Sub-symbolic Re-representation to Facilitate Learning Transfer

Dan Ventura
Computer Science Department
Brigham Young University
ventura@cs.byu.edu

Abstract

We consider the issue of knowledge (re-)representation in the context of learning transfer and present a sub-symbolic approach for effecting such transfer. Given a set of data, manifold learning is used to automatically organize the data into one or more representational transformations, which are then learned with a set of neural networks. The result is a set of neural filters that can be applied to new data as re-representation operators. Encouraging preliminary empirical results elucidate the approach and demonstrate its feasibility, suggesting possible implications for the broader field of creativity.

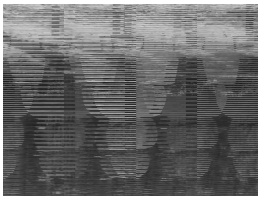
Introduction

Learning transfer is the ability of a system to learn one problem and then to transfer a significant amount of the learned knowledge to a different problem. While the learning transfer itself is often considered a creative act, creativity is additionally required in deciding *which* prior knowledge to use and *how* to use it. Symbolic systems employing some form of analogy have been somewhat successful here, including approaches to analogy (Spellman & Holyoak 1996)(Gentner 1983)(Hofstadter 1995), skill transfer (Detterman & Sternberg 1993), similarity (Vosniadou & Ortony 1989) and metaphor (Ortony 1993); however, these approaches require a significant amount of specialized domain knowledge and do not generalize. We propose the use of sub-symbolic approaches to learning transfer, trading the interpretability of symbolic approaches for representational power and generality.

The idea of sub-symbolic systems that exhibit learning transfer is not new (Pratt 1996) and interesting work has been done on various aspects of the problem, including inducing a "natural" measure of distance between points in input space (Baxter 1998), developing a measure of task relatedness based on closeness of example generating distributions (Ben-David & Schuller 2003), an approach to task clustering (Thrun & O'Sullivan 1998), and kernel-based methods (Evgeniou, Micchelli, & Pontil 2005) (Micchelli & Pontil 2005). However, systems that exhibit *robust* and *substantive* learning transfer have yet to be developed.

One particularly striking example of learning transfer for problem solving is often termed *re-representation*. In essence, this consists of re-encoding the problem at hand in a (usually quite) different way so that it (better) resembles something familiar. For example, most people would have difficulty identifying the subject of the image in Figure 1(a). However, using a simple transformation, the image can be re-represented as the image in Figure 1(b), and the identification problem becomes much easier.

Previous sub-symbolic work related to the idea of re-representation includes developing a low-dimensional *inter-lingua*-type representation (Intrator & Edelman 1996), and modifying a new problem instance via some type of transformation (Thrun & Mitchell 1995) (Miller, Matsakis, & Viola 2000). Re-representation can be the key to learning transfer because it is often necessary in situations in which the problem solver has hit a dead end — the problem may not have a solution for a give representation and therefore, unless a new one is discovered, no progress will be made. Indeed, this process of discovering a useful re-representation has been identified in some theories as the essence of insight (Ohlsson



(a)

(b)

Figure 1: *Re-representation can aid in (for example) identification tasks.* Through a simple transformation, the image on the left can be re-represented as the image on the right. Without the transformation, identifying the subject of the image is difficult.

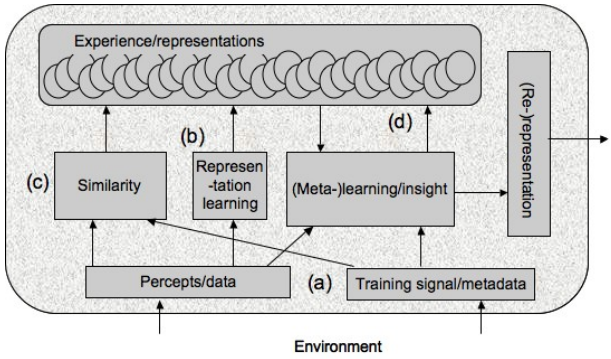


Figure 2: *Logical overview of an intelligent system capable of learning to develop insight for aiding substantive learning transfer.* (a) The system receives percepts and possibly training signals from the environment (b) These data are used by a learning component to construct a representation (c) These data are also used by a similarity component to discover related (and interesting) previously mastered tasks (d) Representations useful for tasks that are related can be transfered to the learning module to provide insight in the form of re-representation, and as new tasks and representations are learned, they are stored in the library of mastered tasks.

1992).

Our long-term goal is to build a system capable of *substantive* learning transfer, incorporating a method for measuring the "transferability" of a pair of tasks and a general (sub-symbolic) mechanism for knowledge representation and transfer (see Figure 2 for a high-level abstraction of such a system.) This work will explore proof-of-concept for several of these ideas, including knowledge representation and task similarity measures, focusing on learning transfer by developing sub-symbolic mechanisms for learning useful representations and examining the correlation between task similarity and the efficacy of learning transfer via re-representation using those mechanisms.

Methodology

We will employ an "image"-based approach to discover a (re-)representation mechanism that is invariant to various transforms. We consider the general case where closed form analytical expressions for such transforms will not be derivable, and propose learning interesting transforms inherent in the data by employing a neural approach as a (hopefully compact) representation. For example, a system exposed to images of people taken from various viewpoints might discover the concept of occlusion.

In many cases, these transformations may occur on a lower-dimensional manifold (that is, lower than the intrinsic representational dimension), and we will have to discover that surface in order to produce an accurate (re-)representation (in the form of a neural filter). Combining (nonlinear) manifold learning with a sub-symbolic transform representation will allow us to discover interesting transforms that can be used to (re-)represent data in a way that facilitates learning transfer.

Figure 3 demonstrates the idea. In the process of learning to recognize the letter A, we collect data in the form of examples. The explicit representation of this data (as pixels) may not be an informative representation or it may contain problem specific information that we would like to generalize away. Learning the implicit manifold on which the data live will often reveal important information about the data. For example, these data live on a 2-dimensional manifold (Figure 3(a)) whose axes naturally correspond to the two important invariant transforms implicitly encoded in our examples: rotation and scaling. Building neural models of these transforms provides a convenient and powerful way to learn these representations and naturally facilitates transfer. The solution to the puzzle shown in Figure 3(b) requires transferring both types of learning. To summarize, this implementation of knowledge (re-)representation is accomplished in two steps:

1. The discovery of the relevant manifold to reveal the transformation(s) to be learned, and
2. Learning the transformations.

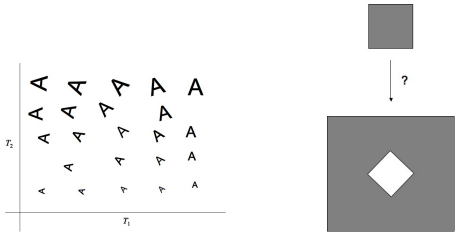
More formally, given a set $\mathcal{T}$ of learning task instances that are related through some transformation ξ, we require a clustering process, p that (at least) induces an ordering on $\mathcal{T}$, $p : \mathcal{T} \rightarrow \mathbb{N}$, such that

$$p(t) < p(u) \Rightarrow t \overset{*}{\underset{\xi}{\Longrightarrow}} u, \forall t, u \in \mathcal{T}$$

where $\overset{*}{\underset{\xi}{\Longrightarrow}}$ represents the (iterative) application of ξ.

Given such an ordering, we can construct a training set S for learning ξ as follows. Choose $t_0 \in \mathcal{T}$ such that

$$p(t_0) \leq p(t), \forall t \in \mathcal{T}$$



(a) (b)

Figure 3: *Transformations that live on the manifold can be discovered and encoded as useful knowledge representations.* On the left, a 2-D manifold reveals two high-level concept transforms: rotation and scaling. On the right, the puzzle problem requires the application of both transformations for its solution.

Figure 5: *A (sub-sampled) one-dimensional manifold of figures reduced from 2500 dimensions.* The manifold clearly shows two disparate tasks and neighboring points on the manifold correctly encode the transform.

and choose $t_i \in \mathcal{T}$ such that

$$p(t_{i-1}) < p(t_i)$$
$$p(t_k) < p(t_i) \Rightarrow p(t_k) < p(t_{i-1}), \forall i, k > 0$$

Then:

$$S = \left\{ \begin{array}{ccc} t_0 & \rightarrow & t_1 \\ t_1 & \rightarrow & t_2 \\ & \vdots & \\ t_{n-1} & \rightarrow & t_n \end{array} \right\}$$

where $n = |\mathcal{T}|$. The quality of S may be significantly improved if the range of p is $\Re$ rather than $\mathbb{N}$ with p inducing not only an ordering on $\mathcal{T}$ but also a distance measure, so that $p(t)$ represents the distance of t from some (arbitrary) origin.

For our implementation of p, we apply an iterative manifold learning algorithm (Gashler, Ventura, & Martinez 2007) to reduce our data to a single dimension whose ordering of the data (hopefully) faithfully represents the transform ξ to be learned. (For example, from Figure 3, the A's should be ordered from largest to smallest or from smallest to largest in the scaling dimension.) Then, neighbors on the manifold act as input/output training patterns in S. For learning S, we employ a standard multi-layer perceptron trained with backpropagation.

The result is a neural filter ζ that (hopefully) closely approximates the transform ξ and can be applied to new tasks to facilitate their solution. In the puzzle example, the task of recognizing A's is represented with multiple instantiations. The manifold learner orders the A's from largest to smallest and produces a set of training pairs that encode this scaling transform. These data are used to train a neural network that learns to scale its input. Similarly, the rotational transform can also be learned as a neural filter. When the puzzle is encountered, the scaling and rotational transforms can be used to (re-)represent the large square as a small diamond, facilitating the puzzle's solution.

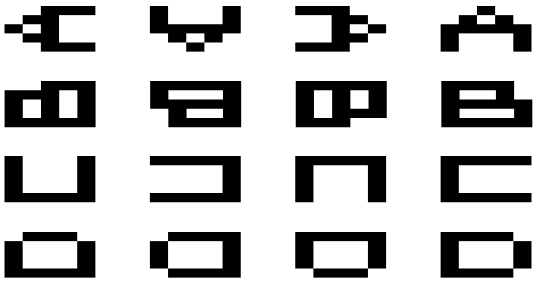


Figure 4: *A small set of examples that exhibit a counter-clockwise rotation in* 90° *increments.*

Empirical Results

As a proof-of-concept, we consider a simple data set consisting of grayscale images of the block letters A, B, C, D and E and collect samples of the first four letters in various attitudes of rotation (some of which are shown in Figure 4). Given these data, we must first discover the (1-dimensional) manifold on which they live, and then learn the relationship that exists between neighbors on that manifold.

Figure 5 shows the (sub-sampled) results of manifold learning on a set of 50x50 images of A's and B's, rotated in 1-degree increments from 0° to 180°. Notice that the A's and B's are well-seperated on the manifold (in reality, they were much more separated, but for visualization purposes the line was scaled in a non-linear manner to bring the two clusters closer together), giving an indication that these are two separate tasks. Notice also that the A's (B's) are ordered in descending (ascending) order of rotation, nicely revealing the transformation to be learned. A training set for learning the transform can now be constructed as $\{a_1 \rightarrow a_2, a_2 \rightarrow a_3, \ldots, a_{n-1} \rightarrow a_n, b_1 \rightarrow b_2, \ldots\}$, with a_{i+1} being the ordinal neighbor of a_i on the manifold, and the large gap between A's and B's making it clear that they are different examples of the transform (and thus $a_n \rightarrow b_1$ is *not* included in the training set).

It should be noted that in general discovering the "correct" dimensionality of the manifold is non-trivial and for comparison Figure 6 shows the same data represented on a 2-dimensional manifold. While the neighbors still correctly encode the transformation, the extra dimension introduces an obfuscating degree of freedom that makes the construction a training set much more difficult.

Given the ability to discover useful transforms via manifold learning, the next step is to learn those transforms, encoding them in a such a way that they can be used for re-representation. Using the small set of examples shown in Figure 4 as a training set, a multi-layer perceptron was trained with backpropagation to act as a neural encoding of the transformation. Figure 7 shows the results of applying the learned transform to E's of various rotations — the system has not only learned to represent the rotational transform encoded in the data of Figure 4 —given a single example of an E, it can employ that transform to (re-)represent that E in var-

ious poses, facilitating recognition (transferring learning from related recognition tasks rather than learning from seeing example E's). This is particularly interesting considering the extremely gross subsampling of the transform examples (90° increments) and the small number of related tasks (4) used for training.

Typically, the greater the number of related tasks to which a learner has been exposed, the more likely it is that the learner will have generalized sufficiently to allow for useful learning transfer. Figure 8 demonstrates this nicely. Each line in the graph corresponds to a letter recognition task (for example, the black line with black triangles indicates performance on recognizing the letter C). Each point on the line represents the average performance on the task given learning transfer from some number of related tasks (indicated on the x-axis). As expected, for most recognition tasks, the greater the number of related tasks, the better the system does at transferring useful knowledge. The obvious exception is the A recognition task.

The key, of course, is that the task(s) from which transfer is attempted must be suitably related to the target task; otherwise, transfer may not only not be beneficial, it could, in fact, be detrimental. To illustrate, consider a simple measure of similarity for the letter recognitions tasks. Given a set of (oriented) candidate tasks $\mathcal{R}$ and an (oriented) new task t, define the similarity between $\mathcal{R}$ and t as $\sigma(\mathcal{R}, t) = 1 - d(\mathcal{R}, t)$ with

$$d(\mathcal{R}, t) = \frac{1}{N\,|\mathcal{R}|} \sum_{r \in \mathcal{R}} Hamming(r, t)$$

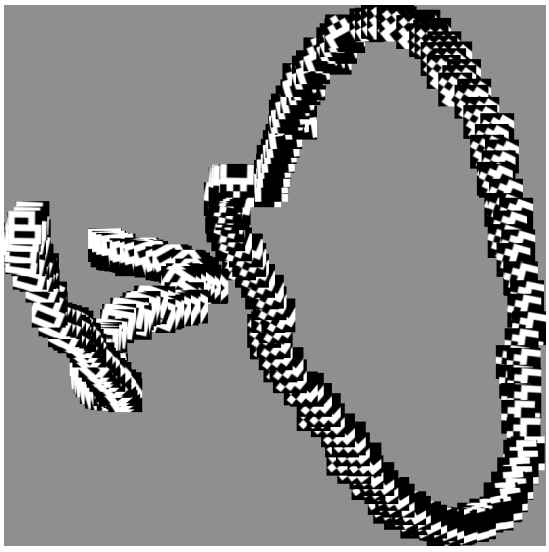


Figure 6: *A two-dimensional manifold of figures reduced from 2500 dimensions.* Although the neighboring points on the manifold do reveal the transformation, the extra degree of freedom makes it less obvious. It is also not clear that, in fact, two distinct tasks are represented.

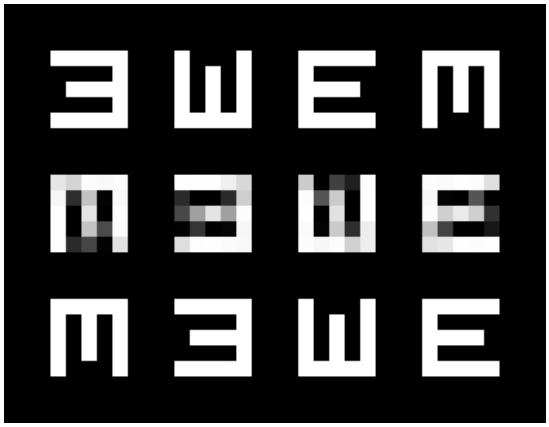


Figure 7: *Transferring rotation to E (from A,B,C,D).* The system is trained with the A, B, C and D examples from Fig. 4. It is then given the stimuli shown on the top row. The second row shows the response activation of the neural field with 256 levels of gray. The third row is the result of applying a hard binary threshold to these activation levels.

where N is the number of pixels in the image and $Hamming(r, t)$ is the number of pixels that differ between r and t. The function d calculates a normalized average Hamming distance between t and the members of $\mathcal{R}$. The normalization constant ensures that the similarity function σ has a range of [0...1]. Interestingly,

$$\begin{aligned}
\sigma(\{B, C, D, E\}, A) &= 0.51 \\
\sigma(\{A, C, D, E\}, B) &= 0.75 \\
\sigma(\{A, B, D, E\}, C) &= 0.70 \\
\sigma(\{A, B, C, E\}, D) &= 0.67 \\
\sigma(\{A, B, C, D\}, E) &= 0.73
\end{aligned}$$

Notice the marked difference between A's similarity to its candidate set and each of the other tasks' similarities to theirs.

Figure 9 further confirms this phenomenon of "transferability". Each point represents an instance of learning transfer in the case of $|\mathcal{R}| = 1$, and there is a point for each such unique transfer scenario: from B to A, from C to A, ..., from D to E. The y-value of each point is the transfer accuracy (how well the transfered learning applied — in this case, how accurately a rotation was performed). The x-value of each point is the similarity σ between the two tasks. Clearly, in this case, tasks with greater similarity enjoy more efficacious learning transfer.

Of course, if the transformation learned was really a rotation, the re-representation would work perfectly for any of the letters. If an E is rotated correctly, why isn't an A? In fact, the transformation that is learned is not a pure rotation, but a rough approximation of one, as encoded by the few task examples used for training. This can be seen in Figure 10, which visualizes a partial anal-

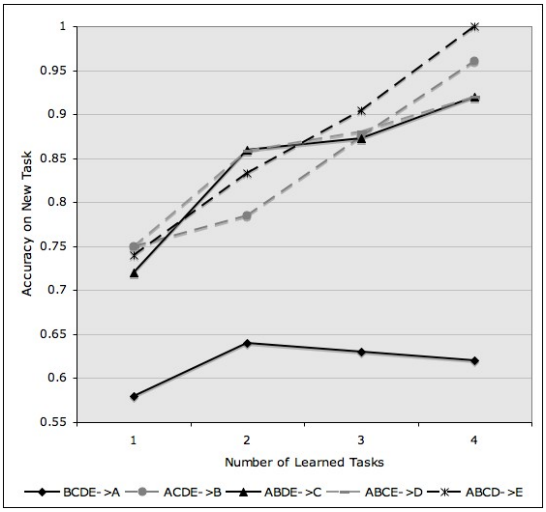


Figure 8: *Transfer improves with the acquisition of additional (related) tasks.* Each line on the graph shows the efficacy of learning transfer for a given letter. Each point on the line represents an average accuracy for that task given a number of previously learned tasks. In four of the five cases (B, C, D and E), accuracy improves with the number of previously learned tasks. In the exceptional case (A), the similarity between the target and previously learned tasks was significantly lower than in the other cases.

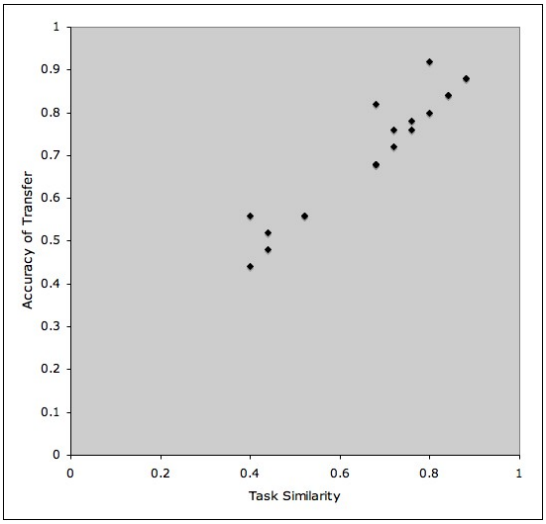


Figure 9: *Task similarity is strongly correlated with success of the transfer.* The scatter plot compares 1-to-1 letter recognition task similarities (x-axis) against learning transfer accuracy (y-axis), indicating a strong correlation.

ysis of the learned transform. Figures 10(a) and 10(c) show a localized stimulus applied to the neural receptor field, and Figures 10(b) and 10(d) show the corresponding activation response. If the learned transformation was truly a rotation, these activation responses should be as localized as the input and centered around the '*' in each figure. Instead, both stimuli produce a rather distributed response. However, interestingly, in both cases the largest contiguous area of output activity does occur in the target locality, suggesting that even given the paucity of training data, some notion of rotation has been acquired by the system. And, to the extent that that approximation applies to new tasks, transfer will be beneficial.

Finally, we consider a much more difficult transform: a non-affine warping of the image that "pinches" the four sides towards the middle. A few examples of As, Bs, Cs and Ds warped to varying degrees were used to train a neural filter for implementing the transform. The trained network was then applied to the letter E, and the result is shown in Figure 11. The transform is effective, though transfer is not completely beneficial, as the middle horizontal stroke is overly compressed.

Discussion

We have demonstrated the feasibility of a sub-symbolic approach to the re-representation problem, showing that we can discover and learn useful transforms in a sub-symbolic form that will facilitate creativity in a complete system. The full development of creative problem solving in sub-symbolic systems will require significant additional research in task similarity metrics, knowledge representation, meta-learning and knowledge transfer mechanisms. A successful prosecution of such research will result in a well-grounded (sub-symbolic) computational explanation for several aspects of creativity: *analogy*, *re-representation* and *insight*.

Though this work does not directly address insight, it does suggest a computational explanation for insight as the solution to the problem of deciding, given a set of learned transforms, which will be useful. Given a set Ξ of such transforms, we can approach the question of which to use as a meta-learning problem, the solution to which will result in the system becoming better at having useful insights with experience. One possible meta-learning approach that makes sense in this context is that of landmarking (Pfahinger, Bensusan, & Giraud-Carrier 2000): first build a small set of *landmark transforms*, $\Lambda \subset \Xi$; second, given a set of learning tasks $\mathcal{T}$, quantify how useful each landmark transform is as a re-representation for each learning task, and discover (perhaps by brute-force initially) which of the known transforms is the most useful for each learning task; third, with this information construct a (meta-)training set of the form $\{\upsilon_i, \xi_i\}$, where i runs over all the learning tasks, υ_i is a vector of landmark utilities for task i and $\xi_i \in \Xi$ is the most appropriate re-representation for task i; fourth, train a (meta-)learner to predict ξ_i given υ_i; finally, given a new task r to solve, try each landmark transform $\lambda_j \in \Lambda$ explicitly to discover the utility of doing so, constructing the utility vector υ_r to use

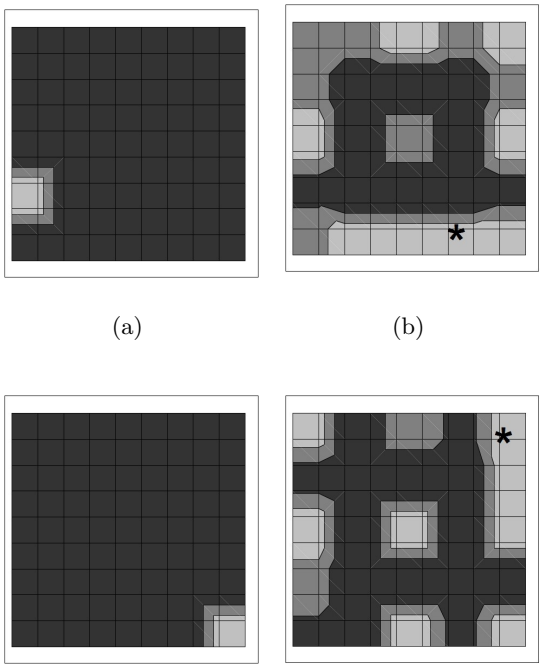


Figure 10: *Sampling the neural map of the "rotation" transformation.* On the left are two different, localized stimuli and on the right, the activation response of the neural field. A canonical rotation would preserve the localized nature of the stimulus and center it on the marked regions of the output field. The pathologically distributed activations indicate that the neural filter is computing a gross approximation of rotation.

as input to the meta-learner, which returns it's prediction for the best re-representation for the new task. As more tasks and more representations are experienced, the system's insight into which (re-)representation will be most useful will improve.

It is interesting to note that this thesis admits both discriminative and generative models. For example, we might construct a (discriminative) model that learns to recognize disguised voices by transferring learning about music transposition. Given various examples of transposed music, the system can learn a manifold that represents transposition, and then learn a sub-symbolic transform that implements it. Later, when asked to recognize a disguised voice, the system can discover that the two tasks are related and apply the (inverse of the) transposition transform to the disguised voice, producing something similar to a known voice.

We also might construct a (generative) model that creates unique aircraft designs by transferring learning about avian anatomy. Given various examples of birds, the system learns a manifold with dimensions representing concepts like wing size, center of mass relative to

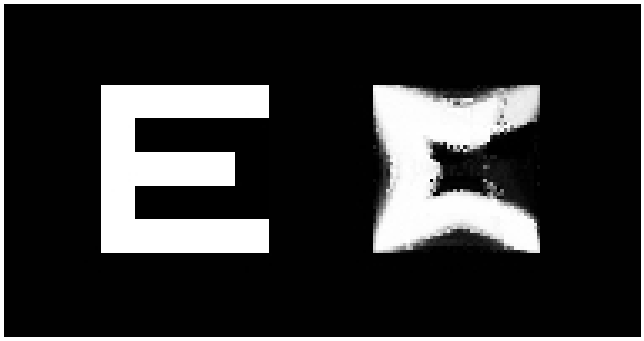


Figure 11: *Transferring warping to E (from A, B, C, D).* The system is trained with A, B, C and D examples of a non-affine transformation that pinches the four sides of the image toward its center. The image on the left is the "canonical" E and the image on the right is the result of applying several iterations of the learned transform.

head, length of tail, feather type, etc., and each of these may be learned as a sub-symbolic transform. Later, when faced with the task of aircraft design, the system discovers that planes and birds are similar and, given a basic prototype design, can generate novel variations to it by applying the various learned transforms. Indeed, given this approach the generative/discriminative dichotomy may be elucidated by whether we are applying learned transforms or their inverses.

Summarizing, we hypothesize that knowledge discovery can be accomplished via manifold learning, knowledge representation can be accomplished via learning transforms implicit in the manifold dimensions, insight can be facilitated by meta-learning that matches transforms to new tasks, and re-representation occurs through applying learned transforms to the new task, resulting in learning transfer (and a system that creatively solves problems using insight and analogy).

Acknowledgements

Thanks to Mike Gashler for helpful discussions on manifold learning for ordering task samples and for producing Figures 5 and 6.

References

Baxter, J. 1998. The canonical distortion measure for vector quantization and function approximation. In Thrun, S., and Pratt, L., eds., *Learning to Learn.* Kluwer Academic Publishing. 159–179.

Ben-David, S., and Schuller, R. 2003. Exploiting task relatedness for multitask learning. In *Proceedings of Computational Learning Theory*, 567–580.

Detterman, D., and Sternberg, R. 1993. *Transfer on Trial: Intelligence, Cognition and Instruction.* Ablex.

Evgeniou, T.; Micchelli, C. A.; and Pontil, M. 2005. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research* 6:615–637.

Gashler, M.; Ventura, D.; and Martinez, T. 2007. Iterative non-linear dimensionality reduction with manifold sculpting. *Advances in Neural Information Processing Systems* 19.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7:155–170.

Hofstadter, D. 1995. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought.* Basic Books.

Intrator, N., and Edelman, S. 1996. Making a low-dimensional representation suitable for diverse tasks. *Connection Science* 8(2):205–224.

Micchelli, C. A., and Pontil, M. 2005. Kernels for multitask learning. In *Proceedings of Neural Information Processing Systems 18*, 921–928.

Miller, E.; Matsakis, N.; and Viola, P. 2000. Meta-learning by landmarking various learning algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 464–471.

Ohlsson, S. 1992. Information-processing explanations of insight and related phenomenon. *Advances in the Psychology of Thinking* 1–44.

Ortony, A. 1993. *Metaphor and Thought Second Edition.* Cambridge University Press.

Pfahinger, B.; Bensusan, H.; and Giraud-Carrier, C. 2000. Meta-learning by landmarking various learning algorithms. In *Proceedings of the International Conference on Machine Learning*, 743–750.

Pratt, L. 1996. A survey of connectionist network reuse through transfer. *Connection Science* 8(2):163–184.

Spellman, B. A., and Holyoak, K. J. 1996. Pragmatics in analogical mapping. *Cognitive Psychology* 31:307–366.

Thrun, S., and Mitchell, T. M. 1995. Learning one more thing. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1217–1225.

Thrun, S., and O'Sullivan, J. 1998. Clustering learning tasks and the selective cross-task transfer of knowledge. In Thrun, S., and Pratt, L., eds., *Learning to Learn.* Kluwer Academic Publishing. 235–257.

Vosniadou, S., and Ortony, A. 1989. *Similarity and Analogical Reasoning.* Cambridge University Press.