On the Readability of Deep Learning Models: the role of Kernel-based Deep Architectures

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

Abstract

English. Deep Neural Networks achieve state-of-the-art performances in several semantic NLP tasks but lack of explanation capabilities as for the limited interpretability of the underlying acquired models. In other words, tracing back causal connections between the linguistic properties of an input instance and the produced classification is not possible. In this paper, we propose to apply *Layerwise Relevance Propagation* over linguistically motivated neural architectures, namely *Kernel-based Deep Architectures* (KDA), to guide argumentations and explanation inferences. In this way, decisions provided by a KDA can be linked to the semantics of input examples, used to linguistically motivate the network output.

Italiano. *Le Deep Neural Network raggiungono oggi lo stato dell'arte in molti processi di NLP, ma la scarsa interpretabilitá dei modelli risultanti dall'addestramento limita la comprensione delle loro inferenze. Non é possibile cioé determinare connessioni causali tra le proprietá linguistiche di un esempio e la classificazione prodotta dalla rete. In questo lavoro, l'applicazione della* Layerwise Relevance Propagation *alle* Kernel-based Deep Architecture*(KDA) é usata per determinare connessioni tra la semantica dell'input e la classe di output che corrispondono a spiegazioni linguistiche e trasparenti della decisione.*

1 Introduction

Deep Neural Networks are usually criticized as they are not epistemologically transparent devices, i.e. their models cannot be used to provide explanations of the resulting inferences. An example can be neural question classification (QC) (e.g. (Croce et al., 2017)). In QC the correct category of a question is detected to optimize the later stages of a question answering system, (Li and Roth, 2006). An epistemologically transparent learning system should trace back the causal connections between the proposed question category and the linguistic properties of the input question. For example, the system could motivate the decision: *"What is the capital of Zimbabwe?" refers to a* `Location`, with a sentence such as: *Since it is similar to "What is the capital of California?"* which also refers to a `Location`. Unfortunately, neural models, as for example Multilayer Perceptrons (MLP), Long Short-Term Memory Networks (LSTM), (Hochreiter and Schmidhuber, 1997), or even Attention-based Networks (Larochelle and Hinton, 2010), correspond to parameters that have no clear conceptual counterpart: it is thus difficult to trace back the network components (e.g. neurons or layers in the resulting topology) responsible for the answer.

In image classification, *Layerwise Relevance Propagation* (LRP) (Bach et al., 2015) has been used to decompose backward across the MLP layers the evidence about the contribution of individual input fragments (i.e. pixels of the input images) to the final decision. Evaluation against the MNIST and ILSVRC benchmarks suggests that LRP activates associations between input and output fragments, thus tracing back meaningful causal connections.

In this paper, we propose the use of a similar mechanism over a linguistically motivated network architecture, the Kernel-based Deep Architecture (KDA), (Croce et al., 2017). Tree Kernels (Collins and Duffy, 2001) are here used to integrate syntactic/semantic information within a MLP network. We will show how KDA input nodes correspond to linguistic instances and by applying the LRP method we are able to trace back causal associations between the semantic classification and such instances. Evaluation of the LRP algorithm is based on the idea that explanations improve the user expectations about the correctness of an answer and shows its applicability in human computer interfaces.

In the rest of the paper, Section 2 describes the KDA neural approach while section 3 illustrates how LRP connects to KDAs. In section 4 early results of the evaluation are reported.

2 Training Neural Networks in Kernel Spaces

Given a training set $o \in D$, a kernel $K(o_i, o_j)$ is a similarity function over D^2 that corresponds to a dot product in the implicit kernel space, i.e., $K(o_i, o_j) = \Phi(o_i) \cdot \Phi(o_j)$. Kernel functions are used by learning algorithms, such as Support Vector Machines (Shawe-Taylor and Cristianini, 2004), to efficiently operate on instances in the kernel space: their advantage is that the projection function $\Phi(o) = \vec{x} \in \mathbb{R}^n$ is never explicitly computed. The Nyström method is a factorization method applied to derive a new low-dimensional embedding $\tilde{x}$ in a l-dimensional space, with $l \ll n$ so that $G \approx \tilde{G} = \tilde{X}\tilde{X}^\top$, where $G = XX^\top$ is the Gram matrix such that $G_{ij} = \Phi(o_i)\Phi(o_j) = K(o_i, o_j)$. The approximation $\tilde{G}$ is obtained using a subset of l columns of the matrix, i.e., a selection of a subset $L \subset D$ of the available examples, called *landmarks*. Given l randomly sampled columns of G, let $C \in \mathbb{R}^{|D| \times l}$ be the matrix of these sampled columns. Then, we can rearrange the columns and rows of G and define $X = [X_1 \;\; X_2]$ such that:

$$G = \begin{bmatrix} W & X_1^\top X_2 \\ X_2^\top X_1 & X_2^\top X_2 \end{bmatrix} = \begin{bmatrix} C \\ X_2^\top X_1 \end{bmatrix}$$

where $W = X_1^\top X_1$, i.e., the subset of G that contains only landmarks. The Nyström approximation can be defined as:

$$G \approx \tilde{G} = CW^\dagger C^\top \tag{1}$$

where $W^\dagger$ denotes the Moore-Penrose inverse of W. If we apply the Singular Value Decomposition (SVD) to W, which is symmetric definite positive, we get $W = USV^\top = USU^\top$. Then it is straightforward to see that $W^\dagger = US^{-1}U^\top = US^{-\frac{1}{2}}S^{-\frac{1}{2}}U^\top$ and that by substitution $G \approx \tilde{G} = (CUS^{-\frac{1}{2}})(CUS^{-\frac{1}{2}})^\top = \tilde{X}\tilde{X}^\top$. Given an example $o \in D$, its new low-dimensional representation $\tilde{\vec{x}}$ is determined by considering the corresponding item of C as

$$\tilde{\vec{x}} = \vec{c}US^{-\frac{1}{2}} \tag{2}$$

where $\vec{c}$ is the vector whose dimensions contain the evaluations of the kernel function between o and each landmark $o_j \in L$. Therefore, the method produces l-dimensional vectors.

Given a labeled dataset, a Multi-Layer Perceptron (MLP) architecture can be defined, with a specific Nyström layer based on the Nyström embeddings of Eq. 2, (Croce et al., 2017).

Such Kernel-based Deep Architecture (KDA) has an *input layer*, a *Nyström layer*, a possibly empty sequence of non-linear *hidden layers* and a final *classification layer*, which produces the output. In particular, the input layer corresponds to the input vector $\vec{c}$, i.e., the row of the C matrix associated to an example o. It is then mapped to the *Nyström* layer, through the projection in Equation 2. 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 $H_{Ny} = US^{-\frac{1}{2}}$: it corresponds to a pretraining stage based on the SVD. Formally, the low-dimensional embedding of an input example o, $\tilde{\vec{x}} = \vec{c}\, H_{Ny} = \vec{c}\, US^{-\frac{1}{2}}$ encodes the kernel space. Any neural network can then be adopted: in the rest of this paper, we assume that a traditional Multi-Layer Perceptron (MLP) architecture is stacked in order to solve the targeted classification problems. The final layer of KDA is the *classification layer* whose dimensionality depends on the classification task: it computes a linear classification function with a softmax operator.

A KDA is stimulated by an input vector c which corresponds to the kernel evaluations $K(o, l_i)$ between each example o and the landmarks l_i. Linguistic kernels (such as Semantic Tree Kernels (Croce et al., 2011)) depend on the syntactic/semantic similarity between the x and the subset of l_i used for the space reconstruction. We will see hereafter how tracing back through relevance propagation into a KDA architecture corresponds to determine which semantic landmarks contribute mostly to the final output decision.

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

Layer-wise Relevance propagation (LRP, presented in (Bach et al., 2015)) is a framework which allows to decompose the prediction of a deep neural network computed over a sample, e.g. an image, down to relevance scores for the single input dimensions, such as a subset of pixels.

Formally, let $f : \mathbb{R}^d \rightarrow \mathbb{R}^+$ be a positive real-valued function taking a vector $\vec{x} \in \mathbb{R}^d$ as input: f quantifies, for example, the probability of $\vec{x}$ characterizing a certain class. The Layer-wise Relevance Propagation assigns to each dimension, or feature, x_d, a relevance score $R_d^{(1)}$ such that:

$$f(x) \approx \sum_d R_d^{(1)} \tag{3}$$

Features whose score $R_d^{(1)} > 0$ (or d $R_d^{(1)} < 0$) correspond to evidence in favor (or against) 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 $l+1$, then it can be decomposed into messages $R_{i \leftarrow j}^{(l,l+1)}$ sent to neurons i in layer l:

$$R_j^{(l+1)} = \sum_{i \in (l)} R_{i \leftarrow j}^{(l,l+1)} \tag{4}$$

Hence the relevance of a neuron i at layer l can be defined as:

$$R_i^{(l)} = \sum_{j \in (l+1)} R_{i \leftarrow j}^{(l,l+1)} \tag{5}$$

Note that 4 and 5 are such that 3 holds. In this work, we adopted the ϵ-rule defined in (Bach et al., 2015) to compute the messages $R_{i \leftarrow j}^{(l,l+1)}$, i.e.

$$R_{i \leftarrow j}^{(l,l+1)} = \frac{z_{ij}}{z_j + \epsilon \cdot \text{sign}(z_j)} R_j^{(l+1)}$$

where $z_{ij} = x_i w_{ij}$ and $\epsilon > 0$ is a numerical stabilizing term and must be small. Notice that weights w_{ij} correspond to weighted activations of input neurons. If we apply LRP to a KDA it implicitly traces the relevance back to the input layer, i.e. to the landmarks. It thus tracks back syntactic, semantic and lexical relations between a question and the landmark and it grants high relevance to the relations the network selected as highly discriminating for the class representations it learned; note that this is different from similarity in terms of kernel-function evaluation as the latter is task independent whereas LRP scores are not. Notice also that each landmark is uniquely associated to an entry of the input vector $\vec{c}$, as shown in Sec 2, and, as a member of the training dataset, it also corresponds to a known class.

4 Explanatory Models

LRP allows the automatic compilation of justifications for the KDA classifications: explanations are possible using landmarks $\{\ell\}$ as examples. The $\{\ell\}$ that the LRP method produces as the most active elements in layer 0 are semantic analogues of input annotated examples. An *Explanatory Model* is the function in charge of compiling the linguistically fluent explanation of individual analogies (or differences) with the input case. The meaningfulness of such analogies makes a resulting explanation clear and should increase the user confidence on the system reliability. When a sentence o is classified, LRP assigns activation scores r_ℓ^s to each individual landmark ℓ: let $L^{(+)}$ (or $L^{(-)}$) denote the set of landmarks with positive (or negative) activation scores.

Formally, an explanation is characterized by a triple $e = \langle s, C, \tau \rangle$ where s is the input sentence, C is the predicted label and τ is the modality of the explanation: $\tau = +1$ for positive (i.e. acceptance) statements while $\tau = -1$ correspond to rejections of the decision C. A landmark ℓ is *positively activated* for a given sentence s if there are not more than $k-1$ other active landmarks[1] ℓ' whose activation value is higher than the one for ℓ, i.e.

$$|\{\ell' \in L^{(+)} : \ell' \neq \ell \wedge r_{\ell'}^s \geq r_\ell^s > 0\}| < k$$

A landmark is *negatively activated* when: $|\{\ell' \in L^{(-)} : \ell' \neq \ell \wedge r_{\ell'}^s \leq r_\ell^s < 0\}| < k$. Positively (or negative) active landmarks in L_k are assigned to an activation value $a(\ell, s) = +1$ (-1). For all other not activated landmarks: $a(\ell, s) = 0$.

Given the explanation $e = \langle s, C, \tau \rangle$, a landmark ℓ whose (known) class is C_ℓ is *consistent* (or *inconsistent*) with e according to the fact that the following function:

$$\delta(C_\ell, C) \cdot a(\ell, q) \cdot \tau$$

is positive (or negative, respectively), where $\delta(C', C) = 2\delta_{kron}(C' = C) - 1$ and δ_{kron} is the Kronecker delta.

The *explanatory model* is then a function $\mathcal{M}(e, L_k)$ which maps an explanation e, a sub set L_k of the active *and* consistent landmarks L for e into a sentence in natural language. Of course several definitions for $\mathcal{M}(e, L_k)$ and L_k are possible.

[1] k is a parameter used to make explanation depending on not more than k landmarks, denoted by L_k.

A general explanatory model would be:

$$\mathcal{M}(e, L_k) = \begin{cases} \text{“ } s \text{ is } C \text{ since it is similar to } \ell \text{ ”} \\ \forall \ell \in L_k^+ \quad \text{if } \tau > 0 \\ \\ \text{“ } s \text{ is not } C \text{ since it is different} \\ \text{from } \ell \text{ which is } C \text{ ”} \\ \forall \ell \in L_k^- \quad \text{if } \tau < 0 \\ \\ \text{“ } s \text{ is } C \text{ but I don't know why ”} \\ \text{if } L_k = \emptyset \end{cases}$$

where L_k^+,$L_k^-$$\subseteq L_k$ are the partitions of landmarks with positive (and negative) relevance scores in L_k, respectively. Here we provide examples for two explanatory models, used during the experimental evaluation. A first possible model returns the analogy only with the (unique) consistent landmark with the highest positive score if $\tau = 1$ and lowest negative when $\tau = -1$. The explanation of a rejected decision in the Argument Classification of a Semantic Role Labeling task (Vanzo et al., 2016), described by the triple $e_1 = \langle$'*vai in camera da letto*', $\text{SOURCE}_{\text{BRINGING}}, -1\rangle$, is:

*I think ”**in camera da letto**” IS NOT* [Source] of [Bringing] *in ”Vai in camera da letto”* (LU:[*vai*]) *since it's different from ”sul tavolino” which is* [Source] of [Bringing] *in “Portami il mio catalogo sul tavolino”* (LU:[*porta*])

The second model uses two active landmarks: one consistent and one contradictory with respect to the decision. For the triple $e_1 = \langle$'*vai in camera da letto*', $\text{GOAL}_{\text{MOTION}}, 1\rangle$ the second model produces:

I think ”**in camera da letto**” IS [Goal] of [Motion] *in ”Vai in camera da letto”* (LU:[*vai*]) *since it recalls **”al telefono”** which is* [Goal] of [Motion] *in ”Vai al telefono e controlla se ci sono messaggi”* (LU:[*vai*]) *and it IS NOT* [Source] of [Bringing] *since different from ”sul tavolino” which is the* [Source] of [Bringing] *in ”Portami il mio catalogo sul tavolino”* (LU:[*portami*])

4.1 Evaluation methodology

In order to evaluate the impact of the produced explanations, we defined the following task: given a classification decision, i.e. the input o is classified as C, to measure the impact of the explanation e on the belief that a user exhibits on the statement “$o \in C$ is `true`”. This information can be modeled through the estimates of the following probabilities: $P(o \in C)$ that characterizes the amount of confidence the user has in accepting the statement, and its corresponding form $P(o \in C|e)$, i.e. the same quantity in the case the user is provided by the explanation e. The core idea is that semantically coherent and exhaustive explanations must indicate correct classifications whereas incoherent or non-existent explanations must hint towards wrong classifications. A quantitative measure of such an increase (or decrease) in confidence is the Information Gain (IG, (Kononenko and Bratko, 1991)) of the decision $o \in C$. Notice that IG measures the increase of probability corresponding to correct decisions, and the reduction of the probability in case the decision is wrong. This amount suitably addresses the shift in uncertainty $-log_2(P(\cdot))$ between two (subjective) estimates, i.e., $P(o \in C)$ vs. $P(o \in C|e)$.

Different explanatory models $\mathcal{M}$ can be also compared. The relative Information Gain $I_{\mathcal{M}}$ is measured against a collection of explanations $e \in T_{\mathcal{M}}$ generated by $\mathcal{M}$ and then normalized throughout the collection's entropy $\mathcal{E}$ as follows:

$$I_{\mathcal{M}} = \frac{1}{\mathcal{E}} \frac{1}{\mid T_{\mathcal{M}} \mid} \sum_{e \in T_{\mathcal{M}}} I(e)$$

where $I(e)$ is the IG of each explanation[2].

5 Experimental Evaluation

The effectiveness of the proposed approach has been measured against two different semantic processing tasks, i.e. Question Classification (QC) over the UIUC dataset (Li and Roth, 2006) and Argument Classification in Semantic Role Labeling (SRL-AC) over the HuRIC dataset (Bastianelli et al., 2014; Vanzo et al., 2016). The adopted architecture consisted in a LRP-integrated KDA with 1 hidden layers and 500 landmarks for QC, 2 hidden layers and 100 landmarks for SRL-AC and a stabilization-term $\epsilon = 10e^{-8}$.

We defined five quality categories and associated each with a value of $P(o \in C|e)$, as shown in Table 1. Three annotators then independently rated explanations generated from a collection composed of an equal number of correct and wrong classifications (for a total amount of 300 and 64 explanations, respectively, for QC and SRL-AC). This perfect balancing makes the prior probability $P(o \in C)$ being 0.5, i.e. maximal entropy with a baseline IG $= 0$ in the $[-1, 1]$ range. Notice that annotators had no information on the

[2] More details are in (Kononenko and Bratko, 1991)

Category	$P(o \in C\|e)$	$1-P(o \in C\|e)$
V.Good	0.95	0.05
Good	0.8	0.2
Weak	0.5	0.5
Bad	0.2	0.8
Incoher.	0.05	0.95

Table 1: Posterior probab. w.r.t. quality categories

Model	QC	SRL-AC
One landmark	0.548	0.669
Two landmarks	0.580	0.784

Table 2: Information gains for two Explanatory Models applied to the QC and SRL-AC datasets.

system classification performance, but just knowledge of the explanation dataset entropy.

5.1 Question Classification

Experimental evaluations[3] showed that both the models were able to gain more than half the bit required to ascertain whether the network statement is true or not (Table 2). Consider:

I think "What year did Oklahoma become a state ?" refers to a NUMBER *since recalls me "The film Jaws was made in what year ?"*

Here the model returned a coherent supporting evidence, a somewhat easy case as for the available discriminative pair, i.e. *"What year"*. The system is able to capture semantic similarities even in poorer conditions, e.g.:

I think "Where is the Mall of the America ?" refers to a LOCATION *since recalls me "What town was the setting for The Music Man ?" which refers to a LOCATION.*

This high quality explanation is achieved even if with such poor lexical overlap. It seems that richer representations are here involved with grammatical and semantic similarity used as the main information involved in the decision at hand. Let us consider:

I think "Mexican pesos are worth what in U.S. dollars ?" refers to a DESCRIPTION *since it recalls me "What is the Bernoulli Principle ?"*

Here the provided explanation is incoherent, as expected since the classification is wrong. Now consider:

I think "What is the sales tax in Minnesota ?" refers to a NUMBER *since it recalls me "What is the population of Mozambique ?" and does not refer to a* ENTITY *since different from "What is a fear of slime ?".*

Although explanation seems fairly coherent, it is actually misleading as ENTITY is the annotated class. This shows how the system may lack of contextual information, as humans do, against inherently ambiguous questions.

5.2 Argument Classification

Evaluation also targeted a second task, that is Argument classification in Semantic Role Labeling (SRL-AC): KDA is here fed with vectors from tree kernel evaluations as discussed in (Croce et al., 2011). The evaluation is carried out over the HuRIC dataset (Vanzo et al., 2016), including about 240 domotic commands in Italian, comprising of about 450 roles. The system has an accuracy of 91.2% on about 90 examples, while the training and development set have a size of, respectively, 270 and 90 examples. We considered 64 explanations for measuring the IG of the two explanation models. Table 2 confirms that both explanatory models performed even better than in QC. This is due to the narrower linguistic domain (14 frames are involved) and the clearer boundaries between classes: annotators seem more sensitive to the explanatory information to assess the network decision. Examples of generated sentences are:

I think **"con me"** *is NOT the* MANNER *of* COTHEME *in "Robot vieni* ***con me*** *nel soggiorno*? (LU:[*vieni*])" *since it does NOT recall me* **"lentamente"** *which is* MANNER *in "Per favore segui quella persona* ***lentamente*** (LU:[*segui*])". *It is rather* COTHEME *of* COTHEME *since it recalls me* **"mi"** which is COTHEME *in "Segui***mi** *nel bagno* (LU:[*segui*])".

6 Conclusion and Future Works

This paper describes an LRP application to a KDA that makes use of analogies as explanations of a neural network decision. A methodology to measure the explanation quality has been also proposed and the experimental evidence confirms the effectiveness of the method in increasing the trust of a user upon automatic classifications. Future work will focus on the selection of subtrees as meaningful evidences for the explanation, or on the modeling of negative information for disambiguation as well as on more in depth investigation of the landmark selection policies. Moreover, improved experimental scenarios involving users and dialogues will be also designed, e.g. involving further investigation within Semantic Role Labeling, using the method proposed in (Croce et al., 2012).

[3] For details on KDA performance against the task, see (Croce et al., 2017)

References

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

Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, Luca Iocchi, Roberto Basili, and Daniele Nardi. 2014. Huric: a human robot interaction corpus. In *LREC*, pages 4519–4526. European Language Resources Association (ELRA).

Michael Collins and Nigel Duffy. 2001. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02), July 7-12, 2002, Philadelphia, PA, USA*, pages 263–270. Association for Computational Linguistics, Morristown, NJ, USA.

Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2011. Structured lexical similarity via convolution kernels on dependency trees. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1034–1046. Association for Computational Linguistics.

Danilo Croce, Alessandro Moschitti, Roberto Basili, and Martha Palmer. 2012. Verb classification using distributional similarity in syntactic and semantic structures. In *ACL (1)*, pages 263–272. The Association for Computer Linguistics.

Danilo Croce, Simone Filice, Giuseppe Castellucci, and Roberto Basili. 2017. Deep learning in semantic kernel spaces. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 345–354, Vancouver, Canada, July. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Igor Kononenko and Ivan Bratko. 1991. Information-based evaluation criterion for classifier's performance. *Machine Learning*, 6(1):67–80, Jan.

Hugo Larochelle and Geoffrey E. Hinton. 2010. Learning to combine foveal glimpses with a third-order boltzmann machine. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 1243–1251.

Xin Li and Dan Roth. 2006. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(3):229–249.

John Shawe-Taylor and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK.

Andrea Vanzo, Danilo Croce, Roberto Basili, and Daniele Nardi. 2016. Context-aware spoken language understanding for human robot interaction. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*