Hebbian Artificial Neural Network Model for Stroop Effect

Vadim Kulikov

April 24, 2018

Abstract

1 Introduction

This paper introduces a new model for the Stroop effect. As is common in literature I often refer to the Stroop effect as just "Stroop". This is a theoretical model which gives theoretical predictions, but it can also be implemented as an artificial neural network (ANN) thereby giving rise to a computational model. This model is based on the assumption that cognition is in a sense modular: that the ink colour is processed by a visual colour processing module, written words are processed by a visual verbal processing module while spoken words are processed by a motor verbal processing module. The connections between these models are assumed to arise in a Hebbian manner: a connection between "red" in one modality to the "red" in the other is determined by the frequency with which they have co-occured in the past. Philosophically this model is supportive of the view that semantic meanings (of say "red") emerges from the interaction of different projections of "red" to different modalities (Kulikov, 2015).

It will be shown how this model explains a variety of aspects of Stroop which are individually explained by separate models in the literature, but which, to the knowledge of the author, are not all explained simultaneously in any of the existing models. These aspects are

1. **Task asymmetry** When the response is given verbally, i.e. the name of a colour is pronounced, then there is little to no interference by incongruent ink colour in the reading task while there is a significant interference by the incongruent words in the colour naming task (Stroop, 1935).

2. **Facilitation-interference asymmetry** Facilitation of congruent stimuli is smaller than the interference by the incongruent stimuli (as compared to neutral stimuli).

3. **Reverse effect** If instead of verbal the response is given by finger-pointing to a colour patch or placing a card with the stimulus on it into a box signified by a colour patch, then the above described assymtery is reversed: matching a colour word to the colour patch is slowed down significantly if the ink colour is incongruent while matching the ink colour to the reponse becomes almost effortless (Durgin, 2000; Virzi & Egeth, 1985).

4. **Semantic interference** Interference between word and colour can occur even when there is no response incongruency. For example if blue and yellow ink colours are both associated to the same button press response, then "yellow" written in yellow ink will elicit a faster response than "blue" written in yellow ink (van Veen & Carter, 2005).

5. **Response interference** On the other hand, if the situation is as in the above scenario and "red" is written in yellow ink while red ink colour is associated with a different button press response, then the interference is even higher than when "blue" is written in yellow ink. This extra interference is interpreted as *response interference* (van Veen & Carter, 2005).

6. **Influence of the percentage of congruent trials** The more congruent trials there are mixed within the incongruent ones the more interference the incongruent trials elicit (Zajano & Gorman, 1986).

7. **Impact of automaticity** Often various aspects of Stroop are attributed to the idea that some processes are more automatic than others. For example text-recognition is considered to be more automatic than colour recognition which in turn is more automatic than the recognition of newly learned shapes. The Stroop effect for the pair (colour; newly learned shapes) was shown to work in an isomorphic way as for the pair (text; colour). (MacLeod & Dunbar, 1988)

8. **Output modality** There is less interference when the response is manual (button press) than when it is oral (in the setup of item 1 above).

9. **Other modalities** Not only what the word means, but also its orthography, how it sounds and its associations contribute to Stroop. A number of studies has been also conducted with non-visual stimuli such as e.g. auditory (Green & Barber, 1981; Dennis & Newstead, 1981).

The above cited papers mostly concentrate on one or two phenomenona and are not presenting collective explanations of all the phenomena at once. MacLeod (1991) attempted to collect all empricial and theoretical results upto date about Stroop and presents 18 "major empirical results that must be explained by any successful account of the Stroop effect" in the appendix. From the above list all except 4 and 5 are in the list of MacLeod while 4 and 5 are newer observations.

The computational models of Stroop presented in the literature that are closest to the present model are the translational model (Virzi & Egeth, 1985) and the parallel distributed processing (PDP) model (Cohen, Dunbar, & McClelland, 1990). The main aspect of those models, which is also central to the present Hebbian NN model, is that it emphasises the importance of the *translation* between modalities rather than *processing* within a single modality. To illustrate this point, consider the original explanation by Cattell (1886) of the fact that words are read more rapidly than colours are named: He suggests that this is because reading a word is "automatic", because the "association ... has taken place so often" but naming a colour requires a "voluntary effort to choose a name". From his usage of the word "association" and the phrasing of the last quote "...to choose a name" it is evident that he is talking about association between the reading and speaking modalities in the first part and between the colour recognition and speaking modalities in the second part. However, one can also consider processing within a single modality and ask whether colour processing or colour recognition (not colour *naming*!) all by itself is a slower process than word processing or word recognition. There seems to be some confusion between those types of cognitive processes in the context of the Stroop effect and the discussion of "automaticity". Is automaticity referring to *within* or *between* modality processing? For example MacLeod (1991) writes

> . . .the automaticity account, which was rooted in Cattell's (1886) work. . .. Here the basic idea is that processing of one dimension requires much more attention than does processing of the other dimension.

which seems to talk about *within* modality processing and then continues:

> Thus, naming the ink color draws more heavily on attentional resources than does reading the irrelevant word.

But "naming" and "reading" are already *between* modality processes: from visual to verbal. In any circumstances it is clear that colour recognition within the colour domain is *very* automatic, because to say that the processing of colours *within* the colour domain is not "automatic", "voluntary" or "attention requiring" is to say that when you open your eyes, you see in black-and-white unless you specifically pay attention to the colours which is obviously problematic. Thus, I suggest to focus on the "automaticity" or "strength" (Cohen et al., 1990) or between modal

processing or translation as the computational models mentioned above do (Virzi & Egeth, 1985; Cohen et al., 1990). Let me briefly present these two models.

2 Translational Model

The translational model by Virzi and Egeth (1985) assumes that there are separate cognitive systems processing different types of stimuli. For example in the classical Stroop task the linguistic system and visual colour processing systems are assumed to be involved; denote these two systems by L and C for linguistic and colour respectively. The stimulus is processed by both: the word by L and the colour by C. The output is verbal, so it takes place in L. Therefore if the task is to read the word, then there is no need for translation from C to L, the output being in the same modality as the relevant input. If the ink colour needs to be detected, then there is a need for translation from C to L, because the relevant stimulus is in C, but the output needs to be produced in L. This accounts for clause 3, the asymmetry in the Stroop interference, which is then reversed, if the output has to be done in C. Of course, we do not have a natural modality for "colour output", so Virzi and Egeth design an experiment where both outputs have a similar status. Subjects were required to perform a card sorting task. Stimulus printed on the cards was either a colour word printed in black, a string of coloured X's or a colour word printed in a congruent or incongruent ink colour. The task was to sort the cards either according to ink colour or the word meaning into two bins which were either labelled by a colour word in black ink or a colour patch. The results they obtained were symmetric: when the bins were labelled by colour word, then the task where incongruently coloured words needed to be sorted by ink colour was the only task with significantly increased RT's and when the bins were labelled with colour patches, then the task where incongruently coloured words needed to be sorted by word meaning was the only task with significantly increased RT's.

Virzi and Egeth performed three other experiments also confirming the translational hypothesis and included a brief literature review showing that this model can also account for previously obtained data.

Drawbacks The model does not explain the semantic nor task interference (clauses 4 and 5), or at least the difference between them (van Veen & Carter, 2005; Stirling, 1979). Further, it does not explain the interference effect in a situation where both stimulus modalities are non-verbal, but the output is verbal as in the study of MacLeod and Dunbar (1988) nor the fact that the interference is sensitive to the amount of congruent trials within the stimulus sequence (clause 6). This model doesn't account for the facilitation-interference assymetry either (clause 2).

3 Parallel Distributed Processing Model

Cohen et al. (1990) develop a model which can be thought of as an extension of the translational model, although they didn't explicitly state it in the paper (they don't cite Virzi and Egeth (1985)). It is based on *backpropagation*, an idea stemming from machine learning and computational cognitive neuroscience. They call it *parallel distributed processing model* (PDP) because the processes of colour and word recognition are assumed to occur in parallel in a system where information is distributed over multiple units (neurons). Their model is a layered feed-forward artificial neural network (ANN) with three layers. The output-layer contains a neuron for each possible colour. The input layer is divided into three groups of neurons I_{ink}, I_{word} and I_{task} and the hidden layer into the two groups H_{naming} and $H_{reading}$. The group I_{ink} contains a neuron for every possible ink-colour and I_{word} a neuron for all possible colour words. The group I_{task} contains two neurons which specify whether the task is to name the ink colour or to read the word. The neurons in I_{ink} are only connected to neurons in H_{naming} and neurons in I_{word} only to neurons in $H_{reading}$. The neurons in I_{task} are connected to all hidden neurons and all hidden neurons are connected to all output neurons. The ANN is trained to "read" and to "name" colours using backpropagation algorithm: "reading" means that when the neuron in I_{task} corresponding to the reading task is activated together with the neuron corresponding to the word "green" in I_{word}, then the output neuron corresponding to green should also activate after the forward-propagation and so on, Figure 1.

Then, when confronted with the actual Stroop task, the input is administered from both sites, I_{ink} and I_{word} corresponding to the situation where the participant is presented with a word written in some ink colour. Then this input is forward-propagated through the ANN so many times as is needed for one of the output neurons to cross a given threshold. Once this happens, then the output is obtained. The number of needed iterations is supposed to model the reaction time.

They showed that if the network is trained a lot with the reading task, but only a little bit with the colour-naming task, then the standard Stroop effect and the related asymmetric pattern can be replicated.

This model replicates of course also the findings of MacLeod and Dunbar (1988) where different shapes were presented in colours, because this backpropagation model is blind to what are the actual stimuli, because it is a matter of interpretation of the neural activations which makes sense only from the "outside" point of view. The pathway which is trained the most is going to exhibit "more" automaticity or strength than the other. According to this view

> [t]he speed and accuracy with which a task is performed depends on the speed and accuracy with which information flows along the appropriate

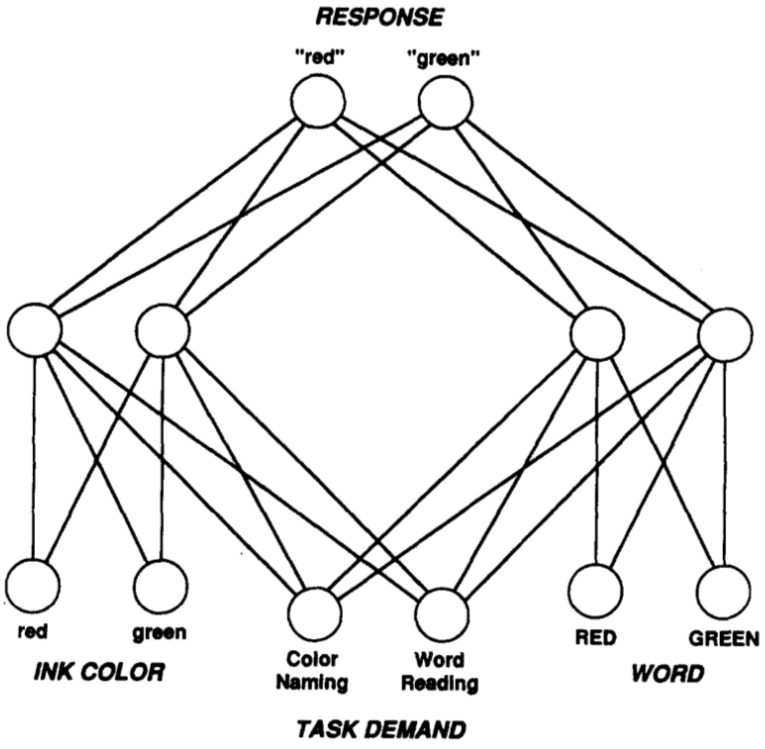

Figure 1: The network architecture of the PDP model of Cohen et al. (1990).

processing pathway. (Cohen et al., 1990)

Drawbacks The drawbacks of the PDP models are the same as of the translational model, to which one might add that such a neural architecture is not biologically nor cognitively very plausible (e.g. Bekolay, 2011).

4 Hebbian ANN Model

We present the model first theoretically without going to those mathematical and computational details that are necessary for simulation and precise calculations. This will be sufficient to derive most predictions of the model.

The model assumes that a fixed number of modalities is involved in the task. For example in the classical Stroop task the modalities would be "text reading", "colour preception" or "colour processing" and "speech". In the modified task where the output is produced by pressing a button, "speech" would be replaced by "button press". Each modality consists of a number of neurons which correspond to the possible inputs in the input modalities or to outputs in the output modalities. Again, in the classical Stroop task, if the colours involved are red, green, blue and yellow, then each modality consists of four neurons each neuron corresponding to one colour. Diagram on Figure 2 shows a version for two colours, green and red.

The connections between the neurons are weighted and these weights signify the strength of the connection where *strength* is as in the graded automaticity theory of Cohen et al. (1990). It can also be viewed as the strength of association. This is why it will be modelled with a simple incremental Hebbian learning, and not with the quite complicated backpropagation as done in the PDP model of Cohen et al..

In the simplest possible situation, the Stroop's original experiment with colours and words as percepts and verbal output modality, the ANN would look like depicted on Figure 2. Notice that apart from the hidden layer it is very similar to the one of the PDP model, Figure 1.

One of the differences to the PDP model is that there is also a connection between the colour processing and visual text processing modalities reflecting the fact that people also learn associations between written words and colours.

4.1 Theoretical Analysis

There are many ways of specifying the model. It is a matter of taste whether to make it into a continuous dynamical system where the states of the neurons can get any values between some limit points with continuous time dynamics or to make it

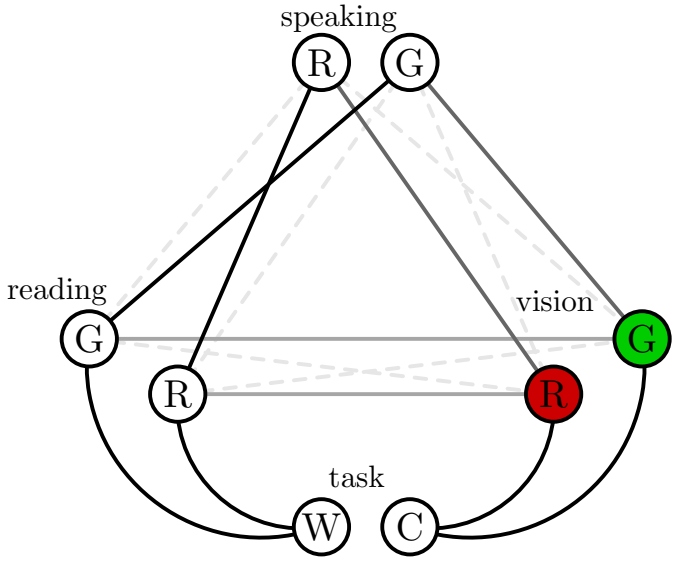

Figure 2: The architecture of the ANN to model the classical Stroop effect. The darker lines correspond to stronger associations while the dashed lines correspond to no or even negative association.

discrete following just a finite number of time steps; or whether to add statistical parameters or not. A robustness claim, however, would be that all of those will give equivalent results. We will begin with a fairly simple model specification which is easy to analyse and then we will proceed to justify the robustness claim and why this simple discrete model is, in fact, giving equivalent predictions to continuous and statistical versions of the model.

Given an experimental setup, we will model it as follows. For each modality involved (usually two input modalities like colour and written text and one output modality like speaking) there is a set of neurons. Each neuron corresponds to a possible intput/output of that modality (usually it is the same set of colours in each modality). The neurons are *linear*; the state of each neuron is the sum of all its inputs and its own previous state, which means that it integrates all the inputs so far. The output is determined in a more complex way, see below. Then there are connections between neurons in different modalities whose strengths correspond to the notion of the connection strength is from Cohen et al. (1990). It can be experimentally measured with an experiment where stimuli from one modality are presented and the answers have to be presented in the other modality. For example the first experiments by Cattell (1886) suggests that the connection between seeing a written word and speaking it is twice as strong as the connection between seeing a colour and speaking it. Additionally there will always be one "relevant" input

modality, namely the one whose inputs have to be reported in the output modality. This is the information received by the participant from the experimenter. All the inputs into the relevant modality are multiplied by the relevance factor r. The model is initiated so that the activation of each neuron is 0. At the first time step the input is delivered to the input modalities. At the following time steps exactly one neuron fires in each modality. It is the neuron whose activation is the highest within that modality and the strength of the output is the difference between its and the second strongest active neuron. For example if there are four neurons in modality A and their activations are $(1, 0, 3, 0)$, then the output is produced by the third neuron and equals to 2, i.e. the output of the modality looks like this: $(0, 0, 2, 0)$. At a given timestep (in the simplest case, after the third one) the reaction time is evaluated as the inverse of the difference between the highest active and the second highest active neurons in the output modality. For example if the neurons in the output modality have activations $(2, 1, 5, 1)$, then the output is produced by the third neuron (whose activation is 5) and the time needed to produce it is $1/3$. This is the whole model. In short:

1. Each neuron's activation is the sum of all its inputs so far.

2. Weights between neurons in different modalities are determined by a Hebbian rule which is consistent with the idea of strength of automaticity as in (MacLeod & Dunbar, 1988) when interpreted as a strength of connection or translation between two modalities and not as automaticity within a modality.

3. A neuron's output is 0 if it is not the most active neuron in its modality. Otherwise it is $\alpha - \alpha'$ where α is its activation and α' is the activation of the second most active neuron in the same modality. This is motivated by the "winner takes it all" dynamics.

4. The model runs for three steps before the predicted RT is calculated,

5. The RT is the inverse of the difference between the activations of the most and the second most active neurons in the output modality. This is motivated by the winner takes it all dynamics too. The winner wins, but it takes time inversely proportionate to *by how much* it won. This is a similar mechanism as the one proposed by Cohen et al. (1990)

6. The input to a neuron is the weighed sum of the activations of those neurons that are connected to it.

7. All the inputs to the relevant modality are additionally multiplied by the *relevance parameter* r.

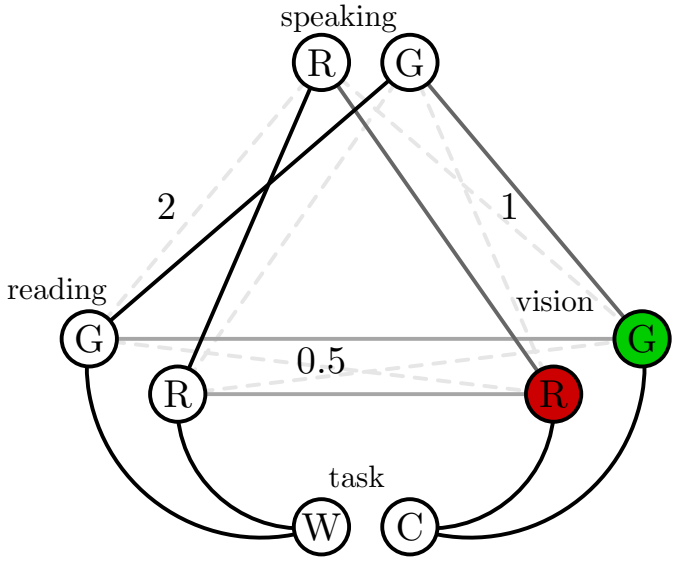

Figure 3: Same as before, but with weights.

8. we assume that the output modality doesn't back-propagate information, i.e. the connections from an input modality to the output modality are one-directinal.

This is the simplest possible way of defining a neural network. One may wish to add continuity, noise and non-linearity, but it is all unnecessary to make the predictions we want. Additionally, as explained in Section **??**, these additions will hardly change the predictions.

Let us analyse the simplest case: there are two input modalities W and C and one output modality O. Each contains 4 neurons and the strength between the neurons in C and O is 2, the one between W and C is 1 and the one between W and O is 4. So the network looks like the one on Figure 2. The assumption that the strength between written word modality and speaking is twice as strong as the one between ink colour modality and speaking is already traced back to Cattell's results. The assumption that the connection between written word and colour is half as strong as the one between colour and spoken word is currently based on the author's intuition but is something that can be verified (or falsified) experimentally. Assuming a Hebbian architecture this should follow from (the other assumption) that written colour words and colours co-occure rarer than than spoken colour words and colours. Also let us fix $r = 3$. Figure 2 shows the architecture for 2 colours. In this case the outcome of the model is not dependent on the number of neurons in each modality.

We will analyse what happens here. In fact we will calculate precisely the predicted reaction times for this model. We denote by $(g, r)_w$ the activation of

green and red in the written word modality and similarly $(g, r)_c$ and $(g, r)_s$ for ink colour and spoken word modalities.

Congruent input, word recognition.

Step 1 A congruent input arrives with an amplification by the factor of 3 in the written word modality, so the activations are $(0, 3)_w$, $(0, 1)_c$ and $(0, 0)_o$.

Step 2 Signals from the input modalities propagate to output weighed by the connection strengths, so we get $(0, 7)_o$. Also the input modalities propagate to each other, so in the colour modality we get $(0, 2\frac{1}{2})_c$. The input to the written word modality is multiplied by 3 so we get $(0, 4\frac{1}{2})$.

Step 3 The output modality's red neuron receives $2 \cdot 4\frac{1}{2} + 1 \cdot 2\frac{1}{2} = 11\frac{1}{2}$, so the result is $(0, 18\frac{1}{2})_o$. This implies a reaction time of

$$2/37$$

Neutral input, word recognition.

Step 1 A neutral input arrives with an amplification by the factor of 3 in the written word modality, so the activations are $(0, 3)_w$, $(0, 0)_c$ and $(0, 0)_o$.

Step 2 The only signal from the written word modality propagates both to the output and to the colour modality, so we obtain $(0, 6)_o$ and $(0, 1\frac{1}{2})_c$. The written word modality remains unchanged.

Step 3 The output modality's red neuron receives $2 \cdot 3 + 1 \cdot 1\frac{1}{2} = 7\frac{1}{2}$, so the result is $(0, 13\frac{1}{2})_o$. This implies a reaction time of

$$2/27$$

Incongruent input, word recognition.

Step 1 An incongruent input arrives with an amplification by the factor of 3 in the written word modality, so the activations are $(0, 3)_w$, $(1, 0)_c$ and $(0, 0)_o$.

Step 2 Propagation to the output results in $(1, 6)_o$. Propagation between input modalities results in $(1, 1\frac{1}{2})_c$ and $(1\frac{1}{2}, 3)_w$. In both cases the red neuron dominates.

Step 3 The output modality receives now more activation to the red neuron: $2 \cdot 1\frac{1}{2} + 1 \cdot \frac{1}{2} = 3\frac{1}{2}$, so the result is $(1, 9\frac{1}{2})_o$. This implies a reaction time of

$$1/8\frac{1}{2} = 2/17$$

Congruent input, colour recognition.

Step 1 A congruent input arrives with an amplification by the factor of 3 in the ink colour modality, so the activations are $(0,1)_w$, $(0,3)_c$ and $(0,0)_o$.

Step 2 Signals from the input modalities propagate to output weighed by the connection strengths, so we get $(0,5)_o$. Also the input modalities propagate to each other, so in the colour modality we get $(0,4\frac{1}{2})_c$ and in the written word modality $(0,2\frac{1}{2})_w$.

Step 3 The output modality's red neuron receives $2\cdot 2\frac{1}{2}+1\cdot 4\frac{1}{2}=9\frac{1}{2}$, so the result is $(0,14\frac{1}{2})_o$. This implies a reaction time of

$$2/29$$

Neutral input, colour recognition.

Step 1 A neutral input arrives with an amplification by the factor of 3 in the colour modality, so the activations are $(0,3)_c$, $(0,0)_w$ and $(0,0)_o$.

Step 2 The only signal from the colour modality propagates both to the output and to the written word modality, so we obtain $(0,3)_o$ and $(0,1\frac{1}{2})_w$. The colour modality remains unchanged.

Step 3 The output modality's red neuron receives $2\cdot 1\frac{1}{2}+1\cdot 3=6$, so the result is $(0,9)_o$. This implies a reaction time of

$$1/9$$

Incongruent input, colour recognition.

Step 1 An incongruent input arrives with an amplification by the factor of 3 in the colour modality, so the activations are $(1,0)_w$, $(0,3)_c$ and $(0,0)_o$.

Step 2 Propagation to the output results in $(2,3)_o$. Propagation between input modalities results in $(1\frac{1}{2},3)_c$ and $(1,1\frac{1}{2})_w$. In both cases the red neuron dominates.

Step 3 The output modality receives now more activation to the red neuron: $2\cdot\frac{1}{2}+1\cdot 1\frac{1}{2}=2\frac{1}{2}$, so the result is $(2,5\frac{1}{2})_o$. This implies a reaction time of

$$2/7$$

So we get the following table of predicted reaction times (the columns correspond to the two tasks naming a word or a colour and the rows to the congruent, neutral and incongurent trials):

	Word	Colour
Con	$\frac{2}{37} = 0.054$	$\frac{2}{29} \approx 0.069$
Neu	$\frac{2}{27} \approx 0.074$	$\frac{1}{9} \approx 0.111$
Inc	$\frac{2}{17} \approx 0.118$	$\frac{2}{7} \approx 0.286$

We see task assymetry (**1**): in the word recognition task the difference between congruent and incongruent is

$$\frac{2}{17} - \frac{2}{37} \approx 0.063$$

while in the colour recognition task it is

$$\frac{2}{7} - \frac{2}{29} \approx 0.216$$

which is 3.4 times bigger. The same effect is preserved if one looks at the quotient instead of difference:

$$\frac{2}{17} : \frac{2}{37} \approx 1.88 < 4.14 \approx \frac{2}{7} : \frac{2}{29}.$$

We also see the facilitation-interference asymmetry (**2**): in the word-recognition task the difference between neutral and congruent is

$$\frac{2}{27} - \frac{2}{37} \approx 0.02$$

and between neutral and incongruent is

$$\frac{2}{17} - \frac{2}{27} \approx 0.04$$

while in the colour-recognition task those numbers are

$$\frac{1}{9} - \frac{2}{29} \approx 0.04$$

and between neutral and incongruent is

$$\frac{2}{7} - \frac{1}{9} \approx 0.17.$$

These effects are preserved too if instead of differences we look at quotients.

If we were to replace the output modality by another one, we should modify the weights of the connections. Note that the connection weights can be tested experimentally in a separate experiment, so this gives a robust way to test the model. For example if the output modality is pointing to a colour patch, I would expect the weights to be reversed: $w_2 = 2$ and $w_1 = 1$ and otherwise all the same. In this situation, obviously, the task asymmetry would be reversed accounting for the reverse effect (**3**) and has the potential of accounting for other input and output modalities (**8**,**9**). This implicitly also accounts for the automaticity effect as modelled by the connection weights (**7**).

Let us now look at how this model also explaines the discrepancy between semantic and response interference (**4**,**5**). In this case the neural network will look like in Figure 4. The outputs are given by button presses and two different colours are being mapped to each button. Suppose that ink-recognition is the task. During the training phase the connection from ink colours to the appropriate buttons becomes strong. However, that's not all. At each training trial the activation of an ink-neuron propagates to the corresponding word-neuron and so the activation of those also correlates with the button presses, so a connection between words and buttons gets trained as well, although a weaker one. Based on this analysis, let us assume that the weight between written words and ink is 1 as before, then connection from ink to button presses is $1/2$ and the connection from written words to button presses is $1/4$, see Figure 4.

Let's see what happens now. We amplify again the input to the ink-modality by the factor of 3, because that's the relevant modality. Suppose we have a congruent input, say red-red. Then in the first step there is activation 3 in the red neuron on the right and 1 on the red neuron on the left. In the next step these activations propagate to the output modality adding up to $\frac{1}{4} \cdot 1 + \frac{1}{2} \cdot 3 = 1\frac{3}{4}$. The signal also propagates from word to colour and vice versa resulting in the activations $1 + 3 = 4$ and $3 + 3 \cdot 1) = 6$ whose linear combination $\frac{1}{4} \cdot 4 + \frac{1}{2} \cdot 6 = 4$ adds directly up to the output creating the activation of $5\frac{3}{4}$ in the output and the reaction time becomes $4/23 \approx 0.174$. Suppose the input is responce-congruent, but semantically incongruent. Say we have the activation of green on the left side (written word) and the activation of red on the right side (ink colour). Denote by $(x, y)_w$ the

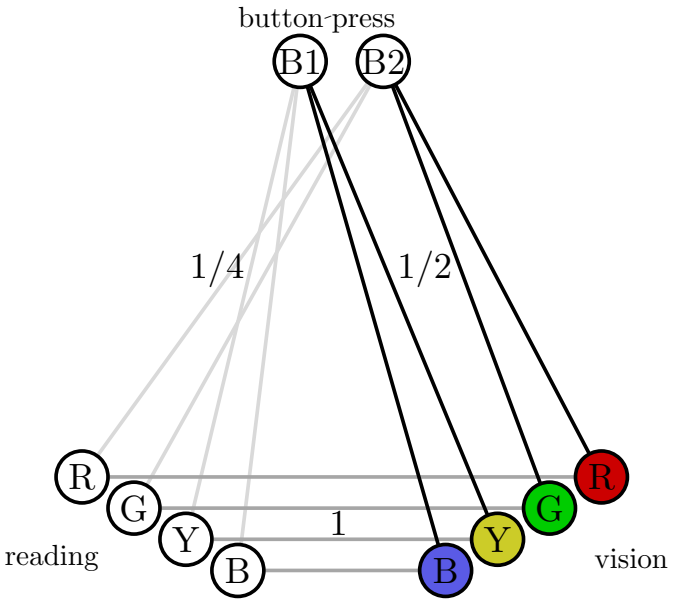

Figure 4: The architecture of the ANN to model the classical Stroop effect. The darker lines correspond to stronger associations while the dashed lines correspond to no or even negative association.

activation of the pair (green, red) on the word modality side and $(x, y)_c$ the same for the colour modality. So we have activations $(1, 0)$ and $(0, 3)$. After one step we will have $\frac{1}{4} + \frac{3}{2} = 1\frac{3}{4}$ in the output modality, as well as $(1, 3)_w$ and $(1, 3)_c$. So in the next step the colour and word modalities propage both a value of $3 - 1 = 2$ to the output button whose activation which becomes

$$1\frac{3}{4} + \frac{1}{4} \cdot 2 + \frac{1}{2} \cdot 2 = 3\frac{1}{4}$$

which results in the reaction time $4/13 \approx 0.308$. Suppose we have a semantically incongruent input, say blue and red. denote the activation in the word and colour modalities similary as above replacing green by blue and now for the output modality we have a pair notation $(x, y)_o$ too, because both buttons can get activated in this case. After first step we have $(\frac{1}{4}, \frac{3}{2})_o$ and $(1, 3)_w$ and $(1, 3)_c$. Again both propagate a 2 to the same output button obtaining

$$\frac{3}{2} + \frac{1}{4} \cdot 2 + \frac{1}{2} \cdot 2 = 3$$

and so we have $(\frac{1}{4}, 3)_o$ which makes the reaction time $1/(3 - \frac{1}{4}) = 4/11 \approx 0.364$.

Here is the table of the predicted reaction times for congruent (Con), semantically incongruent responce congruent (SI) and responce incongruent (RI) trials:

Con	0.174
SI	0.308
RI	0.364

This is consistent with the empirical data obtained by van Veen and Carter (2005).

References

Bekolay, T. (2011). *Learning in large-scale spiking neural networks*. University of Waterloo. (Thesis)

Cattell, J. M. (1886). The time it takes to see and name objects. *Mind*, *11*(41), 63-65.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological review*, *97*(3), 332.

Dennis, I., & Newstead, S. E. (1981). Is phonological recoding under strategic control? *Memory & Cognition*, *9*(5), 472–477.

Durgin, F. H. (2000). The reverse Stroop effect. *Psychonomic Bulletin & Review*, *7*(1), 121–125.

Green, E. J., & Barber, P. J. (1981). An auditory stroop effect with judgments of speaker gender. *Perception & Psychophysics*, *30*(5), 459–466.

Kulikov, V. (2015). Framework theory: A theory of cognitive semantics. In T. R. Besold & K.-U. Kühnberger (Eds.), *Proceedings of the workshop on neural-cognitive integration (NCI@KI2015)* (p. 8-18).

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological bulletin*, *109*(2), 163-203.

MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 126.

Stirling, N. (1979). Stroop interference: An input and an output phenomenon. *Quarterly Journal of Experimental Psychology*, *31*, 121-132.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643-662.

van Veen, V., & Carter, C. S. (2005). Separating semantic conflict and response conflict in the Stroop task: a functional mri study. *Neuroimage*, *27*(3), 497–504.

Virzi, R. A., & Egeth, H. E. (1985). Toward a translational model of Stroop interference. *Memory & Cognition*, *13*(4), 304–319.

Zajano, M. J., & Gorman, A. (1986). Stroop interference as a function of percentage of congruent items. *Perceptual and Motor Skills*, *63*(3), 1087–1096.