Vannes, France
The 6th of June, 2020

Nicolas Courty
Full Professor
University of Bretagne Sud / IRISA
Campus de Tohannic 56000 Vannes, France
ncourty@irisa.fr
tél. + 33 (0)2 97 01 72 13

IRISA UBS

Report on the PhD Thesis of Warith Harchaoui entitled
'Réseaux de neurons et transport optimal pour l'apprentissage de représentations'

Warith Harchaoui PhD document is about *representation* learning with deep neural networks and optimal transport. The problem of learning representations is ubiquitous in modern statistical learning, and encompasses several traditional learning problems, such as unsupervised/supervised learning, feature selection or predictions under uncertainty. The objectives of this PhD are to revisit those themes through the prism of deep learning, for its inherent capacity to model complex functions as compositions of simple functions, and optimal transport for its principled ways of comparing probability distributions. The contributions of this thesis are new methods and algorithms for clustering, feature selection and associating predictions with an uncertainty measure. The manuscript is written in English, and is organized in 5 Chapters.

The first chapter presents the global background of the thesis. It introduces the fundamental concepts and notations needed to apprehend the rest of the document, as well as state-of-the-art references over which the candidate builds his contributions. It starts with a very general introduction to machine learning and the related problems. Deep neural networks are then presented, and optimal transport. A discussion on the importance of representations in high dimensional contexts follows, before an outline of the different contributions of the thesis. The general introduction to machine learning problems is pleasant to read. I was impressed (and this comment is valid for the rest of the document) by the very wide spectrum of references, that shows the candidate strong interest for the field, and his good knowledge of the related issues, as well as the state-of-the-art in this fast-moving area of research. Coming to optimal transport, which is one the core aspect of the thesis, I believe a more formal and complete introduction could have been given, while the focus is mostly given to the Kantorovich-Rubinstein duality and its applications to generative adversarial networks through the work of Arjovsky and colleagues. Maybe this part could have been enhanced by showing other aspects of applications of optimal transport in the field of machine learning. Regarding the presentation of the thesis contributions, the candidate mixes the general statement of his contributions with some elements of state-of-the-art on the corresponding problems. This part is hard to read, and my opinion misses the point of separating the contributions of the thesis with a general presentation of the research thematic. Some paragraphs are disconnected with the rest of the introduction (e.g. presenting again optimal transport in section 1.6.1) and I feel a better organization for the content of this part could have been proposed, either by dispatching its content in the related section and/or focusing more specifically on the contributions of the thesis.

Chapter 2 proposes the core contribution of the thesis: novel clustering techniques based on Wasserstein distances and deep generative models. The first algorithm, called GeWaC, for Generative Wasserstein Clustering, is based on the principle that samples can be efficiently clustered as a mixture of Gaussians in some latent space. This latent space is attained through a non-linear operator (an encoder), that admits an inverse operation (the decoder) that maps this latent space to the original input space. Interestingly, the candidate considers a Wasserstein distance to enforce that the decoded distribution looks alike the original sample distributions, that the distribution in the latent space is similar to a mixture of Gaussians. An algorithm to solve for the corresponding optimization problem is proposed, based on the dual formulation of the Wasserstein distance. For this method, some qualitative results are presented on the MNIST dataset, and quantitative results on four datasets, against which the proposed method works remarkably. The idea of imposing a mixture of Gaussians structure inside the latent space is interesting, and has also been investigated in the literature (such as *Gaussian mixture models with Wasserstein distance* by Gaujac and colleagues, or the *Wasserstein Auto-encoders* by Tolstikhin et al., for its use of the reparametrization trick). Maybe those connections with existing papers could have been more carefully examined, to better assess the originality of this work.
The second algorithm, named DiWaC for 'Discriminative Wasserstein Clustering', pushes further the idea by imposing that clusters are far apart in the sense of the Wasserstein distance (hence the term 'discriminative'). The proposed method is sound, and shows a good mastery of the underlying concepts. Again, the empirical results are good, and compare favorably with the selected state-of-the-art competitors. My only concern relates to the model selection part. If there are held-out data samples (with the associated correct labels), as exposed in Equation 56, why not using them in a semi-supervised learning scenario? I guess that running such a model selection is unrealistic in real world cases, where one has no access to labeled data. Maybe a sensitivity analysis related to the impact of this model selection could better expose the qualities of the proposed method. Yet and overall, this Chapter proposes valuable contributions to the domain of coupling machine learning and optimal transport.

Chapter 3 presents a contribution in the direction of feature selection, and can be related to the notion of disentanglement, where one wants to characterize the meaningful part of the data, in a possibly unsupervised way. The candidate proposes a novel approach: Inwamadi, for 'Infinitesimal Wasserstein Maximal Distortion'. As discussed in the introduction, this part of the manuscript gathers mostly some preliminary discussion on this topic, and does not conclude on the practical utility of the proposed algorithm. The subject is interesting per se, as studying/learning the metric in optimal transport problems is a very active topic of research; but the treatment here lacks rigor, and it is difficult to understand exactly what the candidate is trying to achieve, or more precisely what is the optimization problem under examination. For example, the output of Algorithm 5 is unclear, so as how it could be used in real world problem. It remains that some of the presented ideas are interesting, and further investigations could reveal parallels with Sobolev norm regularization for instance.

Chapter 4 deals with the problem of prediction with uncertainty and constitutes the last contribution of the PhD document. After re-establishing existing links between classical losses in deep neural networks, likelihood maximization and minimization of statistical divergences such as Kullback-Leibler divergence, Warith proposes to revisit the classical empirical risk minimization principle by substituting the output predicted scalar value by an expected value over a smooth function within an interpretable class of functions. As clearly stated by the candidate, the goal is not to improve the quality of prediction or classification of the neural network, but rather to augment it with a capacity to provide how much the prediction is reliable or uncertain. This idea is nice and sound. I really believe it has some connections with robust optimization, where one wants to enforce that the prediction is consistent in the vicinity of a sample (usually within a norm ball around the sample). This last topic has notably inspired adversarial regularization techniques, that could have been discussed further, but I acknowledge that the goals are different in the sense that the method is not tailored to enhance the generalization capacity of the network. Among other existing related works, the recent MixUp framework acts similarly as Eq. 121, again with different objectives. Results are presented over a toy dataset and a real-world example consisting of a classification of dogs versus wolves images. Results are interesting and show promising use of the technique in a context of interpretable/explainable AI, though I can regret here that no comparisons with respect to the existing state-of-the-art methods were performed, that could have been useful to publish this work in a machine learning venue.

Chapter 5 ends the document with a short conclusion. It is followed by a technical annex related to (pre-)training of WGANs, which is valuable and highlights the technical difficulties associated with learning such neural networks.

Summary of the analysis:
Overall, the manuscript shows the good knowledges of the domain challenges acquired by the candidate during his PhD. The spectrum of techniques used is large, and some of the contributions are valuable. The main works of the thesis have been published through communications in French venues, such as Statlearn or Data Science summer schools, and would have deserved to be published in good machine learning conferences, provided that more thorough analysis would have been conducted with respect to existing related works.

To summarize, I consider that Warith Harchaoui has made valuable contributions to the field of machine learning, and I agree that his thesis ought to be defended.

Nicolas Courty
Full Professor in Computer Science
University of Bretagne Sud