Deep Learning vs Spectral Clustering into an active clustering with pairwise constraints propagation

Nicolas Voiron[1], Alexandre Benoit[1], Patrick Lambert[1] and Bogdan Ionescu[2]
[1]LISTIC, Universit Savoie Mont Blanc, 74940, Annecy le Vieux, France
{*nicolas.voiron, alexandre.benoit, patrick.lambert*}*@univ-smb.fr*
[2]LAPI, University Politehnica of Bucharest, 061071, Bucharest, Romania
bionescu@alpha.imag.pub.ro

***Abstract*—In our data driven world, categorization is of major importance to help end-users and decision makers understanding information structures. Supervised learning techniques rely on annotated samples that are often difficult to obtain and training often overfits. On the other hand, unsupervised clustering techniques study the structure of the data without disposing of any training data. Given the difficulty of the task, supervised learning often outperforms unsupervised learning. A compromise is to use a partial knowledge, selected in a smart way, in order to boost performance while minimizing learning costs, what is called semi-supervised learning. In such use case, Spectral Clustering proved to be an efficient method. Also, Deep Learning outperformed several state of the art classification approaches and is interesting in this context. In this paper, we firstly introduce the concept of Deep Learning into an active semi-supervised clustering process and compare it with Spectral Clustering. Secondly, we introduce constraint propagation and demonstrate how it maximizes partitioning quality while reducing annotation costs. Experimental validation is conducted on two different real datasets. Results show the potential of the clustering methods.**

I. Introduction

In our anywhere anytime connected world, the amount of available multimedia information explodes. One has to rely now on automatic tools to index and categorize these huge amount of data in order to provide users with efficient searching and browsing capabilities. In this context, our work is focused on automatic categorization which is a critical point for enabling the management of large databases. In challenging situations, clustering, as a non-supervised approach, generally provides unsatisfactory or even inappropriate results. Classification may solve this problem by using a fully annotated data subset (training dataset). But this labelling requires costly human expertise. It is therefore interesting to consider an intermediate approach that uses only a partial knowledge i.e., the semi-supervised techniques.

Jain [1] describes two main types of partial knowledge:

- the partially labelled knowledge given by absolute class annotations only known on a subset of the whole training set. Furthermore, unlabelled data is used in the classification process;
- the partially constrained knowledge that provides similarity pairs annotations between multimedia objects. Commonly known as "Must Link" and "Cannot Link", it simply indicates if two objects belong or not to the same class.

Regarding last option, such pairwise constraints are generic enough and can be provided via external knowledge, e.g., user input, user studies, etc. They can also be deduced from the absolute class annotations. Furthermore, these are actually similarity annotations that are easier to obtain compared to an absolute class annotation.

In an online interactive process, semi-supervised clustering can be enrolled into an iterative process in order to become interactive. At each iteration, a supervision provides some knowledge. In the case of a similarity based one, supervision provides "Must Link" and "Cannot Link" constraints. Furthermore, if a pair selection strategy is taken into account, semi-supervised clustering is made active. In such a framework, state of the art approaches propose different semi-supervised clustering methods to be compared. Selection strategies can be random as in [2] or focus on specific pairs as in [3].

In view of this idea, it is of major interest to optimize the constraints (i.e., annotated pairs) thus to maximize clustering quality while minimizing the costs of user knowledge acquisition. One of the most common strategies consists in using a pairwise constraint automatic propagation approach [4]. This will be one of the two aspects of this paper.

Clustering literature is rich [5], [6] and encompass classical convex data clustering such as the simple k-means algorithm to more complex approaches such as mixture-resolving, mode-seeking approaches or artificial neural networks that are able to cope with more difficult cluster representations.

One particular category of clustering relies on Spectral Clustering Graph Cut techniques [7], that belong to manifold learning. Such methods are preferred when dealing with non-convex data clusters. However, standard Spectral Clustering remains unsupervised and cannot benefit from external user knowledge. Recent advances [3], [2] have shown the benefits of introducing pairwise constraints to guide the clustering procedure and provide some robust semi-supervised Spectral Clustering algorithms. Such approaches gets closer to classical supervised techniques such as Support Vector Machines (SVM) but manage cheaper annotations (similarities) than absolute class names.

978-1-4673-8695-1/16/$31.00 © 2016 IEEE

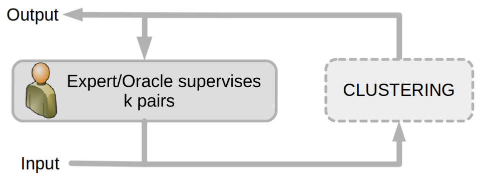

Fig. 1. Interactive semi-supervised clustering.

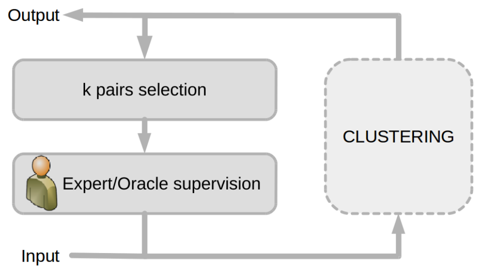

Fig. 2. Active semi-supervised clustering.

Another way to perform clustering is to rely on the recently popularized deep learning approaches. More specifically, Chopra et al [8] showed that it is possible to use stacks of parallel neuron layers trained with sample pairs and similarity ground truth. This allows a similarity metric to be learned and inputs to be projected in a lower dimensional space. In the projected space a simple clustering step such as k-means allows the different classes to be identified. Such solution naturally competes with the Spectral Clustering process in order to address large scale clustering scenarios. In parallel, in an online training use case, Bengio et al [9] showed that it is possible to train a neural model only relying on the last observed examples and still optimizing the generalization error. Then, when dealing with large scale experiments involving similarities, online Deep Learning approaches relying on a siamese architecture sounds appealing.

Then, the second aspect of this paper will aim at comparing Deep Learning and Spectral Clustering into an active semi-supervised clustering schema in a bi-class context. We also get an overview of the benefits provided by constraint propagation for the two clustering methods. Experiments are conducted on two datasets. The first one is provided by CMLIS from the University of California at Irvine (UCI). The second one comes from a video genre classification dataset [10].

The rest of the paper is organized as follows. Section II presents current state of the art on semi-supervised clustering, pairwise constraint propagation, Deep Learning and Spectral Clustering. Section III present the architecture of our active clustering process and how clustering methods are involved. Section IV presents experimental results while Section V concludes and discusses future work.

II. State of the Art

A. Semi-supervised clustering framework

Semi-supervised clustering, potentially being part of an iterative process, may be used in an interactive way. As shown in figure 1, semi-supervision can be introduced by alternating clustering and expert feedback retrieval. An expert (sometimes called Oracle) usually has to observe object pairs and has to assess one of the binary similarity labels, “Must Link” (ML, same classe) or “Cannot Link” (CL, different classes) [11].

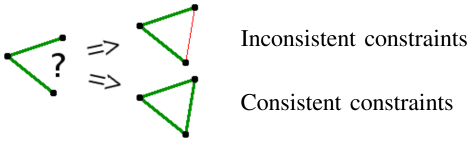

Fig. 3. Critical constraints. ML are pictured as bold green lines and CL constraints as thin red lines.

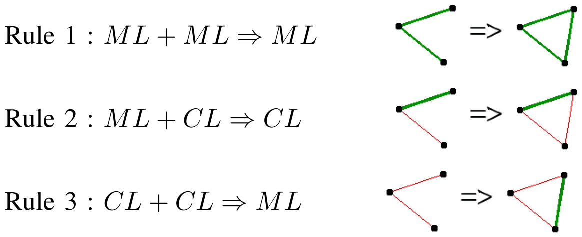
Rule 1 : $ML + ML \Rightarrow ML$
=>
Rule 2 : $ML + CL \Rightarrow CL$
=>
Rule 3 : $CL + CL \Rightarrow ML$
=>

Fig. 4. Pairwise constraints propagation in a 2-classes problem.

This interactive processing can be completed by an automatic selection of the pairs to be submitted to the expert. No choice is then left on the pairs to supervise. Resulting architecture is depicted in figure 2. This is the so called active semi-supervised clustering [12].

Automatic selection can involve different strategies:

- by focusing on the most ambiguous pairs in the aim to get a more informative decision [3]. However, the definition of an “ambiguous pair” must be defined;
- by randomly selecting object pairs. This strategy is generally slightly less efficient. However it allows a fair comparison between different semi-supervised clustering approaches as shown by Rangapuram [2].

B. Pairwise Constraint propagation

In such an active semi-supervised clustering, one has to rely on expert annotations. However, those may produce inconsistent configurations as shown in figure 3. In order to be robust against inconsistent configurations, a simple solution consists in not facing the expert with the critical cases such as the one depicted in figure 3. Furthermore, those critical configurations can be automatically deduced using coherence rules thus enriching the annotated pairs dataset without resorting to the expert. This inference is called “automatic propagation of constraints” [13].

In a 2-classes dataset, this automatic propagation is composed of three rules as illustrated in figure 4.

The annotation framework being described, the clustering problem has now to be considered.

C. Semi-supervised Deep Learning Clustering

Deep learning has been driving most of the attention since 2012 with the breakthrough performances obtained on large

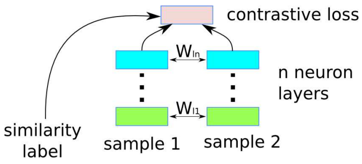

Fig. 5. Typical siamese architecture [8]

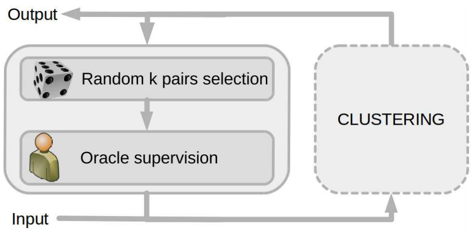

Fig. 6. Active semi-supervised clustering process with random pair selection

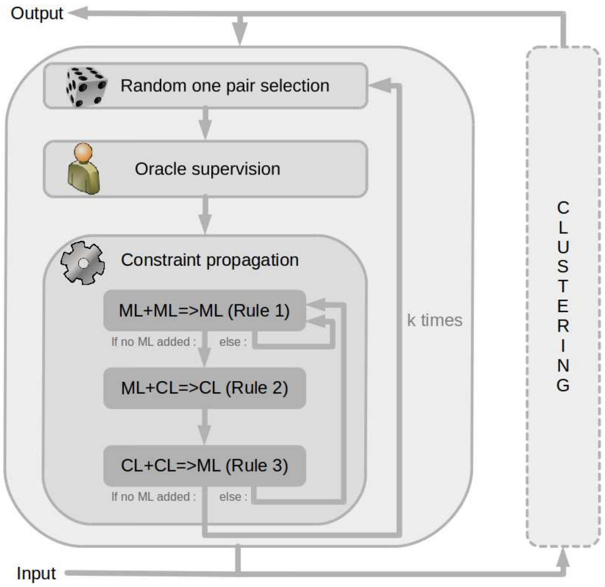

Fig. 7. Active semi-supervised clustering process with random pair selection and pairwise constraints propagation

scale image classification problems [14]. However neural networks revival was initiated years before and several architectures where already proposed to fulfil today needs. Regarding our topic, an efficient Deep Learning architecture dedicated to metric learning was proposed in [8]. As shown in figure 5, it consists in a duplicated feed-forward neural net stack with shared weights that is fed by two data samples, one per branch. Each layer can be convolutional, fully connected with state of the art non linearities and pooling methods as used for classical feed forward architecture. Training is ensured by the contrastive loss function shown in eq. 1 with binary similarity s provided as ground-truth, d the (euclidian) distance between pair elements in the projected space, b the number of sample pairs per training batch and the separating $margin$ generally set to 1. Such loss is minimal when the distance in the projected space of the last network layer corresponds to the semantic distance given by ground truth. Once training has been done, clustering can be performed using a simple convex clustering algorithm such as k-means.

$$E = \frac{1}{2b}\sum_{i=1}^{b} s(i)\times d(i)^2+(1-s(i))\times max(margin-d(i),0)^2 \quad (1)$$

D. Semi-supervised Spectral Clustering

Spectral Clustering algorithms [7] are composed of three steps: first a similarity graph between objects is built and permits the computation of an adjacency matrix; then, a projection is performed on a spectral space where clusters are easier to identify; finally, a standard convex clustering is performed in this new spectral space.

In order to introduce semi-supervision, different strategies allow pairwise constraints to be taken into account at each of the two first steps of the Spectral Clustering:

- during the *similarity graph computation*. For example, in the Active Clustering (AC) [3], inspired from Spectral Learning (SL) [15], constraints are introduced in the adjacency matrix by setting values 1 for ML and 0 for CL. However, despite being low computational cost, there is no guarantee for the constraints to be taken into account;
- during the *spectral graph computation* [16] [17] [18]. In our work, we chose the approach proposed by Rangapuram [2] denoted “Constrained One-Spectral Clustering” (COSC) where constraints are introduced in the convex optimization problem based on a gradient descent algorithm. Such method directly generates a 2-classes partition and avoids the use of a final clustering technique such as k-means. This approach can be extended to multi-partition situations by recursive calls. In a 2-classes problem, COSC error rate systematically reaches down to zero and ensures all the constraints to be taken into account contrary to other semi-supervised Spectral Clustering methods. However, the drawback is processing time that remains significantly higher than Spectral Learning.

III. Use case

A. Active semi-supervised clustering with random pair selection

From the framework described in section II-A, we build a benchmarking architecture that enables clustering methods comparison. Two test benches are proposed:

- The first one is shown in figure 6. It consists of an active semi-supervised clustering that allows, for each loop, k not annotated pairs to be randomly selected and to be submitted to the Oracle. Once done, each loop ends with a clustering step;
- The second test bench is shown in figure 7. In this case, only one pair is randomly selected from all the not annotated pairs and is submitted to the Oracle. Next,

the automatic constraint propagation described in [4] is applied. This step guaranties that the maximum number of supervision loops to perform equals the number of considered individuals of the dataset. In the end, each loop also ends with a clustering step.

B. Deep Learning Clustering used

We use a deep siamese architecture trained in an online way. Clustering task is performed after each training step using k-means on the projection of the input data at the final network layer output. The basic idea is to update the model iteratively at each clustering step of the proposed framework as long as Oracle supervision and propagation provide new annotated pairs. Compared to full retraining, the model is trained only with the last annotated pairs thus reducing computational cost while still ensuring global training error optimization. In such context, the main challenge is the choice of the network architecture and learning parameters that allows the model to be trained reliably with respect to the available quantity of training sample at each training period. More detailed, we chose to make some of the parameters dependant on the batch size b that defines how many samples are sent to the model in a single training iteration. First the minimum number of new annotated pairs that allows a clustering step to be launched is defined as $b \times a$. Regarding the framework proposed in the previous section, the clustering step can thus be skipped if too few pairs are available. By experimentally setting $b = 20$ and a in range [20;100], we allow a reasonable number of supervision steps before allowing for a single clustering step. Lower a, b values favour overfitting while high values would inquire too many supervision steps before retraining. Next, the target number of epoch per clustering step is experimentally set to 20 in order to limit overfitting. Finally we apply early stopping when training error cannot get lower after a period of $1/2$ of the target number epoch and we allow the system to skip one training sessions if overfitting is observed. All those presented cautions actually reveal the sensitivity of the training parameters that allow for an efficient deep net training session. Regarding the experimented deep architectures, we rely on a cascade of 4 convolutional layers with 50 neurons each and kernel size 3. ReLU non linearities finalize the process of each convolution. Layer stride values switch between 1 and 3 alternating signal projection and pooling objectives. The final layer is a fully connected one which number of neurons is set to 2 to ensure a 2-D projection on the data. We report here the architecture and the parameters configuration providing the best results (fully connected architecture works worst). This architecture experimented on two different datasets allows some conclusions to be drawn when comparing Deep Learning and Spectral Clustering while simplifying system description.

C. Spectral Clustering used

Regarding semi-supervised Spectral Clustering, we experiment with the supervision options presented in section II-D. The first method called “Spectral Clustering 1” relies on the COSC algorithm that introduces constraints in the spectral problem. Constraints are then considered during the spectral graph computation. The second method called “Spectral Clustering 2” also relies on the COSC algorithm. However constraints are directly injected into the adjacency matrix thus forcing the algorithm to consider constraints during the graph computation as with Spectral Learning methods. Such strategy allows the two types of constraint management to be compared using the same low level clustering algorithms.

For both methods, we use the standardized normal distribution of all values on each attribute. We then construct the similarity matrix using a search of the k-nearest neighbours $k \approx log(n)$ (n being the number of individuals of the dataset), and using a Gaussian weighting.

IV. Experimental results

Experiments are conducted on two bi-class datasets. The first one is “Sonar” provided by the Center for Machine Learning and Intelligent Systems (CMLIS) from the University of California at Irvine (UCI). It consists of 208 objects described by vectors of size 60 normalized between 0 and 1. The second one is a collection of videos from the dataset Blip10000 [10]. It consists of 2431 video of two genres “music and entertainment” and “technology”. Each video is described from its audio channel by a non standardized real-valued 196 attributes long vector described in [19].

To assess clustering quality, we use the standard adjusted Rand index [20] which measures the similarity between two clustering results, i.e. ground truth and our systems clustering results. Its main advantage is to consider its values between -1 and 1. The 1 value stands for identical partitions while 0 indicates untied partitions. -1 refers to contradictory partitions which does not generally occur in practice in this context.

Experimental results are presented in the four figures 8, 9, 10 and 11 where the x-axis corresponds to the number of pairs supervised by the Oracle and the y-axis indicates the normalized Rand index. The different curves represent the evolution of the three methods (Deep Learning and Spectral Clustering 1 or 2) throughout the iterations of the active semi-supervised process with random pairs selection with or without constraint propagation. All curves are the average of at least 5 executions per method.

Figure 8 shows the results of the first test bench without propagation using the Sonar dataset. We can note that the three methods have highly different behaviour. Spectral Clustering 1, which fully respects the constraints, reaches perfect clustering after 700 supervised pairs. In this case, it is the best clustering method. Spectral Clustering 2 converges less quickly (after 800 supervised pairs). The weaker performance can be explained by the fact that this method cannot take all constraints into account: they are injected into the adjacency matrix and they weakly constrain clustering. Regarding the Deep Learning approach, its curve starts only after several hundred annotations once enough annotated pairs are available. Next it has a slower convergence than the other two methods. We can assume that this lowest performance comes from the classical need of a

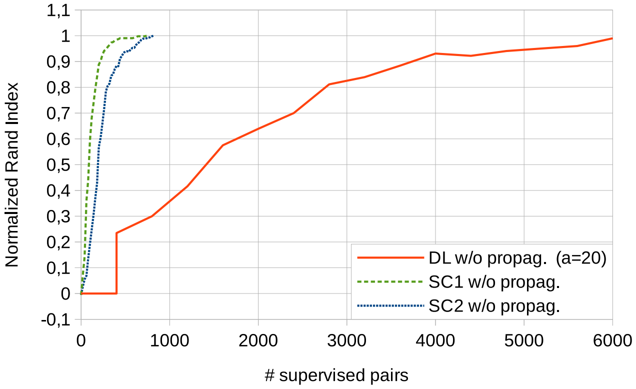

Fig. 8. Active semi-supervised clustering without propagation on Sonar dataset (208 objects - 2 classes).

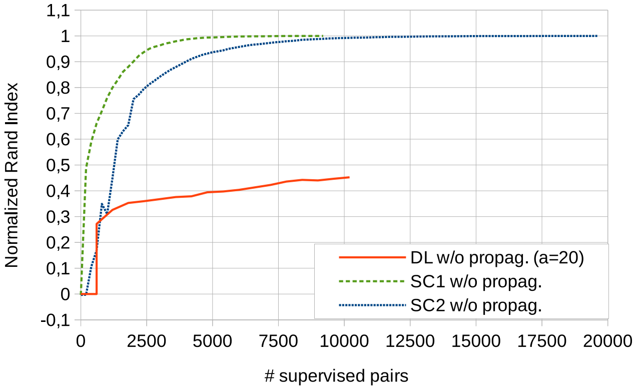

Fig. 10. Active semi-supervised clustering without propagation on Technology and Music genres of the MediaEval dataset (2431 videos - 2 classes).

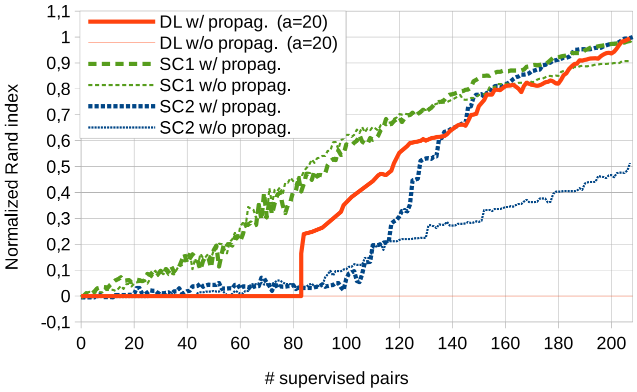

Fig. 9. Active semi-supervised clustering with (and without) propagation on Sonar dataset (208 objects - 2 classes).

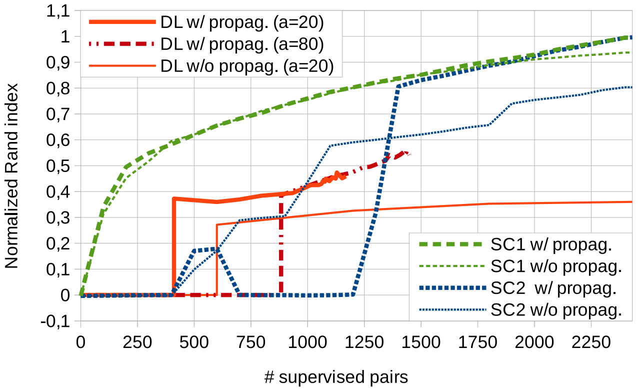

Fig. 11. Active semi-supervised clustering with (and without) propagation on Technology and Music genres of the MediaEval dataset (2431 videos - 2 classes).

large quantity of annotated data for efficient Deep Learning training.

Figure 9 shows the results of the second test bench with propagation on the same Sonar dataset. We can note the scale change of the x-axis (Oracle supervision steps) that is a consequence of the pairwise automatic propagation. Thin curves of figure 8 are represented again to compare the two test benches. The number of annotated pairs due to propagation is dramatically higher than the number of supervised pairs: after 208 supervised pairs, we obtain the annotation of all of the 21,528 existing pairs. Regarding Spectral Clustering methods, the two methods allow a perfect clustering at the end of those 208 supervisions. The performances are also better after the iterations 120 or 140. This shows that propagation boosts the two methods. Regarding the Deep Learning approach, one can observe that its performances now reach an intermediate level between the two Spectral Clustering methods. Thanks to the huge increase of annotated data and a network architecture that fits with the data, clustering converges to a significantly faster extent. In a real use case scenario such method would allow Oracle supervision to be stopped much earlier while obtaining a more comfortable clustering quality.

Moving to the second dataset, figure 10 shows the results of the first test bench without propagation using the Technology and Music genres of the MediaEval dataset. Similarly to the Sonar dataset, Spectral Clustering 1 obtains perfect clustering after 9,000 Oracle supervision. Spectral Clustering 2 reaches the goal after 20,000 supervised pairs. However, regarding the Deep Learning approach, results strongly differs from the first observations. First, with $a = 20$, clustering starts when 600 supervised pairs are obtained and reaches a clustering performance of 0.27. Next, performance increase keeps slow until 10,000 supervised pairs. 0.45 normalized Rand index is reached but remains below the other clustering methods. Finally, deep net training interrupts because of gradient explosion during training.

Figure 11 shows the results of the second test bench with constraint propagation on the same MediaEval dataset. Thin curves of figure 10 are represented again to compare the two test benches. Spectral Clustering methods show similar behaviours compared to the ones obtained with the Sonar dataset and both ensure perfect clustering with 2431 supervised pairs. The Deep Learning approach presents the same

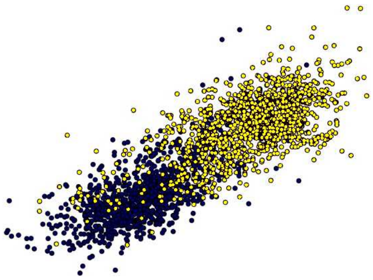

Fig. 12. Deep Learning embedding of the MediaEval samples (2431 videos - 2 classes). Colors encode ground truth labels Technology and Music genres.

early interruption observed without propagation. However, in early supervision steps, the proposed neural architecture quality reach still an intermediate level between the two Spectral Clustering methods. Nevertheless, early clustering interruption remains a problems. Different tests showed that, despite testing with different parameters including learning rate, the proposed network architecture is probably not adapted to this clustering context. As an illustration, two tests are shown in figure 11 with two different values for parameter a. In case (1), $a = 20$ allows for earlier clustering start and stop while case (2) with $a = 80$ delays start and stop with increased partition quality. Batch normalisation could be experimented to avoid gradient explosion but architectural network changes should be preferred to better adapt to the data.

Figure 12 shows the last valuable obtained 2-D projection of the data samples with ground truth labels. One can observe that embedding could allow for a moderate quality clustering but the final k-means cannot do miracles. All in all, this shows the difficulty of the use of a deep net system, from the architectural design task to the parameters optimisation steps to adapt to specific data. On the contrary Spectral Clustering is almost parameter free. However in large scale configuration, due to its similarity matrix, Spectral Clustering becomes computationally untrackable.

V. Conclusion

This paper presents a comparison between Deep Learning and Spectral Clustering in two active semi-supervised clustering test benches. It allows both clustering methods to be compared with and without automatic pairwise constraint propagation. We experiment with two real-world datasets. The first contribution of this paper addresses the comparison of Deep Learning versus Spectral Clustering. It shows the potential interest of Deep Learning siamese based approaches and Spectral Clustering methods. However, this is conditioned by an accurate design and parameter setup of the deep architecture. Second, this paper shows the impressive improvements provided by propagation for both clustering methods. We recommend the use of propagation since it strongly reduces the cost of constraints acquisition and facilitates clustering quality much faster. This is particularly of interest with Deep Learning that always benefits from large data amounts for efficient training.

Further work will address Deep Learning with pairwise constraint propagation. An immediate perspective is to adjust siamese architecture to several multimedia datasets. An other perspective is to compare such approach to state of the art supervised classification techniques in challenges such as MediaEval and study complementarity between Deep Learning and Spectral Clustering.

References

[1] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recogn. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[2] S. S. Rangapuram and M. Hein, "Constrained 1-spectral clustering," in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, April 21-23, 2012*, 2012, pp. 1143–1151.

[3] C. Xiong, D. M. Johnson, and J. J. Corso, "Active clustering with model-based uncertainty reduction," *CoRR*, vol. abs/1402.1783, 2014.

[4] N. Voiron, A. Benoit, A. Filip, P. Lambert, and B. Ionescu, "Semi-supervised spectral clustering with automatic propagation of pairwise constraints," in *13th International Workshop on Content-Based Multimedia Indexing, CBMI 2015, Prague, Czech Republic, June, 2015*.

[5] J. Camargo and F. Gonzlez, "Visualization, summarization and exploration of large collections of images: State of the art," in *Latin-American Conference On Networked and Electronic Media. LACNEM*, 2009.

[6] I. Witten, E. Frank, and M. Hall, "Data mining: Practical machine learning tools and techniques," *Morgan Kaufmann Publishers*, 2011.

[7] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[8] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 539–546.

[9] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 437–478.

[10] S. Schmiedeke, P. Xu, I. Ferrané, M. Eskevich, C. Kofler, M. Larson, Y. Estève, L. Lamel, G. Jones, and T. Sikora, "Blip10000: A social video dataset containing spug content for tagging and retrieval," *ACM Multimedia Systems Conference*, 2013.

[11] K. Wagstaff and C. Cardie, "Clustering with instance-level constraints," in *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, 2000, pp. 1103–1110.

[12] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM-04)*, April 2004.

[13] P. K. Mallapragada, R. Jin, and A. K. Jain, "Active query selection for semi-supervised clustering," in *19th International Conference on Pattern Recognition (ICPR 2008), December 8-11, 2008, Tampa, Florida, USA*.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[15] S. D. Kamvar, D. Klein, and C. D. Manning, "Spectral learning," in *In IJCAI*, 2003, pp. 561–566.

[16] X. Wang and I. Davidson, "Flexible constrained spectral clustering," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, 2010, pp. 563–572.

[17] L. Xu, W. Li, and D. Schuurmans, "Fast normalized cut with linear constraints," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 2866–2873, 2009.

[18] Z. Li, J. Liu, and X. Tang, "Constrained clustering via spectral regularization," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 2009, pp. 421–428.

[19] I. Mironica, B. Ionescu, P. Knees, and P. Lambert, "An in-depth evaluation of multimodal video genre categorization," in *IEEE International Workshop on Content-Based Multimedia Indexing*, 2013.

[20] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.