Target-to-distractor similarity can help visual search performance

Vencislav Popov (vencislav.popov@gmail.com)
Lynne Reder (reder@cmu.edu)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

We found an unexpected *positive* effect of target-to-distractor similarity (TD) in a visual search task, despite overwhelming evidence in the literature that TD similarity hurts visual search performance. Participants with no prior knowledge of Chinese performed 12 hour-long sessions over 4 weeks, where they had to find a briefly presented target character among a set of distractors. At the beginning of the experiment, TD similarity hurt performance, but the effect reversed during the first session and remained positive throughout the remaining sessions. We present a simple connectionist model that accounts for that reversal of TD similarity effects on visual search and we discuss possible theoretical explanations.

Keywords: visual search; learning; similarity; connectionist model; neural network

Introduction

Intuitively, the more similar two objects are to each other, the more difficult it should be to say whether they are the same object or not. Research with the visual search task has confirmed this intuition repeatedly – when a target is more similar to distractors in the search array accuracy decreases and response times (RTs) increase (Treisman & Gormican, 1988; Duncan & Humphreys, 1989; Treisman, 1991), and more errant saccades are made to the highly similar distractors (Bichot & Schall, 1999). Despite the ubiquity of this negative target-to-distractor (TD) similarity effect, in a recent experiment that explored how frequency of exposure affects a variety of tasks, including a visual search task, we discovered by accident a positive TD similarity effect in visual search (Reder, Xiaonan, Keinath & Popov, 2016). We found that greater TD similarity eventually lead to greater accuracy and faster RTs.

The visual search task was performed with Chinese characters over 12 hour-long sessions and the participants were US undergraduates with no previous knowledge of Chinese characters. Interestingly, during the initial stages of the visual search task we observed a negative TD similarity effect, as is expected from prior research, but this effect reversed quickly. After a single training session, higher TD similarity lead to better performance. Since this result was not reported in Reder et al. (2016), we will first describe the experiment and the key results with respect to frequency and similarity.

Method

Participants

Twenty U.S. college students with no prior experience learning Chinese participated in this experiment.

Materials

The stimuli for the visual search task were 64 Chinese characters. We grouped the characters based on their visual similarity in 16 sets of four characters. Characters within a set had a higher similarity with each other compared to characters from other sets. This was determined by a native Chinese speaker and was subsequently confirmed by analyzing orthographic vector representations of the characters (Xing et al, 2004; Yang et al 2009)[1]. We used highly similar distractors in order to force participants to encode the entire character rather than a subset of diagnostic features. For each participant, half of the sets were randomly assigned to the high-frequency condition and were presented 20 times more often during the visual search task.

Procedure

The visual search task was performed over 12 sessions. There were three session per week and each lasted for about 1 hour. Each trial began with a sample character presented in the middle of the screen for 2 seconds. The sample character was followed by a display of 3 to 5 characters. On half of the trials, the display included the target character and participants were to respond whether the target was present. Three of the characters were from the same similarity set as the target character. Additionally, 0-2 characters from different sets of the same frequency class as the target were also present as distractors. After participants made their response, they received immediate accuracy feedback.

Results and Discussion

We analyzed the accuracy data via logistic mixed-effects regressions and RTs via linear mixed-effects regressions, both with participants and items[2] as random intercept effects. All effects discussed below were significant ($p < .05$) as determined by likelihood ratio tests that compared alternative regression models with and without each effect. Most results concerned with effects of frequency are described in Reder et al. (2016; see also Reder et al., 2007); here we focus primarily on the role of similarity.

[1] We thank Xiaonan Liu for pointing us to these representations

[2] i.e., trials with the same target regardless of distractors

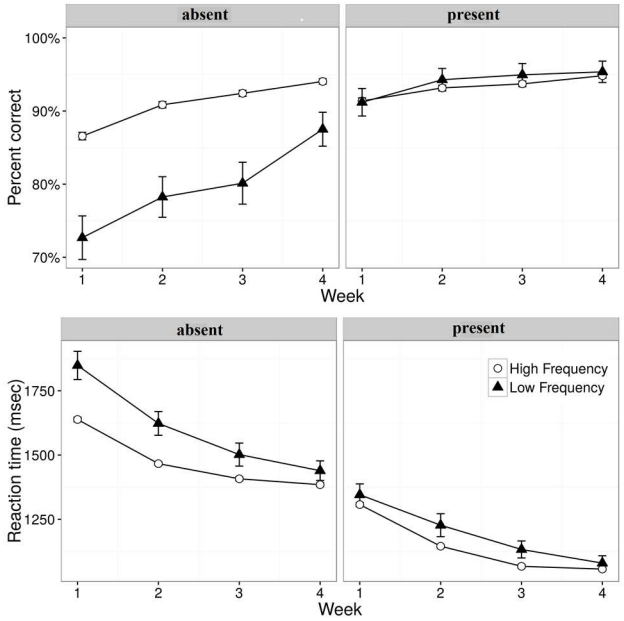

Fig 1. Accuracy and RTs to search the display as a function of target presence or absence, week of training and frequency of exposure.

Figure 1 shows the effect of frequency of exposure on accuracy and RTs for finding the target character. Overall, accuracy was greater and RTs faster for characters from high frequency sets. There was a two-way interaction between frequency and whether the target was present or absent. For accuracy, the effect of frequency was evident only when the target was absent. In Reder et al. (2015), we proposed that frequent exposure facilitates the development of unitized representations of each character. That is, a character seen less often has a weaker chunk representation and is more likely to be encoded as a configuration of some of its features rather than as a single higher-level unit. Thus, when a participant is searching for a LF character the probability of partially matching some of the target's features with features of the distractors is much greater compared to HF characters. This leads to more false alarms in the absent condition, but does not affect the present condition. The interaction was also evident in RTs, although there was still a small effect of frequency in the present trials, likely reflecting the differential efficiency of encoding high and low frequency characters.

A number of previous versions of this experiment had failed to show the hypothesized frequency effects. In those experiments, the distractors in each search array were chosen at random and thus were not very similar to the targets on each trial. In contrast, in the current experiment we ensured that targets were paired with highly similar distractors. We believe that the discrimination required in the prior versions of our visual search task was too easy and as a result, participants were able to perform the task by noting and remembering individual features that distinguish the target from its distractors. As a result, participants did not have to develop stable chunks for each character.

If that is the correct interpretation, then we expected to see an analogous effect within this experiment based on the discriminability (similarity) of the target character to its distractors in the search array. We should see that greater TD similarity leads to a better performance over time, because the increased difficulty in discriminating the target from the distractors forces people to develop stronger and more stable representations of each character as a whole unit/chunk. Note that this prediction is contrary to an intuitive and classic result in the visual search literature – usually, the more similar a target to its distractors, the more difficult it is to perform the task (Duncan & Humphreys, 1989).

TD similarity was calculated based on vector representations obtained from Yang et al. (2009). Each character was represented as a vector of 270 binary features for five dimensions – simple features, shapes, structure, position and strokes. These vector representations are based on an orthographic analysis of the characters and prior behavioral work (Xing et al, 2004). These representations have been already used successfully to model print-to-sound mappings in Chinese (Yang et al, 2009) with a connectionist model similar to those used in modeling English print-to-sound mappings (Harm & Seidenberg, 1999). For each search array, we calculated the mean Euclidean distance between the target and each distractor. Low and high similarity groups were defined as being below or above one SD around the mean similarity of all search arrays.

Figure 2 shows that our prediction was confirmed. During the first session, initially greater TD similarity lead to slower RTs. However, by the end of the first session the effect had reversed and throughout the remaining sessions high similarity between the target and distractors lead to faster RTs and greater accuracy.

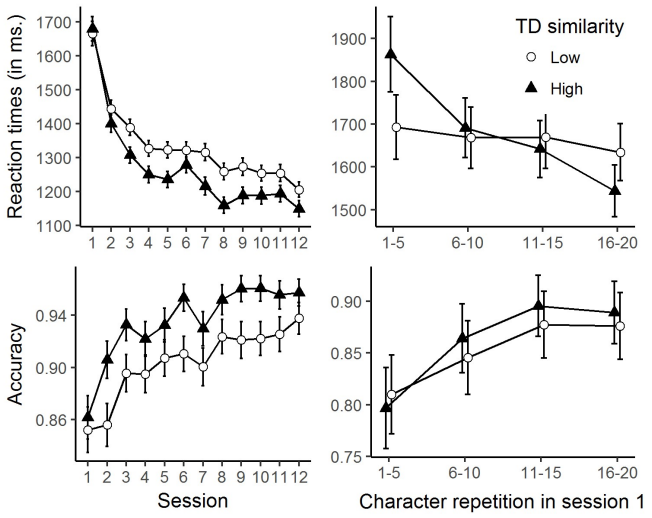

Fig 2. Accuracy and RTs for visual search as a function of similarity of target to distractors within the search array. Right panel shows performance over time during the first session. Left panel shows performance over all 12 sessions.

Contrary to our findings, visual search tasks usually show that high TD similarity leads to lower accuracy and slower response times. Why is it that we found exactly the opposite? A trivial explanation would be that TD similarity in our study was confounded with distractor-to-distractor similarity. The latter consistently shows positive similarity effects. We discounted this explanation by showing that the positive effect of target-to-distractor similarity remained even after controlling for distractor-to-distractor similarity in the regression model.

A theoretical explanation is that most visual search studies use simple stimuli that have pre-existing representations in long-term memory and no additional learning is required. Our study instead used Chinese characters, which are a complex configuration of features for participants who do not know Chinese. Since these characters did not have preexisting representations, participants had to develop them while doing the visual search task. We suggest that those representations were influenced by the demands of the task – to make highly similar patterns more distinct from one another so as to be better suited to support future performance. In essence, we argue that over time when the target is presented along highly similar distractors, the cognitive system builds more distinctive and stable representations of these targets.

Additional support from this argument comes from the fact that (as it was with frequency) the similarity effect is mostly observed in the absent condition (Figure 3, left panel). That is, the benefit from gaining more distinct and stable representations is mostly to prevent the partial matching of shared features between the target and distractors on absent trials.

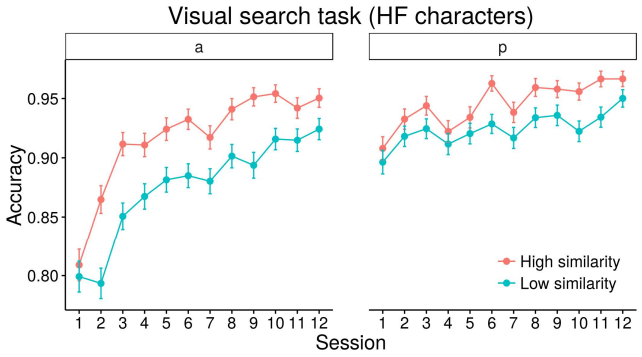

Fig. 3. Effect of TD similarity on visual search accuracy as a function of training session and whether the target was absent (left panel) or present (right panel).

A connectionist model

The reversal of the similarity effect is something of a challenge from a modeling perspective. How exactly are more distinct representations built over time and what mechanism drives that differentiation?

In order to capture the reversal of similarity effects, we decided to apply a novel connectionist model that will be able to perform a visual search task while continuously modifying its internal representations of the stimuli. Connectionist models that represent stimuli as distributed patterns of activity are well suited for exploring the time-dependent changes in the structure of conceptual representations that consist of multiple features. In line with our theoretical interpretation of the data, we expected that in the beginning of training, the model will behave similarly to our human participants and will make more errors for highly similar distractors. However, since this initial behavior would lead to more errors, over time the error-dependent learning might cause the model to alter its internal representations of each stimulus so as to make them more distinct from one another.

In this way, the problem that a connectionist model of this task has to solve is akin to the XOR problem. Specifically, how should the representation of the input patterns be transformed so that similarity is reversed through the transformation? One possibility is for our model to have at least one layer that intervenes between the input layer and a layer that computes similarity between patterns. After a number of failed attempts using a single hidden layer we tried two separate hidden layers, which allowed the network to more gradually change the similarity structure in the input.

The visual search task here requires that participants are able to initially encode the sample character and keep it active in short term memory while comparing it in turn with each candidate character in the search array. To model the task as fully as possible, a model was implemented with a single input layer that can send activation through two different pathways – either to a working memory (WM) module (implemented as a kind of a long short-term memory module), or directly to the comparison layer. This dual pathway represents two ways to use information coming through the senses. One pathway can store the representation of the target in short-term memory and then manipulate it in the absence of the stimulus itself. The other pathway can directly use the incoming information (i.e. the candidates for comparison in the search array).

We assumed that the visual search is performed serially, because the RTs increased linearly with the search array size, and because the slope in the absent condition was twice as large as the slope in the present condition (Treisman & Gelade, 1980). As a simplification, the model presented below will deal only with this serial search case.

Architecture

The network consisted of the following layers (the architecture is presented in Figure 4):

- Input: 20 units
- Hidden1: 15 sigmoid units
- Hidden2: 10 sigmoid units
- LSTM module 1
    - LSTM_Input: 10 linear units
    - LSTM_Buffer: 10 linear units
    - LSTM_Context: 10 linear units
    - LSTM_Output: 10 linear units
    - LSTM_Input_gate: 1 input unit
    - LSTM_Context_gate: 1 input unit
    - LSTM_Output_gate: 1 input unit

- Direct_output: 10 linear units
- Direct_output_gate: 1 linear unit
- Comaprison: 20 units
- Response: 2 softmax units (output layer)

The input was connected in a feedforward manner to hidden1, which in turn was connected to hidden2. We expected the two hidden layers to progressively extract higher order features of the input. Initial weights between these layers were randomized with a mean of 0 and sd of 0.5. Each unit of the hidden2 layer was connected with the corresponding unit in the input layer of the LSTM module, as well as with the direct_output layer with a frozen weight of 1. The same applied to the connections from LSTM_Input to LSTM_Buffer and from LSTM_Buffer to LSTM_Output. Thus, the output of the hidden2 layer was copied forward to the output layer of the LSTM module, and to the direct_output layer. The LSTM module also had a recurrent context layer that was connected bi-directionally to the LSTM_Buffer layer.

The purpose of the four gates was to control the flow of activation through these two modules. There was a fixed negative bias of -1 to the LSTM_input and LSTM_output layers, and a fixed positive connection of 1 with their corresponding gates. Since they were all linear units that were cropped at 0 and 1, when a gate was off, no activity was copied to corresponding and subsequent layers. When a gate was on, it negated the bias and the layer copied the output of the preceding layer and passed it forward.

Network functioning

Each example trial was composed of four events (i.e., presenting different input patterns):

1. Presentation of the target
2. First candidate from the search array
3. Second candidate from the search array
4. Final candidate from the search array

When the target was presented to the input, only the LSTM_input gate was on. Thus, the activity in hidden2 layer that corresponds to the sample input was copied to the LSTM1_input, LSTM_buffer and LSTM1_context layer. All other gates were off, thus preventing the sample input from transferring to the direct_output layer.

When each candidate from the search array appeared, the LSTM_input gate was off, preventing the candidate representation from entering LSTM module. All the other gates were on. This meant several things happened. The candidate representation on hidden2 was copied to the direct_output layer. The representation of the sample that was encoded in the LSTM_context layer on the previous time step was transfered back on to the LSTM_buffer, and from there it was transfered to the LSTM_output. At this point, the network had the hidden2 representation of the sample instantiated on the LSTM_output layer, while the hidden2 representation of the first candidate was active on the direct_output layer.

Both the LSTM_output and the direct_output layers were connected with free random weights with sd 0.1 to the comparison layer, which integrated the representations of the sample and the first candidate input. The comparison layer was connected to the response layer, which consisted of two units - 1 for responding that the two representations are the same, and the other for responding that they are different.

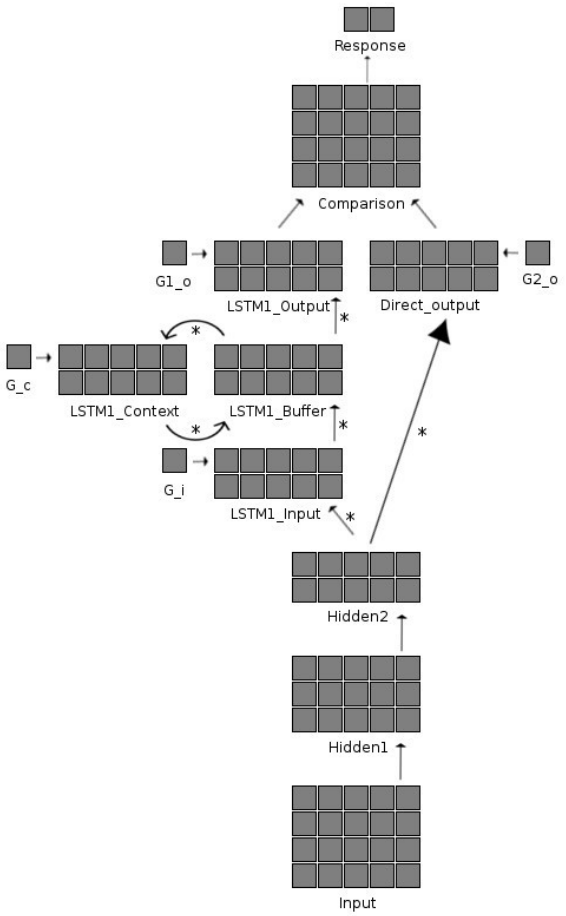

Fig 4. The network architecture. Arrows with stars (*) represent *copy* connections, where each unit in the sending layer is connected with a single connection with fixed strength 1 to the corresponding unit in the receiving layer.

Training

To mimic the experiment's stimuli, 64 input patterns of length 20 were created with binary values that were grouped into 16 sets, which had greater similarity within sets than between sets. On average, 50% of the features in each input vector were "on". The randomization and conditions were equivalent to those in the experiment.

Mean similarity in a set was calculated using Euclidean distance. The groups in the lower 25% quantile of the distance distribution were designated as "Low distance / High similarity" sets, while groups in the higher 25% quantile were designated as "High distance / Low similarity" sets.

When a distractor was present, the network was trained to activate the "mismatch" response unit, while when a target was present it was trained to activate the "match" response unit. Therefore, the goal of the network was to discover a suitable combined representation in the comparison layer such that it will be able to discriminate when the LSTM_output and the direct_output layers had the same or different patterns of activation.

We used a back-propagation training algorithm with a learning rate of 0.01 and a momentum descent with a momentum rate of 0.9. The network was trained for 4000 passes through the training set and the weights were updated at the end of each pass. After every 100 updates, we recorded the output activation of the hidden1, hidden2 and the response layer.

Results and discussion

Frequency effects. The main results of the simulation are presented in Figures 5 and 6, which show the activation of the "match" response unit over training time. Since the response layer had softmax units this value can be directly interpreted as the proportion of "match" responses the network would give in response to a pattern. In Figure 5 we can see that the training patterns that were presented more often lead to greater accuracy.

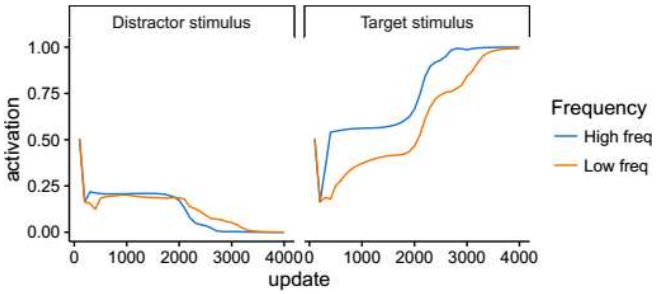

Fig 5. Activation of the *Match* output unit as a function of training time, stimulus type and frequency of the input pattern.

Several things should be noted about this pattern. We can see that initially the network deactivated the match response unit for both types of stimuli. We can also see that overall the effect of frequency was much greater on target stimuli compared to distractor stimuli, which is exactly the opposite effect than the one we found in the behavioral data. This was probably because there were 5 times more distractor items than target items (3 in absent conditions, 2 in the present conditions). In the actual experiment, this too was the case, but participants got feedback only for their final response, thus they had equal amounts of "present" and "absent" feedbacks. On the other hand, the network was trained as if each individual comparison required a response, which causes the discrepancy between distractor and target stimuli.

Thus, while the network captures the overall effect of frequency, its training regimen causes it to miss the specific pattern of frequency for different types of stimuli. This could possibly be solved by considering the current response layer to be an internal response, reflecting whether there is a match or not. Then a secondary motor response layer can be added which outputs a 'present' response if the internal match response is higher than a threshold, or stays inactive until all candidates have been compared. If by the last one none of them had elicited a match response, it produces an 'absent' response. In this way the network would reflect the actual behavior more closely, and weight updating would be affected only by the final response in each example.

Similarity effects. As can be seen from Figure 6, initially the network performance is better for input stimuli that are less similar to their distractors. This is a normal behavior of connectionist networks, and it is also what is expected by previous behavioral data from the visual search paradigm (Duncan & Humphreys, 1989). However, after about 2300 weight updates the effect reverses and stimuli that are closer to each other in the input space lead to better performance. Importantly, this reversal happens very shortly after the behavior of the network starts to approximate the behavioral result levels (~70% accuracy), which is exactly the pattern we have seen from the behavioral session - greater similarity impairs performance during the first session of training, but the effect reverses by the end of that session. Indeed, if we limit our attention to the window between updates 2200 and 3000, which is immediately after the pre-training, and before the performance saturates at ceiling, there is a close correspondence between the network performance both in terms of frequency and similarity structure.

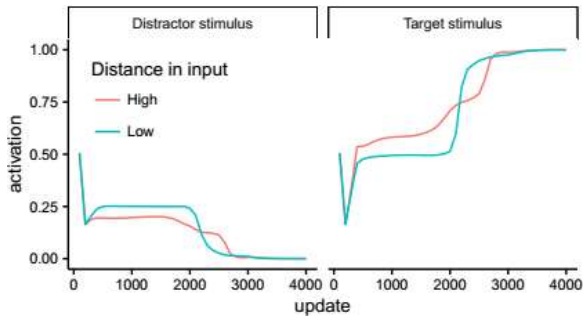

Fig 6. Activation of the 'Match' output unit as a function of training time, type of stimulus and Euclidean distance between the target and distractor input in each array.

What could be causing this reversal of the distance/similarity effect? A possible answer comes from examining the input-output mappings, as well as the hidden representations the network develops during training.

If we split the candidate input patterns into targets, similar distractors and dissimilar distractors, then the network is supposed to produce the following outputs. For targets, which are identical to the sample item (thus 0 distance or perfect correlation) the network has to produce a match response, but for distractors that are highly similar as well, it has to produce mismatch responses. Thus, a major conflict during training comes from the fact that when distance is high, the network has to produce only one type of response, but when it is low,

it has to either respond with a match or a mismatch. One way to achieve these contradictory goals would be to develop such hidden representations of the input that cause highly similar patterns to be represented as less similar to each other.

To test this explanation we looked at the distance between the sample item and its distractors in activation patterns in each of the two hidden layers, split by the distance in the input layer. In right panel of Figure 7, we can see that in the first hidden layer the distance structure in the input has been preserved. In the second hidden layer, however, in the beginning there is no difference in distance due to the two layers of random weights and the sigmoid nature of the stimuli. As training progresses, stimuli that were low in distance in the input and the first hidden layer become more distant to one another, compared to stimuli that were highly distinct to begin with.

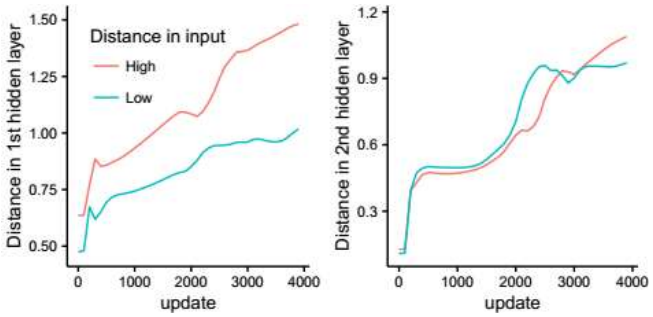

Fig 7. Distance between the hidden layers representations of the target and the distractors in each training set as a function of training time. Left panel shows distance in the first hidden layer, right panel shows the second hidden layer.

General Discussion

The current paper present preliminary data on a novel counter-intuitive finding that the usual target-to-distractor similarity effect in visual search reverses after a short training with previously unfamiliar Chinese characters. Namely, while targets that are highly similar to distractors in a search array are usually more difficult to detect, when the stimuli are complex visual objects, this effect reverses after about 20 repetitions of each object as a target. We propose that visual discrimination and learning interact in such a way that greater difficulty in discriminating the stimuli causes the development of more distinct and stable representations.

To test this idea of differentiation in the character representation over time, we fit a novel connectionist model. When it comes to frequency, the network successfully captured the overall effect that more frequently exposed stimuli led to better performance (although see the preceding discussion for some limitations). Theoretically, this was presumably because low frequency made it more likely that people depend on representing the characters as a configuration of features, rather than on its weak chunked representation. This caused them to be more likely to partial match constituent features and confuse distractors with targets. In contrast, the network showed exactly the opposite effect, because distractors were present 5 times more than targets and had a greater influence over the weight updates.

The most interesting aspect of the model is that it was able to successfully capture the reversal of the similarity effect on visual search performance. It achieved this by transforming the input through multiple hidden layers, which allowed it to change the similarity structure in the input so that highly similar distractors became more and more differentiated in the second hidden layer as training progressed.

This explanation was further supported by a model that involved direct connections from the input to the comparison layer without hidden layer representations (not shown here). This model did not show the similarity reversal effect. This model is analogous to performing the task without having to develop novel representations. One novel prediction from the comparison of these two models and task versions would be that people who learned the Chinese characters under a visual search task would rate highly similar characters as less similar after the training.

Finally, while we simulated the input patterns in this model to resemble as closely as possible how our stimuli were structured, the simulation results might be specific to the interaction between the model architecture and the generated stimuli. Initial modeling results using the actual 270-length vector representations of the Chinese characters show the same pattern as the simplified model presented here.

References

Bichot, N. P., & Schall, J. D. (1999). Effects of similarity and history on neural mechanisms of visual selection. *Nature Neuroscience*, 2(6), 549–554.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, 96(3), 433–458.

Reder, L. M., Paynter, C., Diana, R. A., Ngiam, J., & Dickison, D. (2007). In B. Ross & A. S. Benjamin (Eds.), *The Psychology Of Learning And Motivation* (pp. 271–312). New York, NY: *Academic Press*.

Reder, L. M., Liu, X. L., Keinath, A., & Popov, V. (2016). Building knowledge requires bricks, not sand: The critical role of familiar constituents in learning. *Psychonomic Bulletin & Review*, 23(1), 271–277.

Treisman, A. (1991). Search, similarity, and integration of features between and within dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 17(3), 652–676.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95(1), 15–48.

Xing, H. B., & Li, P. (2004). The acquisition of Chinese characters: Corpus analyses and connectionist simulations. *Journal of Cognitive Science*, 5(1), 1–49.

Yang, J., McCandliss, B. D., Shu, H., & Zevin, J. D. (2009). Simulating Language-specific and Language-general Effects in a Statistical Learning Model of Chinese Reading. *Journal of Memory and Language*, 61(2), 238–257.