The Role of Input and Target Similarity in Assimilation

Brian T. Bartell † **Garrison W. Cottrell †** **Jeffrey L. Elman ‡**

† Department of Computer Science and Engineering
‡ Center for Research in Language
University of California, San Diego
San Diego, CA 92093

Abstract

We investigate the situation in which some target values in the training set for a neural network are left unspecified. After training, unspecified outputs tend to *assimilate* to certain values as a function of features of the training environment. The roles of the following features in assimilation are analyzed: similarity between input vectors in the training set, similarity between target vectors, linearity versus non-linearity of the mapping, training set size, and error criterion. All are found to have significant effects on the assimilation value of an unspecified output node.

Introduction

We consider here the case in which the target vectors in the training set for a neural network are not completely specified. That is, for certain output units on some input patterns, the desired (target) values may be indeterminate or irrelevant (*don't-care*). During Back Propagation [Rumelhart *et al.*, 1986] the error at these don't-care outputs is zero, so weight changes do not include terms from these nodes.

We define *assimilation in neural networks* to be the act of a don't-care output unit taking on (assimilating to) a value after training. The focus of this work is on examining the characteristics of the training environment which determine a node's assimilation value.

Assimilation in neural networks is interesting because it has been used to model the articulation of words or phrases occurring in natural language [Jordan, 1986] [Hare, 1990]. Our ability to evaluate these models and apply them to other domains relies heavily on our level of understanding of the underlying assimilation phenomenon in the model.

Our interest in assimilation also derives from its involvement with other more general neural network learning issues. It involves an interesting form of generalization, in which the input patterns for the don't-care targets have been seen by the net during training, rather than having been reserved for a separate test phase. How a don't-care node responds depends on how the representational resources of the hidden units have been employed to perform the task. This, of course, would seem to depend on the similarity structure in the input and target domains and other characteristics of the mapping. The value which an output assimilates to may also suggest values which would be easier for the output to learn if it were required to. Thus, understanding assimilation may offer insights on the general learning process.

The Assimilation Effect

We briefly review two works which use assimilation to model articulation data from linguistics [Jordan, 1986] [Hare, 1990]. Each work offers an alternative hypothesis to explain the assimilation results which were witnessed. We will present these two explanations, and offer counter-examples demonstrating that neither is accurate in all cases.

Both works use the same Jordan network model[1] trained to generate a sequence of phonemes representing the articulation of a word or phrase. Phonemes are represented as feature vectors, and some of the features for certain phonemes are left unspecified. For example, the nasal feature for the phoneme subsequence */ria/* might be left unspecified in the articulation of *freon* (phoneme sequence */frian/*) [Jordan, 1986]. After training, this

[1] The network can be viewed as a feed forward network except that a portion of the input vector (the "state") is a function of the previous outputs: the state at time t equals the state at $t-1$ times a decay parameter $\mu, (0 \leq \mu \leq 1)$ plus the output at time $t-1$. See [Jordan, 1986] for a detailed treatment of the Jordan Network architecture.

Time	Input/State							Target						
1	.5	.5	.5	.5	.5	.5	.5	1	0	0	0	0	1	1
2	1.3	.3	.3	.3	.3	1.3	1.3	0	0	0	1	0	0	1
3	.8	.2	.2	1.2	.2	.8	1.8	*/0.9	0	0	0	0	1	1

Table 1: Jordan's hypothesis based on input similarity predicts that the don't-care output ("*") will assimilate to value 0. This is because the third input pattern is more like the second input than like the first input, and the don't-care output at time $t = 2$ is 0. However, assimilation is to 1.

feature might take on intermediate values (between low /f/ and high /n/), indicating that the model predicts anticipation of the nasal feature for the phonemes before the /n/.

Input Similarity

Coarticulation in speech is a phenomenon in which the pronunciation of two phonemes overlaps in time. That is, there can be a blurring of articulation features between phonemes which are nearby temporally.

In Jordan's assimilation model of this phenomenon, don't-care outputs also assimilate to nearby articulated features. Jordan's explanation for this observed assimilation effect in the network is that, typically, similar inputs generate similar outputs. We may justify this *input similarity* hypothesis by reasoning that the function computed by a network is a continuous mapping; therefore, two similar inputs will generate two similar outputs, unless trained otherwise. Additionally, Jordan notes that the sequence of states traversed by the network tends to be continuous in time (that is, nearby states in time will tend to be more similar than arbitrary pairs). He calls this the *continuity property* of the next-state function. This, in combination with the hypothesized input similarity effect[2], predicts the observed results: that don't-care outputs will assimilate to specified outputs that are nearby in time.

Although this hypothesis is useful in explaining Jordan's results, a counter-example can be provided. Table 1 depicts such a case (due to Hare[3]). A don't-care target is denoted by a "*" in the table, followed by the mean assimilation value after training. In this case, an input similarity explanation predicts assimilation to 0 since the third input is more like the second than like the first (by euclidean distance). However, assimilation is actually to 1 ($mean = 0.9$, $stddev = 0.05$, 50 samples).

Target Similarity

In the Hungarian vowel harmony system, suffix vowels will alternate in backness in order to agree with the last vowel of the root. For example, the suffix vowel *a* in *pugo + nak* is a back vowel in order to agree with the back vowel *o*. However, in certain exceptional cases, the last vowel of the root will be *transparent* to the assimilation process. In this case, the suffix vowel will agree for backness with a non-final root vowel, ignoring the final transparent vowel. An example is *taxi + nak*, in which *i* is transparent and the suffix vowel alternates to agree with the back vowel *a*.

In Hare's assimilation model of this phenomenon, the don't-care outputs assimilate to the last vowel of the root in the general case, and also correctly assimilate to the back vowel in the exceptional case. This linguistic data is satisfied without the stipulation of transparent vowels by the model.

Hare's explanation for this observed assimilation effect in the network is that high similarity between target patterns can override the effect of input similarity. Thus, although a suffix vowel will usually assimilate with another vowel that has the most similar input (which, according to Jordan, will be the most recent vowel), the suffix vowel will override input similarity in the case when the suffix and another vowel share many of the same target features. In this case, assimilation is to the similar target, rather than input. A possible justification for this *target similarity* hypothesis is that similar outputs will constrain the hidden unit representation of their inputs to nearby regions in activation space. This is plausible since the hidden unit representation must be linearly separable for the mapping to be learned. Therefore, two hidden unit representations generating very similar outputs may be relatively similar, causing a don't-care output in one pattern to take on the value of the specified output in the other.

As in the case with Jordan's input similarity hypothesis, we can also construct a contra-

[2]Note that in this model, the "state" acts as the input for the mapping performed at each time step. Thus, similar states also generate similar outputs.

[3]Our assimilation results in this case (averaged over 50 samples) differ somewhat from the results reported in [Hare, 1990], although this may be attributable to different initial weights or a larger sample size. Results are achieved with initial state = 0.5, μ (state decay) = 0.6, η (learn rate) = 0.1, and α (momentum) = 0.0.

Time	Input/State							Target						
1	.5	.5	.5	.5	.5	.5	.5	0	0	0	0	0	0	0
2	.3	.3	.3	.3	.3	.3	.3	1	1	1	1	1	1	1
3	1.2	1.2	1.2	1.2	1.2	1.2	1.2	*/0.0	1	1	1	1	1	1

Table 2: Hare's hypothesis based on target similarity predicts that the don't-care output will assimilate to value 1. This is because the target pattern at time $t = 3$ is very similar to the target at $t = 2$, and the output at $t = 2$ for the don't-care node is 1. However, assimilation is to 0.

diction to Hare's target similarity explanation. This case is illustrated in Table 2. Although the target at the third time step is exactly like the second target but unlike the first, assimilation of the don't-care output is to the first output ($mean = 0.0$, $stddev = 0.03$, 25 samples).

Some Factors Affecting Assimilation

We have presented two possible explanations for the assimilation effect in neural networks: assimilation based on similar inputs, and assimilation based on similar targets. Both hypotheses have intuitive justifications and have been used to explain network behavior in the literature. Nevertheless, the counter-examples suggest that other factors are also involved.

We turn now to a set of experiments which examine the effects on assimilation of the following factors:

- input and target similarity,
- linearity versus non-linearity of the mapping,
- the error criterion to halt training, and
- the size of the training set (TS), for fixed hidden layer size.

These factors are considered because they are conspicuous features of the simulations in [Hare, 1990], and therefore may contribute to the observed assimilation effects. Specifically, in Hare's work, input and target similarity are not independently varied while testing the assimilation effect (see section V in [Hare, 1990] – varying the first target in the sequence also varies all subsequent states). Also, the networks tended to be highly trained and have twice as many hidden units as there are patterns to learn. Lastly, all mappings were linear.

We restrict our attention to feed forward binary mappings[4]. In the following experiments, η (learn rate) = 0.2 and α (momentum) = 0.9.

Input and Target Similarity

A 3 x 3 x 3 x 11 experiment was performed in order to test the independent and interactive effects of input similarity, target similarity, training set size, and error criterion, on the assimilation process. For each of the $3 \cdot 3 \cdot 3 \cdot 11 = 297$ cells in the experiment, 50 trials were performed. In each trial, a randomly initialized network was trained on a parameterized randomly initialized training set.

Design We use a strategy in which certain patterns (called Key) in a training set serve as potential "magnets" or sources of assimilation for other patterns (called Don't-Care) which contain unspecified units. The training set is parameterized by the input similarity and target similarity factors (each are one of LOW, MEDIUM, or HIGH), and the training set size (one of 4, 8, or 12). Table 3 depicts an example training set for factor values input=HIGH, target=LOW, and TS size=4. The training set is constructed as follows: 1) random binary vectors are generated for the Don't-Care pattern, 2) the Key pattern is generated such that its input and target vectors are as similar to the Don't-Care pattern as is specified by the input and target similarity factors, 3) Neutral patterns are added to make the training set the size of the TS size factor, and 4) the extra target unit for the Don't-Care pattern is made a don't-care ("*"), the target unit for the Key pattern is made a 1, and for all Neutral patterns it is made a 0[5].

The network is a 3-layer feed forward architecture with 10 inputs, 4 hidden units, and 11 outputs. The network is trained until the error criterion factor is reached (one of 11 mean squared error levels between 2.0 and 0.0002 in roughly a logarithmic progression)[6]. Once trained, the Don't-Care input is presented, and the don't-care output unit is examined. A value near 1 indicates high assimilation to the Key pattern; lower values indicate a decreasing

[4]Note that Hare's simulations can be posed as static feed forward tasks, because all inputs are determined by the initial state and target vectors.

[5]There is a bias introduced by making the extra target unit for all of the Neutral patterns a 0. With many Neutral patterns in the training set, the don't-care output is biased towards 0 more than with fewer Neutral patterns.

[6]A trial is aborted if the network takes many cycles without reaching the error criterion. Therefore, there are not 50 samples in each test cell.

Pattern	Input (10)										Target (10 + 1)										
Don't-Care	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	*
Key	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
Neutral-1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1	1	0	1	1	1	0
Neutral-2	1	1	1	0	0	0	1	0	0	1	0	0	0	1	1	0	0	1	1	1	0

Table 3: An example training set. The Key pattern has HIGH input similarity with the Don't-Care pattern, but LOW target similarity. The 2 Neutral patterns all have NEUTRAL input and target similarity with the Don't-Care pattern.

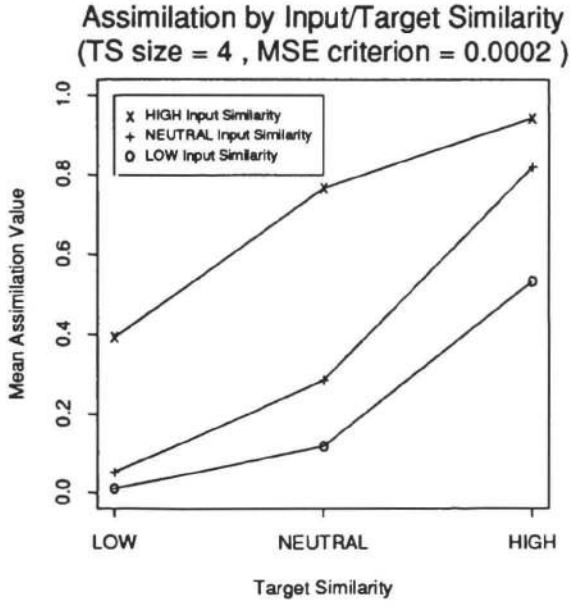

Figure 1: Mean assimilation results by input and target similarity for an underconstrained (TS size = 4) and highly trained (MSE = 0.0002) network.

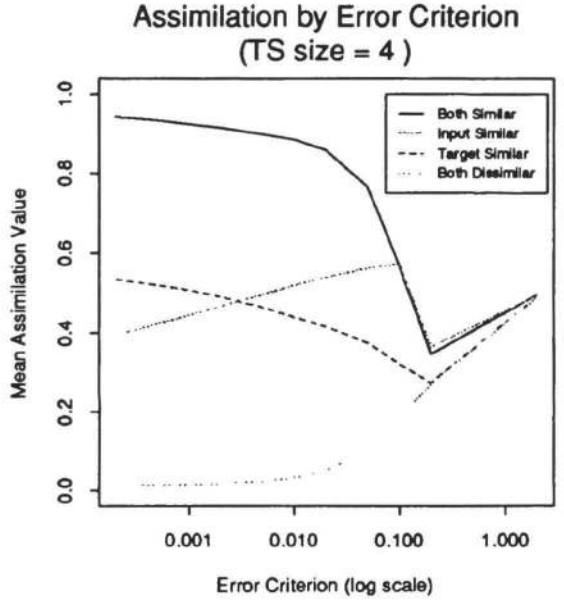

Figure 2: Mean assimilation results by input and target similarity as the MSE criterion varies between 2.0 and 0.0002. High MSE values (corresponding to networks trained for a short time) are on the right; low MSE values (for highly trained networks) are on the left.

assimilatory effect of the Key pattern.

Results All four factors examined as well as all possible interactions between factors have significant effects on assimilation (with $p < 0.001$ for all factors, except size x input and size x input x target with $p < 0.01$; the size=12 training set case is excluded from the analysis of variance because at this level, too few of the networks respond in all variations of the other factors). The single largest effect is due to target similarity ($F = 293.9$), followed by interaction between target similarity and error criterion ($F = 197.2$), input similarity ($F = 99.9$), error criterion ($F = 58.0$), and training set size ($F = 52.0$).

Figure 1 displays mean assimilation values by the 3 x 3 combinations of input and target similarity. Note that assimilation is high whenever one of the input or target patterns is highly similar and the other is neutral or similar. The assimilation response is also above the *a priori* rate (0.33) whenever either input or target similarity is high, regardless of the similarity of the other vector. The assimilation response is actually suppressed (below the *a priori* rate) when one of the input or target patterns is dissimilar and the other is neutral or dissimilar.

Figure 2 displays mean assimilation values as the MSE criterion varies between 2.0 and 0.0002. Initially, input similarity alone has a higher mean assimilation value than target similarity alone, and generally mimics the case in which both input and target patterns are similar. However, below MSE= 0.1 input similarity steadily decreases, and is eventually overtaken by the increasing target similarity case just above MSE= 0.001. Thus, a high degree

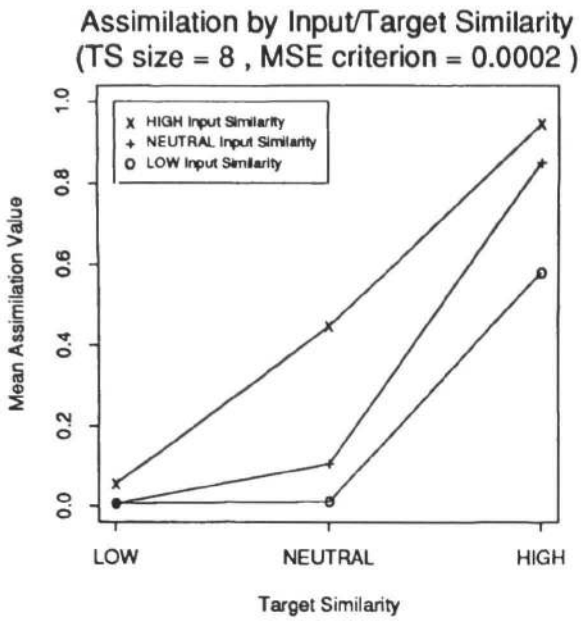

Figure 3: Mean assimilation results by input and target similarity when the training set increases in size (TS size = 8) relative to a fixed number of hidden units (4).

of training seems to accentuate the role of target similarity over input similarity.

Figure 3 displays mean assimilation values by input and target similarity for the case in which the training set has increased in size to 8. In this case, with high target similarity there is some assimilation response regardless of the level of input similarity (and the response gets larger as input similarity increases, as would be expected). However, the same is not true for high input similarity. When input similarity is high but target similarity is low, there is no assimilation response. Thus, the target similarity effect appears to be somewhat robust to an increase in the training set size, whereas input similarity appears less robust.

Non-linear Task Effect

The relationship between assimilation and the non-linearity of the mapping task is tested using the 2, 3, and 4 bit parity problems with varying numbers of hidden units. Only the 2 hidden unit XOR case is reported; more hidden units and 3 and 4 bit parity yield qualitatively similar results.

Ten output units are used. The 10 targets are identically the XOR of the 2 inputs, except that a single output unit is left unspecified for one of the input patterns (00) or (11). See Table 4 for an example. For 50 trials, a network

Input		Hidden		Output (10 units)				
0	0	0	0	0	0	0	...	0
0	1	0	1	1	1	1	...	1
1	0	0	1	1	1	1	...	1
1	1	1	1	*/1	0	0	...	0

Table 4: An example of the XOR mapping, including the hidden unit representations learned during one trial. Although target similarity predicts assimilation to 0, the don't-care output never sees the "boxed" hidden units (11) during training and instead assimilates to 1.

with random weights is trained on this task. 17 of the 50 samples did not solve the problem at the end of 500 epochs and were discarded.

Based on target similarity, we would predict that the don't-care value should take on a value 0. This is because the target vector containing the don't care output is all 0's, and this is identical to another target vector in the training set which is also all 0's and specified on all units. In fact, the opposite occurs. In 30 of the 33 trials (91%), the don't-care output assimilates to 1 rather than 0.

These results can be understood by analyzing the internal representations learned to solve the task. The representations learned during one trial, in which assimilation is to 1, are depicted in Table 4. The don't-care output is never trained on the fourth internal state corresponding to input = (11). Thus, it never sees the first hidden unit on. Because of this, the don't-care output simply learns to be on whenever the second hidden unit is on. All other output units learn to be on if the second hidden unit is on *and* the first hidden unit is off.

An interesting fact occurs in the XOR case: assimilation to similar targets can be encouraged by adding another layer for possible recodings. To demonstrate this, an additional hidden layer with 4 units is added between the first hidden layer and the output layer. Of 74 sampled networks which learn the mapping, 42 (57%) assimilate to 0. Recall 0 is the value which would be expected if target similarity was having an effect on assimilation. 22 (30%) assimilate to 1 and 10 take on intermediate values. Thus, "target similarity" based assimilation can occur in the non-linear case, but it requires an extra layer for recoding.

Non-linear Interaction

It is interesting to examine what happens if similar and dissimilar inputs and targets are placed in a *single* training set. Our experiment to test this has a design quite similar to the 3 x 3 x 3 x 11 design, except that there are no

Neutral patterns, and there are four Key patterns corresponding to the four combinations of similar and dissimilar inputs and targets. The Don't-Care pattern also has four don't-care bits instead of just one, in order to examine the assimilation values in the four cases in a single training set.

Non-linear interactions are observed when similar and dissimilar patterns are grouped in this way. For example, in a small training set (4 Key patterns and 1 Don't-Care pattern), the response when both input and target are similar in the same pattern is larger than the sum of the responses for similar input alone and similar target alone. As the training set size increases (to 12 Key patterns 3 sets of similar and dissimilar inputs and targets), the effect is magnified tremendously. An assimilation response is only achieved when both input and target are similar in the same pattern, and the response is very strong and consistent. In the other cases, the assimilation response is suppressed to 0.0.

Discussion

The results presented here suggest the following framework for understanding assimilation in feed forward nets on random boolean tasks:

- both input and target similarity have an effect on assimilation,
- in a non-linear mapping, positive target assimilation responses are minimal unless an extra hidden layer is provided,
- the assimilation response for the target similarity case appears robust to an increase in the number of patterns in the training set; the input similarity response is less so,
- when the error criterion is high, input similarity has a stronger assimilation response; for low error criteria, target similarity is stronger, and
- when similar and dissimilar inputs and targets are in the same training set, non-linear interactions can occur: in our simulations both input similarity and target similarity together are necessary for any assimilation response (for a large training set).

It is possible to reexamine Hare's work in the context of our results. In particular, the work corresponds to the linear, highly trained, small training set case. We can predict that, had the mappings learned in that work been non-linear, similar target assimilation results would have been achieved only using a second layer of hidden units. Our results also suggest that training a single Jordan Network to generate multiple sequences will result in a minimized role for target similarity or input similarity individually on the assimilation value. Rather, assimilation will tend to coincide with patterns in which both the input *and* the target are similar to the assimilation pattern.

Further work is needed before we can provide a complete explanation of the assimilation effect in networks. The current focus on binary mappings and feed forward networks should be widened. Also, although our training set construction is a convenient formulation, it does not cover the range of possible similarity relationships between vectors in a training set. For a truly complete account, we further need an understanding of the effect of individual training patterns on the *learning process*, perhaps through analyzing the induced error surface.

Conclusion

We have examined two hypothesized explanations from the literature for the effects of similarity on assimilation in networks with don't-care outputs. Both are demonstrated to be incomplete by counter-example. Experimental evidence is provided which suggests factors to be included in a more comprehensive account. These factors are: the similarity of the don't-care pattern to other input and target vectors, the non-linearity of the mapping, the amount of training performed, and the size of the training set for a fixed number of internal units. Results in the literature can be reexamined in light of the current findings. This provides an alternate descriptive framework, and allows predictions of a model's behavior in novel training environments.

Acknowledgements: The authors wish to thank Mary Hare and Mark St. John for useful discussions. The first author is supported by a Cubic Fellowship and Peregrine Systems, Inc., Carlsbad CA., and can be reached at *bbartell@cs.ucsd.edu*.

References

[Hare, 1990] Mary Hare. The role of similarity in hungarian vowel harmony: a connectionist account. *Connection Science*, 2(1), 1990.

[Jordan, 1986] Michael I. Jordan. Serial order: A parallel distributed processing approach. Technical Report ICS Report 8604, Institute for Cognitive Science, May 1986.

[Rumelhart *et al.*, 1986] David E. Rumelhart, Geoffrey E. Hinton, and Ronald Williams. Learning internal representations by error propagation. *Nature*, 323:533–536, 1986.