In Proceedings of the Conference on Integration of Elementary Functions into Complex Behavior, Zentrum fur interdisziplinare Forschung, Universitat Bielefeld, Bielefeld, Germany, July 12-15, 1994.

Combining Uni-Modal Classifiers to Improve Learning[†]

Virginia de Sa*

Computer Science Department
University of Rochester
Rochester, NY 14627-0226
(email:desa@cs.rochester.edu)

Abstract

One of the key ideas in both robotics and neuroscience is that complex behaviour can arise from the interaction of many cooperating simple agents or modules. In this paper we suggest that this idea can be extended; just as combining simple agents may be important for complex behaviour, combining tasks is important for learning the parts themselves. In particular we show that combining classifications across different modalities can help solve the teaching signal dilemma and allow the development of task relevant classifications without external supervision. We recap some psychophysical and neurobiological data supporting the idea that information from different modalities can assist (or interfere) with classification in another modality and describe a neural network algorithm that is able to take advantage of the structure between the pattern distributions to different sensory modalities to eliminate the need for a teaching signal during training of each network. The algorithm is demonstrated on the problem of learning to recognize speech both acoustically and visually. Simultaneous presentation of moving mouth images and emanating sound waves allows the development of lip-reading and acoustic speech classifiers. The resulting classifiers approach the performance of supervised classifiers without requiring hand-labeling of the training patterns.

One of the key ideas in both robotics and neuroscience (and emphasized in this volume) is that complex behaviour can arise from the interaction of many cooperating simple agents or modules. However, while intelligent behaviour is thought to be obtained from combining modules, it has been common practice to study the development of individual modules or systems in isolation. In this paper we argue that combining modules during learning can help solve the teaching signal dilemma and allow the system to learn without an external teaching signal.

Learning appropriate boundaries in different spaces for classifying objects is a ubiquitous and non-trivial task. While an informative boundary may exist within a modality (or set of related dimensions), learning this boundary from unsupervised pattern presentations may be difficult or impossible. For modeling or designing intelligent behaviour, often an external teaching or labeling signal is required to learn the desired mapping. Various arguments are given to justify providing the required teaching signal. Often in Artificial Intelligence(AI) projects, assumptions are made that other tasks have been solved when constructing algorithms or solutions for the particular task of interest. This becomes especially circular when the referenced solutions have assumed that the first task, or some task dependent on the first, was already solved.

Similarly cognitive models frequently fall prey to the same traps. Thus providing a 'cow' label with cow images in a model learning to recognize animals may be justified with the statement that "an infant is told 'cow' when shown a cow." This of course ignores the point that the spoken word 'cow' is not a useful teaching signal until the auditory system has learned to correctly parse and classify speech signals. (This is immediately apparent to those who have tried to build a machine speech recognition system or even observed sound spectrograms of spoken words.) Similarly, as computer vision researchers are well aware, the cow picture is not a useful teaching signal for the 'cow' acoustic signal until the visual system has correctly learned to recognize cows.

In order to more fully understand development as well as to produce autonomous learning agents, it is necessary to study learning without assuming that other learning tasks have been previously learned — to look at learning in the whole integrated system. We argue that rather than making the problem bigger and more difficult, the greater information available from looking at two or more problems together enables task-relevant solutions without requiring explicit labels during training. Experience is naturally multi-sensory

[†]This research was supported by a grant from the Human Frontier Science Program and by a Canadian NSERC 1967 Science and Engineering Scholarship.

*Current address: Department of Computer Science, University of Toronto, 6 King's College Rd., Toronto, Canada M5S 1A4

and we propose that its multi-modal nature is important not only for dealing with cross-modal integration but for developing single modality abilities.

In this paper we concentrate on the problem of learning to classify signals from two or more sensory modalities. We have developed a simple algorithm that takes advantage of the global structure present in the environment in order to improve classification in the individual sensory modalities. The algorithm uses the structure between signals from two or more modalities to assist in the development of a piecewise-linear classifier within each modality. The structure in natural environments leads to signals that are correlated between the different sensory modalities. For example, hearing "mooing" and seeing cows tend to occur together. So, although the sight of a cow does not come with an internal prescient "cow" label, it does co-occur with an instance of a "moo". The key is to process the "moo" sound to obtain a self-supervised label for the network processing the visual image of the cow and vice-versa. This idea is schematized in Figure 1. Note that this is fundamentally different than running two separate supervised learning systems. The networks actually develop together. Before they have developed into good classifiers, they cannot provide good labels to each other. This algorithm is also different from that in [BH92] as the resulting rule is specifically designed for classification, seems to perform significantly better in this area [dS94c, LSF94][Steeg and de Sa, unpublished data], and does not require backwards propagation of high dimensional vectors (the only feedback information is the output class from the other network).

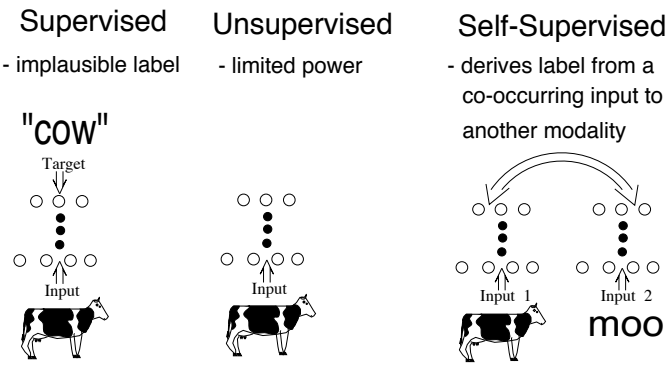

Figure 1: **The idea behind the self-supervised algorithm.**

Most of the paper will deal with the neural network algorithm simulation results, but we start with some biological motivation behind the idea of using the cross-modal structure.

1 Cross-Modal Learning in Humans

1.1 Psychophysics

While we have not been able to find psychophysical experiments that specifically address the issue of cross-modal information changing classification boundaries in the individual modalities, there is a lot of evidence that it is used in ongoing classification decisions. People are very sensitive to correlations between the inputs to different modalities and changing input to one modality has a powerful effect on the classification decisions in another.

One example of cross-modal integration is the improved speech recognition performance when a speaker's face (particularly the lips) is visible [SP54]. The visual signal from the motion of the lips, jaw and tongue help the auditory system to understand the speech. This is particularly useful in noisy rooms where the acoustic information alone is deficient.

The effect of lip movement on speech recognition is even more prominent when the stimuli are experimentally manipulated so that the visual and acoustic signals are discordant. In the experiments of [MM76, MM78] subjects are presented with acoustic stimuli of various consonant-vowel pairs and simultaneously shown images of faces speaking a different consonant with the vowel. Thus for example when presented acoustically with a /ba/ syllable and visually with a face speaking /ga/, 98% of adult subjects hear /da/ [MM76]. The result is very striking and not subject to conscious control. It shows that visual and auditory stimuli are able to interact to produce a unified percept, different from the stimuli actually given to either modality.

By four and a half months of age infants are able to recognize that particular lip motions go with particular sounds. Kuhl and Meltzoff showed that infants looked significantly longer [KM84] at the matching face when presented with the sound /a/ or /i/. Their preference was specific to the actual speech information as they did not show this effect when the speech signals were replaced with tones that followed the duration, amplitude envelope and onset/offset timing of the original speech sounds [KM84]. Furthermore it seems that the ability of visual signals to influence acoustic classification is at least partially learned. Pre-school and school children show significantly less of the McGurk effect than do adults [MM76].

Probably the best example of the type of cross-modal learning discussed below is demonstrated in a novel experiment in [Dur95]. Durgin shows that consistently pairing different tones with different density patterns results in differences in perceived density in the presence of the two tones. The experiment involved repeated brief presentations of random dot patterns in two rectangular areas of a screen. On each presentation, one of the two areas received 25 dots/deg^2 and the other 2 dots/deg^2. The visual presentations were paired with auditory tone stimuli such that the pitch of the tone was perfectly correlated with the side of the denser dot pattern. After 180 flashed presentations, a staircase procedure was used to determine the perceived density equivalence (for test patterns with dot densities between the two trained densities) between the two areas when presented with each of the two tones. The experiment showed that there was a significant effect of the tone on the perceived density relationship between the patterns in the two areas. The simultaneous presentation of the tone associated with a denser texture in one area during training, lead to an impression of greater dot density in that area during testing. To match a constant density, the difference between the density required in the presence of the high pitch and that with a low pitch was 10% [Dur95].

1.2 Neurobiology

The previous section examined results showing that information from different sensory modalities is combined in determining our perception. Often, as for the McGurk effect mentioned above, the combination is not subject to conscious control. It is as if the results are not simply being combined at a high-level output stage but are able to influence each other in the individual processing stages. This is corroborated by neurophysiological studies which have found responses of cortical cells in primary sensory areas that respond to features from other sensory modalities. For example [SSB68] found sound frequency specificity in cells in primary *visual* cortex of the cat and [FM73] found that these bimodal cells tend to be clustered together. As support for the unified percept observed in psychophysical studies, the stimuli are able to affect the same cell. In fact acoustic responses in a single cell could be inhibited by inhibitory visual stimuli [Mor72]. As there are no direct afferent (feed-forward) connections from one input modality to another, the information from other modalities could either be coming bottom-up from shared subcortical structures such as the superior colliculus or alternatively top-down from the multi-sensory integration areas such as entorhinal cortex and other limbic polymodal areas. This idea is suggested in [Rol89] and seems to be supported by the evidence from visual cortex. As stated by [SSB68]

> non-visual stimuli affect the activity of ganglion cells only minutely [SPW65, SW66, Spi67]; they affect that of the geniculate cells to a greater extent [MCBS65] and very markedly affect cortical cells [MCyR65]. Even more interaction appears to be present in prestriate cortex [BB59].

2 The Classifier

Motivated by the data presented in the last section, we hypothesize that one way in which animals may compensate for a lack of teaching signals is through using information in other modalities to assist in learning to classify within another modality. Following the anatomical evidence and under the assumption

that it is infeasible to have neurons receiving input from the sensory transducers of all modalities, we propose an architecture such as that shown in Figure 1 in which each modality has its own processing stream (or classification network) but access to each other's output at a high level. This information can reach the lower levels within each processing stream through feedback within each stream. This feedback, as described below, is simply the output of the other network and does not require implausible propagation of information backwards along connections as is required in algorithms using back-propagation [RHW86] learning.

Each modality is modeled as a piecewise linear classifier and objects are represented as n-dimensional pattern vectors. The piecewise linear classifier is defined in terms of *codebook vectors* which are also vectors in the space of the input patterns. Besides a position in the input space, each codebook vector also has a class associated with it. Patterns are classified as belonging to the class (given by the label) of the closest codebook vector—the classification boundaries are given by the segments of the Voronoi tessellation between codebook vectors of different classes. The learning problem thus is to determine appropriate positions for the codebook vectors so that they define good classification boundaries.

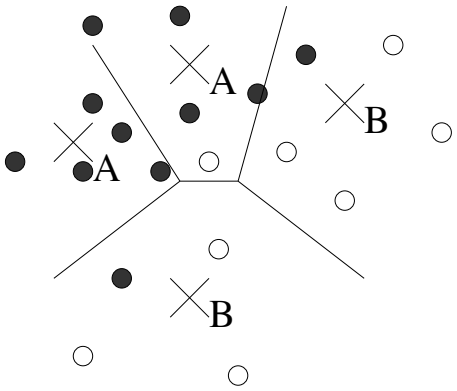

Figure 2: **A Piecewise-Linear Classifier.** The circles represent data samples. The filled circles denote data from one class and the open circles those from another. The X's represent the weight vectors of competitive neurons. The input space is divided between the weight vectors such that data points are assigned to the closest weight vector. The boundaries drawn in this way correspond to the Voronoi tessellation of the weight vectors.

Unsupervised methods of moving the codebook vectors are the Competitive learning and related Kohonen feature mapping algorithms [Gro76, Koh82, RZ86]. In Competitive learning the codebook vectors move to minimize the distance from every input pattern to their closest codebook vector. This tends to move the codebook vectors to the centres of clusters. Kohonen feature mapping is similar except that the codebook vectors have specified positional relationships and nearby ones move similarly. These algorithms work well for clustering and cheaply encoding data but are not optimal for forming classification boundaries.

If the class labels are given with the training patterns, a supervised algorithm can be used to move the codebook vectors to give more appropriate borders for classification. It can be shown [dSB93] that (a slight variant of) Kohonen's supervised LVQ2.1 [Koh90] algorithm minimizes the number of misclassifications in the resultant classifier. However for autonomous robotic applications and models of human learning, a learning algorithm that learns without a supervisory signal is required. The Minimizing-Disagreement (M-D) algorithm [dS94b, dS94a, dS94c]allows the benefit of a task-related labeling signal without requiring an external labeler or backwards propagation of detailed error signals.

The algorithm learns from paired presentations to the two modality nets. Patterns from the same class are presented to each network. For example presentation of a cow image to a visual network would co-occur with a presentation of a moo sound to an auditory network. The outputs from the two modalities provide label signals to each other. This allows the whole system to bootstrap itself.

The initial codebook vectors are randomly picked from the respective input spaces. The algorithm then consists of two stages (which correspond to the two properties of codebook vectors—labels and positions). In the first stage the initial codebook vectors are given labels. This initial labeling stage essentially runs a competitive learning algorithm on the vectors over the codebook vector layer. Codebook vectors that tend to be activated together will increase their connections to the same output neuron. After several iterations the

codebook vectors are given the arbitrary label of the output neuron to which they have the strongest weight. The second and most important stage of the algorithm is moving the codebook vector positions in order to form better classification boundaries (The labels are also updated as required in this stage). As we can't directly monitor the number of misclassified patterns we can't minimize this measure directly as we can in a supervised algorithm. Instead the algorithm monitors the number of disagreements between the outputs of the two networks. The mathematical derivation in [dS94b, dS94a, dS94c] shows that the Minimizing-Disagreement (M-D) algorithm performs the (modified) LVQ2.1 algorithm except instead of using externally provided labels each network uses the label given by the other network (for the co-occurring pattern). The supervision has been replaced by the co-occurrence of patterns from the same class to the two modalities.

3 Algorithm Simulations

As an example of the idea and algorithm we used the algorithm to learn to recognize consonant-vowel utterances both visually and acoustically. We show that by learning both together with no external labels during training, we can do almost as well as supervised trained networks that receive labels with each pattern presentation.

3.1 The Dataset

Data were collected from 5 male English speakers as they spoke 26 iterations [2] of /ba/ /va/ /da/ /ga/ /wa/. Each set of 10 utterances (twice through the set) was preceded by a clap using a clapboard arrangement similar to that used in commercial movie production for matching the visual and auditory signals. The camera recorded 30 frames a second and was positioned to view the tip of the nose through chin of the speaker. The audio was recorded through a cardioid microphone positioned approximately 5 inches from the speaker's mouth.

The acoustic data were low-pass filtered and segmented automatically (using time-domain wave magnitude) using the ESPS software from Entropic Research Laboratory, Inc. Each utterance was taken from 50msec before the automatically detected utterance start to 50msec after[3]. These utterances were then encoded using a 24 channel mel code[4] over 20msec windows overlapped by 10msec. This gave a (24*9 = 216) dimension auditory code for each utterance.

The visual data was processed using software designed and written by Ramprasad Polana [Pol94]. The visual frames were digitized as 64×64 8 bit gray-level images using the Datacube MaxVideo system. The video and auditory tracks were aligned using the clapboard arrangement. Visual detection of the clap was performed manually which allowed alignment to within 1 video frame (1/30 second). (For an example of a video sequence showing a clap frame see Figure 3). The frame of the clap was matched to the time of the acoustically detected clap allowing the automatic segmentation obtained from the acoustic signal to be used to segment the video. The segments were taken as 6 frames before the acoustically determined utterance onset and 4 after. The normal flow was computed using differential techniques between successive frames. Each pair of frames was then averaged resulting in 5 frames of motion over the 64×64 pixel grid. The frames were then divided into 25 equal areas (5×5) and the motion magnitudes within each frame were averaged within each area. This gave a final visual feature vector of dimension (5 frames * 25 areas)= 125.

Figure 3: **Example frames showing the clap detection.** The clap is in the centre frame.

[2] A few speakers spoke more iterations.
[3] This was to ensure that all the consonantal information was retained.
[4] linear spacing below 1000Hz and logarithmic above 1000Hz

Figure 4: **Example /ba/ utterance.**

Figure 5: **Example /va/ utterance.**

3.2 Experimental Results

The data for one speaker were unusable due to a problem with the video tape. The training set was made up of each of the other speakers' first 20 cycles through the utterances (minus a few cycles that could not be used due to lost frames during digitization). The test set was made up of the next 6 cycles [5].

We first benchmarked the dataset by running the supervised LVQ2.1 algorithm. Using 30 codebook vectors for the auditory patterns we achieved an accuracy of 99% on the training set and 97% on the test set. Using 60 codebook vectors for the visual patterns the performance was 83% on the training set and 60% on the test set.

Note that the unsupervised competitive learning (or Kohonen feature mapping) algorithm does not give codebook labels but in order to compare the appropriateness of the placement of the codebook vectors for classification we can determine the optimal labels for the codebook vector positions and measure performance of the resulting classifier. This can be considered a hybrid unsupervised-supervised algorithm and reveals the best that one can do given codebook vectors positioned using no external labels. The hybrid algorithm gave accuracies of 84% and 55% on the two training sets.

Ideally we would like to test the M-D algorithm by presenting to the auditory and visual networks only the pairs of patterns that occurred together. However, to get a good covering of the spaces, many utterances need to be collected. Due to the time involved in the current method of synchronizing the audio and video (they are processed separately and synchronized manually through the visual clap detection) it was decided to run preliminary experiments that artificially increase the dataset by matching each auditory pattern of one utterance with each visual pattern of that utterance in the training set. For example, an individual acoustic pattern from a /ba/ utterance is randomly paired with the visual sample from a randomly chosen /ba/ utterance. This technique makes the assumption that within an utterance class the exact auditory and visual patterns are independent and thus each auditory pattern can be paired with each visual pattern from the same class (not just the one that it actually co-occurred with) because with enough data collection that combination would be possible.

For these experiments, the Minimizing-Disagreement algorithm was applied to codebook vectors resulting from the unsupervised Kohonen learning algorithm instead of randomly initialized ones in the respective spaces [6]. The initial labeling algorithm on these codebook vectors resulted in 72% on the training set and 68% on the test set for the auditory network and 48% (training) and 36% (test) for the visual network. These results reflect the ability of the unsupervised labeling algorithm to give appropriate labels to the initial

[5] For some speakers there were a few extra cycles that were also included.

[6] Initial experiments suggested that this might provide better results but later tests indicated there was not much difference.

Figure 6: **Example /da/ utterance.**

Figure 7: **Example /ga/ utterance.**

Figure 8: **Example /wa/ utterance.**

codebook vector positions and can be directly compared with the optimal supervised labeling algorithm given above (84%,55%) as they were both applied to codebook vectors positioned by the same algorithm.

The real benefit of the multi-modal approach requires moving the codebook vectors. With this, the Minimizing-Disagreement stage was able to greatly increase the classification performance from this initial state to 97% and 72% for the auditory network and 82% and 58% for the visual network. The performance results are summarized in Figure 9.

While the previous results were encouraging, it was important to demonstrate the algorithm in the fully unsupervised way, making no assumptions about independence between the modalities, and using only the cross-modality information sampled from the environment. In order to accurately sample the space, we restricted the problem to that of a single speaker. This speaker repeated 120 cycles of /ba/, /va/,/da/./ga/,/wa/. The first 100 cycles (minus two that lost frames during the digitization) were used as the training set and the last 20 were used as the test set. Again for this cross-modal dataset the M-D algorithm was applied using the results of the Kohonen learning algorithm as the initial codebook vector positions. The algorithm achieved accuracies of 92% (train), 92% (test) on the auditory data and 91% (train), 78% (test) on the visual data. For comparison the supervised LVQ2.1 algorithm, as well as the M-D algorithm using full-pairings as above, were also run on this dataset. The supervised results were 99% (aud-train), 95% (aud-test) and 96% (vis-train), 82% (vis-test). The M-D algorithm using the artificially increased pairing resulted in 98% (aud-train), 95% (aud-test) and 95% (vis-train), 80% (vis-test). These results are displayed in Figure 10. The results demonstrate that for one speaker, the natural lip-sound co-occurrences were enough to give performance within 7% percent of the supervised results on the training set and within 4% of the supervised results on the test set. More data collection is needed to determine if the fully unsupervised algorithm would work on the multi-speaker problem.

4 Conclusions and Hypotheses

We know from psychophysical studies that information from different modalities is combined and that information from one modality can assist or interfere with classification in another. The physiological evidence supports this finding in showing that input to other modalities can influence processing in another sensory pathway. This, combined with the anatomical evidence that shows no direct input from one modality's transducers to another pathway, suggests that perhaps this information is coming top-down through feedback pathways from multi-sensory areas.

We suggest that this multi-sensory integration may be doing more than affecting the properties of developed systems but may play an important role in the learning process itself. Just as lip-reading is a learned classification ability, correlations between inputs to different sensory modalities may affect other classification learning in the individual modalities. The simulation results support this idea that using the global information can be helpful for learning within individual modalities. Performance of the M-D algorithm was within 4% (on the test set) of the related supervised algorithm and significantly better than performance achieved by the related unsupervised algorithm in which the codebook vectors were positioned without using information from the other modality (but labels were obtained optimally).

From a biological perspective the algorithm offers an explanation for why cells in one sensory area also

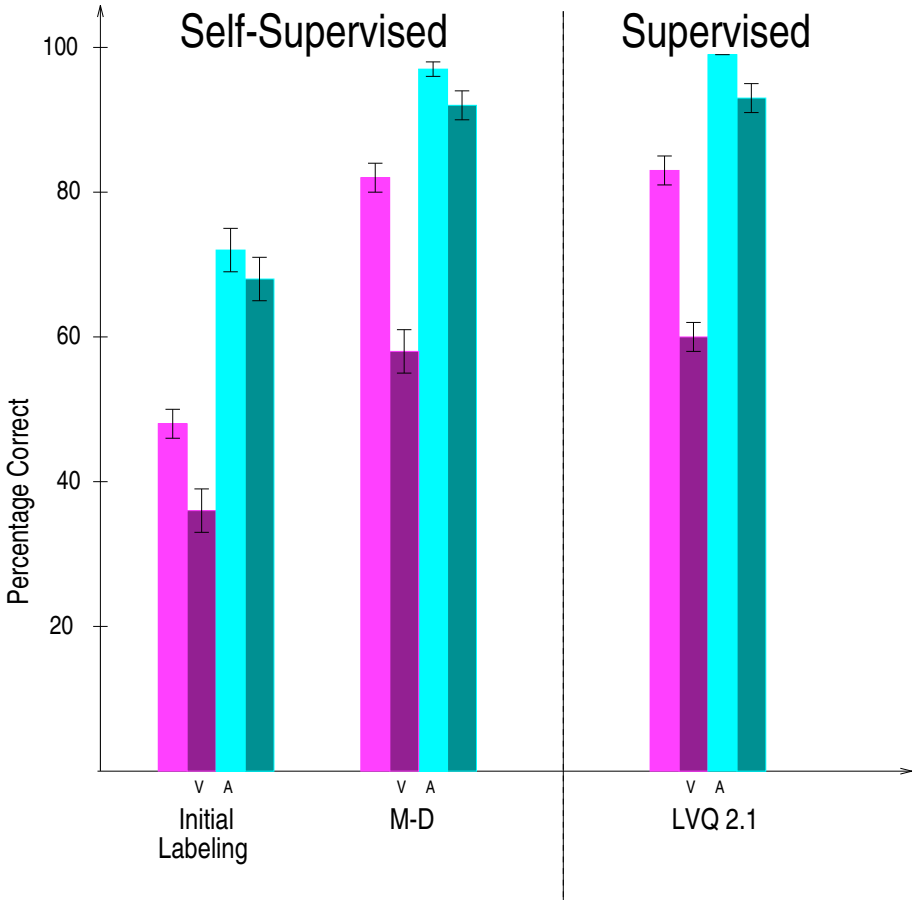

Figure 9: **Results on the preliminary cross-modal dataset.** The two leftmost bars in each set of four give the performance of the visual network and the rightmost bars show the auditory network's performance. Within the two bars for each modality, the lighter and leftmost bar represents performance on the training set. The darker, rightmost bars give results on the test set. The error bars represent 1 standard deviation.

respond to inputs to another sensory modality. We have shown that without connecting neurons to all sensory input we can still take advantage of the greater structure available in the higher dimensional total space of inputs. This occurs through integrating the modalities at a higher level and using feedback connections. This may provide an explanation for the ubiquitous back projections in cortex whose purpose is not yet well understood. In fact [SS82] found that inactivating V2 had little effect on the response properties of cells in V1, indicating that back-projections, or at least the V2-V1 projection has little purpose during regular processing. We suggest that these back-projections are important for the organization of the incoming sensory stimuli during learning.

From a machine-learning point of view, the M-D algorithm enables trainers to avoid the costly hand-labeling of training data, provided that the learning system has access to information from two or more modalities, sub-modalities or points in time, that are providing redundant but not identical information. At a minimum it would allow the learning system to train on patterns labeled verbally by an observer.

Acknowledgments

I am very grateful for helpful comments from Dana Ballard, Peter Dayan, Jeff Dean and Jeff Schneider. Also thanks to Ramprasad Polana and Ramesh Sarukkai for their assistance in obtaining the cross-modal data.

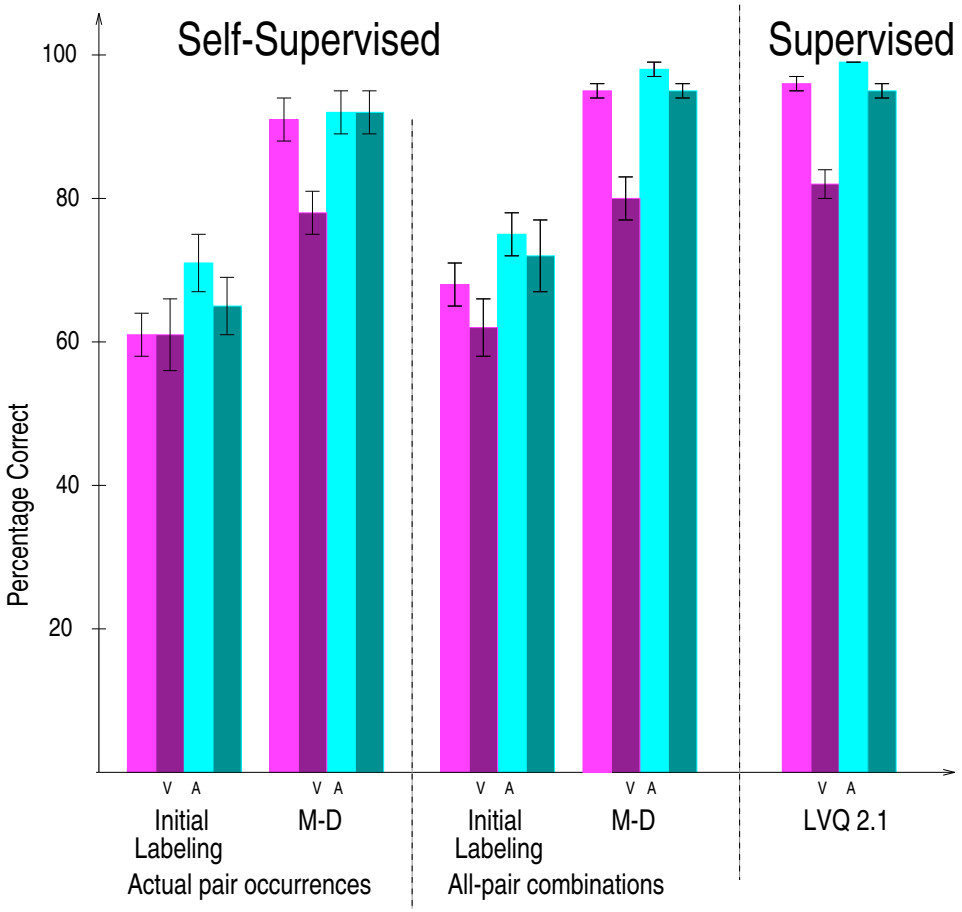

Figure 10: **Results on the single-speaker cross-modal dataset.** The two leftmost bars in each set of four give the performance of the visual network and the rightmost bars show the auditory network's performance. Within the two bars for each modality, the lighter and leftmost bar represents performance on the training set. The darker, rightmost bars give results on the test set. The error bars represent 1 standard deviation.

References

[BB59] P. Buser and P. Borenstein. Responses somesthesiques, visuel et auditives, recuellies, au niveau du cortex "associatif" infrasylvien chez le chat curarise non anesthesie. *Electroencephalog. Clin. Neurophysiol.*, 11:285–304, 1959.

[BH92] S. Becker and G. E. Hinton. A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161–163, January 1992.

[dS94a] Virginia R. de Sa. Learning classification with unlabeled data. In J.D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 112—119. Morgan Kaufmann, 1994.

[dS94b] Virginia R. de Sa. Minimizing disagreement for self-supervised classification. In M.C. Mozer, P. Smolensky, D.S. Touretzky, J.L. Elman, and A.S. Weigend, editors, *Proceedings of the 1993 Connectionist Models Summer School*, pages 300—307. Erlbaum Associates, 1994.

[dS94c] Virginia R. de Sa. *Unsupervised Classification Learning from Cross-Modal Environmental Structure*. PhD thesis, Department of Computer Science, University of Rochester, 1994.

[dSB93] Virginia R. de Sa and Dana H. Ballard. Self-teaching through correlated input. In *Computation and Neural Systems 1992*, pages 437—441. Kluwer Academic, 1993.

[Dur95] Frank H. Durgin. *Contingent aftereffects of texture density: Perceptual learning and contingency.* PhD thesis, Department of Psychology, University of Virginia, 1995.

[FM73] Mark C. Fishman and Charles R. Michael. Integration of auditory information in the cat's visual cortex. *Vision Research*, 13:1415–1419, 1973.

[Gro76] Stephen Grossberg. Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23:121–134, 1976.

[KM84] Patricia K. Kuhl and Andrew N. Meltzoff. The intermodal representation of speech in infants. *Infant Behavior and Development*, 7:361—381, 1984.

[Koh82] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[Koh90] Teuvo Kohonen. Improved versions of learning vector quantization. In *IJCNN International Joint Conference on Neural Networks*, volume 1, pages I–545–I–550, 1990.

[LSF94] Alan Lapedes, Evan Steeg, and Robert Farber. Neural network definitions of highly predictable protein secondary structure classes. In J.D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 809—816. Morgan Kaufmann, 1994.

[MCBS65] M. Meulders, J. Colle, and N. Biosacq-Schepens. Macro and microelectrode studies of somatic responses in the lateral geniculate body. In *Proceedings, XXIII International Congress of Physiological Sciences*, page 364, Tokyo, 1965.

[MCyR65] K. Murata, H. Cramer, and P. Bach y Rita. Neuronal convergence of noxious, acoustic and visual stimuli in the visual cortex of the cat. *Journal of Neurophysiology*, 28:1233–1239, 1965.

[MM76] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, December 1976.

[MM78] J. MacDonald and H. McGurk. Visual influences on speech perception processes. *Perception & Psychophysics*, 24(3):253–257, 1978.

[Mor72] Frank Morrell. Visual system's view of acoustic space. *Nature*, 238:44–46, July 1972.

[Pol94] Ramprasad Polana. *Temporal Texture and Activity Recognition.* PhD thesis, Department of Computer Science, University of Rochester, 1994.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In David E. Rumelhart, James L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, pages 318–364. MIT Press, 1986.

[Rol89] Edmund Rolls. The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampus. In Durbin, Miall, and Mitchison, editors, *The Computing Neuron*, chapter 8, pages 125–159. Addison-Wesley, 1989.

[RZ86] D. E. Rumelhart and D. Zipser. Feature discovery by competitive learning. In David E. Rumelhart, James L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, pages 151–193. MIT Press, 1986.

[SP54] W.H. Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, March 1954.

[Spi67] D.N. Spinelli. Receptive field organization of ganglion cells in the cat's retina. *Experimental Neurology*, 19:291–315, 1967.

[SPW65] D.N. Spinelli, K.H. Pribram, and M. Weingarten. Centrifugal optic nerve responses evoked by auditory and somatic stimulation. *Experimental Neurology*, 12:303–319, 1965.

[SS82] Julie H. Sandell and Peter H. Schiller. Effects of cooling area 18 on striate cortex cells in squirrel monkey. *Journal of Neurophysiology*, 48(1):38—48, July 1982.

[SSB68] D.N. Spinelli, Arnold Starr, and Terence W. Barrett. Auditory specificity in unit recordings from cat's visual cortex. *Experimental Neurology*, 22:75–84, 1968.

[SW66] D.N. Spinelli and M. Weingarten. Afferent and efferent activity in single units of the cat's optic nerve. *Experimental Neurology*, 15:347–362, 1966.