Learning a Multi-Modal Similarity Metric with Application to 2D-3D Matching

Raia Hadsell Bogdan Matei Harpreet Sawhney
Sarnoff Corporation
Princeton, NJ 08540
(rhadsell, bmatei, hsawhney)@sarnoff.com

In recent years, learned similarity metrics have been found to be very useful for matching, recognition, retrieval, and verification problems in the image domain [4, 1, 2, 6, 5]. Distances measured from image to image using pixel values are dominated by variations in lighting, registration, and viewpoint, which are often precisely the transformations that are irrelevant in the context of semantic similarity and recognition questions. Similarity metrics are trained to be invariant to such irrelevant transformations by learning same/different labels on pairs or triples of data samples. The result is a linear or non-linear mapping of the input image to an output space in which Euclidean distances can be used to solve the original problem.

Unfortunately, this paradigm does not work when the desired similarity metric is between inputs that are fundamentally different, such as data from different sensor types (e.g., matching color images with rendered 3D point clouds) or images containing different types of content (e.g., matching images of objects and images of scenes). A similarity metric is trained to be invariant to some differences in the inputs while enhancing certain other similarities between the inputs, but if there are no meaningful similarities in the input domain, then this becomes exceedingly difficult. Therefore, we propose a new, more powerful learning architecture to handle such inter-domain similarity problems. Instead of learning a single transformation of the input, we propose to train multiple functions which map to the same output space and are trained simultaneously using similarity labels. By disconnecting the mapping functions of different data types but enforcing meaningful distances in the shared output space, the individual mapping functions can preserve more information content.

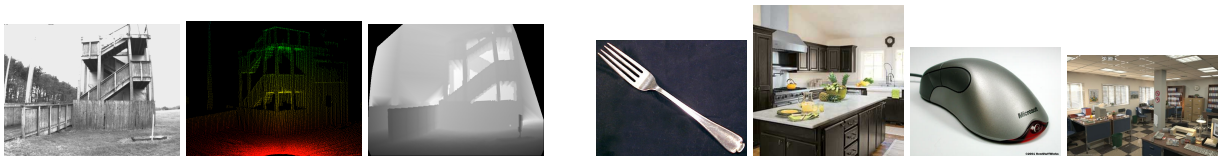
The DrLIM approach is used as a starting point for this work [3]. In the DrLIM approach, a single non-linear function is trained on pairs of inputs supervised by a binary similarity label, using a loss function L that penalizes similar pairs that are far apart in the output space and dissimilar pairs that are close in the output space:

$$L(w, S, X_1, X_2) \tag{1}$$

$$= (1 - s)\frac{1}{2}||F_w(x_1) - F_w(x_2)||_2^2 + \frac{s}{2}\max(0, m - ||F_w(x_1) - F_w(x_2)||_2)^2 \tag{2}$$

where S is the similarity label and m is a margin used to prevent the output space from endlessly expanding.

Topic: visual processing and pattern recognition
Preference: poster



a b c d e f g

Figure 1: Examples of datasets with inter-domain input pairs. Images a, b, and c show the same building in a color image, a rendered point cloud, and a stereo depth map. Images d, e, f, and g show objects paired with their usual environment: fork+kitchen, mouse+office. Mapping these examples to an output space such that their semantic similarity is preserved as a simple distance using a single function is near impossible.

The architecture of the new, proposed approach is given in Figure 2. Instead of a single function F_W, two (or more) functions F and G are trained. The functions need not be related in any way except the outputs must be of the same dimension and both functions must be differentiable with respect to their parameter vectors. The loss function is still the squared-squared loss:

$$L(w_1, w_2, X, Y, S) \tag{3}$$

$$= (1-s)\frac{1}{2}||F_{w_1}(x) - G_{w_2}(y)||_2^2 + \frac{s}{2}\max(0, m - ||F_{w_1}(x) - G_{w_2}(y)||_2)^2 \tag{4}$$

The gradients are computed with respect to each of the component functions so that the entire system can be trained with gradient-based optimization.

The proposed approach is tested on a correspondence problem between 2D image data and 3D point cloud data. The point cloud is rendered to produce a depth map and the similarity metric is trained on roughly aligned patches in the image and the depth map. To create the training data, 3D laser range data and video are collected simultaneously using calibrated sensors on a mobile robot. Registered patches are given a label of 0 (similar), and non-registered patches are given a label of 1 (dissimilar).

References

[1] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 539–546, 2005. 1

[2] J. Goldberger, S. T. Roweis, G. E. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems (NIPS)*, 2004. 1

[3] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1735–1742, 2006. 1

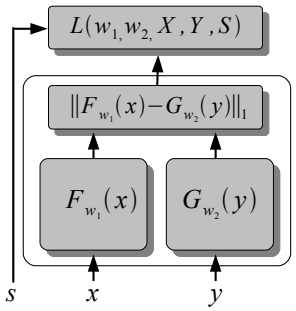


Figure 2: An architecture to learn a similarity metric between different data input types by training two functions to map their inputs to the same feature space.

[4] H. Mobahi, R. Collobert, and J. Weston. Deep learning from temporal coherence in video. In *Proc. of International Conference on Machine Learning (ICML)*, page 93, 2009. 1

[5] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems (NIPS)*, 2005. 1

[6] S. A. J. Winder and M. Brown. Learning local image descriptors. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 1