Auditing Deep Learning processes through Kernel-based Explanatory Models

Danilo Croce and Daniele Rossini and Roberto Basili
Department of Enterprise Engineering
University of Roma, Tor Vergata
{croce, basili}@info.uniroma2.it
rossini.danie@gmail.com

Abstract

While NLP systems become more pervasive, their accountability gains value as a focal point of effort. Epistemological opaqueness of non-linear learning methods, such as deep learning models, can be a major drawback for their adoptions. In this paper, we discuss the application of Layerwise Relevance Propagation over a linguistically motivated neural architecture, the Kernel-based Deep Architecture, in order to trace back connections between linguistic properties of input instances and system decisions. Such connections then guide the construction of argumentations on the network’s inferences, i.e., explanations based on real examples that are semantically related to the input. We also propose here a methodology to evaluate the transparency and coherence of analogy-based explanations modeling an audit stage for the system. Quantitative analysis on two semantic tasks, i.e., question classification and semantic role labeling, shows that the explanatory capabilities (native in KDAs) are effective and they pave the way to more complex argumentation methods.

1 Introduction

AI systems are currently used in a wide variety of applications, with several levels of societal impact, and are expected to be soon deployed in safety-critical fields, e.g., autonomous driving. The definition of codes of conduct in the development of AI applications to ensure their ethical sustainability across dimensions, such as fairness, reliability and beneficialness (Kroll et al., 2016; Garfinkel et al., 2017; Dignum, 2017) is becoming a crucial issue. Hence, a natural need for the ethical accountability of such systems is gaining importance.

A central issue lies in designing systems whose decisions are transparent (Ribeiro et al., 2016; Doshi-Velez et al., 2017), i.e., they must be easily interpretable by humans, as users must be able to suitably weight and trust the assistance of such systems.

Deep neural networks are clearly problematic in this regard: their high non-linearity, despite allowing for state-of-the-art performances in several challenging problems, amplifies the epistemological opaqueness of the decision-flow and limits its interpretability. The concept of transparency of a machine learning model spans multiple definitions, focusing on different aspects, from the simplicity of the model, e.g., the number of nodes in a decision tree, to the intuitiveness of its parameters and computations (Chakraborty et al., 2017).

In this context, an important capability of an AI system is the ability to provide post-hoc explanations in terms of evidences supporting the produced decisions: although they usually do not formally elucidate how a model works, they often have the property of being quite intuitive, conveying useful information also to end-users without any AI or machine learning expertise (Lipton, 2018). In semantic inference tasks (e.g., text classification), an explanation model generating post-hoc explanations should hence be able to trace back connections between the output categories and the semantic and syntactic properties of the input texts. Such models should have three desired properties: semantic transparency, informativeness w.r.t. the system decision and effectiveness in enabling auditing processes against the system.

In this work we focus on a specific post-hoc mechanism, which is to provide, along with the prediction, a comparison with one or more other examples, namely landmarks, that share task-relevant linguistic properties with the input. From an argument theory perspective, this corresponds to supporting decisions through an “argument by analogy” schema (Walton et al., 2008): a user ex-
posed to such a kind of argument will endow a different level of trust into the machine decision according to the linguistic plausibility of the analogy. He will implicitly gauge the evidence from the linguistic properties shared between the input sentence (or its parts) and the one(s) used for comparison as well their importance with respect to the output decision. Let us consider, for example, the following prediction in a question classification (QC) task (Li and Roth, 2006): “What is the capital of Zimbabwe?” refers to a Location. We would like the system to motivate its decision with an argument such as: ...since it recalls me of “What is the capital of California?” which also refers to a Location. Notice that explanation of a decision is quite different from sentence or document ranking in Information Retrieval so that semantic similarity plays only a minor role: clear and trustful analogies may exist with semantically different training examples that imply similar relationships between the input and the decision.

Recent work has been inspired by efforts in improving model’s interpretability in image processing tasks, in particular by the Layerwise Relevance Propagation (LRP) (Bach et al., 2015). In LRP the classification decision of a deep neural network is decomposed backward across the network layers and evidence about the contribution to the final decision brought by individual input fragments (i.e., pixels of the input image) is gathered.

In this paper, we propose to extend the LRP application to the linguistically motivated network architectures, known as Kernel-Based Deep Architectures (KDAs) (Croce et al., 2017), which frames semantic information captured by linguistic Tree Kernel methods (Collins and Duffy, 2001) within the neural-based learning paradigm. The result is a mechanism that, for each system’s prediction such as in question classification, generates an argument-by-analogy explanation based on real training examples, not necessarily similar to the input.

We also propose a novel approach to evaluate numerically the interpretability of any explanation-enriched model applied in semantic inference tasks. By defining a specific audit process, we derive a synthetic metric, i.e. Audit-ing Accuracy, that takes into account the properties of transparency, informativeness and effectiveness. The evaluation of the proposed methodology shows the meaningful impact of LRP-based explanation models: users faced with explanations are systemmatically oriented to accept (or reject) the system decisions, so that post hoc judgments may even improve the overall application accuracy.

In the rest of the paper, Section 2 reports related works, while Section 3 describes the LRP and its extension to KDAs. In Section 4, we propose three explanation models and illustrate a novel evaluation methodology, commenting on the audit process and deriving quantitative notions such as the Auditing Accuracy measure. Section 5 presents and discusses the system effectiveness against two semantic tasks, i.e., question classification and frame-based argument classification in a semantic role labeling chain. Finally, in Section 6 conclusions are derived.

2 Related Work

In recent years, research communities showed great interest in improving neural models’ interpretability, as testified by the effort of defining the concept of interpretability itself and the development of a variety of approaches to the problem. In (Chakraborty et al., 2017) and (Lipton, 2018), the authors examine the different notions of interpretability found in literature and categorize techniques according to the transparency properties they confer to decision models. Common approaches to improve the readability of a neural model in image-related tasks are based on backward algorithms that reuse arc weights to propagate the prediction down to the input (Erhan et al., 2010; Zeiler and Fergus, 2013), thus leading to the re-creation of meaningful patterns in the input space. Typical examples are deconvolution heatmaps, used to approximate through Taylor series the partial derivatives at each layer (Simonyan et al., 2013), or the so-called Layer-wise Relevance Propagation (LRP), that redistributes back positive and negative evidence across the layers (Bach et al., 2015).

Local explanation approaches focus on highlighting a handful of crucial features (Baehrens et al., 2010) or deriving simpler, more readable models from a complex one, e.g., a binary decision tree (Frosst and Hinton, 2017), or by local approximation with linear models (Ribeiro et al., 2016). However, although they can explicitly show the representations learned in the specific hidden neurons (Frosst and Hinton, 2017), these approaches base their effectiveness on the user ability to study
the quality of the reasoning and of the account-
ability as a side effect of the quality and the co-
herence of the features selection: this can be very
hard in tasks where boundaries between classes
are not well defined. Another strategy is pair-
ing the decision model with a generative model
to produce verbose explanations (Krenning et al.,
2017). Sometimes explanations are associated to
vector representations as in (Ribeiro et al., 2016),
i.e., bag-of-words in case of text classification, that
are clearly weak at capturing significant linguistic
abstractions, such as the involved syntactic rela-
tions. In this work, we systematically extend the
model presented in (Croce et al., 2018) which al-
 lows to provide explanations that are easily inter-
pretable even by non-expert users, as they are ex-
pressed in natural language. Moreover, the inves-
tigated approach is is computationally affordable,
as it roughly corresponds to a forward pass across
the network. In addition, we also provide a sys-
tematic way to evaluate the provided explanations
with a methodology able to support the audit of the
targeted AI systems.

3 Layer-wise Relevance Propagation in
Kernel-based Deep Architectures

In this section, we will review the Layer-wise Re-
levance Propagation technique (LRP, as in (Bach
et al., 2015)), usually applied in image process-
ing, and show how it can be naturally extended to
Kernel-based Deep Architectures (KDA, as in
(Croce et al., 2017)) in order to select real exam-
ple useful to support the network decisions.

LRP is mainly a framework which allows to de-
compose the prediction of a deep neural network
computed over a sample, usually an image, down
to relevance scores for the single input dimensions
of the sample, such as sub-pixels of the image it-
self. More formally, let \( f : \mathbb{R}^d \rightarrow \mathbb{R}^+ \) be a func-
tion that quantifies, for example, the probability
of \( x \in \mathbb{R}^d \) being in a certain class. The Layer-wise
Relevance Propagation assigns to each dimension,
or feature, \( x_d \) a relevance score \( R_d^{(l)} \) such that
\[
 f(x) \approx \sum_d R_d^{(l)}.
\]
Features whose score is \( R_d^{(l)} > 0 \) or \( R_d^{(l)} < 0 \) correspond to evidence in favor
or against, respectively, the output classification.
In other words, LRP allows to identify fragments
of the input playing key roles in the decision, by
propagating relevance backwards. Let us suppose
to know the relevance score \( R_j^{(l+1)} \) of a neuron \( j \)
at network layer \( l + 1 \). This can be decomposed
into messages \( R_{i\rightarrow j}^{(l+1)} \) sent from \( j \) to neurons \( i \)
in layer \( l \) according to \( R_j^{(l+1)} = \sum_{i \in D} R_{i\rightarrow j}^{(l+1)} \).
Then \( t_i \) directly follows that the relevance of a neu-
ron \( i \) at layer \( l \), that is the quantity of informa-
tion travelling through \( i \), can be defined as
\[
 R_i^{(l)} = \sum_{j \in (l+1)} R_{i\rightarrow j}^{(l+1)}.
\]
In this work, we adopted the \( \epsilon \)-rule defined in (Bach et al., 2015) to compute the
messages \( R_{i\rightarrow j}^{(l+1)} = \frac{z_{ij}}{\sum \sigma(z_{ij})} R_{i}^{(l+1)} \)
where \( z_{ij} = x_i w_{ij} \) and \( \epsilon > 0 \) is a small numerical sta-
bilizing term. The informative value is justified by
the fact that the weights \( z_{ij} \) are linked to the ac-
tivation weights \( w_{ij} \) of the input neurons.

Given the capability of computing relevance scores for input dimensions, we now summa-
rize the KDA to motivate how LRP can be ap-
plied also to tasks other than image classification.
In a nutshell, the KDA is a neural network
trained in low-dimensional spaces which approxi-
mate a generic Reproducing Kernel Hilbert Space
(RKHS) (Shawe-Taylor and Cristianini, 2004).
These low-dimensional approximations are de-
\[
 & \Phi(O) \approx \Phi(O_{\text{training}}) + \Phi(O_{\text{landmarks}}).
\]
Given the capability of computing relevance scores for input dimensions, we now summa-
rize the KDA to motivate how LRP can be ap-
plied also to tasks other than image classification.
In a nutshell, the KDA is a neural network
trained in low-dimensional spaces which approxi-
mate a generic Reproducing Kernel Hilbert Space
(RKHS) (Shawe-Taylor and Cristianini, 2004).
These low-dimensional approximations are de-
\[
 & \Phi(O) \approx \Phi(O_{\text{training}}) + \Phi(O_{\text{landmarks}}).
\]
Given the capability of computing relevance scores for input dimensions, we now summa-
rize the KDA to motivate how LRP can be ap-
plied also to tasks other than image classification.
In a nutshell, the KDA is a neural network
trained in low-dimensional spaces which approxi-
mate a generic Reproducing Kernel Hilbert Space
(RKHS) (Shawe-Taylor and Cristianini, 2004).
These low-dimensional approximations are de-
\[
 & \Phi(O) \approx \Phi(O_{\text{training}}) + \Phi(O_{\text{landmarks}}).
\]
landmarks

Figure 1: Kernel-based Deep Architecture.

function $\tilde{x} = \bar{c} US^{-\frac{1}{2}}$. Therefore, the method produces $l$-dimensional vectors$^1$.

In (Croce et al., 2017), the Nyström representation $\tilde{x}$ has been used to map semantically annotated grammatical trees to the linear input of a Multi-Layer Perceptron (MLP). In fact, given a dataset $\mathcal{L}$, with $o \in \mathcal{L}$ denoting a generic instance, the MLP architecture is defined with a specific Nyström input layer based on the Nyström embeddings. The resulting Kernel-based Deep Architecture (KDA) includes an input layer, the Nyström layer, a sequence of hidden layers and the final classification layer, which produces the output. The input layer corresponds to the input vector $\vec{c}_i$, i.e., the row of the $C$ matrix associated to an example $o_i$. The input layer is mapped to the Nyström layer, through the Nyström projection. Notice that the embedding provides also the proper weights, defined by $US^{-\frac{1}{2}}$, so that the mapping can be expressed through the Nyström matrix $H_N = US^{-\frac{1}{2}}$. The resulting $\tilde{x}$ is the input to one or more hidden layers. Clearly, the first hidden layer receives in input $\tilde{x} = \bar{c} H_N y$. Finally, the classification layer computes a linear classification function with a softmax operator, as shown in Figure 1. A KDA optimizes the standard cross-entropy function with $L_2$ regularization.

It is worth recalling that the network is triggered by an input vector $\vec{c}$ expressing the kernel evaluations $K(x,l_i)$ between the example and the landmarks. When using linguistic kernels (such as Semantic Tree Kernels, (Croce et al., 2011)), this measure corresponds to the grammatical and lexical semantic similarity between $x$ and the subset of landmarks. The expected explanation is obtained from the network output by applying LRP to revert the propagation process, thus linking the output back to the input. In a KDA that models linguistic instances, LRP implicitly traces back the syntactic, semantic and lexical relations between the example and the landmarks across the Nyström layer: the side effect is to select those real examples that mostly influenced the identification of the predicted structure in the sentence.

4 Generating explanations in Kernel-based Deep Architectures

Justifications for the KDA emissions can be obtained by exploiting landmarks $\{\ell\}$ as the evidence in favour or against a class. The idea is to select those $\{\ell\}$ that the LRP highlights as the most active elements in layer 0. Once such active landmarks are detected, an Explanatory Model is a function in charge to compile a linguistically fluent explanation by comparing the input case with such selection.

The semantic expressiveness of such analogies makes the resulting explanation clear and increases the user confidence on the system reliability. When a sentence $s$ is classified, LRP assigns activation scores $r^s_{\ell}$ to each individual landmark $\ell$: let $\mathcal{L}^{(+)}$ (or $\mathcal{L}^{(-)}$) denote the set of landmarks with positive (or negative) activation scores. Formally, each explanation is characterized by a triple $e = \langle s, C, \tau \rangle$ where $s$ is the input sentence, $C$ is a target label and $\tau$ is the modality of the explanation: $\tau = +1$ for positive (i.e., acceptance) statements while $\tau = -1$ correspond to rejections of $C$. A landmark $\ell$ is positively activated for a given sentence $s$ if there are at most $k - 1$ other active landmarks $\ell'$ with activation value higher than the one for $\ell$, i.e.,

$$|\{\ell' \in \mathcal{L}^{(+)} : \ell' \neq \ell \land r^s_{\ell'} \geq r^s_{\ell} > 0\}| < k$$

Similarly, a landmark is negatively activated when:

$$|\{\ell' \in \mathcal{L}^{(-)} : \ell' \neq \ell \land r^s_{\ell'} \leq r^s_{\ell} < 0\}| < k$$

where $k$ is a fixed parameter used to make the explanation depending on not more than $k$ landmarks, denoted as a set by $\mathcal{L}_k$. Positively (or negative) active landmarks in $\mathcal{L}_k$ are assigned an activation value $a(\ell, s) = +1 (-1)$, while $a(\ell, s) = 0$ for not active landmarks. Given the explanation $e = \langle s, C, \tau \rangle$, a landmark $\ell$, whose known class is $C_\ell$, is called consistent (or inconsistent) with $e$ if the function

$^1$Note that, while any sampling policy for landmarks is admissible, in (Kumar et al., 2012) it is demonstrated that uniform sampling without replacement achieves results comparable with alternative, more resource-consuming policies.
\(\delta(C, C') \cdot a(\ell, q) \cdot \tau\) is positive (or negative, respectively), where \(\delta(C', C) = 2\delta_{\text{ kron}}(C', C) - 1\) and \(\delta_{\text{ kron}}\) is the Kronecker delta.

We can thus partition such landmarks into the set of positively consistent landmarks \(L^c_+ \subseteq L^c_+ \subseteq L^c\) and negatively consistent ones \(L^c_- \subseteq L^c_- \subseteq L^c\), with \(L^c_+ \cup L^c_- = L^c\), that aggregates all the consistent landmarks.

An explanatory model is then a function \(M(e, L^c_k)\) which maps an explanation \(e\) and the set \(L^c_k\) for \(e\) into a sentence \(f\) in natural language. Of course several definitions of \(M(e, L^c_k)\) are possible, e.g.,

\[
M(e, L^c_k) = \begin{cases} 
  "s is \(C\) since it recalls me of \(\ell\) & \forall \ell \in L^c_+ \text{ if } \tau > 0 \\
  "s is not \(C\) since it does not recall me of \(\ell\) which is \(C\)" & \forall \ell \in L^c_- \text{ if } \tau < 0 \\
  "s is \(C\) but I don’t know why" & \text{if } L^c \equiv \emptyset 
\end{cases}
\]

Here we introduce three explanatory models used during experimental evaluation:

- **Singleton Model.** The first model is the simplest as it returns a single analogy with the consistent landmark with the highest positive score, if \(\tau = 1\), or lowest negative score, when \(\tau = -1\). As an example, the explanation of an accepted decision in a semantic argument classification task, described by the triple \(e_1 = \langle \text{PUT this plate in the center of the table}, \text{THEME PLACING}, 1\rangle\), would be mapped by the model into: \(I \text{ think “this plate” is THEME of PLACING in “Robot PUT this plate in the center of the table” since it recalls me of “the soap” in “Can you PUT the soap in the washing machine?” and “my coat in the closet in the bedroom”}.\)

- **Conjunctive Model.** In a second model, denoted as Conjunctive, the system makes reference to up to \(k_1 \leq k\) analogies with positively (or negatively) active and consistent landmarks. Given the above explanation \(e_1\), and \(k_1 = 2\), it would return: \(I \text{ think “this plate” is THEME of PLACING in “Robot PUT this plate in the center of the table” since it recalls me of “the soap” in “Can you PUT the soap in the washing machine?” and also of “my coat” in “HANG my coat in the closet in the bedroom”}.\)

- **Contrastive Model.** The last proposed model is more complex since it returns both a positive and a negative analogy by selecting, respectively, the most positively relevant and the most negatively relevant consistent landmark. Given \(e_1\), it would return: \(I \text{ think “this plate” is THEME of PLACING in “Robot PUT this plate in the center of the table” since it recalls me of “the soap” in “Can you PUT the soap in the washing machine?” and it is not the GOAL of PLACING since it does not recall me of “on the counter” in “PUT the plate on the counter”}.\)

In case no active and consistent landmark can be found, the models return a phrase stating only the predicted class, with no explanation.

### 5 Experimental Evaluation

Evaluating the explanatory quality of an inductive model is still an open problem and universally recognized gold standards are not available for comparative analysis. In order to rely on a quantitative analysis, we assume that an explanation to be effective should assist a human user to ascertain whether the proposed classification is correct or not. Plausible and coherent explanations should thus be generated from correct system’s decisions, while bad decisions should correspond to ambiguous or plainly fallacious arguments.

Hence, the evaluation of an explanatory model should reflect the model’s adherence to three desired properties: **semantic transparency**, i.e., argument’s linguistic grounding should be clear and straightforward, requiring as less knowledge on the system’s functioning and on the specific task as possible; **informativeness** with respect to the system’s decision, i.e., the explanation’s generating process should be highly dependent on how the system processes input information; **effectiveness w.r.t an audit against the system**, i.e., the explanation should convey enough meaningful information so that a human can correctly decide whether to trust the system prediction or not.

Consequently, we define a auditing task in which annotators are required to judge if a proposed explanation would commit them to trust the system decision. This judgment is discretized...
within five possible labels: Very Good if the analogy is strongly convincing and linguistically clear; Good if the explanation is still accepted but the pertinence is slacker; Uncertain if the annotator gains no meaningful information from the explanation or no explanation is provided at all; Bad if some connections can be detected between the input sentence and the one used as a comparison but they are so ambiguous that the explanation is rejected; Incoherent if the argument appears totally inconsistent and meaningless. Given the nature of the argument by analogy schema (Walton et al., 2008), it follows that annotators assigning a Very Good or Good label to an explanation are also implicitly accepting the system decision as correct, whereas they are rejecting it as wrong in the other cases.

Given an explanation dataset \( E = \{(e, c, x_c)\} \) where \( e \) is an explanation, \( c \in \{1, -1\} \) expresses if the explanation was generated from a correct \( (c = 1) \) or incorrect \( (c = -1) \) classification, and \( x_c \) is the numerical value corresponding to one of the five labels categories \( C \) above\(^2\), we can define the set \( A_e \) of accepted explanations generated from correct predictions and the set \( R_{nc} \) of rejected explanations generated from not correct predictions as follows:

\[
A_c = \{(e, c, l) \in E | c = 1 \land x_C > 0 \}
\]

\[
R_{nc} = \{(e, c, l) \in E | c = -1 \land x_C \leq 0 \}
\]

Accordingly, the Audit Accuracy (AuAcc)

\[
AuAcc = \frac{|A_c| + |R_{nc}|}{|E|}
\]

measures the ratio between correct acceptance/rejection decisions and the total number of decisions made by the human auditor.

Additionally, the Pearson Correlation between the system classification accuracy and the human judgment of an explanation can be interpreted as a concrete measure of the consistency of an explanatory model: an ideal model should map correct classifications to convincing explanations and incorrect classification to implausible explanations. It will be thus exploited to compare alternative explanatory models. To test an explanation approach as well as of the proposed evaluation metrics, we will address two different semantic processing tasks, i.e., question classification (QC) and argument classification (AC) in semantic role labeling.

**Experimental Setup.** The Nyström projection has been implemented in the KeLP framework\(^3\), while the LRP-integrated KDA in Tensorflow, with 1 and 2 hidden layers, respectively, whose layer-size is equal to the number of randomly selected Nyström landmarks (500 and 200, in QC and AC respectively). For both tasks, training have been executed in 500 epochs, using the Adam optimizer and adopting early-stop and dropout strategy while selecting the best model according to performances over the development set. We conducted preliminary evaluations on small samples of the dataset and set the parameter \( k = 5 \), which defines the cardinality of the active landmarks \( E_k \). The remaining hyper-parameters were tuned via grid-search.

A group of human annotators was asked to rate each explanation with one out of the five labels described early in this section, basing their judgment only on the perceived level of trust w.r.t. the explanations. Each annotator was exposed to explanations derived from a perfectly balanced set of correct and incorrect classifications, so that annotators are not biased by the (possibly high) quality of the classifier when judging the explanations.

### 5.1 Question Classification

We replicated the experiments reported in (Croce et al., 2017) with respect to the question classification task, using the UIUC dataset (Li and Roth, 2006), including a training and test set of 5452 and 500 questions, respectively, organized in 6 coarse-grained classes (as ENTITY or HUMAN). We generated Nyström representation of the Compositionally Smoothed Partial Tree Kernel function (Annesi et al., 2014) consistently with (Croce et al., 2017). Using 500 landmarks, the KDA accuracy was 93.6%, which is comparable with state-of-the-art neural models, as discussed in (Croce et al., 2017). The audit manual task was independently performed by 3 annotators\(^5\): each annotator

---

\(^2\)We assigned values \( x_C \in [1, -1] \): \( x_{\text{Very Good}} = 1 \), \( x_{\text{Good}} = 0.8 \), \( x_{\text{Uncertain}} = 0 \), \( x_{\text{Bad}} = -0.8 \), \( x_{\text{Incoherent}} = -1 \)

\(^3\)http://www.kelp-ml.org, presented in (Filice et al., 2018).

\(^4\)No particular differences have emerged in the generation of explanations when slightly different values of \( k \) where considered.

\(^5\)Annotators have different levels of fluency in English (even though no one is a native speaker) and different levels of expertise: one has no specific technical knowledge, one is a graduate student with advanced knowledge in NLP, while
evaluate 300 explanations (100 for each model), reaching an inter-annotation agreement if 0.82 on these data.

Results in Figure 2 suggest that the annotators were able to properly discriminate correct from incorrect decisions, just through the exposure to the explanations: in both acceptance or rejection cases, all models tend to assign positive labels (Very Good and Good) to explanations of correct decisions and negative ones (Uncertain, Bad and Incoherent) to explanations of incorrect decisions instead. Note that an explanation rejecting a class should be labeled as positive, if the landmark used for negative analogy is actually not recalling the input sentence. The graphical intuition in Figure 2 is confirmed by the metrics: the Singleton, Conjunctive and Contrastive models reach an Audit Accuracy of 89.3%, 84.7% and 86.3%, respectively. The Pearson Correlation between acceptance and correctness is 78.9%, 69.4% and 72.8%, while if we measure the correlation between the explanation quality score and the decision correctness, the Pearson coefficients become 76.1%, 71.2% and 77.2%: these are slightly lower basically for the lower reward assigned to $x_{\text{Good}}$ w.r.t. $x_{\text{Very Good}}$. Small numerical differences among models emerge: it seems that the Conjunctive and Contrastive models are not always able to retrieve meaningful additional information, while the Singleton model is simpler and more direct. An example of output analogies is given by I think “How many Admirals are there in the U.S. Navy?” refers to a NUMBER since it recalls me of both “How many words are there in the Spanish language?” and “How many sides does an obelisk have?”, generated by the Conjunctive model. Here the semantic hint corresponds to the discriminative fragment “How many”. However meaningful connections between the input and landmarks are also traced against poor overlaps in syntactic and lexical information as in: I think “Where is the Mall of the America?” refers to a LOCATION since it recalls me of “What town was the setting for The Music Man?”.

Table 1 reports question-explanation pairs with similarity estimates based on the adopted CSPTK kernel function. It is clear from the examples that similarity alone is not able to correlate with classification decisions: questions in different classes (e.g. first two rows in the table) may have very high similarity scores. Second, landmarks correlate with decisions in interesting ways that do not depend on strict lexical and grammatical similarity. Conceptually more grounded associations seem to emerge: e.g., explaining “What was J.F.K.’s wife’s name” by the analogy with “What was Darth Vader’s son named?” is abstracted across a conceptual relation (e.g. has_name) and the derived analogy is quite clear. Notice that active landmarks are independent from similar questions, as landmarks triggered by similar questions are not similar to each other.

Interestingly, explanations of ambiguous instances are harmonic with human uncertainty. The explanation I think “What is the sales tax in Minnesota?” refers to a NUMBER since it recalls me of “What is the population of Mozambique?” and does not refer to a ENTITY since it does not recall me of “What is a fear of slime?” is convincing, but incorrect. Here, the lack of context impacts on the disambiguation of two plausible interpretations that are (1) the definition of the notion of “sales tax” (ENTITY), w.r.t (2) its current value (NUMBER): the gold standard suggests ENTITY as the correct category.

5.2 Argument Classification

Semantic role labeling (SRL (Palmer et al., 2010)) consists in detecting the semantic arguments associated with the predicate of a sentence and their
classification into their specific roles ((Fillmore, 1985)). For example, given the sentence “Bring the fruit onto the dining table”, the task would be to recognize the verb “bring” as evoking the bringing frame, with its roles, Theme for “the fruit” and Goal for “onto the dining table”. Argument classification corresponds to the subtask of assigning labels to the sentence fragments spanning individual roles.

As proposed in (Moschitti et al., 2008), SRL can be modeled as a multi-classification task over each parse tree node \( n \), where argument spans reflect sub-sentences covered by the tree rooted at \( n \). Consistently with (Croce et al., 2011), in our experiments the KDA has been empowered with a Smoothed Partial Tree Kernel, operating over Grammatical Relation Centered Trees (GRCT) derived from dependency grammar. The reference benchmark, i.e., the HuRIC dataset related to an Interactive Robotics task (Bastianelli et al., 2016), includes about 650 annotated transcriptions of spoken robotic commands, organized in 18 frames and about 60 arguments. Individual arguments extracted amount to 1,300 examples. Experimental setup was similar to that of Section 5.1, but due to the limited data size we applied 10-fold cross-validation, optimizing network hyper-parameters via grid-search for each fold. We generated the Nyström representations of a SPTK function with default parameters \( \mu = \lambda = 0.4 \) as in (Croce et al., 2011). With these settings, the KDA accuracy was 96.1%. Due to the slightly higher complexity of the task, w.r.t. QC, in the case the two independent auditors had at least graduate-level knowledge in NLP. They were requested to judge about 700 generated explanations with an inter-annotation agreement of 0.89. For the Singleton, Conjunctive and Contrastive model, respectively, the Audit Accuracy is 91.6%, 93.4%, 88.4% while Pearson Coefficients between acceptance/rejection and correctness are 83.3% (80.1% for quality-correctness correlation), 86.9% (81.9%), 77.3% (78.7%); this suggests an higher annotator’s sensitivity to the explanations’ plausibility, as reflected also by the charts in Figure 3, probably due to the task itself being more challenging for humans.

As in QC, the system can convey semantically transparent and useful information without relying on lexical similarity alone; e.g., consider I think “is hot” is Desired State of Inspecting in “Robot check whether the oven is hot?” since it recalls me of “is empty” in “See if the washing machine is empty”. In this task, the Contrastive model could also to produce explanations exemplifying differences between separate roles in the same frame, for example: I think “to me” is not

### Table 1: Examples of semantically similar questions in the same or different classes, with the corresponding landmarks activated during the classification.

<table>
<thead>
<tr>
<th>Class</th>
<th>Questions ((q_i))</th>
<th>(k(q_1, q_2))</th>
<th>Activated Landmarks ((l_i))</th>
<th>(k(l_1, l_2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOC</td>
<td>“What is the capital of Ethiopia?”</td>
<td>0.98</td>
<td>“What is the name of David Letterman’s dog?”</td>
<td>0.61</td>
</tr>
<tr>
<td>NUM</td>
<td>“What is the population of Nigeria?”</td>
<td>0.97</td>
<td>“What is the name of David Letterman’s dog?”</td>
<td>0.61</td>
</tr>
<tr>
<td>HUM</td>
<td>“What was J.F.K.’s wife’s name?”</td>
<td>0.90</td>
<td>“What is the name of David Letterman’s dog?”</td>
<td>0.61</td>
</tr>
<tr>
<td>HUM</td>
<td>“What is the Ohio state bird?”</td>
<td>0.83</td>
<td>“What was the first TV set to include a remote control?”</td>
<td>0.61</td>
</tr>
<tr>
<td>NUM</td>
<td>“What was the first satellite to go into space?”</td>
<td>0.73</td>
<td>“What is average salary of restaurant manager in United States?”</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Figure 3: Results of the audit process in the SRL task for the three explanatory models: system acceptance (top), rejection (bottom).
Goal of Bringing in “Can you go to the kitchen find a glass and bring it to me?” since it does not recall me of “to the bedroom” in “Bring the phone to the bedroom” and it’s the Beneficiary of Bringing since it recalls me of “to me” in “can you please search the book and bring it to me”.

6 Conclusions

This paper proposes a quantitative evaluation of the automatic generation of epistemologically transparent and linguistically fluid explanations for neural inferences. The proposed approach applies LRP to a Kernel-based Deep Architecture (KDA) that redistributes the prediction value to training entries (i.e., annotated landmarks). The resulting sentence exploits analogies with training instances, according to different explanatory strategies. Given that KDAs (based on Nyström embeddings) can be flexibly adopted in neural learning for NLP, we show how the auditing mechanism outlined in the paper is epistemologically very effective and emphasizes the neural embeddings with a strong impact on explainability. First, language semantics is promoted by design and associations generated between input instances and decisions are obtained without ever leaving the language level. Second, different and mathematically solid models for different levels of language semantics can be obtained by modifying the adopted kernel formulations. In this way, a unique general auditing mechanism is able to support fine tuning towards very different tasks, without changes in the learning architecture. Finally, as Table 1 shows, explanations are strictly dependent on the induced neural model and are not just triggered by text similarity metrics: they are epistemologically principled evidences about the neural learning stages, based on the observed examples and the selected landmarks.

Empirical investigations carried out against the QC and AC tasks also confirm that the good explanatory models strongly correlate with consistent decisions and effectively contribute to increase the user confidence in the neural inference consistency. This makes an auditing activity for human users viable. On one side, it allows to limit the impact of machine mistakes, in a natural and portable manner. Moreover, it can also serve as a novel comparative evaluation paradigm. The reachable auditing accuracy thus measures the explanatory power of different models and can be employed as a comparative benchmark. While the methods proposed in this paper stem just from early explorations, the ways activated landmarks can be made useful to meaningful explanations stimulate further research, involving feature based analysis such as suggested in (Ribeiro et al., 2016) or the application of LRP to architectures more complex than a MLP. Argumentation theory, applied to the active landmark semantics and the source input example as captured by the kernel, provides a very rich framework to design future and more complex justification mechanisms.

References


