On the Impact of Linguistic Information in Kernel-based Deep Architectures

Danilo Croce, Simone Filice and Roberto Basili
{croce,filice,basili}@info.uniroma2.it

ABSTRACT: *Kernel methods enable the direct usage of structured representations of textual data during language learning and inference tasks. On the other side, deep neural networks are effective in learning non-linear decision functions. Recent works demonstrated that expressive kernels and deep neural networks can be combined in a Kernel-based Deep Architecture (KDA), a common framework that allows to explicitly model structured information into a neural network. This combination achieves state-of-the-art accuracy in different semantic inference tasks. This paper investigates the impact of linguistic information on the performance reachable by a KDA by studying the benefits that different kernels can bring to the inference quality. We believe that the expressiveness of data representations will play a key role in the wide spread adoption of neural networks in AI problem solving. We experimentally evaluated the adoption of different kernels (each characterized by a growing expressive power) in a Question Classification task. Results suggest the importance of rich kernel functions in optimizing the accuracy of a KDA.*

Kernel methods

- ✓ State-of-the-art results in many NLP tasks
- ✓ Allow decoupling the learning algorithm from the representations (e.g., reflecting explicitly linguistic structures)
- ✓ Operate directly on complex and discrete structures (e.g., trees or graphs)

- X Scalability issues in the learning phase, i.e., complexity is almost $O(n^2)$
- X Scalability issues in the classification phase, i.e., complexity is O(#SV)

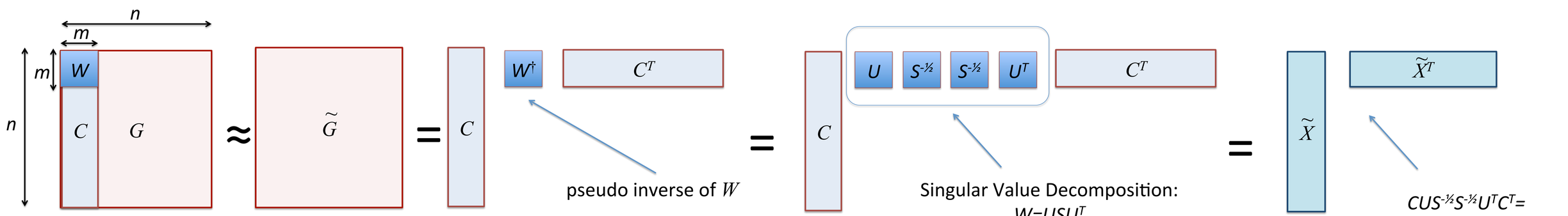
Deep Neural Networks

- ✓ State-of-the-art results in many NLP tasks
- ✓ Highly accurate against non-linear phenomena
- ✓ Highly scalable both in training and classification
- ✓ Enable representation learning

- X Standard architectures require inputs to be modeled via vectors or tensors
- X No common design practice to treat different complex data structures

Nystrom method to scale up Kernel machines

Scope: given a kernel *K* and a dataset of *n* instances, it creates a projection function that embed the *n* instances into a $m << n$ dimensional vector space so that the dot product in this space is the best approximation $\widetilde{G}$ of the Gram matrix *G* associated to the kernel *K* .



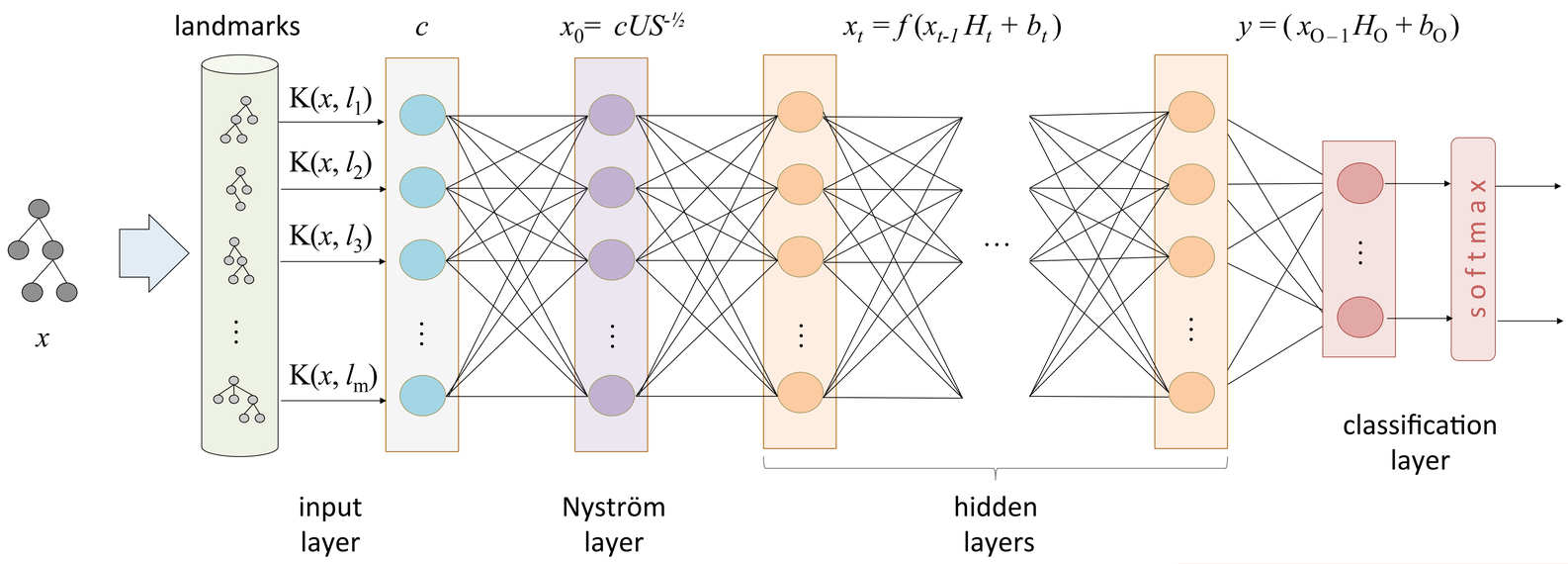
Original Gram matrix (*m* random landmarks in the first positions)

Approximated Gram matrix

Each row of $\widetilde{X}$ is a m-dimensional embedding → we can project a new instance o by computing $cUS^{-½}$

A joint learning model: Kernel-based Deep Architecture

- A general approach to combine the above paradigms based on the *Nystrom method*
  - It approximates the Gram Matrix generated by *any* kernel function
  - It enables the *projection* of input examples into dense *low-dimensional spaces*
- A KDA is a mathematically justified solution *for enabling Kernel-based Learning through Neural Networks*
  - the Nystrom based embeddings are used as input of (deep) neural networks where …
  - … standard back-propagation can be applied
  - The NN can learn non-linear functions in Kernel Spaces!



Impact of different Linguistic Information in the Kernel-based Deep Architecture

In Statistical Language Learning, several kernels operating on linguistic structures have been largely investigated:

- **Tree Kernels**, such as the **Partial Tree Kernel** (Moschitti, 2006), capture structural analogies directly from syntactic parse trees
- **Smoothed Partial Tree Kernels** (**SPTKs**) (Croce et al., 2011) exploit Distributional Semantic Models (DMs) for a "Semantic Smoothing" over the recursive tree kernel function.
- **Compositionally Smoothed Partial Tree Kernel** (**CSPTK**) (Annesi et ai., 2014) considering Semantic Compositionality over the tree structure

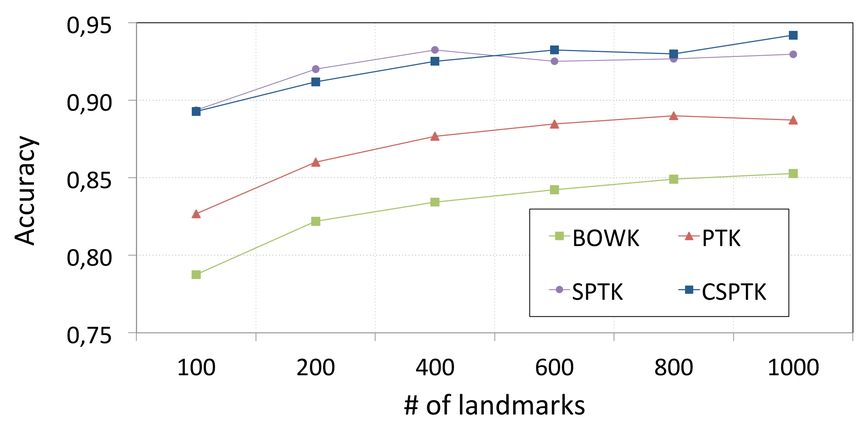
The major hypothesis underlying a KDA is the ability of the Nystrom reconstruction to capture the semantics of the input linguistic data expressed in the kernel feature space to guide the inductive inference.

Are increasingly expressive representations able to achieve better generalization?

Is the improvement in performance invariant with respect to one given KDA?

Is the kernel representation correlated with the accuracy reachable by a given KDA?

- Question classification consists in assigning a question to class reflecting its intention
  - "*What is the width of a football field*?" → `Number`
- The CSPTK achieves state-of-the-arte results by directly operating over the syntactic parte tree (Annesi et at, 2014)
- QC Dataset (6 classes)
  - 5,452 training examples, 500 test examples



D. Croce, S. Filice, G. Castellucci and R. Basili.
Deep Learning in Semantic Kernel Spaces, ACL 2017.