The Explanation Game:
Towards Prediction Explainability through Sparse Communication

Marcos V. Treviso
Instituto de Telecomunicações
Instituto Superior Técnico
University of Lisbon, Portugal
marcos.treviso@tecnico.ulisboa.pt

André F. T. Martins
Instituto de Telecomunicações
LUMLIS (Lisbon ELLIS Unit)
Instituto Superior Técnico & Unbabel
Lisbon, Portugal
andre.t.martins@tecnico.ulisboa.com

Abstract

Explainability is a topic of growing importance in NLP. In this work, we provide a unified perspective of explainability as a communication problem between an explainer and a layperson about a classifier's decision. We use this framework to compare several explainers, including gradient methods, erasure, and attention mechanisms, in terms of their communication success. In addition, we reinterpret these methods in the light of classical feature selection, and use this as inspiration for new embedded explainers, through the use of selective, sparse attention. Experiments in text classification and natural language inference, using different configurations of explainers and laypeople (including both machines and humans), reveal an advantage of attention-based explainers over gradient and erasure methods, and show that selective attention is a simpler alternative to stochastic rationalizers. Human experiments show strong results on text classification with post-hoc explainers trained to optimize communication success.

1 Introduction

The widespread use of machine learning to assist humans in decision making brings the need for explaining models' predictions (Doshi-Velez, 2017; Lipton, 2018; Rudin, 2019; Miller, 2019). This poses a challenge in NLP, where current state-of-the-art neural systems are generally opaque (Goldberg and Hirst, 2017; Peters et al., 2018; Devlin et al., 2019). Despite the large body of recent work (reviewed in §7), a unified perspective modeling the human-machine interaction—a *communication* process in its essence—is still missing.

Many methods have been proposed to generate explanations. Some neural network architectures are equipped with built-in components—attention mechanisms—which weigh the relevance of input features for triggering a decision (Bahdanau et al., 2015; Vaswani et al., 2017). Top-k attention weights provide plausible, but not always faithful, explanations (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019). Rationalizers with hard attention are arguably more faithful, but require stochastic networks, which are harder to train (Lei et al., 2016; Bastings et al., 2019). Other approaches include gradient methods (Li et al., 2016a; Arras et al., 2017), querying the classifier with leave-one-out strategies (Li et al., 2016a; Feng et al., 2018), or training local sparse classifiers (Ribeiro et al., 2016).

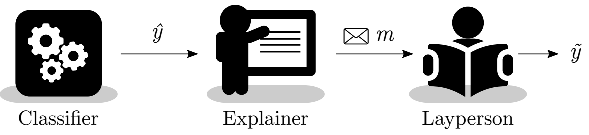

Figure 1: Our framework to model explainability as communication. Predictions $\hat{y}$ are made by a classifier C; an explainer E (either embedded in C or operating post-hoc) accesses these predictions and communicates an explanation (a message m) to the layperson L. Success of the communication is dictated by the ability of L and C to match their predictions: $\tilde{y} \overset{?}{=} \hat{y}$. Both the explainer and layperson can be humans or machines.

How should these different approaches be compared? Several diagnostic tests have been proposed: Jain and Wallace (2019) assessed the explanatory power of attention weights by measuring their correlation with input gradients; Wiegreffe and Pinter (2019) and DeYoung et al. (2020) developed more informative tests, including a combination of comprehensiveness and sufficiency metrics and the correlation with human rationales; Jacovi and Goldberg (2020) proposed a set of evaluation recommendations and a graded notion of faithfulness. Most proposed frameworks rely on correlations and counterfactual simulation, sidestepping the main practical goal of prediction explainability—the ability to *communicate* an explanation to a human user.

Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 107–118
Online, November 20, 2020. ©2020 Association for Computational Linguistics

In this work, we fill the gap above by proposing a unified framework that regards explainability as a **communication problem**. Our framework is inspired by human-grounded evaluation through **forward simulation/prediction**, as proposed by Doshi-Velez (2017, §3.2), where humans are presented with an explanation and an input, and must correctly simulate the model's output (regardless of the true output). We model this process as shown in Figure 1, by considering the interaction between a *classifier* (the model whose predictions we want to explain), an *explainer* (which provides the explanations), and a *layperson* (which must recover the classifier's prediction). We show that different configurations of these components correspond to previously proposed explanation methods, and we experiment with explainers and laypeople being both humans and machines. Our framework also inspires two new methods: embedded explainers based on **selective attention** (Martins and Astudillo, 2016; Peters et al., 2019), and **trainable explainers** based on emergent communication (Foerster et al., 2016; Lazaridou et al., 2016).

Overall, our contributions are:

- We draw a link between recent techniques for explainability of neural networks and classic feature selection in linear models (§2). This leads to new embedded methods for explainability through selective, sparse attention (§3).
- We propose a new framework to assess explanatory power as the communication success rate between an explainer and a layperson (§4).
- We experiment with text classification, natural language inference, and machine translation, using different configurations of explainers and laypeople, both machines (§5) and humans (§6).

2 Revisiting Feature Selection

A common way of generating explanations is by highlighting *rationales* (Zaidan and Eisner, 2008). The principle of parsimony ("Occam's razor") advocates simple explanations over complex ones. This principle inspired a large body of work in traditional feature selection for linear models. We draw here a link between that work and modern approaches to explainability.

Table 1 highlights the connections. Traditional feature selection methods (Guyon and Elisseeff, 2003) are mostly concerned with **model interpretability**, i.e., understanding how models behave globally. Feature selection happens *statically* during model training, after which irrelevant features are permanently deleted from the model. This contrasts with **prediction explainability** in neural networks, where feature selection happens *dynamically* at run time: here explanations are input-dependent, hence a feature not relevant for a particular input can be relevant for another. Are these two worlds far away? Guyon and Elisseeff (2003, §4) proposed a typology for traditional feature selection with three classes of methods, distinguished by how they model the interaction between their main two components, the *feature selector* and the *learning algorithm*. We argue that this typology can also be used to characterize various explanation methods, if we replace these two components by the *explainer* E and the *classifier* C, respectively.

- **Wrapper methods**, in the wording of Guyon and Elisseeff (2003), "utilize the learning machine of interest as a black box to score subsets of variables according to their predictive power." This means greedily searching over subsets of features, training a model with each candidate subset. In the dynamic feature selection world, this is somewhat reminiscent of the leave-one-out method of Li et al. (2016b), the ablative approach of Serrano and Smith (2019), and LIME (Ribeiro et al., 2016), which repeatedly queries the classifier to label new examples.
- **Filter methods** decide to include/exclude a feature based on an importance metric (such as feature counts or pairwise mutual information). This can be done as a preprocessing step or by training the model once and thresholding the feature weights. In dynamic feature selection, this is done when we examine the gradient of the prediction with respect to each input feature, and then select the features whose gradients have large magnitude (Li et al., 2016a; Arras et al., 2016; Jain and Wallace, 2019),[1] and when thresholding softmax attention scores to select relevant input features, as analyzed by Jain and Wallace (2019) and Wiegreffe and Pinter (2019).
- **Embedded methods**, in traditional feature selection, embed feature selection within the learning algorithm by using a sparse regularizer such as the ℓ_1-norm (Tibshirani, 1996). Features that receive zero weight become irrelevant and can

[1] In linear models this gradient equals the feature's weight.

	Static selection (model interpretability)	**Dynamic selection** (prediction explainability)
Wrappers	Forward selection, backward elimination (Kohavi and John, 1997)	Input reduction (Feng et al., 2018), representation erasure (leave-one-out) (Li et al., 2016b; Serrano and Smith, 2019), LIME (Ribeiro et al., 2016)
Filters	Pointwise mutual information (Church and Hanks, 1989), recursive feature elimination (Guyon et al., 2002)	Input gradient (Li et al., 2016a), layerwise relevance propagation (Bach et al., 2015), top-k softmax attention
Embedded	ℓ_1-regularization (Tibshirani, 1996), elastic net (Zou and Hastie, 2005)	Stochastic attention (Xu et al., 2015; Lei et al., 2016; Bastings et al., 2019), sparse attention (**this paper**, §3)

Table 1: Overview of static and dynamic feature selection techniques.

be removed from the model. In dynamic feature selection, this encompasses methods where the classifier produces rationales together with its decisions (Lei et al., 2016; Bastings et al., 2019). We propose in §3 an alternative approach via **sparse attention** (Martins and Astudillo, 2016; Peters et al., 2019), where the selection of words for the rationale resembles ℓ_1-regularization.

In §4, we frame each of the cases above as a communication process, where the explainer E aims to communicate a short message with the relevant features that triggered the classifier C's decisions to a layperson L. The three cases above are distinguished by the way C and E interact.

3 Embedded Sparse Attention

The case where the explainer E is embedded in the classifier C naturally favors faithfulness, since the mechanism that explains the decision (the *why*) can also influence it (the *how*).

Attention mechanisms (Bahdanau et al., 2015) allow visualizing relevant input features that contributed to the model's decision. However, the traditional softmax-based attention is *dense*, i.e., it gives *some* probability mass to every feature, even if small. The typical approach is to select the top-k words with largest attention weights as the explanation. However, this is not a truly embedded method, but rather a filter, and as pointed out by Jain and Wallace (2019) and Wiegreffe and Pinter (2019), it may not lead to faithful explanations.

An alternative is to embed in the classifier an attention mechanism that is inherently **selective**, i.e., which can produce sparse attention distributions natively, where some input features receive exactly zero attention. An extreme example is hard attention, which, as argued by DeYoung et al. (2020), provides more faithful explanations "by construction" as they discretely extract snippets from the input to pass to the classifier. A problem with hard attention is its non-differentiability, which complicates training (Lei et al., 2016; Bastings et al., 2019). We consider in this paper a different approach: using end-to-end differentiable sparse attention mechanisms, via the **sparsemax** (Martins and Astudillo, 2016) and the recently proposed **1.5-entmax** transformation (Peters et al., 2019), described in detail in §A. These sparse attention transformations have been applied successfully to machine translation and morphological inflection (Peters et al., 2019; Correia et al., 2019). Words that receive non-zero attention probability are *selected* to be part of the explanation. This is an embedded method akin of the use of ℓ_1-regularization in static feature selection. We experiment with these sparse attention mechanisms in §5.

4 Explainability as Communication

We now have the necessary ingredients to describe our unified framework for comparing and designing explanation strategies, illustrated in Figure 1.

Our fundamental assumption is that explainability is intimately linked to the ability of an explainer to **communicate** the rationale of a decision in terms that can be understood by a human; we use the success of this communication as a criterion for how plausible the explanation is.

4.1 The Classifier-Explainer-Layperson setup

Our framework draws inspiration from Lewis' signaling games (Lewis, 1969) and the recent work on emergent communication (Foerster et al., 2016; Lazaridou et al., 2016; Havrylov and Titov, 2017). Our starting point is the classifier $C : \mathcal{X} \to \mathcal{Y}$ which, when given an input $x \in \mathcal{X}$, produces a prediction $\hat{y} \in \mathcal{Y}$. This is the prediction that we want to explain. An explanation is a **message** $m \in \mathcal{M}$, for a predefined message space $\mathcal{M}$ (for example, a rationale). The goal of the explainer E is to compose and **successfully communicate** messages m to a layperson L. The success of the communication is dictated by the ability of L to reconstruct $\hat{y}$ from m with high accuracy. In this paper, we experiment with E and L being either humans or machines. Our framework is inspired by human-grounded evaluation through forward simulation/prediction, as proposed by Doshi-Velez (2017, §3.2). More formally:

- The **classifier** C is the model whose predictions we want to explain. For given inputs x, C produces $\hat{y}$ that are hopefully close to the ground truth y. We are agnostic about the kind of model used as a classifier, but we assume that it computes certain internal representations h that can be exposed to the explainer.

- The **explainer** E produces explanations for C's decisions. It receives the input x, the classifier prediction $\hat{y} = C(x)$, and optionally the internal representations h exposed by C. It outputs a message $m \in \mathcal{M}$ regarded as a "rationale" for $\hat{y}$. The message $m = E(x, \hat{y}, h)$ should be simple and compact enough to be easily transmitted and understood by the layperson L. In this paper, we constrain messages to be bags-of-words (BoWs) extracted from the textual input x.

- The **layperson** L is a simple model (e.g., a linear classifier)[2] that receives the message m as input, and predicts a final output $\tilde{y} = L(m)$. The communication is successful if $\tilde{y} = \hat{y}$. Given a test set $\{x_1, \ldots, x_N\}$, we evaluate the **communication success rate** (CSR) as the fraction of examples for which the communication is successful:

$$\mathrm{CSR} = \frac{1}{N}\sum_{n=1}^{N} [[C(x_n) = L(E(x_n, C(x_n)))]], \tag{1}$$

where $[[\cdot]]$ is the Iverson bracket notation.

Under this framework, we regard the communication success rate as a quantifiable measure of explainability: a high CSR means that the layperson L is able to replicate the classifier C's decisions a large fraction of the time when presented with the messages given by the explainer E; this assesses how informative E's messages are.

Our framework is flexible, allowing different configurations for C, E, and L, as next described. In §5, we show examples of explainers and laypeople for text classification and natural language inference tasks (additional experiments on machine translation are described in §G).

Relation to filters and wrappers. In the wrapper and filter approaches described in §2, the classifier C and the explainer E are separate components. In these approaches, E works as a *post-hoc explainer*, querying C with new examples or requesting gradient information.

Relation to embedded explanation. By contrast, in the embedded approaches of Lei et al. (2016) and the selective sparse attention introduced in §3, the explainer E is directly *embedded* as an internal component of the classifier C, returning the selected features as the message. This approach is arguably more faithful, as E is directly linked to the mechanism that produces C's decisions.

4.2 Joint training of explainer and layperson

So far we have assumed that E is given beforehand, chosen among existing explanation methods, and that L is trained to assess the explanatory ability of E. But can our framework be used to *create* new explainers by training E and L jointly? We will see how this can be done by letting E and L play a cooperative game (Lewis, 1969). The key idea is that they need to learn a communication protocol that ensures high CSR (Eq. 1). Special care needs to be taken to rule out "trivial" protocols and ensure plausible, potentially faithful, explanations. We propose a strategy to ensure this, which will be validated using human evaluation in §6.[3]

Let E_θ and layperson L_ϕ be **trained models** (with parameters θ and ϕ), learned together to optimize a multi-task objective with two terms:

- A **reconstruction term** that controls the information about the classifier's decision $\hat{y}$. We use a cross-entropy loss on the output of the layperson L, using $\hat{y}$ (and not the true label y) as the ground truth: $\mathcal{L}(\phi, \theta) = -\log p_\phi(\hat{y} \mid m)$, where m is the output of the explainer E_θ.

- A **faithfulness term** that encourages the explainer E to take into account the classifier's

[2] The reason why we assume the layperson is a simple model is to encourage the explainer to produce simple and explanatory messages, in the sense that a simple model can learn with them. A more powerful layperson could potentially do well even with bad explanations.

[3] Other approaches, such as Lei et al. (2016) and Yu et al. (2019), develop rationalizers from cooperative or adversarial games between generators and encoders. However, those frameworks do not aim at explaining an external classifier.

decision process when producing its explanation m. This is done by adding a squared loss term $\Omega(\theta) = \|\tilde{h}(E_\theta), h\|^2$ where $\tilde{h}$ is E's prediction of C's internal representation h.

The objective function is a combination of these two terms, $\mathcal{L}_\Omega(\phi, \theta) := \lambda\Omega(\theta) + \mathcal{L}(\phi, \theta)$. We used $\lambda = 1$ in our experiments. This objective is minimized in a training set that contains pairs $(x, \hat{y})$. Therefore, in this model the message m is latent and works as a "bottleneck" for the layperson L, which does not have access to the full input x, to guess the classifier's prediction $\hat{y}$—related models have been devised in the context of emergent communication (Lazaridou et al., 2016; Foerster et al., 2016; Havrylov and Titov, 2017) and sparse autoencoders (Trifonov et al., 2018; Subramanian et al., 2018).

We minimize the objective above with gradient backpropagation. To ensure end-to-end differentiability, during this joint training we use sparsemax attention (§3) to select the relevant words in the message. One important concern in this model is to prevent E and L from learning a trivial protocol to maximize CSR. To ensure this, we forbid E from including stopwords in its messages and during training we use a linear schedule for the probability of the explainer accessing the predictions of the classifier ($\hat{y}$), which are hidden otherwise. At the end of training, the explainer will access it with probability β. In our experiments, we set β to 20% (chosen on the validation set as described in §F.2).

5 Experiments

We experimented with our framework on two NLP tasks: text classification and natural language inference. Additional experiments on machine translation are reported in §G, with similar conclusions.

We used 4 datasets (SST, IMDB, AgNews, Yelp) for text classification and one dataset (SNLI) for NLI, with statistics and details in Table 5 (§B).

Classifier C. For text classification, the input $x \in \mathcal{X}$ is a document and the output set $\mathcal{Y}$ is a set of labels (e.g. topics or sentiment labels). The message is a bag of words (BoW) extracted from the document. As in Jain and Wallace (2019) and Wiegreffe and Pinter (2019), our classifier C is an RNN with attention. For NLI, the input x is a pair of sentences (premise and hypothesis) and the labels in $\mathcal{Y}$ are entailment, contradiction, and neutral. We let messages be again BoWs, and we constrain them to be selected from the premise (and concatenated with the full hypothesis). We used a similar classifier as above, but with two independent BiLSTM layers, one for each sentence. We used the additive attention of Bahdanau et al. (2015) with the last hidden state of the hypothesis as the query and the premise vectors as keys.

CLASSIFIER	SST	IMDB	AGN.	YELP	SNLI
BoW (L)	82.54	88.96	95.62	68.78	69.81
RNN, softmax (C)	86.16	**91.79**	96.28	**75.80**	78.34
–,1.5-entmax (C_{ent})	86.11	91.72	96.30	75.72	79.20
–, sparsemax (C_{sp})	**86.27**	91.52	96.37	75.72	78.78
Bernoulli (C_{bern})	81.99	86.99	95.68	70.12	79.24
HardKuma (C_{hk})	84.13	91.06	**96.38**	74.36	**85.49**

Table 2: Accuracies of the original classifiers on text classification and natural language inference.

We also experimented with RNN classifiers that replace softmax attention by 1.5-entmax (C_{ent}) and sparsemax (C_{sp}), and with the rationalizer models of Lei et al. (2016) (C_{bern}) and Bastings et al. (2019) (C_{hk}). Details about these classifiers and their hyperparameters are listed in §D. Table 2 reports the accuracy of all classifiers used in our experiments. The attention-based models all perform very similarly and generally better than the rationalizer models, except for SNLI, where the latter use a stronger model with decomposable attention. As expected, in general, all these classifiers outperform a bag-of-words model which is the model we use as the layperson.

Layperson L and explainer E. We used a simple linear BoW model as the layperson L. For NLI, the layperson sees the full hypothesis, encoding it with a BiLSTM. The BoW from the explainer is passed through a linear projection and summed with the last hidden state of the BiLSTM.

We evaluated the following explainers:

1. **Erasure**, a wrapper similar to the leave-one-out approaches of Jain and Wallace (2019) and Serrano and Smith (2019). We obtain the word with largest attention, zero out its input vector, and repass the whole input with the erased vector to the classifier C. We produce the message by repeating this procedure k times.

2. **Top-k gradients**, a filter approach that ranks word importance by their "input × gradient" product, $|\frac{\partial \hat{y}}{\partial \mathbf{x}_i} \cdot \mathbf{x}_i|$ (Ancona et al., 2018; Wiegreffe and Pinter, 2019). The top-k words are selected as the message.

		SST		IMDB		AGNEWS		YELP		SNLI	
CLF.	EXPLAINER	CSR	ACC$_L$	CSR	ACC$_L$	CSR	ACC$_L$	CSR	ACC$_L$	CSR	ACC$_L$
C	Random	69.41	70.07	67.30	66.67	92.38	91.14	58.27	53.06	75.83	68.74
C	Erasure	80.12	81.22	92.17	88.72	97.31	95.41	78.72	68.90	77.88	70.04
C	Top-k gradient	79.35	79.24	86.30	83.93	96.49	94.86	70.54	62.86	76.74	69.40
C	Top-k softmax	84.18	82.43	93.06	89.46	**97.59**	95.61	81.00	70.18	78.66	71.00
C_{ent}	Top-k 1.5-entmax	**85.23**	**83.31**	93.32	89.60	97.29	**95.67**	82.20	70.78	80.23	73.39
C_{sp}	Top-k sparsemax	**85.23**	81.93	93.34	89.57	95.92	94.48	82.50	70.99	**82.89**	**74.76**
C_{ent}	Selec. 1.5-entmax	83.96	82.15	92.55	89.96	97.30	95.66	81.38	70.41	77.25	71.44
C_{sp}	Selec. sparsemax	**85.23**	81.93	93.24	89.66	95.92	94.48	83.55	71.60	82.04	73.46
C_{bern}	Bernoulli	82.37	78.42	91.66	86.13	96.91	94.43	84.93	66.89	76.81	69.65
C_{hk}	HardKuma	85.17	80.40	**94.72**	**90.16**	97.11	95.45	**87.39**	**71.64**	74.98	71.48

Table 3: CSR and layperson accuracy (ACC$_L$) for several explainers. For each explainer, we indicate the corresponding classifier from Table 2; in all cases the layperson is a BoW model. Only explainers of the same classifier can be compared in terms of CSR. Top rows report performance for random, wrapper and filter explainers, for fixed k-word messages (the values of k for the several datasets are $\{5, 10, 10, 10, 4\}$, respectively). Bottom rows correspond to embedded methods where k is given automatically via sparsity. The average k obtained by 1.5-entmax, sparsemax, Bernoulli and HardKuma are: SST: $\{4.65, 2.59, 6.10, 4.82\}$; IMDB: $\{28.23, 12.94, 39.40, 24.18\}$; AGNEWS $\{5.65, 4.14, 4.01, 9.68\}$; YELP: $\{60.61, 23.86, 9.15, 33.18\}$; SNLI: $\{12.96, 8.27, 15.04, 6.40\}$.

3. **Top-k and selective attention:** We experimented both using attention as a *filter*, by selecting the top-k most attended words as the message, and *embedded* in the classifier C, by using the selective attentions described in §3 (1.5-entmax and sparsemax).

4. **The rationalizer models of Lei et al. (2016) and Bastings et al. (2019).** These models compose the message by stochastically sampling rationale words, respectively using Bernoulli and HardKuma distributions. For SNLI, since these models use decomposable attention instead of RNNs, we form the message by selecting all premise words that are linked with any hypothesis word via a selected Bernoulli variable.

We also report a **random** baseline, which randomly picks k words as the message. We show examples of messages for all explainers in §I.

Results. Table 3 reports results for the communication success rate (CSR, Eq. 1) and for the accuracy of the layperson (ACC$_L$). For each explainer, we indicate which classifier it is explaining; note that the CSR is only comparable across explainers that use the same classifier. The goal of this experiment is to answer the following questions: (i) How do different explainers (wrappers, filters, embedded) compare to each other? (ii) Are selective sparse attention methods effective? (iii) How is the trade-off between message length and CSR?

The first thing to note is that, as expected, the random baseline is much worse than the other explainers, for all text classification datasets.[4] Among the non-trivial explainers, **the attention and erasure outperform gradient methods**: the erasure and top-k attention explainers have similar CSR, with a slight advantage for attention methods. Note that the attention explainers have the important advantage of requiring a single call to the classifier, whereas the erasure methods, being wrappers, require k calls. The worse performance of top-k gradient (less severe on AGNEWS) suggests that the words that locally cause bigger output changes are not necessarily the most informative ones.[5]

Regarding the different attention models (softmax, entmax, and sparsemax), we see that **sparse transformations tend to have slightly better ACC$_L$**, in addition to better ACC$_C$ (see Table 2). The embedded sparse attention methods achieved communication scores on par with the top-k attention methods without a prescribed k, while producing, by construction, more faithful explanations. Both our proposed models (sparsemax and 1.5-entmax) seem generally more accurate than the Bernoulli model of Lei et al. (2016) and comparable to the HardKuma model of Bastings et al. (2019), with a much simpler training procedure,

[4] This is less pronounced in SNLI, as the hypothesis alone already gives strong baselines (Gururangan et al., 2018).

[5] A potential reason is that attention directly influences C's decisions, being an inside component of the model. Gradients and erasure, however, are extracted after decisions are performed. The reason might be similar to filter methods being generally inferior to embedded methods in static feature selection, since they ignore feature interactions that may jointly play a role in model's decisions.

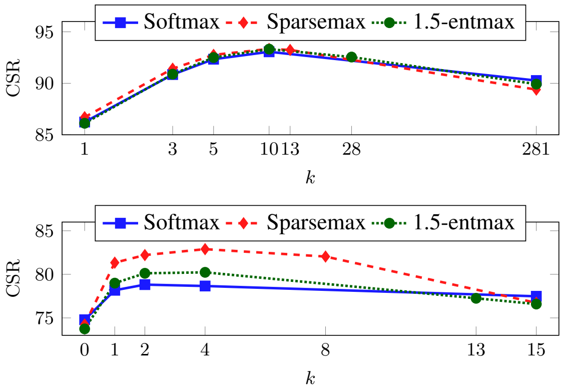

Figure 2: Message sparsity analysis for IMDB (top) and SNLI (bottom). For SNLI, $k = 0$ corresponds to a case where the layperson only sees the hypothesis. The rightmost entry represents an explainer that simply passes forward all words to the layperson.

not requiring gradient estimation over stochastic computation graphs.

Finally, Figure 2 shows the trade-off between the length of the message and the communication success rate for different values of k both for IMDB and SNLI (see Figure 4 in §G for the IWSLT experiments, with similar findings). Interestingly, we observe that **CSR does not increase monotonically with k.** As k increases, CSR starts by increasing but then it starts dropping when k becomes too large. This matches our intuition: in the two extreme cases where $k = 0$ and where k is the document length (corresponding to a full bag-of-words classifier) the message has no information about how the classifier C behaves. By setting $k = 0$, meaning that the layperson L only looks at the hypothesis, the CSR is reasonably high (∼74%), but as soon as we include a single word in the message this baseline is surpassed by 4 points or more.

6 Human Evaluation

To fully assess the quality of the explanations in a more realistic forward simulation setting, we performed human evaluations, where the layperson L is a human instead of a machine.

Joint training of E and L. So far we compared several explainers, but what happens if we train E and L jointly to optimize CSR directly, as described in §4.2? We experiment with the IMDB and SNLI datasets, comparing with using humans for either the layperson, the explainer, or both.

Human layperson. We randomly selected 200 documents for IMDB and SNLI to be annotated by humans. The extracted explanations (i.e. the selected words) were shuffled and displayed as a cloud of words to two annotators, who were asked to predict the label of each document when seeing only these explanations. For SNLI, we show the entire hypothesis as raw text and the premise as a cloud of words. The agreement between annotators and other annotation details can be found in §H.

Human explainer. We also consider explanations generated by humans rather than machines. To this end, we used the e-SNLI corpus (Camburu et al., 2018), which extends the SNLI with human rationales. Since the e-SNLI corpus does not provide highlights over the premise for neutral pairs, we removed them from the test set.[6]

We summarize our results in Table 4. We observe that, also with human laypeople, top-k attention achieves better results than top-k gradient, in terms of CSR and ACC, and that the ACC of erasure, attention models, and human explainers are close, reinforcing again the good results for these explainers. Among the different attention explainers, we see that selective attention explainers (§3) got very high ACC$_H$, outperforming top-k explainers for SNLI. We also see that the joint explainer (§4.2) outperformed all the other explainers in ACC$_L$ and CSR$_L$ and achieved very high human performance on IMDB, largely surpassing other systems in CSR$_H$ and ACC$_H$. This shows the potential of our communication-based framework to develop new post-hoc explainers with good forward simulation properties. However, for SNLI, the joint explainer had much lower CSR$_H$ and ACC$_H$, suggesting that for this task more sophisticated explainers are required.

7 Related Work

There is a large body of work on analysis and interpretation of neural networks. Our work focuses on *prediction explainability*, different from transparency or model interpretability (Doshi-Velez, 2017; Lipton, 2018; Gilpin et al., 2018).

Rudin (2019) defines explainability as a plausible reconstruction of the decision-making process, and Riedl (2019) argues that it mimics what humans do when rationalizing past actions. This inspired our post-hoc explainers in §4.2 and their use of the faithfulness loss term.

[6] Note that the human rationales from eSNLI are not explanations about C, since the humans are explaining the gold labels. Therefore, we have CSR=ACC always.

CLF.	EXPLAINER	IMDB					SNLI				
		k	CSR$_H$	CSR$_L$	ACC$_H$	ACC$_L$	k	CSR$_H$	CSR$_L$	ACC$_H$	ACC$_L$
C	Erasure	5.0	89.25	94.00	86.25	90.00	4.0	72.50	73.50	83.50	70.00
C	Top-k gradient	5.0	73.50	84.50	73.00	80.50	4.0	65.75	72.50	76.75	68.00
C	Top-k softmax	5.0	89.25	93.00	88.25	88.00	4.0	72.00	76.50	82.75	71.50
C_{ent}	Top-k 1.5-entmax	5.0	89.25	92.50	85.75	86.50	4.0	70.00	81.50	80.50	76.50
C_{sp}	Top-k sparsemax	5.0	89.00	89.50	87.50	88.00	4.0	68.25	88.00	80.25	77.00
C_{ent}	Selec. 1.5-entmax	27.2	86.50	92.50	84.00	89.50	12.9	75.25	77.00	87.00	77.00
C_{sp}	Selec. sparsemax	12.8	87.75	92.50	86.75	89.00	8.0	72.25	82.00	85.00	79.00
C_{bern}	Bernoulli	39.4	79.00	93.50	75.00	87.00	15.2	74.50	76.00	86.75	69.50
C_{hk}	HardKuma	24.3	83.75	93.50	80.75	89.00	6.4	79.25	71.50	**87.50**	68.50
C	Joint E and L	2.7	**96.75**	**98.50**	**89.25**	**91.50**	2.8	58.00	**93.50**	70.00	78.50
-	Human highlights	-	-	-	-	-	2.8	**83.25**	83.50	83.25	**83.50**

Table 4: Results of the human evaluation. Reported are average message length k, human layperson CSR$_H$/ACC$_H$, and machine layperson CSR$_L$/ACC$_L$. Only explainers of the same classifier can be compared in terms of CSR.

Recent works questioned the interpretative ability of attention mechanisms (Jain and Wallace, 2019; Serrano and Smith, 2019). Wiegreffe and Pinter (2019) distinguished between faithful and plausible explanations and introduced several diagnostic tools. Mullenbach et al. (2018) use human evaluation to show that attention mechanisms produce plausible explanations, consistent with our findings in §6. None of these works, however, considered the sparse selective attention mechanisms proposed in §3. Hard stochastic attention has been considered by Xu et al. (2015); Lei et al. (2016); Alvarez-Melis and Jaakkola (2017); Bastings et al. (2019), but a comparison with sparse attention and explanation strategies was still missing.

Besides attention-based methods, many other explainers have been proposed using gradients (Bach et al., 2015; Montavon et al., 2018; Ding et al., 2019), leave-one-out strategies (Feng et al., 2018; Serrano and Smith, 2019), or local perturbations (Ribeiro et al., 2016; Koh and Liang, 2017), but a link with filters and wrappers in the feature selection literature has never been made. We believe the connections revealed in §2 may be useful to develop new explainers in the future.

Our trained explainers from §4.2 draw inspiration from emergent communication (Lazaridou et al., 2016; Foerster et al., 2016; Havrylov and Titov, 2017). Some of our proposed ideas (e.g., using sparsemax for end-to-end differentiability) may also be relevant to that task. Our work is also related to sparse auto-encoders, which seek sparse overcomplete vector representations to improve model interpretability (Faruqui et al., 2015; Trifonov et al., 2018; Subramanian et al., 2018). In contrast to these works, we consider the non-zero attention probabilities as a form of explanation.

Some recent work (Yu et al., 2019; DeYoung et al., 2020) advocates *comprehensive* rationales. While comprehensiveness could be useful in our framework to prevent trivial communication protocols between the explainer and layperson, we argue that it is not always a desirable property, since it leads to longer explanations and an increase of human cognitive load. In fact, our analysis of CSR as a function of message length (Figure 2) suggests that shorter explanations might be preferable. This is aligned to the "explanation selection" principle articulated by Miller (2019, §4): *"Similar to causal connection, people do not typically provide all causes for an event as an explanation. Instead, they select what they believe are the most relevant causes."* Our sparse, selective attention mechanisms proposed in §3 are inspired by this principle.

8 Conclusions

We proposed a unified framework that regards explainability as a communication problem between an explainer and a layperson about a classifier's decision. We proposed new embedded methods based on selective attention, and post-hoc explainers trained to optimize communication success. In our experiments, we observed that attention mechanisms and erasure tend to outperform gradient methods on communication success rate, using both machines and humans as the layperson, and that selective attention is effective, while simpler to train than stochastic rationalizers.

Acknowledgements

This work was supported by the European Research Council (ERC StG DeepSPIN 758969), by the P2020 program MAIA (contract 045909), and by the Fundação para a Ciência e Tecnologia through contract UID/50008/2019. We are grateful to Thales Bertaglia, Erick Fonseca, Pedro Martins, Vlad Niculae, Ben Peters, Gonçalo Correia and Tsvetomila Mihaylova for insightful group discussion and for the participation in human evaluation experiments. We also thank the anonymous reviewers for their helpful discussion and feedback.

References

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in NLP. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 1–7, Berlin, Germany. Association for Computational Linguistics.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark. Association for Computational Linguistics.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 2015 International Conference on Learning Representations*.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 2–14.

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. 2019. Adaptively sparse transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2174–2184, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Shuoyang Ding, Hainan Xu, and Philipp Koehn. 2019. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 1–12, Florence, Italy. Association for Computational Linguistics.

Been Doshi-Velez, Finale; Kim. 2017. Towards a rigorous science of interpretable machine learning. In *eprint arXiv:1702.08608*.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. 2015. Sparse overcomplete word vector representations. In *Proceedings* *of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1491–1500, Beijing, China. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems 29*, pages 2137–2145. Curran Associates, Inc.

L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89.

Yoav Goldberg and Graeme Hirst. 2017. *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(null):1157–1182.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3):389–422.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems 30*, pages 2149–2159. Curran Associates, Inc.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894, International Convention Centre, Sydney, Australia. PMLR.

Ron Kohavi and George H. John. 1997. Wrappers for feature subset selection. *Artif. Intell.*, 97(1–2):273–324.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

David K. Lewis. 1969. Convention: A philosophical study.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016a. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. volume 48 of *Proceedings of Machine Learning Research*, pages 1614–1623, New York, New York, USA. PMLR.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1 – 38.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1 – 15.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Ben Peters, Vlad Niculae, and André F. T. Martins. 2019. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519, Florence, Italy. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. 2018. Spine: Sparse interpretable neural embeddings. In *Proceedings of the Thirty Second AAAI Conference on Artificial Intelligence (AAAI).*

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Valentin Trifonov, Octavian-Eugen Ganea, Anna Potapenko, and Thomas Hofmann. 2018. Learning and evaluating sparse interpretable sentence embeddings. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 200–210, Brussels, Belgium. Association for Computational Linguistics.

Constantino Tsallis. 1988. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics.*

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language*

Processing, pages 31–40, Honolulu, Hawaii. Association for Computational Linguistics.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.