Rationale-based Human-in-the-Loop via Supervised Attention

Teja Kanchinadam, Keith Westpfahl, Qian You, Glenn Fung
(tkanchin,kwestpfa,qyou,gfung)@amfam.com
American Family Insurance
Madison, WI

ABSTRACT

In machine learning, "Human-in-the-Loop" alludes to algorithms that incorporate human interaction in the learning process to either improve algorithm performance or to complement the information provided by the data. The most recent research on text classification focus on human annotation at the instance level. Our work describes a simple and effective approach to incorporate human annotated text highlights as rationale's to their instance level annotations as a form of auxiliary human feedback that can significantly complement data during training. This process can be seen as a supervised attention mechanism coupled with an active learning process. Specifically, we add a light-weight attention mechanism to our feed-forward neural network classifier that is computationally inexpensive. This design is simple and especially appropriate for active learning models that require regular retraining. Experiments on several publicly available datasets empirically show that our model outperforms other baseline approaches by a significant margin. We also show experiments on an insurance domain dataset where we achieve better classification performance given the same or smaller labeling budget.

ACM Reference Format:
Teja Kanchinadam, Keith Westpfahl, Qian You, Glenn Fung. 2020. Rationale-based Human-in-the-Loop via Supervised Attention. In *Proceedings of KDD 2020 (DaSH@KDD).* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

There has been renewed research interest in minimizing the amount of human feedback needed to train machine learning models. Given the high cost of acquiring so many labeled data, an increasing body of research has been exploring and adapting human-in-the-loop approaches within the context of the latest advancements in machine learning and deep learning. For instance, the active learning community has been exploring variant forms of annotators' feedback to optimize the human-in-the-loop learning process, therefore improving the underlying model performance. Particularly, there have been a few efforts [18, 21, 28] in which the human's highlights (i.e. rationale) in the text are captured as additional information to the bag-of-words features to enhance text classifiers. However, little work has been done to exploit human highlights as a highly informative form of supervision for active learning framework.

More recently, the creation and applications of attention mechanisms [1, 14, 23, 26] have not only led to breakthrough performances of deep learning models; they have also opened up opportunities for narrowing the gap between machine learning processes and human cognition.

DaSH@KDD, August 24, 2020, Virtual Conference

In this work, we propose two novel effective human-in-the-loop algorithms for text classification: rationale-based active learning via supervised attention (RALSA) and rationale-based active learning via linear model (RALM). In addition to the instance level labels, RALSA and RALM incorporate text rationale's, or explanations, into the traditional active learning to make the process more efficient. See Figure 1 for our proposed human in the loop active learning process. Our main focus is to explore the interaction between human-provided rationale and supervised attention in an active learning setting. We are aware that active learning is a vast area and a more detailed exploration of the synergy between rationale and more complex active learning algorithms and query strategies is left for future work.

RALM uses combined embedding representations of both document and rationale, then trains a linear model. RALSA uses a simple light-weight attention mechanism coupled to a shallow neural network classifier that is computationally efficient, hence it is appropriate for the active learning paradigm that retrains the underlying model often.

We rigorously tested RALM and RALSA on open source datasets and one insurance domain data set. In order to run those experiments, we use Mechanical Turk and internal labeling tool to curate three new text classification datasets with ground truth and annotated rationale. We also evaluated RALM and RALSA against their counterparts which do not use rationale information. The results have shown the rationale-based active learning, especially coupled with supervised attention can outperform their counterparts by a big margin. In summary our work provides the following

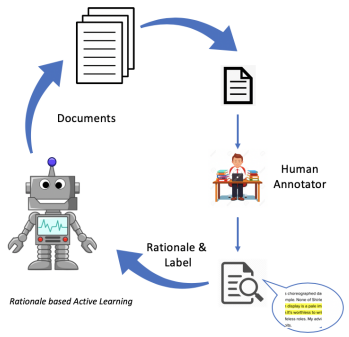

Figure 1: Proposed Human-in-the-loop Rationale-based active learning approach

contributions:

(1) Both the linear convex combination approach in RALM and the light-weighted attention mechanism in RALSA are novel ways to incorporate human highlight/rationale as a form of auxiliary human feedback into an underlying machine learning model.

(2) Incorporating human in the loop rationale especially in the form of supervised attention in active learning results in faster convergence of classifiers. At the same time, our proposed supervised attention model is computationally efficient and only requires little extra effort from annotators, which make this human-in-the-loop process simple and budget aware.
(3) We have created three new rationale-annotated labeled datasets that will be shared with the research community (there were very few in existence prior to this work)

We also qualitatively inspected a few machine generated rationale's from learnt attention weights, and then compared generated rationale to the human rationale. We found that the machine generated rationale are both syntactically and semantically similar to human annotator ones.

2 RELATED WORK

2.1 Active Learning

Active learning aims to develop label-efficient algorithms by sampling the most representative queries to be labeled by an oracle, which usually is a human annotator. Many sampling strategies have been developed over the past decades [3, 7, 16, 17, 24]. The most effective and commonly used pool-based active learning is probably uncertainty sampling [5, 20, 30]. Recently developed deep active learning also research on how to adapt new model architectures to uncertainty sampling [8, 22, 27]. Although deep models can out-perform classic uncertainty sampling, they are usually computationally inefficient.

2.2 Rationale-based Active Learning

Adding human annotated rationale's proves to be effective in improving a number of NLP tasks including text classification [29] and question and answers [12]. It is natural for the human in the loop active learning process to exploit and leverage the human annotation at the same time when the labeling oracle is provided. Few researchers [21, 28] in the active learning community explored the effectiveness of using human annotated highlights i.e. rationale. However, they only adopted preliminary feature representations to model rationale and use those representations to enhance the classification feature space. The result is interactions between rationale and the original machine learning task are not fully exploited.

2.3 Rationale and Attention

The proposal and dissemination of the standard attention mechanism [1, 14, 23, 26] have led to some of the most successful and recently-proposed language models [6, 19]. Attention, in its first introduction in [1], is a context vector which enables the models to attend to certain hidden states in the model decoding phase. And the attention mechanism is a standalone module and can be coupled with simple sequence to sequence models [14, 26], or complex encoder-decoder architectures [23].

Lately, various researchers explored different forms and structure of attention formulation to model attention that mimics human rationale's i.e. an interpretation for the machine learning class [2, 10, 13]. More sophisticated neural models are proposed to incorporate human rationale into models via attentions [12, 29]. Among them, [29] is the closest to the research in this paper. It uses human annotation as "supervision" to augment a CNN document classifier by minimizing the categorical cross entropy between model attention and human rationale. However, it trains in two steps which will probably violate the latency required by human in the loop process such as active learning. [12] generates answers and supporting evidence from a clinical trial report, given a prompt i.e. which medical treatments work. Although this paper developed multiple variants of attention to leverage prompts (i.e. manually created question) and evidence (i.e. manually created text span), their data curation process is expensive and extensive. Therefore, it is unclear if their modeling approach can be adapted to the budget constrained active learning settings.

As far as we know, we are the first active learning paper which uses attention as the form to model human rationale as supervision to the underlying text classification model. In our human-in-the-loop process, the supervised attention can use human annotators' rationale to update the model in near real time.

3 METHODOLOGY

3.1 Active Learning

The typical setup for pool-based active learning for classification is as follows: a pool of unlabeled examples $\mathcal{U}$, a pool $\mathcal{L}$ of labeled example-label pairs (x, y_x), an oracle - usually a human annotator that can supply the label of any $x \in \mathcal{U}$, and a query strategy that selects which example $x^* \in \mathcal{U}$ the oracle should label such that $\mathcal{L}^* = \mathcal{L} \cup \{(x^*, y_{x^*})\}$ yields the maximum information gain versus $\mathcal{L}$. Uncertainty Sampling is an effective and widely used query strategy. It captures the classifier's (θ) uncertainty about the class of x, and can be given as: $x^* = argmax_{x \in \mathcal{U}}(\mu(x))$, where $\mu(x)$ can be defined as the Shannon Entropy in a classification setting as $\mu(x) = -\sum_{c \in C} p_c(x) \log p_c(x)$, where (C) are our possible classes, and $p_c(x)$ is the probability that our classifier assigns to x having class c.

3.2 Rationale-based Active Learning

We base our discussion of rationale-based active learning in the context of text classification. The human annotators are asked to not only label the sample but also to highlight the *rationale* r_{x^*} for documents x^* behind their decisions. Algorithm 1 below is a simple setting of rationale based active learning.

Algorithm 1 Rationale-based Active Learning

1: **input:** $\mathcal{U}$ - unlabeled documents, $\mathcal{L}$ - labeled documents, θ - underlying classification model
2: **while** $|\mathcal{U}| > 0$ **or budget not reached do**
3: $x^* = argmax_{x \in \mathcal{U}}(\mu(x))$
4: request label and rationale
5: $\mathcal{L} \leftarrow \mathcal{L} \cup \{(x^*, y_{x^*}, r_{x^*})\}$
6: $\mathcal{U} \leftarrow \mathcal{U} \backslash \{(x^*)\}$
7: **end while**

Previously rationale-based active learning research [21, 28] showed adding rationale's can improve the classification performance. However most of them use word frequency representation (bag-of-words). In contrast, our approach transforms both textual samples

and human annotations into low-dimensional dense representation which not only allow us to exploit the correlation between documents and human annotated sentences in one semantic space, but also provide us the flexibility to leverage more advanced models such as neural networks and attention.

3.3 Rationale-based Active Learning via Linear Model

The underlying classification model θ discussed in the Algorithm 1 can be as simple as a linear classifier of the form: $\theta(X) = W'X - \gamma$, obtained by solving: $\min L(\theta(X), Y) + \nu Reg(W)$, where x_i is the i^{th} row of a matrix $X \in R^{m \times n}$ containing m number of samples, n is the feature dimension; $x_i \in R^n$ is a vectorial representation or embedding for the ith sample; ν is a regularization parameter; Y is a vector of labels/classes and is +1 and −1 respectively; $Reg(W)$ is a regularization term on W.

The vectorial representation or embedding at the sentence, text or document for a given sample d_i is given by :

$$x_i = \varphi_\vartheta(d_i) \tag{1}$$

where φ_ϑ can be any model which can map the input to a fixed length vectorial representation. Throughout this work, we use pre-trained transformer based Universal Sentence Encoder described in [4] to map input to vector representation.

Since we have *rationale* r_{ij} for every rational sentence j in document i, we enrich the representation of input as follows:

$$x_i^f = \lambda * x_i + (1 - \lambda) * \sum_j r_{ij} \tag{2}$$

where λ is a parameter in the range $[0, 1]$; j denotes the number of rationale; r_{ij} and x_i can be obtained from Equation 1 and the resulting x_i^f are the enriched features used only for training. Note that since θ is linear we have that:

$$\theta(x_i^f) = \lambda * \theta(x_i) + (1 - \lambda) * \sum_j \theta(r_{ij}) \tag{3}$$

This convex combination aims to blend the entire instance representation with the rationale representation encoded on the parameters of the model. The "optimal" value of λ depends on the specific problem and it's a parameter to be learned by tuning during training.

It is important to note that since there can be multiple rationale's for a single document, we average over their representations before the enrichment. Also, the document representation x_i obtained from Equation 1 is an average of sentence vectors. During testing, since we don't have to access to rationale we only use x_i.

3.4 Rationale-based Active Learning with Supervised Attention

The underlying classification model θ discussed in Algorithm 1 can also be a neural network based model, and is given as: $\theta_n = f(X, Y, R)$, where, $f(.)$ refers to *Rationale-based Supervised Attention* model and we will discuss this model more in the following sections; x_i is a document with a sequence of sentences/words and $x_i \in [1, s]$ where s is the maximum number of sentences/words, now $X \in R^{m \times s \times n}$ is a three dimensional tensor with m documents, s sentences/words and n is the embedding or vector representation of each sentence/word; Y is a vector of labels/classes and is +1 and −1; $R \in R^{m \times s}$ is *rationale* for the corpus of m documents where r_i is the rationale for document x_i and $r_i \in [1, s]$ is a vector of 0's and 1's where 1 indicates that the sentence/word is a *rationale* and 0 otherwise.

Attention mechanism is designed in a way to enable the neural network to magnify attention weights or "pay more attention" to certain phrases or words than others.

3.4.1 Attention neural networks. We first design a simple attention based neural network and later extend it to support for supervised attention in next section. Inspired by the attention mechanism proposed in [25], we design a simple attention based neural network $f(.)$ as follows:

Sigmoid Attention Layer: Assume that $X \in R^{m \times s \times n}$ is an input matrix where each sample is a sequence of s sentences/words represented in a n dimensional vector. Let's define:

$$u = tanh(WX + b),\ \alpha = \frac{1}{1 + exp^{-u^T u_s}},\ v = \sum_t \alpha X \tag{4}$$

where u_s is a *context vector* and is learned during training. and α are the attention weights, v is the document vector. In contrast to most of the attention mechanisms, we use a sigmoid function instead of softmax. Because multiple (part of) sentences can be valid rationale, and we do expect human annotators to highlight multiple rationale's of which they think are important. Therefore, we chose sigmoid cause it allows for multiple selection in its structure in contrast to softmax which only allows for a single selection.

Classification: The document vector v is now a high level representation of the document as it summarizes all of the input into one fixed length representation, we further extend this representation by adding a fully connected layer as follows:

$$h = \text{ReLU}\left(W_f v + b_f\right), o = \text{Softmax}\left(W_s h + b_s\right), \mathcal{L}_t = -\sum_i p_i \log o_i \tag{5}$$

h are used as features for classification in the Softmax layer and $\mathcal{L}_t$ is the cross entropy loss. , p_i is the class label for i^{th} document and o_i is the corresponding network prediction.

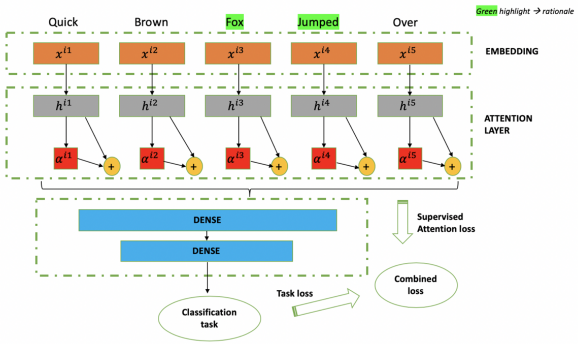

Figure 2: The architecture of Rationale-based Active Learning with Supervised Attention (RALSA).

3.4.2 Supervised Attention. We now incorporate human annotators' rationale into our network and specifically, we want the network to pay more attention towards the *rationale R*. Hence, we enforce this via a loss function as follows:

$$\mathcal{L}_a = -\sum_i \sum_j (r_{ij} \log \alpha_{ij} + (1 - r_{ij}) \log(1 - \alpha_{ij})) \tag{6}$$

where α_{ij} is the attention weight for the i^{th} document at j^{th} sentence and similarly r_{ij} is the rationale information (1 if sentence is a rationale and 0 otherwise) for the i^{th} document at j^{th} sentence.

The total loss can now be formulated as a convex combination of $\mathcal{L}_t$ and $\mathcal{L}_a$ and is as follows:

$$\mathcal{L} = \lambda \mathcal{L}_t + (1 - \lambda)\mathcal{L}_a \tag{7}$$

where λ is a value in the range $[0, 1]$

The architectural view of the proposed RALSA is shown in Figure 2. We also attempted to adapt uncertainty sampling by calculating entropy of attention weights as part of uncertainty estimation. However this newer version of uncertainty sampling did not yield better results than the generic version.

4 EXPERIMENTS AND RESULTS

We use binary classification to evaluate the performance of our algorithms by comparing the following methods: Linear model (LM), Rationale-based Active Learning with Linear Model (RALM), Attention Neural Network (AN) and Rationale-based Active Learning with Supervised Attention (RALSA). We incrementally build complexity into our active learning experiments to evaluate the performance improvement brought in by adding rationale's and the supervised attention.[1] Note that some of the baseline approaches don't use rationale's as supervision. Our intention with this work is both to confirm that supervised attention can be used to accelerate learning and propose an algorithm to do so in an active learning setting.

4.1 Datasets

To simulate the active learning process, we curated four datasets with different sample sizes, lengths and topics of content for binary text classification tasks. (see Table 1). Since we require all documents to have ground truth labels as well as human annotated rationale's, we found IMDB small as the only available open source data set [28] that meets these requirements. We have used Amazon Mechanical Turk to curate another dataset with rationale's. We also developed an internal active learning labeling and annotation tool, and recruited a human annotator for creating the rest of the datasets.

IMDB small: This is an IMDB movie reviews dataset consisting of 1000 positive reviews and 1000 negative reviews. These reviews were labeled and annotated by human annotators and is best described in this work

IMDB large: Inspired by the approach taken by [28], we have sampled around 22,000 IMDB movie reviews to label (including rationale's) via Amazon Mechanical Turk.

[1]All the code and datasets are available here: https://github.com/tkanchin-amfam/ALSupervisedAttention

TREC QA: This is a dataset for question classification consisting of open-domain, fact-based questions divided into broad semantic categories. We have selected the TREC-6 dataset and sampled questions related to the categories *entity* and *numerical*. We have sampled a subset of around 500 samples for labeling and annotation/rationale using an internally developed labeling tool.

Insurance Claims: This data set consists of 18,000 insurance claims that needs to be categorized based on the cause of loss (reason of the claim) events, e.g. vehicle collisions, vehicle theft, vehicle malfunctions, property damages etc. We used rationale-based active learning to train a model that can predict the loss cause given a short description of the claim. In order to obtain labels and rationale's, a team of product analysts labeled and highlighted the sentences and phrases which indicate the (non)existence of certain loss cause verbiage.

4.2 Experimental settings

4.2.1 Rational-based Active Learning with Linear Model. : **Linear Model**: As defined in Section 3.3, we have used a least squares support vector machines (LSSVM) as the underlying linear classification model. For all the datasets, we have used transformer based universal sentence encoder [4] to map input to a fixed length vector representation. The value of λ for the RALM experiment and the hyper-parameters for LSSVM model are chosen via a grid search using a validation set.

Active Learning Process: We initialize the linear model with 2 randomly selected labeled examples to start the active learning simulation. In each iteration, we include one example from the unlabeled set to update the linear model until the unlabeled set is completely exhausted. We also calculate the receiver-operating-characteristic (ROC), or an AUC score of the model against a hold out test set. Metrics are reported by averaging the results from 10 runs and for the simplicity of display, we have not reported standard deviations.

4.2.2 Rational-based Active Learning with Supervised Attention. **Words-level and Sentence-level tokenization**: To achieve a uniform tensor input format, we tokenized the dataset with short documents into words and mapped them to a fixed length vector representation using universal sentence encoder [4]. For the datasets with large documents, we have tokenized the documents into sentences and mapped each sentence to a fixed vector representation. This transformation resulted in an input matrix $X \in R^{m \times s \times n}$ where s refers to the number of words/sentences in a sample and n is the length of the feature vector.

The datasets *TREC QA* and *INSURANCE* consists of short documents and we have selected the value of s as 50 and 75 respectively for these datasets. The datasets *IMDB Large* and *IMDB Small* are long document and we have selected the value of s as 80, 150 and 200 respectively. Documents are padded or truncated respectively based on the value of s.

Neural Network Model: In all of the experiments we have used a batch size of 64 and number of epochs as 100 for the neural network, Attention dimension (W) as 128 and learning rate is set to $\alpha = 10^{-3}$. We used the Adam optimizer [11] with β_1 set to 0.9 and β_2 set to 0.999 and initialized the weights of the network with Xavier initialization [9]. We have used model check-pointing to checkpoint

Table 1: The details of the datasets used in the experiments. The column "Annotation" refers to both labeling and highlighting words or phrases (*rationale* by a human annotator). Two of the IMDB datasets have all the samples annotated. For the remaining datasets we have only selected a smaller subset of samples to annotate and to test our theory.

Dataset	No. of Samples (Annotated)	No. of Words (avg)	No. of Sentences (avg)	No. of Annotation (avg)	Annotation tool
IMDB small	2000 (2000)	224	135	39	[28]
IMDB large	22000 (22000)	288	85	32	Mechanical Turk
Insurance Claims	18000 (800)	52	–	8	Internal tool
TREC QA	5000 (500)	26.74	–	4	Internal tool

the best model using a validation set. All the experiments were conducted on a K80 GPU Amazon Web services instance.

Active Learning Process: We have started our active learning simulation with 5% randomly selected labeled examples due to the input requirements of a neural network. We set the budget size to 2% i.e. at every iteration 2% of examples from unlabeled set are added to the labeled set and the model is retrained. At each iteration, we have calculated the receiver-operating-characteristic (ROC), or a AUC score of the model against a hold out test set. Metrics are reported by averaging the results from 10 runs and for the simplicity of display, we have not reported standard deviations.

4.3 Results

4.3.1 Rational-based Active Learning with Linear Model. In Figure 3, we evaluate whether RALM improves upon LM in both random and uncertainty sampling. We have observed that in both IMDB datasets, RALM with uncertainty sampling consistently outperforms other methods. In the case of Insurance and TREC QA, RALM with uncertainty sampling is comparable to its LM counterparts; we infer that the reason RALM does not improve upon LM in these datasets is due to the fact that the ratio between the total number of words to rationale words is very high and adding in rationale information via a convex combination in this case doesn't help much.

4.3.2 Rational-based Active Learning with Supervised Attention. In 4.3.1, we have observed that the use of rationale can help with the performance of active learning process. Therefore, we have extended our ideas by adding rationale's to a feed forward neural network for active learning. Specifically we compare a neural network with an attention layer (AN) and the same neural network with supervised attention (RALSA) from human rationale. Figure 4, clearly shows RALSA outperforms AN across all four datasets by a significant margin. This strongly suggests that the active learning process converges faster if the neural network is supplied with rationale via supervised attention. In addition, Figure 4 shows RALSA outperforms AN methods almost always from the beginning, suggesting supervised attention can bootstrap the active learning process with only few samples.

4.3.3 Rational-based Active Learning with Linear Model vs Supervised Attention. We also compare the performance of our rationale based attention models (RALSA) with all other methods, and the results are shown in Table 2. We only compare them with the uncertainty sampling since it's one of the most popular AL query strategies and in most cases it outperforms random sampling. Note that the literature on query strategies for AL is extensive and comparing to other more complex query strategies is out of scope for this work. In Table 2, we report AUC at 5%, 10%, 20%, 50% and 90% of the labeling budget. For both IMDB datasets, we can see that RALSA outperforms all other methods at all percentages of the labeling budget. We also observe that for the insurance dataset, RALSA performs very well specially at lower percentages of labeled data. We have empirically observed, that the non-linearity introduced by neural nets seems to help boost early performance, therefore RALSA is consistently better than the other two linear models.

In the case of TREC QA, linear model performs the best for active learning. This can be due to the fact TREC QA data set is smaller and the classification task is relatively simple. Hence, linear models may be sufficient for this task.

5 DISCUSSION

5.0.1 Effectiveness of embedded rationale features. Figure 5 shows human annotated rationale features adds separability into the embedded feature space for the *IMDB Large* dataset. We used t-SNE [15] to map embedding space into a 2D space. We can observe that **document features** shows that the embedded reviews are hard to separate visually however, the rationale's have good visual separation in this space; Hence, **document + rationale feature** shows that after a linear combination of document and rationale features, the samples in the dataset are visually separable again.

5.0.2 Effectiveness of the attention model. Next, we present examples that illustrate what RALSA is learning from the supervised attention architecture. For example, in a positive movie review unseen by RALSA, the three sentences with top attention weights are "***the cast is excellent throughout**, and , rather than singling out anybody , kudos are due to **the fine ensemble acting**,...*", *"...the beauties of the armpit area ,or Ronald Reagan, **the script never seems to run out of hilarious invention.**"* and *"...but really the point is that the foot-massage master hasn't got a monopoly on **plot twists and fast, funny, irreverent lines**..."* (attention in **bold**). Several of these model-generated attention sentences overlap with the human annotation, but there are also some phrases which are not found in original human-provided rationale's, for example: *"...is the essential family state - and you* **better learn to love it.***"*. These observations empirically imply that RALSA is not only able to learn from rationale's syntax but also rationale's semantics.

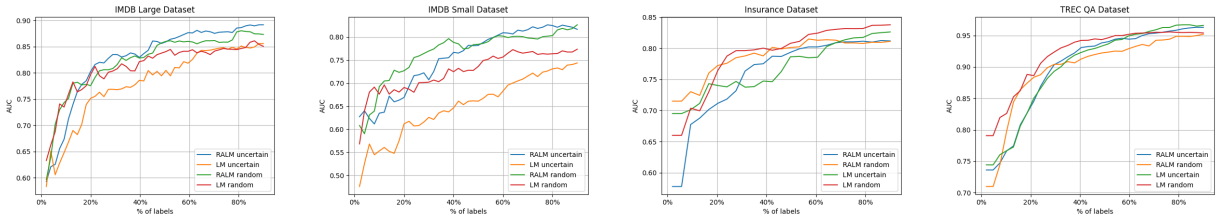

Figure 3: Rationale-based Active Learning with Linear model. RALM refers to Rationale based active learning with Linear model, LM refers to Linear model, uncertain refers to uncertainty sampling and random refers to random sampling. The methods RALM and LM are described in Section 3.3.

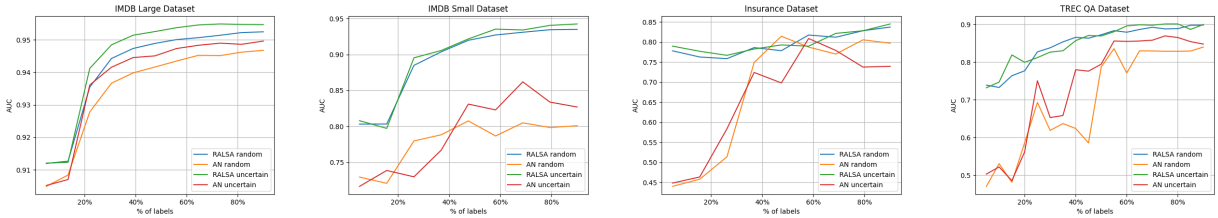

Figure 4: Rationale-based Active Learning with Supervised Attention experiments. RALSA and AN methods are described in Section 3.4. RALSA refers to Rationale based active learning with Supervised Attention, AN refers to Attention neural networks, uncertain refers to uncertainty sampling and random refers to random sampling. The methods RALSA and AN are described in Section 3.4.

Table 2: The performance comparison among RALSA, AN, RALM and LM experiments using uncertainity sampling. In the Table, D refers to dataset and M refers to the method. We have simulated the active learning process on each dataset and method and reported the AUC at 5%, 10%, 20%, 50% and 90% of training labels.

D/M	IMDB Large	IMDB Small	INSURANCE	TREC QA
	<5%, 10%, 20%, 50%, 90%>	<5%, 10%, 20%, 50%, 90%>	<5%, 10%, 20%, 50%, 90%>	<5%, 10%, 20%, 50%, 90%>
RALSA	<**91.2**, **91.6**, **94.1**, **95.2**, **95.4**>	<**80.8**, **81.8**, **83.7**, **92.1**, **94.2**>	<**78.7**, **78.9**, 77.6, 79.2, **84.4**>	<**74.7**, 75.4, 79.9, 86.9, 89.9>
AN	<90.5, 90.7, 93.6, 94.5, 94.9>	<71.6, 71.6, 73.8, 83.1, 82.5>	<44.8, 44.8, 46.3, 69.7, 73.9>	<50.2, 52.1, 56.1, 79.4, 84.7>
RALM	<59.2, 80.2, 83.5, 89.7, 91.1>	<62.7, 66.8, 75.5, 81.1, 84.5>	<57.7, 71.4, 74.7, 80.8, 82.3>	<73.6, **76.6**, **82.6**, **93.2**, 95.7>
LM	<58.3, 75.1, 77.7, 85.7, 88.8>	<47.5, 61.1, 63.7, 74.2, 80.2>	<69.5, 73.9, **77.9**, **81.7**, 83.5>	<74.4, **76.6**, **82.6**, 92.7, **96.2**>

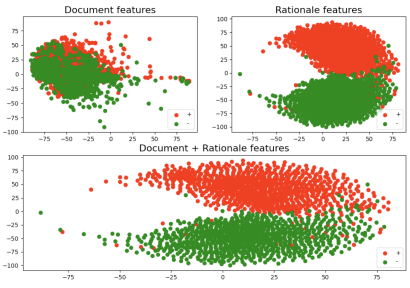

Figure 5: TSNE plot visualizing various feature representations. These visualizations are from IMDB Small dataset (movie reviews) where + (green) refers to positive movie review and - (red) refers to negative movie review.

6 CONCLUSION AND FUTURE WORK

In this work, we have proposed a novel algorithms, rationale-based active learning with supervised attention (RALSA) which encode human annotation as a light-weighted supervised attention mechanism to the underlying neural network. Our approach is consistent with the requirements of a active learning setting where the model has to be retrained often and fast. We also rigorously tested them across multiple datasets with different domains. RALSA achieves best results in 3 out of the 4 datasets, and the linear version RALM is very competitive in the remaining one. We have created 3 new rationale-based labeled datasets that will be shared with the human-in-the-loop community (there are very few in existence prior to this work).

Those results point to future explorations of attention-based rationale approaches to other linguistic tasks , designing attention mechanism with more complex language model architectures and to explore other AL query strategies.

REFERENCES

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[2] Y. Bao, S. Chang, M. Yu, and R. Barzilay. Deriving machine attention from human rationales. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1903–1913. Association for Computational Linguistics, 2018.

[3] M. Bloodgood. Support vector machine active learning algorithms with query-by-committee versus closest-to-hyperplane selection. *CoRR*, abs/1801.07875, 2018.

[4] D. Cer, Y. Yang, S. yi Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil. Universal sentence encoder, 2018.

[5] A. Culotta and A. McCallum. Reducing labeling effort for structured prediction tasks. In *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 746–751, 2005.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[7] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[8] Y. Gal, R. Islam, and Z. Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1183–1192, 2017.

[9] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[10] Y. Kim, C. Denton, L. Hoang, and A. M. Rush. Structured attention networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] E. Lehman, J. DeYoung, R. Barzilay, and B. C. Wallace. Inferring which medical treatments work from reports of clinical trials. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3705–3717. Association for Computational Linguistics, 2019.

[13] T. Lei, R. Barzilay, and T. S. Jaakkola. Rationalizing neural predictions. In J. Su, X. Carreras, and K. Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 107–117. The Association for Computational Linguistics, 2016.

[14] T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421, 2015.

[15] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[16] D. J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

[17] A. McCallum and K. Nigam. Employing EM and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998*, pages 350–358, 1998.

[18] R. Munro. *Human-in-the-Loop Machine Learning*. 2020.

[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[20] M. E. Ramirez-Loaiza, M. Sharma, G. Kumar, and M. Bilgic. Active learning: an empirical study of common baselines. *Data Min. Knowl. Discov.*, 31(2):287–313, 2017.

[21] M. Sharma, D. Zhuang, and M. Bilgic. Active learning with rationales for text classification. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 441–451, 2015.

[22] A. Siddhant and Z. C. Lipton. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2904–2909, 2018.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[24] J. Wu, A. Guo, V. S. Sheng, P. Zhao, Z. Cui, and H. Li. Adaptive low-rank multi-label active learning for image classification. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1336–1344, 2017.

[25] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[26] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy. Hierarchical attention networks for document classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1480–1489, 2016.

[27] D. Yoo and I. S. Kweon. Learning loss for active learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 93–102, 2019.

[28] O. Zaidan, J. Eisner, and C. Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 260–267, 2007.

[29] Y. Zhang, I. Marshall, and B. C. Wallace. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 795–804, Austin, Texas, Nov. 2016. Association for Computational Linguistics.

[30] J. Zhu, H. Wang, T. Yao, and B. K. Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*, pages 1137–1144, 2008.