LING3804: Lecture Notes 3
A Model of Associative Memory

Neural activation patterns can carry a lot of information, but synaptic weights can carry more:

- hundreds of dendrites per neuron,
- hundreds of receptors per dentrite.

Biological neurons can rapidly, durably modify synaptic weights by adding receptor capacity.

LLMs typically don't do this, but it is thought to be part of human language comprehension.

That said, some ideas from associative memory will be useful to understand LLMs.

Contents

3.1 Mental states as patterns of neural activation

Mental states (e.g. from looking at pictures) are associated with active firing of characteristic patterns of neurons [Mitchell et al., 2008].

Activation of neurons in the cortex can be modeled with **vectors** of firing rates for neurons or clusters:

.58	← neuron/cluster #1, say, closest to center of motor cortex
.0	← neuron/cluster #2, second closest to center of motor cortex
.58	← neuron/cluster #3, third closest to center of motor cortex
.0	⋮
.0	← neuron/cluster #5, say, closest to center of auditory cortex
.58	← neuron/cluster #6, second closest to center of auditory cortex
.0	⋮

(The values are typically ‘normalized’ so that the point is always one unit away from the origin.)

This kind of model is called ‘distributed’ because the activation is distributed around the cortex.

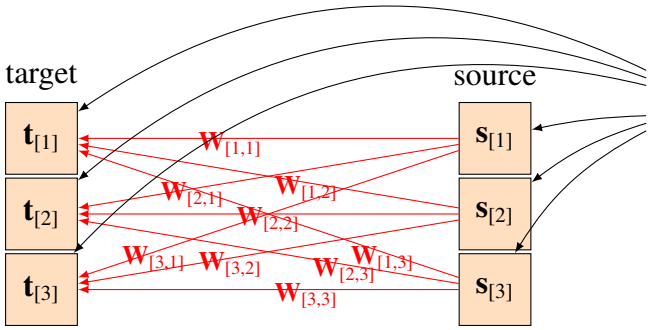
Individual elements of vectors (or characteristic subsets of elements) are called **features**.

An *n*-length vector may also be read as the **coordinates** of a point in an *n*-dimensional space.

3.2 Cued associations as connectivity weights between neurons/clusters

Mental states can be used as **cues** to other associated **target** mental states.

These associations happen by **long-term potentiation** (sensitization) of synapses between pre-synaptic and post-synaptic neurons that are active in the cue and target states, respectively.



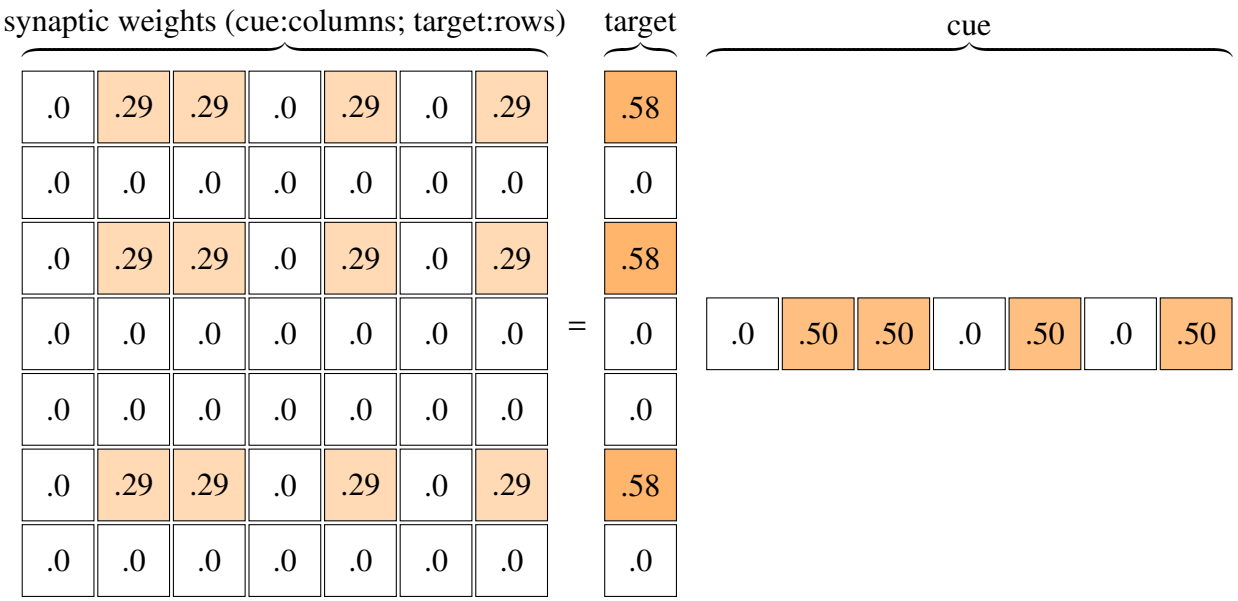
Synaptic weights can be shown in **matrices**, with a row for each target and a column for each cue:

$\mathbf{W}_{[1,1]}$	$\mathbf{W}_{[1,2]}$	$\mathbf{W}_{[1,3]}$
$\mathbf{W}_{[2,1]}$	$\mathbf{W}_{[2,2]}$	$\mathbf{W}_{[2,3]}$
$\mathbf{W}_{[3,1]}$	$\mathbf{W}_{[3,2]}$	$\mathbf{W}_{[3,3]}$

The multi-neuron activation patterns that are associated by these weights are then called **vectors**.

Potentiation among neurons that are active in these patterns can then be modeled using matrices of connections for each pair of neurons in cue and target patterns (specifically it’s an outer product of cue and target vectors, with the cue on the right) [Marr, 1971, Anderson et al., 1977,

Murdock, 1982, Smolensky, 1990, McClelland et al., 1995, Howard & Kahana, 2002]:

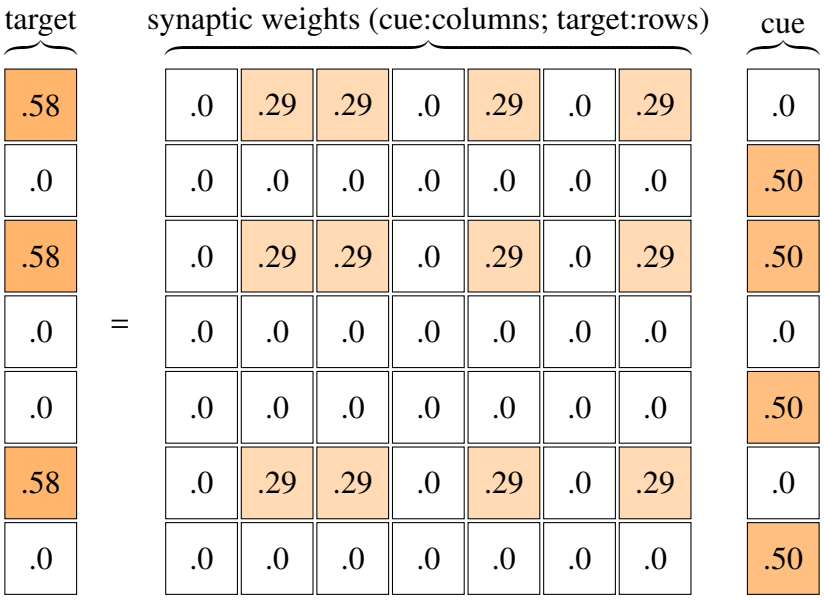


Formally, this is a **matrix product**: the value at row i, column j of the result is the sum of the product of each element in row i of the first factor $\mathbf{M} \in \mathbb{R}^{I \times K}$ (the target) with the corresponding element in column j of the second $\mathbf{N} \in \mathbb{R}^{K \times J}$ (the cue):

$$(\mathbf{M}\,\mathbf{N})_{[i,j]} = \sum_{k=1}^{K} \mathbf{M}_{[i,k]} \cdot \mathbf{N}_{[k,j]}$$

(Here $\mathbb{R}^{K \times J}$ denotes all sets of real numbers arranged in a matrix with K rows and J columns.)

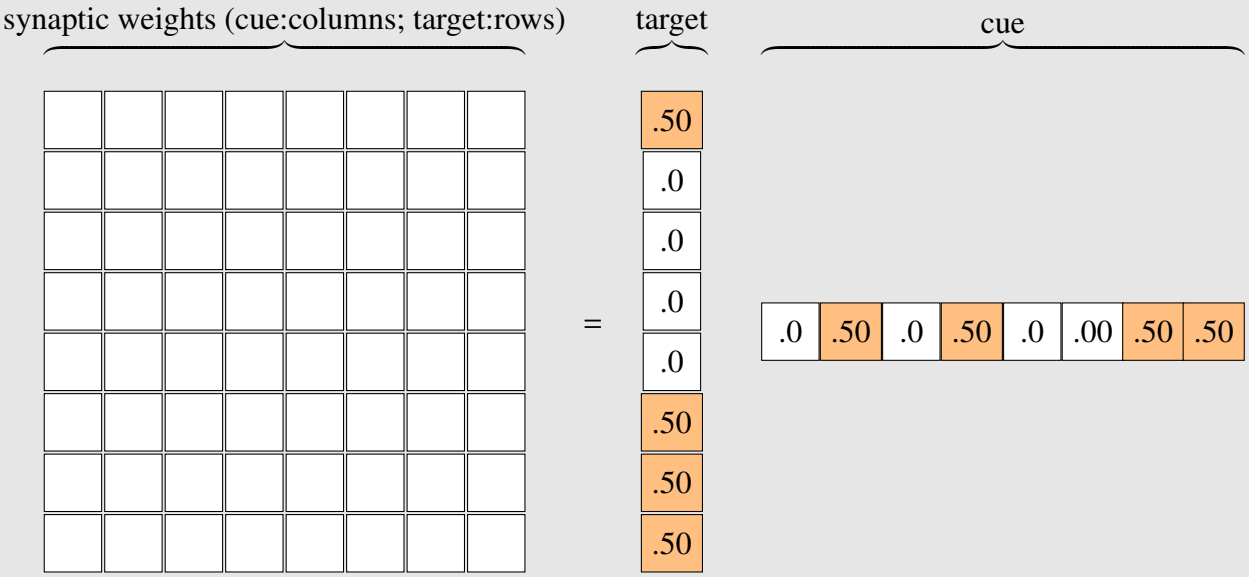
The target is then obtained by applying the association weights to the cue (also a matrix product):



To compute the activation of each post-synaptic neuron (row) in the target vector, the activation of each of the pre-synaptic neurons in the cue is multiplied by the synaptic weight in the memory matrix for that pre-synaptic neuron (column) synapsing with that post-synaptic neuron (row). The contributions of each pre-synaptic neuron are weighted by the corresponding synaptic weight and added together. So, to compute the top element of the target, the four .50's of the 2nd, 3rd, 5th and 7th elements of the cue are multiplied by .29, .29, .29 and .29, respectively (the 2nd, 3rd, 5th and 7th elements of the top row) and added together to give .58, and the other elements in the top row of the matrix are multiplied by zeros in the cue so they don't change anything when they are added in. The same thing happens for each lower row of the matrix, to define each lower element of the target, until you have the result in the figure.

Practice 3.1:

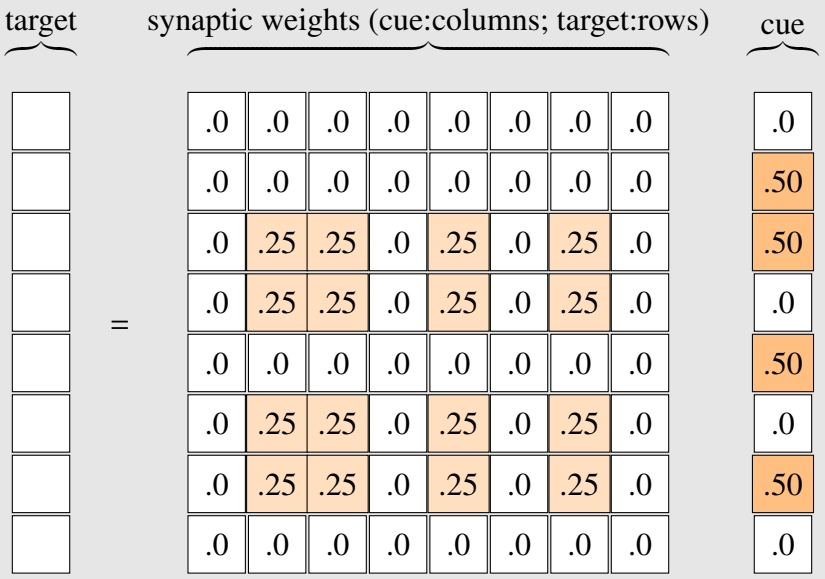
Suppose you have cue and target mental states characterized by the below patterns of cortical activation. What synaptic weights result from long-term potentiation of the cue state immediately followed by the target state:



Practice 3.2:

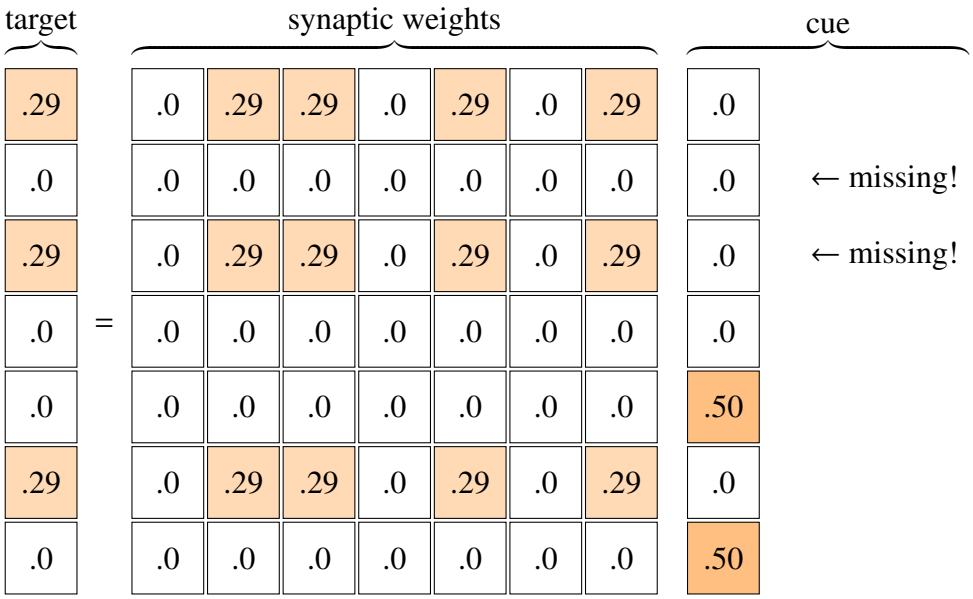
Now suppose you cue the below associative memory matrix of synaptic weights with the below

vector of cortical activations. What will be the result?



3.3 Robustness to incomplete cues ('holographic memory')

Associations from incomplete cues yield complete (but weaker) targets:



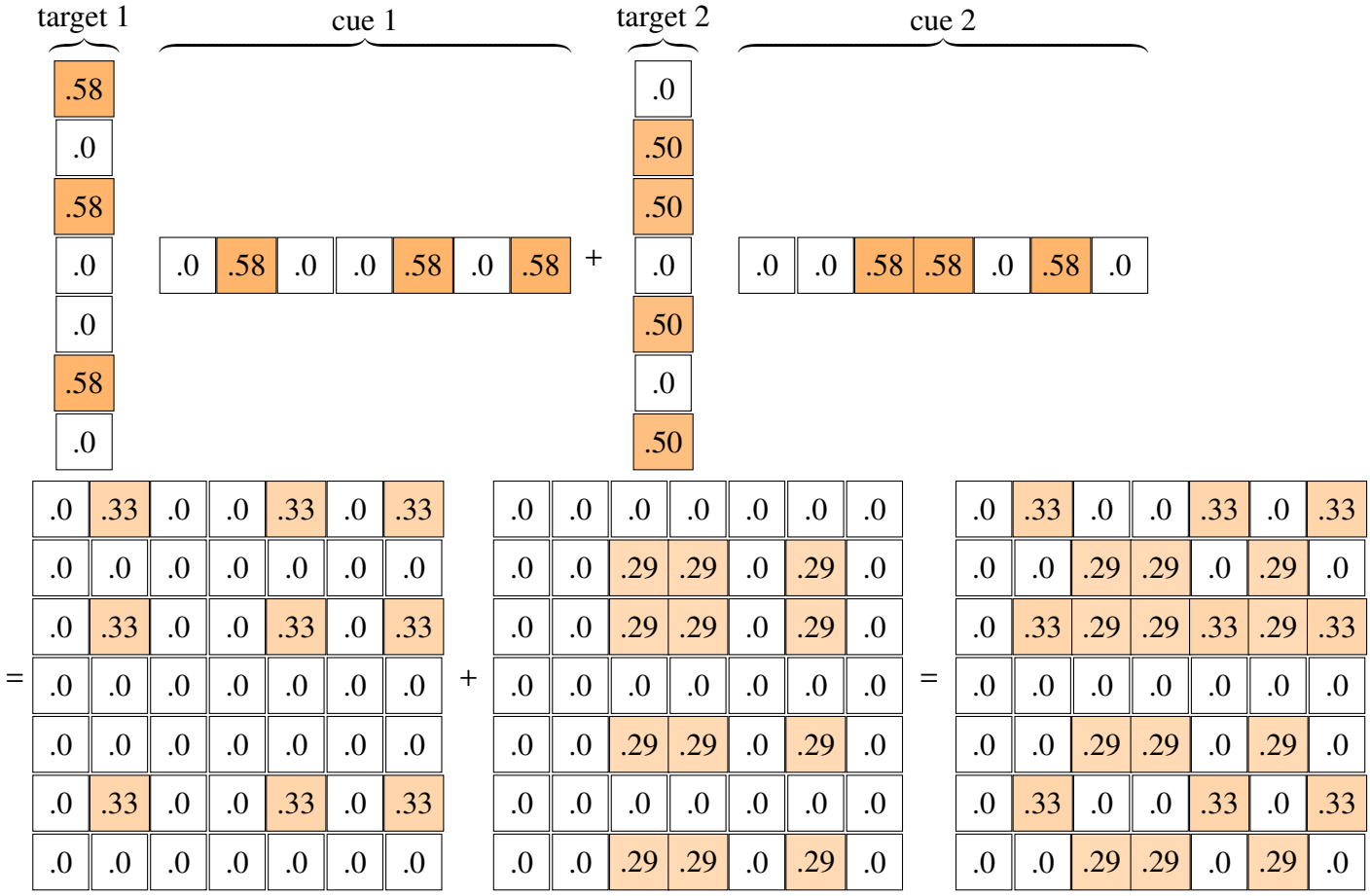
To compute the activation of each post-synaptic neuron (row) in the target vector, the activation of each of the pre-synaptic neurons in the cue is multiplied by the synaptic weight in the memory matrix for that pre-synaptic neuron (column) synapsing with that post-synaptic neuron (row). The contributions of each pre-synaptic neuron are weighted by the corresponding synaptic weight and added together. So, to compute the top element of the target, the two .50's of the 5th and 7th

elements of the cue are multiplied by .29 and .29 (the 5th and 7th elements of the top row) and added together to give .29, and the other elements in the top row of the matrix are multiplied by zeros in the cue so they don't change anything when they are added in. The same thing happens for each lower row of the matrix, to define each lower element of the target, until you have the result in the figure.

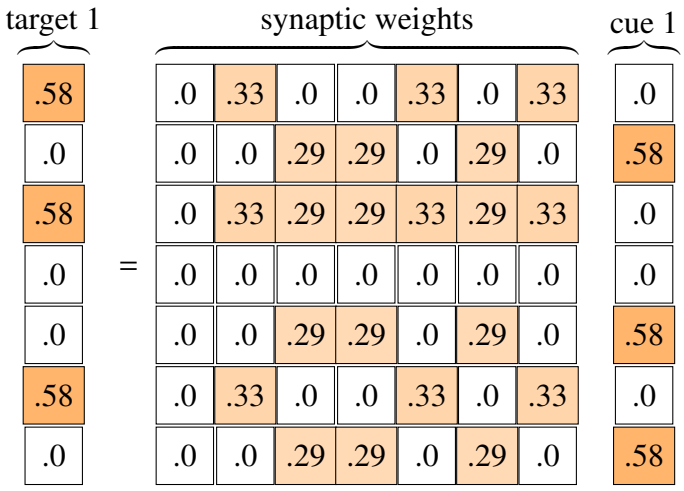
This provides a natural model of brain plasticity following trauma.

3.4 Associations can be combined

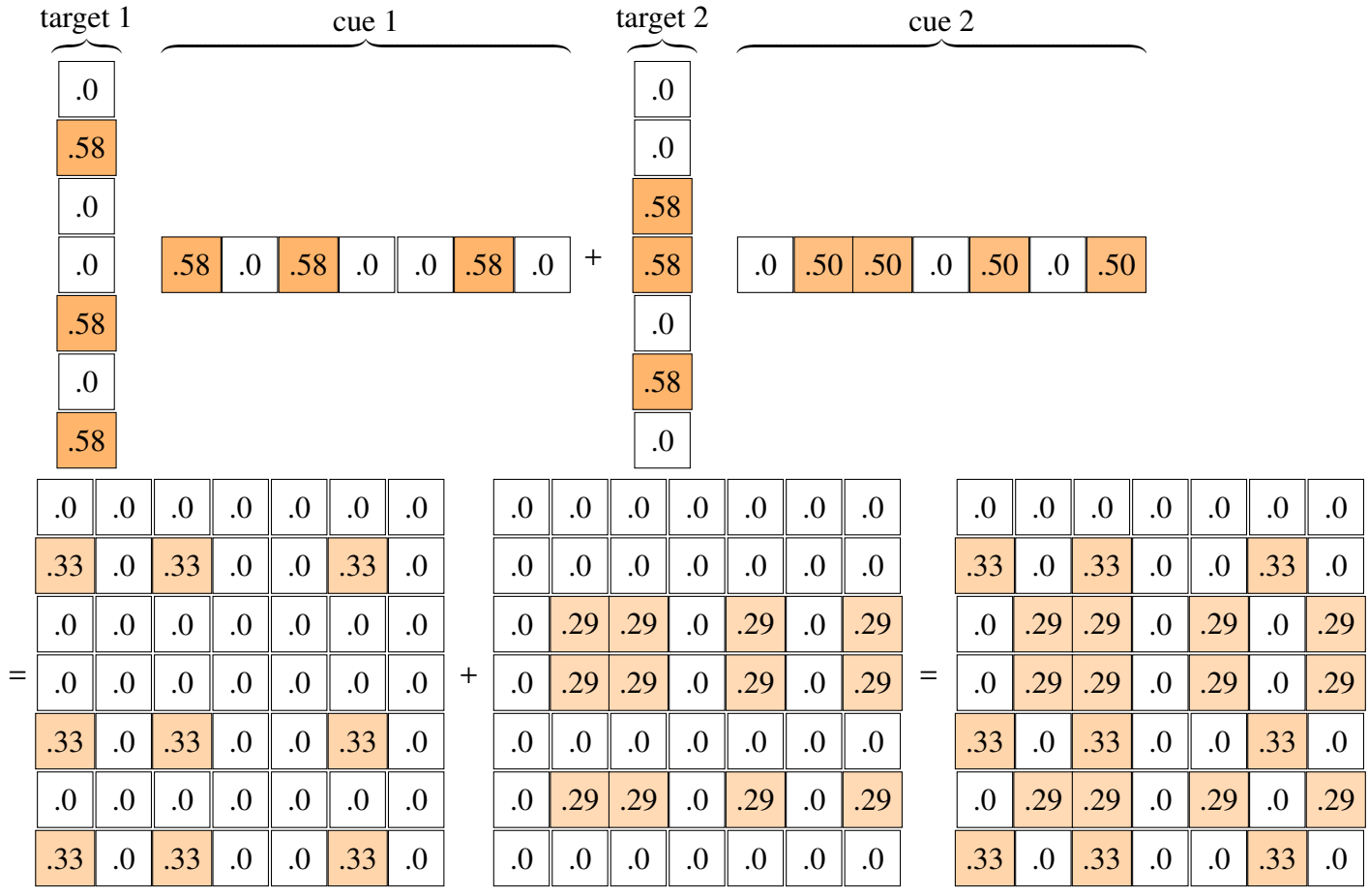
Multiple associations can be combined (stored together) in the same set of synapses:



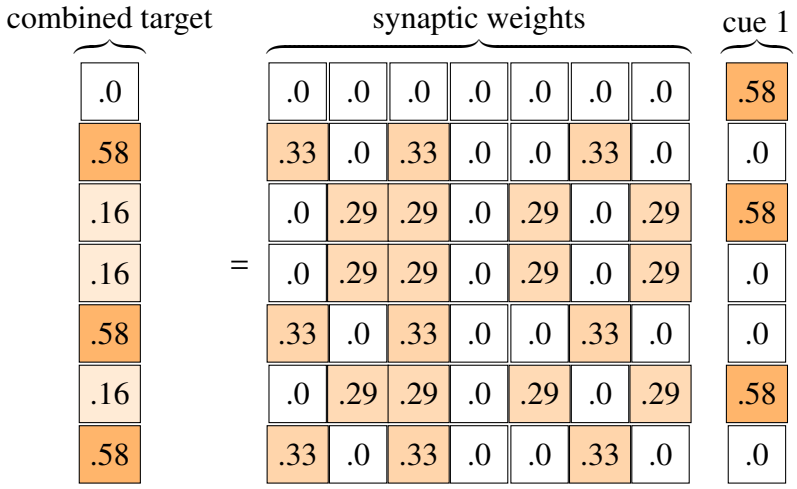
the resulting associations can still be cued:



However, when the stored cues overlap (e.g. the 3rd element in the cues below):



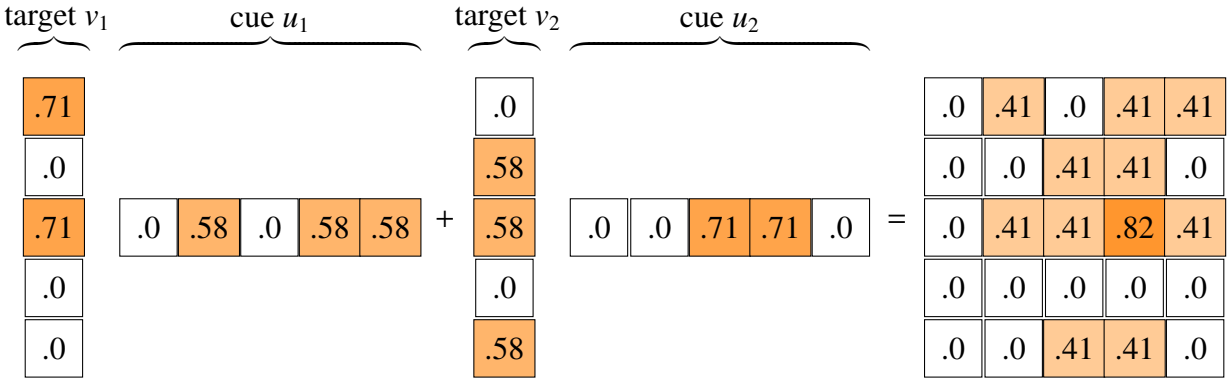
the resulting associations ‘interfere’ with each other when cued, yielding a combined target:



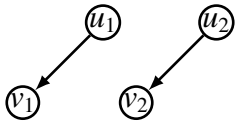
This has been proposed as a process by which forgetting happens [Howard & Kahana, 2002].

3.5 Graphical representations of mental states and cued associations

Recall mental states are coordinates of points in mental space, linked by cued associations:



Cued associations in an associative memory can be represented graphically:



(arbitrarily squashing the n coordinates/dimensions into a two-dimensional figure).

3.6 Multiple associations (multiplexing and tensors)

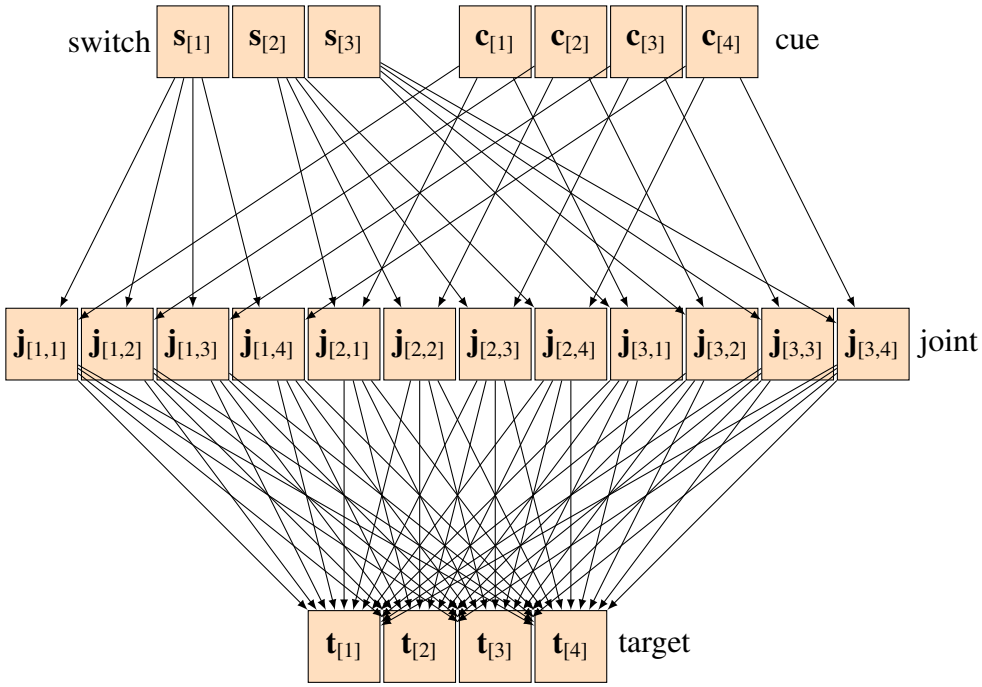
Mental states can cue multiple targets without interference using ‘switching’ elements n.

(Let’s just assume these are the first few elements of each vector.)

Then define a ‘joint’ element for each non-switching element: fire if both it and switch n fire.

Associations M_n may then be cued on these joint features instead of regular elements.

For example, if $n = 3$:

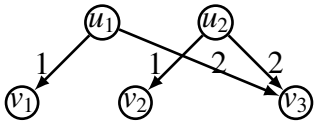


Associations from joint features are modeled using numbered layered matrices (tensors):

target v_1		cue u_1						target v_2		cue u_2						associations M_1				
.71								.0								.0	.41	.0	.41	.41
.0								.58								.0	.0	.41	.41	.0
.71		.0	.58	.0	.58	.58	+	.58		.0	.0	.71	.71	.0	=	.0	.41	.41	.82	.41
.0								.0								.0	.0	.0	.0	.0
.0								.58								.0	.0	.41	.41	.0

target v_3		cue u_1						target v_3		cue u_2						associations M_2				
.58								.58								.0	.33	.74	.41	.33
.58								.58								.0	.33	.74	.41	.33
.0		.0	.58	.58	.0	.58	+	.0		.0	.0	.71	.71	.0	=	.0	.0	.0	.0	.0
.58								.58								.0	.33	.74	.41	.33
.0								.0								.0	.0	.0	.0	.0

Numbered association layers can be represented graphically using edge labels:



Similar '(de-)multiplexing' and has been proposed as a model of the hippocampus [Marr, 1971].

References

[Anderson et al., 1977] Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception and probability learning: Some applications of a neural model. *Psychological Review*, 84, 413–451.

[Howard & Kahana, 2002] Howard, M. W. & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 45, 269–299.

[Marr, 1971] Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society (London) B*, 262, 23–81.

[McClelland et al., 1995] McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102, 419–457.

[Mitchell et al., 2008] Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320, 1191–1195.

[Murdock, 1982] Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89, 609–626.

[Smolensky, 1990] Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2), 159–216.