Neural Models of the Psychosemantics of ‘Most'

Lewis O'Sullivan, Shane Steinert-Threlkeld

MOTIVATION

- A quantifier's semantics consist of a single set of truth conditions. These may be represented in many equivalent ways. Each representation corresponds to a particular verification strategy. Thus, by determining if speakers favour a strategy for a quantifier, we can infer their mental representation for that quantifier.
- Pietroski et al. (2009) found that speakers favour verifying “most” via a cardinality comparison (i.e. most (A) (B) = 1 iff IA∩BI > IA\BI) using representations from the approximate number system (ANS). Register et al. (2018): ANS is usage not universal, but due to a speed accuracy trade off.
- **Our question:** Can neural networks be developed into good cognitive models of the visual verification of “most”? We specifically aim to determine whether networks can quantitatively fit human data well, and be implemented with moveable parameters to generate new predictions.

METHODS

- We used a variation of Pietroski et al's (2009) visual identification task. We trained two types of network to classify a dot matrix stimuli (see header image) according to the truth value of a corresponding statement (“most of the dots are blue”). The total number of dots, dot set ratio and dot arrangement were manipulated between trials. We also manipulated operationalised task duration via the network architectures.
- Two types of neural network were used; Convolutional Neural Networks (CNN) using the VGG architecture (Simonyan and Zisserman 2014) and Recurrent Attention Models (RAM) (Mnih et al. 2014), which process their input via a series of ‘glimpses' akin to saccades and fixations. Four levels of operationalised task duration were used for each network type. The former operationalised task duration via network depth (VGG7, 9, 11 & 13), the latter by number of glimpses (4, 8, 16, 24).
- We selected three “behavioural traces” that networks ought to exhibit if they use a similar verification strategy for “most” to humans. The traces and their associated hypotheses are:

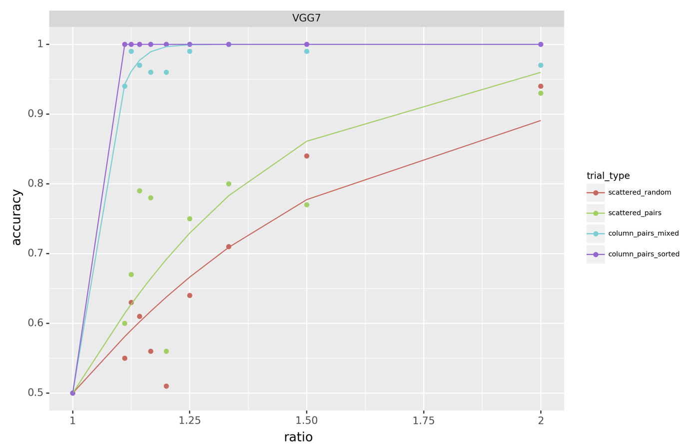
H1. ANS usage: Network accuracy is negatively correlated with the stimulus dot ratio size.
H2. Verification strategy preference: Network accuracy depends the arrangement of the stimulus.
H3. Speed-accuracy trade-off: Network accuracy is positively correlated with an appropriate operationalisation of task duration.

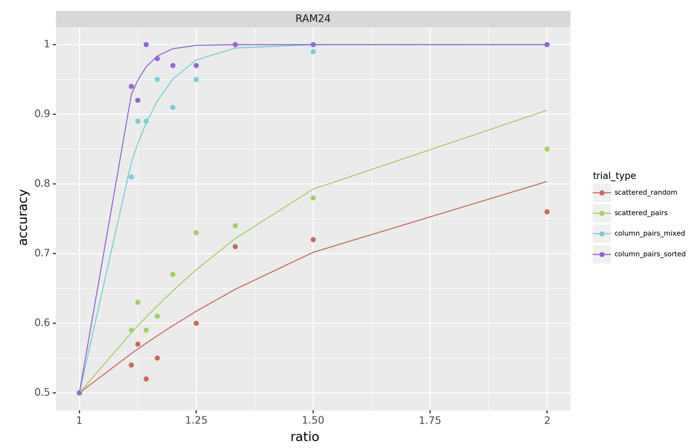
RESULTS

- We fit separate multiple logistic regressions to each network type. The outcome variable was correct label prediction. There were five predictor variables; three related to the hypotheses (image type i.e. dot arrangement, task duration, dot ratio), two as controls (absolute set size difference, total dots). Some variable levels were excluded from the CNN analysis due to response invariance (i.e. near celling accuracy). We interpreted the model results to mean the below for our hypotheses:

	H1	H2	H3
CNN	☑ The log-odds of a correct prediction by the CNNs were significantly reduced as ratios become more balanced. ***Pr(>|z|) < 0.01***	☑ The CNNs were significantly less likely to correctly predict the label of scatter random than scattered paired images ***Pr(>|z|) < 0.001***	☑ VGG7 was significantly less likely than VGG11 to make a correct classification. ***Pr(>|z|) < 0.001*** ☐ No difference was found between VGG9 & VGG11, but this appears to be due to ceiling effects.
RAM	☑ The log-odds of a correct prediction by the RAMs were significantly reduced as ratios become more balanced. ***Pr(>|z|) < 0.001***	☑ The RAMs were significantly less likely (by various degrees) to correctly predict the label of any image type than for column pairs images. ***All levels - Pr(>|z|) < 0.001***	☐ No significant difference was found in the likelihood of the RAM4, 8 or 16 networks correctly predicting a stimuli's label than the RAM24 network.

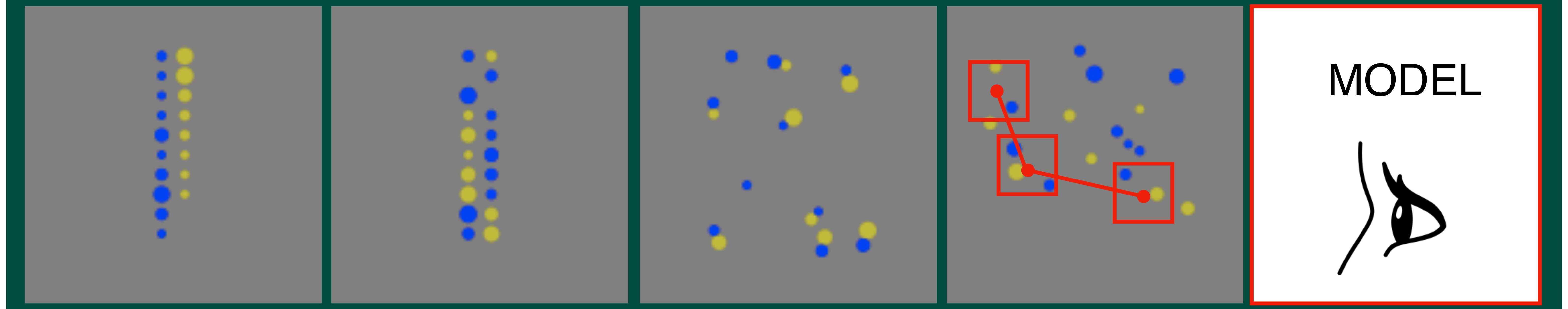
- We also fit a psychophysical model of the ANS to each model's mean accuracy data, broken down by dot ratio and arrangement. VGG7 and RAM 24 can be seen as examples below:





DISCUSSION

- Both network types exhibit qualitatively similar accuracy patterns to humans following manipulations of dot ratio, comparable to human ANS usage. The psychophysical model fits our data well.
- Whilst accuracy varied by image type, it patterns slightly differently to human speakers' (column type trials pattern together for networks), so our networks may use different verification strategies.
- Network depth (CNN) successfully operationalises task duration, but number of glimpses (RAM) does not. This is likely due to the problem of long-term-dependencies.
- In future work, we would like to i) experiment with RAM implementations with the aim of getting glimpse number to affect accuracy (e.g. hyper-parameter tuning, ‘peripheral vision' to help guide glimpses, multi-task learning), ii) use probing tools to infer strategies (e.g. transfer learning, diagnostic classifiers) & iii) test the models against other image types (e.g. with 2+ colour sets).



In a **psychosemantic verification task, neural models of visual attention** exhibit similar accuracy patterns to humans, incl. **sensitivity to set ratio.**

Take a picture to download the full paper

Contact: lewis.osullivan@student.uva.nl

ADDITIONAL MATERIALS & INFORMATION

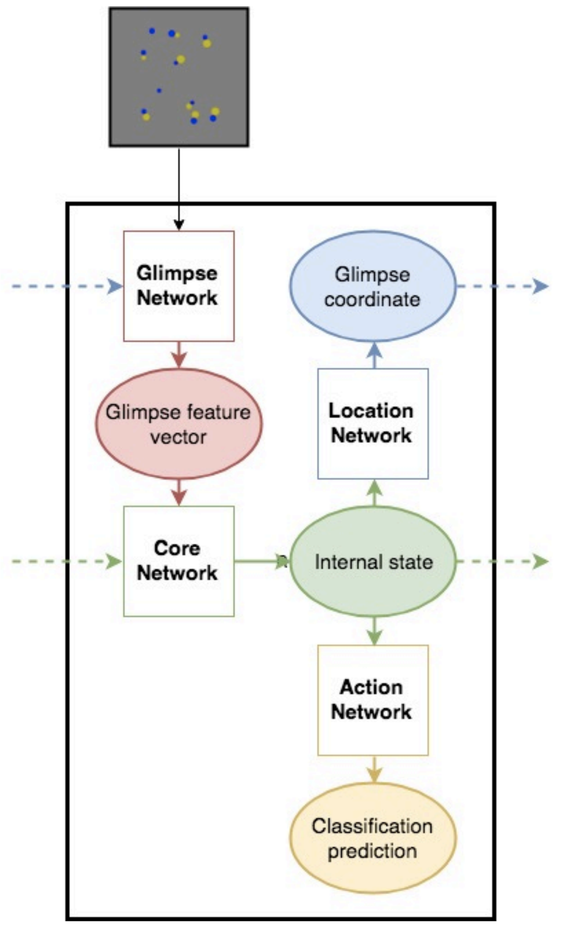
STIMULI:

Like Pietroski et al. (2009), we used up to 22 total dots per image, in ratios from 1:2 to 9:10 in one of four arrangements. Images were 128x128 pixels, converted to grayscale. The stimuli were split into a training set (18000 images), a validation set, and a test set (3600 images each). All three sets were balanced to contain equal proportions of each ratio/image type/truth-value combination.

TRAINING:

The VGG models are trained using the Adam optimizer (Kingma and Ba, 2015). The RAM models used the hybrid supervised learning approach from Mnih et al. (2014), where cross-entropy is back-propagated to train the action, core, and glimpse networks, and the REINFORCE rule (Williams, 1992; Sutton et al., 1999) is used for the location network.

RAM ARCHITECTURE:



RAM is a visual attention model formed of a network of networks. These are:

- The glimpse network, takes the image stimuli and a location coordinate as inputs. At t_0, the coordinate is random. At t1+, it is selected by the location network. Consecutively larger but lower resolution samples centred around on co-ordinate are concatenated into a “glimpse”. This is processed by 3 convolutional layers and one FC ReLU layer generating a “what” vector. In parallel, the coordinate is processed by a FC ReLU layer, generating a “where” vector. The “what” and “where” vectors are point-wise multiplied generating the glimpse feature vector.
- Core network: LSTM cell, 1024 units.
- Location network: FC layer with tanh activation, maps core network state to two values: means of Gaussians (fixed std at 0.03) for the two coordinates; actual are sampled.
- The action network: FC layer, takes the core network's internal state at t as input, outputs a binary image classification. A classification is made at every t, but only the classification decision at the final t is recorded.

Variable	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Image: Scattered pairs (Intercept)	16,20	4,15	3,91	9.42e-05 ***
Image: Scattered random	-0,79	0,09	-8,63	<2e-16 ***
Network: VGG9	-0,73	3,47	-0,21	0,83
Network: VGG7	-12,22	2,45	-4,98	6.37e-07 ***
Dot ratio	-14,90	4,93	-3,02	0.00253 **
Absolute difference	0,17	0,37	0,46	0,64
Total dots	-0,03	0,04	-0,87	0,39
Ratio * Network: VGG9	1,09	4,00	0,27	0,78
Ratio * Network: VGG7	11,81	2,83	4,18	2.97e-05 ***

Table 1: Multiple logistic regression of the CNN trials. Significance: 0 ‘***' 0.001 ‘**' 0.01 ‘*' 0.05 ‘.' 0.1.

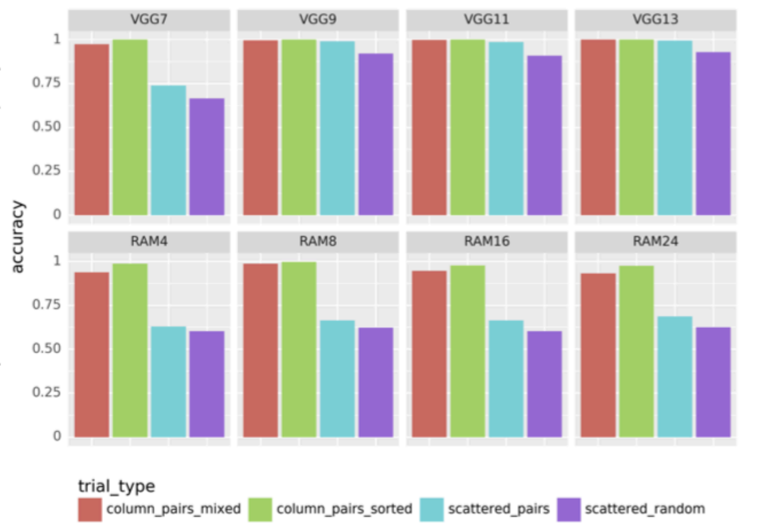


Figure 1. Accuracy by trial type, averaged across ratio.

Variable	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Image: Column pairs sorted (Intercept)	9,57	1,52	6,28	3.41e-10 ***
Image: Column pairs mixed	-1,18	0,16	-7,55	4.37e-14 ***
Image: Scattered pairs	-3,54	0,14	-25,15	< 2e-16 ***
Image: Scattered random	-3,75	0,14	-26,73	< 2e-16 ***
Glimpses: RAM16	-0,32	0,51	-0,63	0,53
Glimpses: RAM8	-0,97	0,51	-1,91	0,06
Glimpses: RAM4	-0,77	0,50	-1,54	0,12
Dot ratio	-6,91	1,84	-3,75	0.000179 ***
Absolute difference	-0,25	0,15	-1,64	0,10
Total dots	0,04	0,02	2,24	0.025427 *
Ratio * Glimpses: RAM16	0,33	0,63	0,52	0,60
Ratio * Glimpses: RAM8	1,34	0,62	2,14	0.032372 *
Ratio * Glimpses: RAM4	0,81	0,62	1,31	0,19

Table 2: Multiple logistic regression on RAM trials.

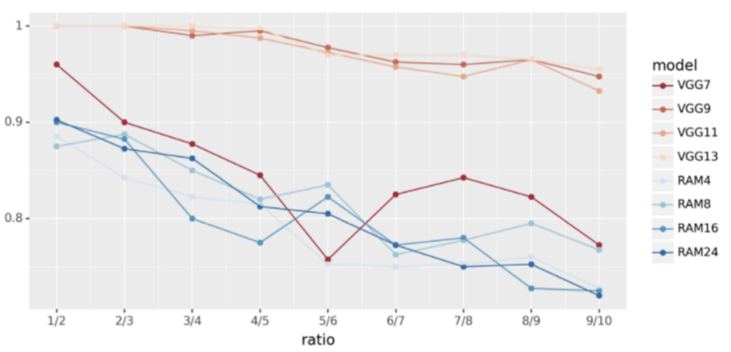


Figure 2. Accuracy by ratio, averaged across trial type.

	VGG7		RAM24		Human subjects (from Pietroski et al. 2009)	
Trial Type	Critical Weber Fraction	R^2	Critical Weber Fraction	R^2	Critical Weber Fraction	R^2
Scattered Random	0.363	0.843	0.524	0.801	0.32	0.9677
Scattered Pairs	0.256	0.581	0.340	0.913	0.33	0.8642
Column Mixed	0.047	0.979	0.078	0.975	0.30	0.9364
Column Sorted	0.012	1.0	0.051	0.984	0.04	0.9806

Table 3: Weber fractions and R^2 for the ANS model.

The authors have received funding from the European Research Council under the European Union's SeventhFramework Programme (FP/2007–2013)/ERC Grant Agreement n. STG 716230 CoSaQ