interpretability is useful in explaining predictions.

To confer interpretability in word representations, we draw inspiration from various feature norming studies [3, 10], where participants were asked to enumerate the properties of several words and concepts. It was observed that participants typically used few *sparse* characteristic properties to describe the words, with limited overlap between different words. For instance, to describe the city of Pittsburgh, one might describe phenomena typical of the city, like erratic weather and large frequent bridges. It is redundant and inefficient to list negative properties, like the absence of the Statue of Liberty. Thus, sparsity and non-negativity are desirable characteristics of representations, that make them interpretable.

We exploited these findings to devise a novel denoising k-sparse autoencoder (see Figure 4) to obtain **SP**arse **I**nterpretable **N**eural **E**mbeddings (**SPINE**), a transformation of input word embeddings [11]. These autoencoders are highly expressive, and facilitate learning of non linear transformations in contrast to linear matrix factorization. Further, they can be seamlessly integrated into a general neural network pipeline. Using our formulation, we attained representations that were twice as interpretable (as measured via intrusion detection tests) when compared to state of the art methods, with no loss in performance.

[Add other projects – Danish].

References

[1] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings*

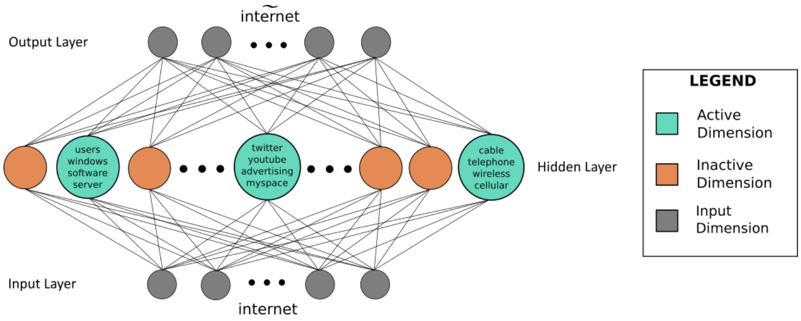

$\widetilde{\text{internet}}$
Output Layer
users windows software server
twitter youtube advertising myspace
cable telephone wireless cellular
Hidden Layer
Input Layer
internet
LEGEND
Active Dimension
Inactive Dimension
Input Dimension

Figure 4: Depiction of our k-sparse autoencoder for an input word ‘internet’. Our variant of the k-sparse autoencoder attempts to reconstruct the input at its output layer, with only a few active hidden units (depicted in green). These active units correspond to an interpretable set of dimensions associated with the word ‘internet’. The rest of the dimensions (depicted in orange) are inactive for this word.

of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pages 1721–1730, 2015.

[2] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718, 2003.

[3] Peter Garrard, Matthew A Lambon Ralph, John R Hodges, and Karalyn Patterson. Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive neuropsychology*, 18(2):125–174, 2001.

[4] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250, 2000.

[5] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

[6] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *18th International Conference on Machine Learning 2001 (ICML 2001)*, 2001.

[7] Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. Inferring which medical treatments work from reports of clinical trials. *arXiv preprint arXiv:1904.01606*, 2019.

[8] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. *Proceedings of the conference on Empirical methods in natural language processing*, 2016.

[9] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

[10] Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559, 2005.

[11] Danish Pruthi, Anant Subramanian, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. Spine: Sparse interpretable neural embeddings. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. Learning to deceive with attention-based explanations. In *Annual Conference of the Association for Computational Linguistics (ACL)*, July 2020. URL https://arxiv.org/abs/1909.07913.

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ” why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[14] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[15] Daniel S Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62(6):70–79, 2019.