Supplementary Material for Learning Continuous Semantic Representations of Symbolic Expressions

Miltiadis Allamanis [1] **Pankajan Chanthirasegaran** [2] **Pushmeet Kohli** [3] **Charles Sutton** [2 4]

1. Synthetic Expression Datasets

Table 1 and Table 2 are sample expressions within an equivalence class for the two types of datasets we consider.

2. Detailed Evaluation

Figure 1 presents the detailed evaluation for our k-NN metric for each dataset. Figure 2 shows the detailed evaluation when using models trained on simpler datasets but tested on more complex ones, essentially evaluating the learned compositionality of the models. Figure 4 show how the performance varies across the datasets based on their characteristics. As expected as the number of variables increase, the performance worsens (Figure 4a) and expressions with more complex operators tend to have worse performance (Figure 4b). The results for UNSEENEQCLASS look very similar and are not plotted here.

3. Model Hyperparameters

The optimized hyperparameters are detailed in Table 3. All hyperparameters were optimized using the Spearmint (Snoek et al., 2012) Bayesian optimization package. The same range of values was used for all common model hyperparameters.

References

Snoek, Jasper, Larochelle, Hugo, and Adams, Ryan P. Practical Bayesian optimization of machine learning algorithms. In *NIPS*, 2012.

Socher, Richard, Huval, Brody, Manning, Christopher D, and Ng, Andrew Y. Semantic compositionality through recursive matrix-vector spaces. In *EMNLP*, 2012.

Work started when M. Allamanis was at Edinburgh. This work was done while P. Kohli was at Microsoft. [1]Microsoft Research, Cambridge, UK [2]University of Edinburgh, UK [3]DeepMind, London, UK [4]The Alan Turing Institute, London, UK. Correspondence to: Miltiadis Allamanis <t-mialla@microsoft.com>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

BOOL8

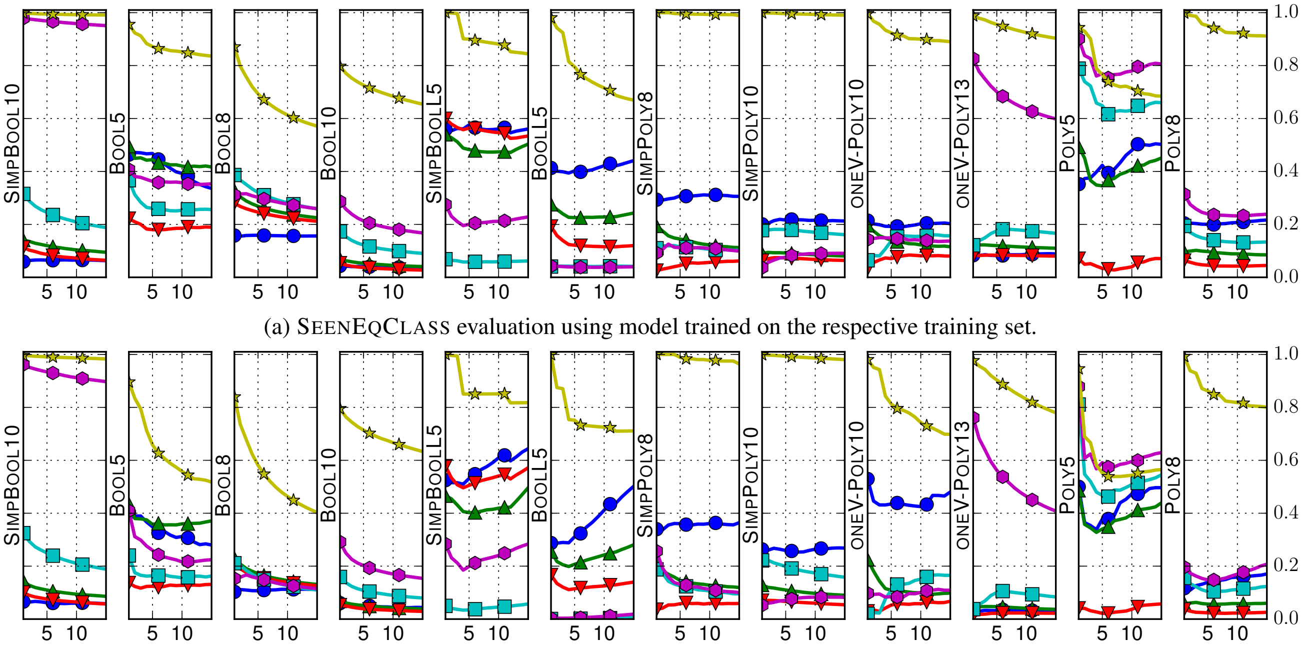
$(\neg a) \wedge (\neg b)$	$(\neg a \wedge \neg c) \vee (\neg b \wedge a \wedge c) \vee (\neg c \wedge b)$	$(\neg a) \wedge b \wedge c$
$a\neg((\neg a) \Rightarrow ((\neg a) \wedge b))$	$c \oplus (((\neg a) \Rightarrow a) \Rightarrow b)$	$\neg((\neg b) \vee ((\neg c) \vee a))$
$\neg((b \vee (\neg(\neg a))) \vee b)$	$\neg((b \oplus (b \vee a)) \oplus c)$	$((a \vee b) \wedge c) \wedge (\neg a)$
$(\neg a) \oplus ((a \vee b) \oplus a)$	$\neg((\neg(b \vee (\neg a))) \oplus c)$	$(\neg((\neg(\neg b)) \Rightarrow a)) \wedge c$
$(b \Rightarrow (b \Rightarrow a)) \wedge (\neg a)$	$((b \vee a) \oplus (\neg b)) \oplus c)$	$(c \wedge (c \Rightarrow (\neg a))) \wedge b$
$((\neg a) \Rightarrow b) \Rightarrow (a \oplus a)$	$(\neg((b \oplus a) \wedge a)) \oplus c$	$b \wedge (\neg(b \wedge (c \Rightarrow a)))$
False	$(\neg a) \wedge (\neg b) \vee (\wedge c)$	$\neg a \vee b$
$(a \oplus a) \wedge (c \Rightarrow c)$	$(a \Rightarrow (\neg c)) \oplus (a \vee b)$	$a \Rightarrow ((b \wedge (\neg c)) \vee b)$
$(\neg b) \wedge (\neg(b \Rightarrow a))$	$(a \Rightarrow (c \oplus b)) \oplus b$	$\neg(\neg((b \vee a) \Rightarrow b))$
$b \wedge ((a \vee a) \oplus a)$	$b \oplus (a \Rightarrow (b \oplus c))$	$(\neg a) \oplus (\neg(b \Rightarrow (\neg a)))$
$((\neg b) \wedge b) \oplus (a \oplus a)$	$(b \vee a) \oplus (x \Rightarrow (\neg a))$	$b \vee (\neg((\neg b) \wedge a))$
$c \wedge ((\neg(a \Rightarrow a)) \wedge c)$	$b \oplus ((\neg a) \vee (c \oplus b))$	$\neg((a \Rightarrow (a \oplus b)) \wedge a)$

Table 1. Sample of BOOL8 data.

POLY8

$-a - c$	c^2	b^2c^2
$(b - a) - (c + b)$	$(c \cdot c) + (b - b)$	$(b \cdot b) \cdot (c \cdot c)$
$b - (c + (b + a))$	$((c \cdot c) - c) + c$	$c \cdot (c \cdot (b \cdot b))$
$a - ((a + a) + c)$	$((b + c) - b) \cdot c$	$(c \cdot b) \cdot (b \cdot c)$
$(a - (a + a)) - c$	$c \cdot (c - (a - a))$	$((c \cdot b) \cdot c) \cdot b$
$(b - b) - (a + c)$	$c \cdot c$	$((c \cdot c) \cdot b) \cdot b$
c	$b \cdot c$	$b - c$
$c - ((c - c) \cdot a)$	$(c - (b - b)) \cdot b$	$(a - (a + c)) + b$
$c - ((a - a) \cdot c)$	$(b - (c - c)) \cdot c$	$(a - c) - (a - b)$
$((a - a) \cdot b) + c$	$(b - b) + (b \cdot c)$	$(b - (c + c)) + c$
$(c + a) - a$	$c \cdot ((b - c) + c)$	$(b - (c - a)) - a$
$(a \cdot (c - c)) + c$	$(b \cdot c) + (c - c)$	$b - ((a - a) + c)$

Table 2. Sample of POLY8 data.



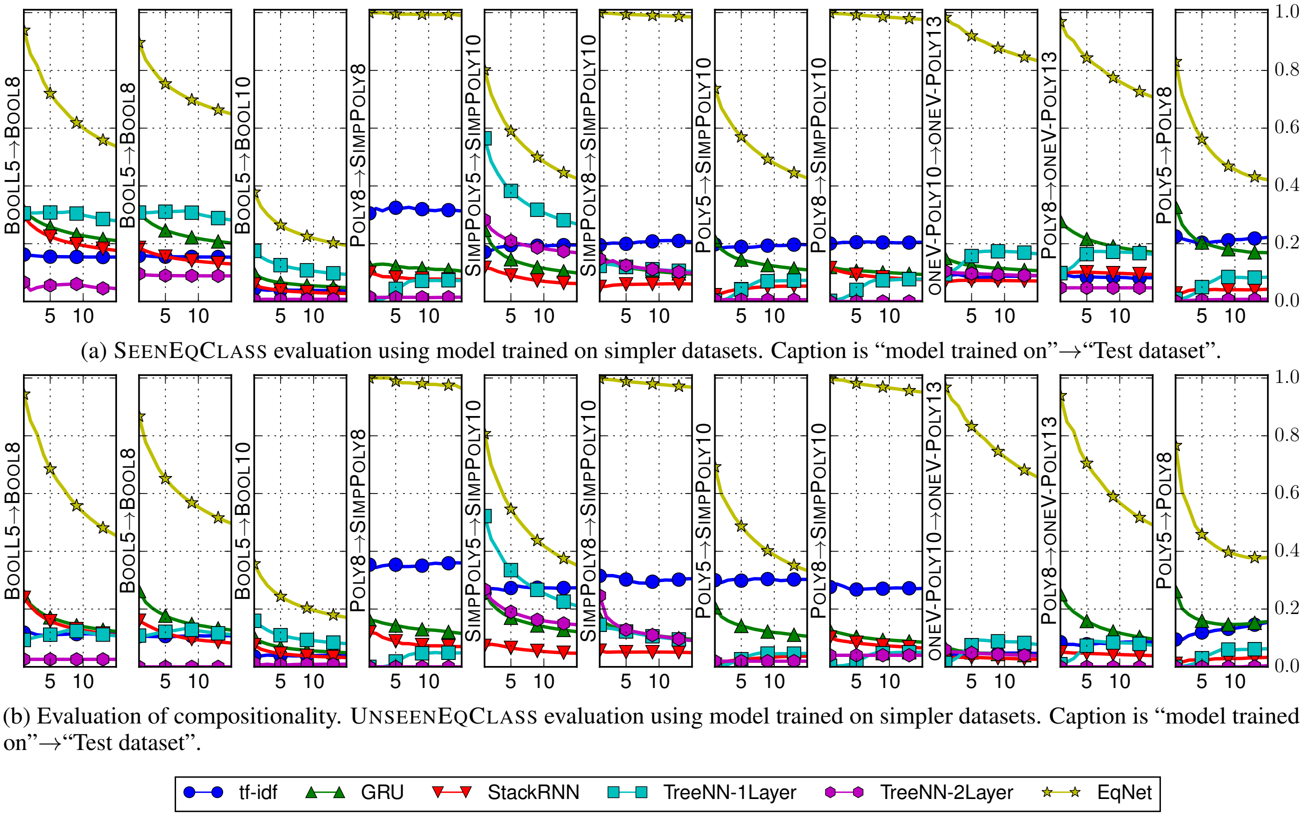
(a) SeenEqClass evaluation using model trained on the respective training set.

(b) UnseenEqClass evaluation using model trained on the respective training set.

Figure 1. Evaluation of $score_x$ (y axis) for $x = 1, \dots, 15$. on the respective SeenEqClass and UnseenEqClass where each model has been trained on. The markers are shown every five ticks of the x-axis to make the graph more clear. TreeNN refers to the model of Socher et al. (2012).

Table 3. Hyperparameters used in this work.

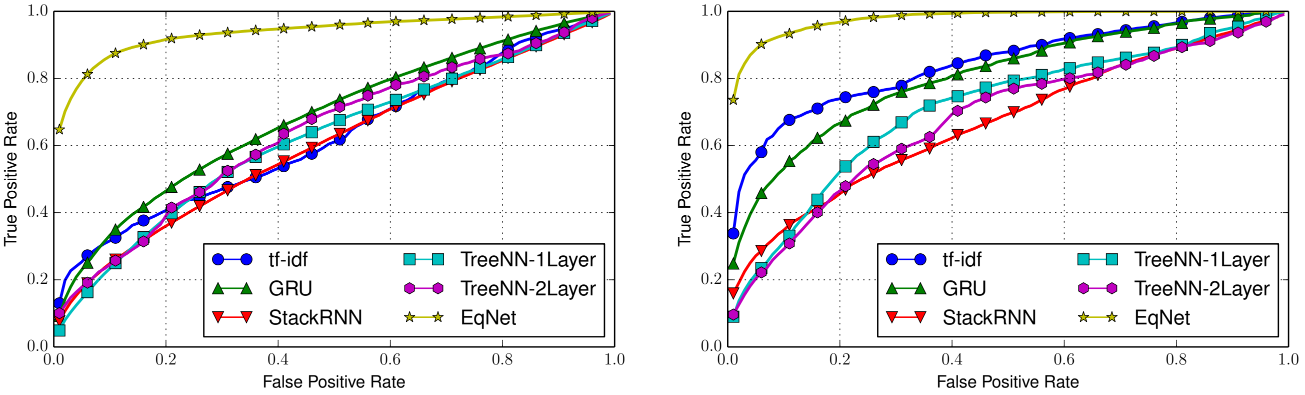
Model	Hyperparameters
EqNet	learning rate $10^{-2.1}$, rmsprop $\rho = 0.88$, momentum 0.88, minibatch size 900, representation size $D = 64$, autoencoder size $M = 8$, autoencoder noise $\kappa = 0.61$, gradient clipping 1.82, initial parameter standard deviation $10^{-2.05}$, dropout rate .11, hidden layer size 8, $\nu = 4$, curriculum initial tree size 6.96, curriculum step per epoch 2.72, objective margin $m = 0.5$
1-layer-TreeNN	learning rate $10^{-3.5}$, rmsprop $\rho = 0.6$, momentum 0.01, minibatch size 650, representation size $D = 64$, gradient clipping 3.6, initial parameter standard deviation $10^{-1.28}$, dropout 0.0, curriculum initial tree size 2.8, curriculum step per epoch 2.4, objective margin $m = 2.41$
2-layer-TreeNN	learning rate $10^{-3.5}$, rmsprop $\rho = 0.9$, momentum 0.95, minibatch size 1000, representation size $D = 64$, gradient clipping 5, initial parameter standard deviation 10^{-4}, dropout 0.0, hidden layer size 16, curriculum initial tree size 6.5, curriculum step per epoch 2.25, objective margin $m = 0.62$
GRU	learning rate $10^{-2.31}$, rmsprop $\rho = 0.90$, momentum 0.66, minibatch size 100, representation size $D = 64$, gradient clipping 0.87, token embedding size 128, initial parameter standard deviation 10^{-1}, dropout rate 0.26
StackRNN	learning rate $10^{-2.9}$, rmsprop $\rho = 0.99$, momentum 0.85, minibatch size 500, representation size $D = 64$, gradient clipping 0.70, token embedding size 64, RNN parameter weights initialization standard deviation 10^{-4}, embedding weight initialization standard deviation 10^{-3}, dropout 0.0, stack count 40



(a) SeenEqClass evaluation using model trained on simpler datasets. Caption is "model trained on"→"Test dataset".

(b) Evaluation of compositionality. UnseenEqClass evaluation using model trained on simpler datasets. Caption is "model trained on"→"Test dataset".

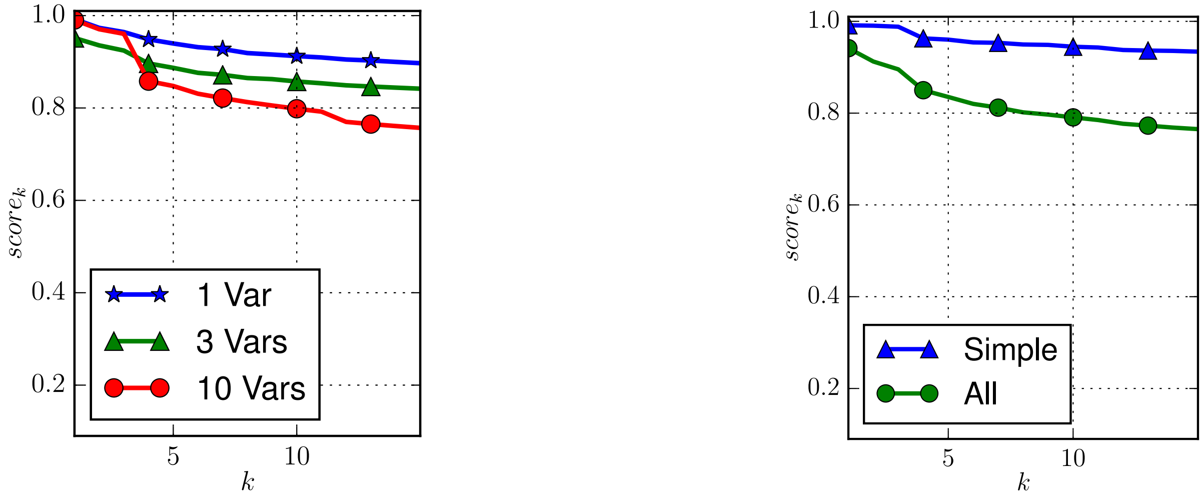
Figure 2. Evaluation of compositionality. Evaluation of $score_x$ (y axis) for $x = 1, \ldots, 15$. The markers are shown every five ticks of the x-axis to make the graph more clear. TreeNN refers to the model of Socher et al. (2012).



(a) SeenEqClass

(b) UnseenEqClass

Figure 3. Receiver operating characteristic (ROC) curves averaged across datasets.



(a) Performance *vs.* Number of Variables

(b) Performance *vs.* Operator Complexity

Figure 4. EQNET performance on SEENEQCLASS for various dataset characteristics