Robust Multi-View Representation Learning (Student Abstract)

Sibi Venkatesan, James K. Miller, Artur Dubrawski
AutonLab, Robotics Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA - 15213, USA

Abstract

Multi-view data has become ubiquitous, especially with multi-sensor systems like self-driving cars or medical patient-side monitors. We propose two methods to approach robust multi-view representation learning with the aim of leveraging local relationships between views.

The first is an extension of Canonical Correlation Analysis (CCA) where we consider multiple one-vs-rest CCA problems, one for each view. We use a group-sparsity penalty to encourage finding local relationships. The second method is a straightforward extension of a multi-view AutoEncoder with view-level drop-out.

We demonstrate the effectiveness of these methods in simple synthetic experiments. We also describe heuristics and extensions to improve and/or expand on these methods.

Introduction

The structural relationship between different views can provide useful insight into the nature of multi-modal data. Further, multiple views can de-noise each other, making learning over them more robust. In this abstract, we approach the problem of robust multi-view representation learning by trying to model and/or utilize local relationships between subsets of views. Beyond just understanding the structure of the data, we can also potentially estimate the "usefulness" of any view within the context of these relationships.

We propose two approaches: the first is closely related to multi-view extensions of Canonical Correlation Analysis (CCA) while the second is a more straightforward adaptation of a multi-view AutoEncoder. We demonstrate their effectiveness on simple experiments, and also describe potential heuristics and extensions to improve upon them.

Notation: The rows of X_i are the data points for each view i of K views. We assume that our data is centered.

Approaches

Multi-view One-vs-Rest CCA

Existing multi-view CCA extensions ((Kakade and Foster 2007), (Rupnik and Shawe-Taylor 2010)) usually find a single set of projections for each view which maximize some overall correlation objective across all views. This forces projections to primarily recover globally shared structure while largely ignoring local relationships. To model all possible local relationships naively, we would need an exponential number of CCA computations for all view-subsets. Even just considering pair-wise relationships would require a quadratic number of CCA computations.

Our approach (MOCCA) reduces this to problem to K one-vs-rest CCA problems, one for each view. In each problem, we try to learn projections for the remaining views (henceforth called "sub-views") to maximize the correlation with the given main view. In essence, these are 2-view CCA computations where the sub-views together form the second "view". However, simple concatenation of the sub-views could disturb the inter and intra-view structure. For this, we enforce a sub-view group-sparse regularization to encourage learning projections which respect this structure.

The intuition here is that by penalizing the contribution of all covariates from a given sub-view together, we allow the optimization to distinguish between existing inter and intra-view relationships. Covariates within a sub-view would tend to either all contribute toward correlation/reconstruction or all get shut down together. This encourages only using the most important sub-views to contribute to the projections.

For each view, we minimize the objective:

$$\min_{P_{ij}} f(X_i P_{ii}, \sum_{j \neq i} X_j P_{ji}) + \lambda \sum_{j \neq i} R_G(P_{ij}) \qquad (1)$$

where $f(A, B) = -A^T B$, R_G is some sparsity regularizer (eg. L_∞ norm). P_{ij} represents the projections to be learned from sub-view j to main view i.

For CCA, we include the usual orthogonality constraints as well and optimize using linearized ADMM as described in (Suo et al. 2017). We can also substitute f for other error metrics like reconstruction error, and potentially include an overall sparsity regularization as well. This objective assumes linear projections but can easily be extended to any

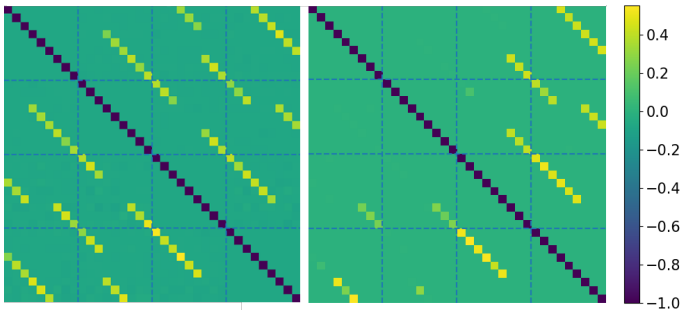

Figure 1: Redundancy matrices as produced by MOCCA without (left) and with (right) group-sparse regularization.

class of projection functions $P_{ij}(X_j)$ which are easy to optimize (eg. deep neural networks).

While group sparsity does not necessarily uncover *all possible* local relationships, it does encourage extracting the more apparent and prominent ones. If we tweak the group-sparse penalty, we could tune the number of represented sub-views in the projections. This would allow us to build a multi-fidelity "relationship" graph between views, with the final "representation" implicitly embedded in it.

Simple Synthetic Experiment We consider a four-view problem where any two views are enough to reconstruct the rest. Our dataset has four underlying feature sets A, B, C, D (eg. independent sensor measurements), with the four views being $X_1 = [BCD]$, $X_2 = [ACD]$, $X_3 = [ABD]$ and $X_4 = [ABC]$ with some noise.

We take a "redundancy matrix" (shown in Figure 1) to be a block matrix where each row i represents the relationship between main view i and all sub-views; i.e. $X_i P_{ii}$ vs. $\sum_{j \neq i} X_j P_{ji}$ where the P_{ij}s are given by the blocks and P_{ii} is taken to be the identity. The figure shows the usefulness of the group-sparse regularization which allows us to find a smaller subset of sub-view projections by exploiting local redundancies and relationships between views.

Heuristic: Greedy Step-wise View Selection Playing with hyper-parameters to modulate the number of sub-views can be tricky, since this is usually data-dependent. We could use a greedy selection method for sub-views to circumvent this. Such an approach would sequentially select the next best view to minimize the residual error. Of course, using this approach to get an ordering over *all* sub-views is counterproductive, since we wanted to avoid a quadratic number of computations to begin with. So, such an approach should only really be used if it is adequate to use only a small subset of projections from a relatively large number of sub-views.

Robust Multi-view AutoEncoder

Typical Multi-view AutoEncoders ((Ye et al. 2016), (Wang et al. 2015)) learn a shared representation exploiting the *intersection* of views. This falls prey to the same pitfalls as before: local relationships are often ignored.

Our proposed method (RMVAE) has the following framework: Every view has its own encoder network; these encodings are then concatenated and fed into a shared encoder to give the final common representation. This representation is decoded back to the original views using individual decoders. For "robustness" of representation, we use view-level dropout; every batch, we drop a random input-view subset while still forcing reconstruction of all views. This encourages the model to exploit redundancy of information across the different views.

Experimental results We conducted simple experiments on similar datasets to the previous approach. Here, we varied the number of available views to reconstruct all output views. The reconstruction error reduces as more views are available, which is the expected trend. We have omitted quantitative results in the interest of space.

Generative Modeling Extension We could replace the common encoder in the RMVAE with a generative equivalent, instead of relying on reconstruction alone. We can follow the overall ideas from flow-based models like RealNVP (Dinh, Sohl-Dickstein, and Bengio 2016) and TANs (Oliva et al. 2018) to learn an invertible encoding of the data into a space where the data distribution is very simple.

Conclusion

In this abstract, we describe two approaches for robust multi-view representation learning: (i) Multi-view One-vs-Rest CCA with group-sparse regularization and (ii) Robust Multi-view AutoEncoder with view-level dropout. These approaches try to uncover local relationships between views, to help better understand the underlying structure of the data.

References

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Kakade, S. M., and Foster, D. P. 2007. Multi-view regression via canonical correlation analysis. In *International Conference on Computational Learning Theory*, 82–96. Springer.

Oliva, J. B.; Dubey, A.; Zaheer, M.; Póczos, B.; Salakhutdinov, R.; Xing, E. P.; and Schneider, J. 2018. Transformation autoregressive networks. *arXiv preprint arXiv:1801.09819*.

Rupnik, J., and Shawe-Taylor, J. 2010. Multi-view canonical correlation analysis. In *Conference on Data Mining and Data Warehouses (SiKDD 2010)*, 1–4.

Suo, X.; Minden, V.; Nelson, B.; Tibshirani, R.; and Saunders, M. 2017. Sparse canonical correlation analysis. *arXiv preprint arXiv:1705.10865*.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015. On deep multi-view representation learning. In *International Conference on Machine Learning*, 1083–1092.

Ye, T.; Wang, T.; McGuinness, K.; Guo, Y.; and Gurrin, C. 2016. Learning multiple views with orthogonal denoising autoencoders. In *International Conference on Multimedia Modeling*, 313–324. Springer.