Deep learning languages: a key fundamental shift from probabilities to weights?

François Coste

Deep learning languages: a key fundamental shift from probabilities to weights?

François Coste
Univ Rennes, Inria, CNRS, IRISA / F-35000 Rennes
francois.coste@inria.fr

Abstract

Recent successes in language modeling, notably with deep learning methods, coincide with a shift from probabilistic to weighted representations. We raise here the question of the importance of this evolution, in the light of the practical limitations of a classical and simple probabilistic modeling approach for the classification of protein sequences and in relation to the need for principled methods to learn non-probabilistic models.

1 Introduction

Probabilistic models have been extremely important for successful language processing, from the first accomplishments in natural language processing (NLP), with n-grams and their successors (Russell and Norvig, 2010, chapters 22 and 23), to the tools routinely used today for the annotation of biological sequences in bioinformatics, based on the profile hidden Markov models (pHMM) or covariance models (CM) (Durbin et al., 1998; Coste, 2016). Yet, after its success in computer vision and pattern recognition, deep learning is now replacing probabilistic models to become the new key player in the NLP field, showing that neural networks based on dense vector representations are very powerful for performing a large variety of NLP tasks. Impact of deep learning is until now limited in biology, but promising advances are being made (see for instance Ching et al. (2018), notably section 3 for the applications to biological sequences) and we can expect progress in NLP to be beneficial for biological sequence analysis once again.

While the main appealing feature of neural networks was initially *non-linearity*, the historical "S"-shape activation functions (such as the $\tanh$ function) are more and more replaced by the Rectified Linear Unit (ReLU) function, reducing non-linearity to its simplest form: a function returning 0 or a weighted linear combination of its input values. The most important feature of deep learning in NLP seems now to be its ability to learn intermediate representations —such as the word vector representations (word embeddings) learned by word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), or the recent deeper contextual representations learned by ELMo (Peters et al., 2018), ULMFiT (Howard and Ruder, 2018), and GPT (Radford et al., 2018a,b)— in a hierarchical manner up to the last layers of the neural networks in charge of their weighted combination for solving the task at hand. Using weighted combination of weight vectors (or matrices) optimized by gradient descent appears to be more and more the core of deep learning, as witnessed for instance by the growing importance of tensor-based libraries and of automatic differentiation (Baydin et al., 2017).

Hierarchical representations are common in formal grammars and gradient descent is not an original approach. The key advantage of deep learning with respect to the inference of classical grammatical representations could come from using and combining *weights* rather than *probabilities*. We think that better understanding and evaluating the importance of this difference is essential for future research in this area. As a first contribution, we present here practical problems illustrating the fundamental limitations of current probabilistic modeling approaches for language learning. The question we raise is whether they could be properly handled with more elaborate probabilistic modeling, or whether they demonstrate the interest of the deep "Probxit" towards weighted models.

2 Modeling with probabilistic grammars

To cope with inherent "noise", "variations" and "errors", languages have long been modelled with probabilistic models. In practice, these models rarely exceed the expressiveness of probabilistic context-free grammars, or even probabilistic automata. We introduce briefly the main related definitions and notations following Vidal et al. (2005a,b) to which we refer the reader for a more detailed presentation of the different probabilistic grammatical models of languages.

A probabilistic context-free grammar (PCFG) G is defined as a tuple $\langle Q, \Sigma, S, R, P \rangle$ where Q is a set of non-terminal symbols, Σ is a finite alphabet, $S \in Q$ is the initial symbol, R is a set of rules $A \to \alpha$ with $\alpha \in (Q \cup \Sigma)^*$ and $P : R \mapsto [0, 1]$ define the probabilities of rules for each $A \in Q$, under the constraint:

$$\forall A \in Q: \sum_{(A \to \alpha) \in R} P(A \to \alpha) = 1. \quad (1)$$

A successful (leftmost) derivation of a sequence x from S by a succession of rules $r_1 \dots r_l \in R^l$ is denoted by $S \overset{r_1 \dots r_l}{\Rightarrow} x$ and its probability is $P(r_1 \dots r_l) = \prod_{i=1}^{l} P(r_i)$. If a sequence x can be successfully derived from S, its probability is $P(x) = \sum_{r \in R^+ : S \overset{r}{\Rightarrow} x} P(r)$, otherwise its probability is 0. Only consistent PCFGs, i.e. satisfying:

$$\sum_{x \in \Sigma^*} P(x) = 1 \quad (2)$$

defining thus a probability distribution on Σ^*, are classically considered.

Probabilistic regular grammars, more often depicted as probabilistic automata (PA), are PCFGs with simpler rules of the form $A \to aB$, $A \to a$, or $A \to \lambda$ with $A, B \in Q$, $a \in \Sigma$ and λ denoting the empty sequence.

To give intuition, we will illustrate some limitations of purely probabilistic approaches on a practical task: modeling a family of protein sequences. This task is classically achieved with pHMMs, which are discrete HMMs with a predefined left-to-right topology. Illustration will focus on the core of this approach, simplified and transposed to PA to remain in the grammatical formalism. As an example of a family of protein sequences, we consider the SH3 domain, a region fairly well conserved among several protein sequences that is known to be important for binding and interacting with other proteins. This example has been realistically covered in the pHMM tutorial by Krogh (1998). We will simply use here a toy "profile PA" built from an excerpt of the alignment of the SH3 regions from some protein sequences, shown in figure 1. The choice of this example was obviously driven by our own domain of expertise, but also because it enables us to illustrate modeling problems with respect to "objective" physico-chemical considerations. The problems presented in the following sections should still be generic enough to be easily transposed to other applications and methods using probabilistic sequence models.

3 Limitations

We present here a list of practical limitations of classical probabilistic language modeling that should not arise when using weighted models. This does not mean that more elaborate probabilistic approaches could not solve these issues, nor that the best way to learn weighted models overcoming these limitations is known. We see these problems as critical testbeds for studying and comparing both approaches in order to better measure the contribution of the shift from probabilities to weights in the success of deep learning. Ideally, we would like these examples to also help developing principled approaches for learning weighted grammatical models, such as those available for probabilistic models. A pragmatic first step in this direction could be the elaboration of well-founded schemes for parameter estimation of "profile weighted automata" outperforming the simple maximum likelihood estimation of pHMMs' probabilities (or, better, the sophisticated Bayesian inference methods used to boost their performances by the addition of pertinent pseudo-counts from Dirichlet mixture priors, as initiated for instance by (Sjölander et al., 1996)).

L1: Probability scattering at each non-deterministic derivation step Comparing the probabilities given by probabilistic grammars to sequences of different lengths is a well-know practical problem. The profile PA in figure 1 will assign smaller probabilities to longer proteins even if they have exactly the same SH3 region, because of the loss of probability in self-loop transitions consuming the rest of the protein sequence. This problem is tackled in bioinformatics by looking at the ratio between the probability of the sequence and its probability according to a null model: these

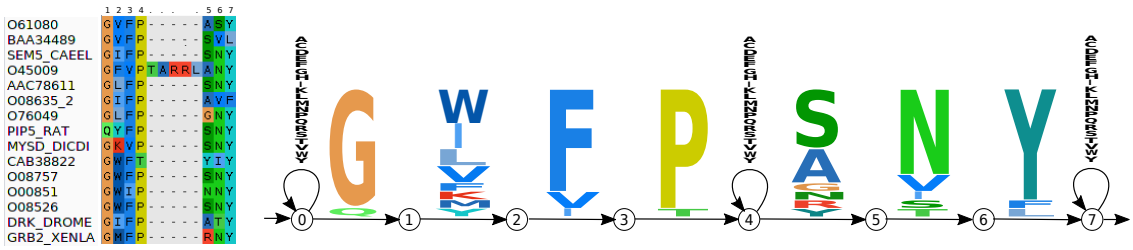

Figure 1: pHMM-like probabilistic automata (right) from a toy alignment of 15 SH3 sequences with 7 conserved columns (left). Height of amino-acid letters is proportional to the probability of the transition labeled by the letter. Here, probability of amino-acids in left-to-right transitions is their observed frequency in the modelled column while they are assumed equiprobable in the self-loop transitions introduced at beginning and end of the PA, to parse complete proteins, and in its middle, to allow insertions between columns 4 and 5 (as in protein O45009).

ratios are usually directly integrated in the profiles by associating log-odds scores to transitions rather than probabilities. Profile are thus already used under the form of weighted models which combine two probabilistic models.

More generally for PCFGs, equation 1 implies a probability loss by a factor strictly smaller than one, at each derivation step involving a non-deterministic rule. Such derivations scatter more and more probabilities between alternative suffix languages, resulting in negligible probabilities for longer derivations. Moreover, they induce a dependence of probability values to the number of non-deterministic derivations, which has often to be corrected or compensated in practical tasks, while this could be avoided by not satisfying equation 1.

L2: Dependence to number of choices for mass probability distribution within alternative rules Another consequence of equation 1 is that the probability mass has to be shared between all alternative rules. Its amount by rule depends then from the number of alternatives. For the characterization of SH3 regions, figure 1 shows the importance of having the amino-acid `G` at the first position. But when we look at the second conserved column, we can see that it contains only hydrophobic amino-acids and it might be as important for SH3 function to have a hydrophobic residue at the second position as a `G` at the first position. This cannot be modeled with PCFGs using the amino-acid alphabet: assuming that there are 14 hydrophobic acids and that they are equiprobable, the probability of transition with these amino-acids will be at most $\frac{1}{14}$ making a small difference with the completely equiprobable $\frac{1}{20}$ null model, in contrast to $\frac{14}{15}$, or any value close to 1, that can be assigned at the first position to a transition with `G`. We could imagine working on a larger alphabets, such as the amino-acids powerset, but the exponential growth of the alphabet size would then complicate training. Considering weights (e.g. reflecting adequacy of particular amino-acids at position, instead of occurrence) could be a more efficient solution in practice.

L3: Identical probability mass for all choice points Because of the normalisation to 1 in equation 1, all choice points have the same total weight in final probability. Yet, some choices can be more important than others. For instance, prior expert knowledge could tell us that amino acid `P` in the fourth position is more important for the function of SH3 than amino-acid `G` in the first position, but this could not be modelled with a PCFG. Authorizing a normalisation to variable values (e.g. reflecting relevance of position modelled with a particular non-terminal) would help here. Note that it could also be an indirect way to overcome limitation L2.

4 Towards language modeling Probxit?

Analysis of limitations Limitations seen above are induced by equation 1 which enables proper PCFGs to satisfy equation 2 and thus define a probability distribution over all sequences. While discarding the constraint of equation 1 is tempting, this constraint is intrinsically needed to distribute the remaining probability mass between all sequences that can be generated after each choice point, and it makes perfect sense to model the probability of sequence occurrence in the language with respect to the uncertainty brought by the non-deterministic rules.

Actually, what is illustrated here is more a misused feature of probabilistic grammars rather than a real limitation of these models: probability of

sequences under constraint of equation 1 specifies the likelihood of their occurrence inside the language, not their probability of being in the language. Considering the SH3 example, probabilities tell us how often we can expect to see a given SH3 sequence comparatively to other SH3 sequences, not to estimate the uncertainty that the sequence can or not perform the SH3 function. If we are interested in membership to language, it is clearly given by the non-probabilistic part of the grammar: a sequence w is in the language L if it can be successfully derived by rules which have non-zero probability. If we were interested in the uncertainty of membership, we should consider learning a probability $P(x \in L)$, with $P(x \in L) + P(x \notin L) = 1$ instead of equation 2. Likelihood of occurrence in the language can thus be considered at most as an imperfect proxy for membership prediction.

Research directions The first research direction to overcome the enumerated limitations is to work on learning the topology of the grammars. Note that this could be completed by learning also a membership probability, but the meaning of this uncertainty, and how it could be learned, would have to be clarified. It could be a measure of the uncertainty of decision with respect to current knowledge (e.g. based on the number of available examples supporting the decision). In the example of SH3, we can also imagine, for instance, that the binding affinity of different SH3 regions could influence their probability of binding to another protein and to perform their function: this could be formalized as a kind of membership probability, but would require information about the probability of performing the function in the training set. We will not consider these extensions here and will continue to focus on membership problem.

Learning a good grammar topology can be hard. In contrast, adding probabilities (and a recognition threshold) to simpler topologies has been shown successful to improve sufficiently their expressiveness for many classification tasks, while maintaining the number of parameters low enough to be estimable from training sets. This strategy has been shown efficient in practice, despite the limitations that we have identified. A second research direction is to study if more elaborate probabilistic approaches could be used more successfully here. The third research direction is to get rid of the constraint of equation 1, which is not really needed for membership prediction, to learn this way weights instead of probabilities of grammar rules.

In this third approach, the weights of sequences could be divided by a partition function (if it exists) to be considered as probabilities taking values in the interval $[0, 1]$ and satisfy equation 2. Besides the problem of the existence and the computation cost of such a partition function, we think that the constraint of equation 2 has also to be discarded, since it restricts the expressiveness of the approach. Indeed, (positive) weights are useful if a threshold greater than zero is chosen so as sequences accepted in the language are those whose weight is above this threshold (otherwise, successful derivation by transitions with non-zero weight is sufficient). By transforming weights into probabilities, the choice of a threshold t would then only enable to define finite languages (with at most $\frac{1}{t}$ sequences since total probability mass is 1), restricting thus unnecessarily the languages that can be represented.

From probabilities to weights Getting rid of both equations means shifting from probabilistic to weighted representations. This move is visible in research on learning weighted grammars (see for instance contrastive estimation from Smith and Eisner (2005), still motivated by mass distribution constraint) and can be considered an intrinsic feature of deep learning approaches. While practical successes show the interest of these approaches, better understanding their fundamental contribution and being able to develop principled approaches for learning weighted grammars, such as those developed for probabilistic grammars, remains a challenge. A key we identified is not (mis)using occurrence probabilities anymore and we exhibited some advantages of using weighted rather than probabilistic models, but theoretical and practical studies are still needed. The problem of training weighted, instead of probabilistic, profile automata on proteins seems a good support for these studies since it is simple and enables objective evaluation of the approaches. To address these questions, we think that we will need first to find the equivalent for weighted grammars of maximum likelihood or Bayesian approaches (especially for the incorporation of prior knowledge, an essential feature of using pHMMs on proteins) and the nice compositional properties of probabilistic approaches, and would really like to discuss this during the workshop.

Acknowledgments

I am grateful to Hugo Talibart and Witold Dyrka for fruitful discussions that motivated this work submitted as a four pages opinion paper to ACL 2019 Workshop "Deep Learning and Formal Languages: Building Bridges".

References

Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. 2017. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18:153:1–153:43.

Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E. Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M. Cofer, Christopher A. Lavender, Srinivas C. Turaga, Amr M. Alexandari, Zhiyong Lu, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K. Wiley, Marwin H. S. Segler, Simina M. Boca, S. Joshua Swamidass, Austin Huang, Anthony Gitter, and Casey S. Greene. 2018. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141).

François Coste. 2016. *Learning the Language of Biological Sequences*, pages 215–247. Springer Berlin Heidelberg, Berlin, Heidelberg.

Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *ACL*. Association for Computational Linguistics.

Anders Stærmose Krogh. 1998. An introduction to hidden markov models for biological sequences. In *Computational methods in molecular biology*. Elsevier.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018a. Improving language understanding by generative pre-training. Technical report.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018b. Language models are unsupervised multitask learners. Technical report.

Stuart J. Russell and Peter Norvig. 2010. *Artificial Intelligence - A Modern Approach (3. internat. ed.)*. Pearson Education.

Kimmen Sjölander, Kevin Karplus, Michael Brown, Richard Hughey, Anders Krogh, I. Saira Mian, and David Haussler. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Computer Applications in the Biosciences*, 12(4):327–345.

Noah A. Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 354–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Enrique Vidal, Franck Thollard, Colin de la Higuera, Francisco Casacuberta, and Rafael C. Carrasco. 2005a. Probabilistic finite-state machines-part i. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(7):1013–1025.

Enrique Vidal, Frank Thollard, Colin de la Higuera, Francisco Casacuberta, and Rafael C. Carrasco. 2005b. Probabilistic finite-state machines-part ii. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(7):1026–1039.