CICM'2018: First Experiments with Neural Translation of Informal to Formal Mathematics

Qingxiang Wang (Shawn)

University of Innsbruck &

Czech Technical University in Prague

August 2018

Overview

- Why Auto-formalization?
- Machine Learning in Auto-formalization
- Deep Learning
- Deep Learning in Theorem Proving
- An Initial Experiment
- Further Experiments
- Discussion

A mathematical paper published in 2001 in *Annals of Mathematics*:

Invariant differential operators and eigenspace representations on an affine symmetric space

By Jing-Song Huang*

Abstract

Let G/H be an affine symmetric space of split rank r. Let $\mathbf{D}$ be a preferred polynomial algebra of G-invariant differential operators on G/H generated by r elements. We show that the space of K-finite joint eigenfunctions of $\mathbf{D}$ on G/H form an admissible $(\mathfrak{g}, K)$-module which is called an eigenspace representation. The main content of this paper is description of the algebras of invariant differential operators and determination of the eigenspace representations on G/H. We also obtain a Poisson transform for τ-spherical eigenfunctions on G/H by Eisenstein integrals.

Gaps were found in 2008. It took 7 years for the author to fixed the proof.

Erratum and Addendum to: Invariant Differential Operators and Eigenspace Representations on an Affine Symmetric Space

Jing-Song Huang

(Submitted on 15 Jul 2017)

The purpose of this erratum and addendum is to correct the errors in [1]. It consists of five components:
1. Lemma 7.1 and Proposition 7.2 are wrong and discarded;
2. A new proof of existence $\lambda(\xi)$ in (7.1) without Proposition 7.2;
3. Definition of a new bijection in Theorem 5.2 and a proof by a new technique;
4. A new proof of Theorem 5.5 based on the new bijection in Theorem 5.2;
5. Correction to the list of exceptional simple pairs in Proposition 3.1.
The main results of [1] remain true as stated. We also add a final remark on generalization.

In 2017, the 16-year old paper was withdrawn:

Erratum and Addendum to: Invariant Differential Operators and Eigenspace Representations on an Affine Symmetric Space

Jing-Song Huang

(Submitted on 15 Jul 2017)

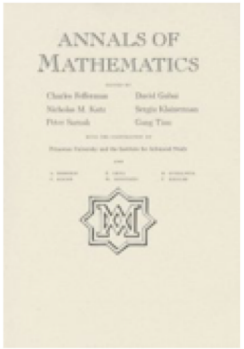
The purpose of this erratum and addendum is to
1. Lemma 7.1 and Proposition 7.2 are wrong and
2. A new proof of existence $\lambda(\xi)$ in (7.1) without
3. Definition of a new bijection in Theorem 5.2 a
4. A new proof of Theorem 5.5 based on the new
5. Correction to the list of exceptional simple pai

The main results of [1] remain true as stated. We

Author "shocked" after top math journal retracts paper

One of the world's most prestigious mathematics journals has issued what appears to be its first retraction.

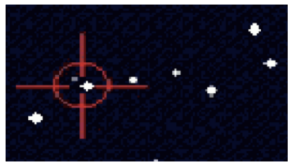
The *Annals of Mathematics* recently withdrew a 2001 paper exploring the properties of certain symmetrical spaces.



Why Auto-formalization

- Formalized libraries.


Coq


Mizar

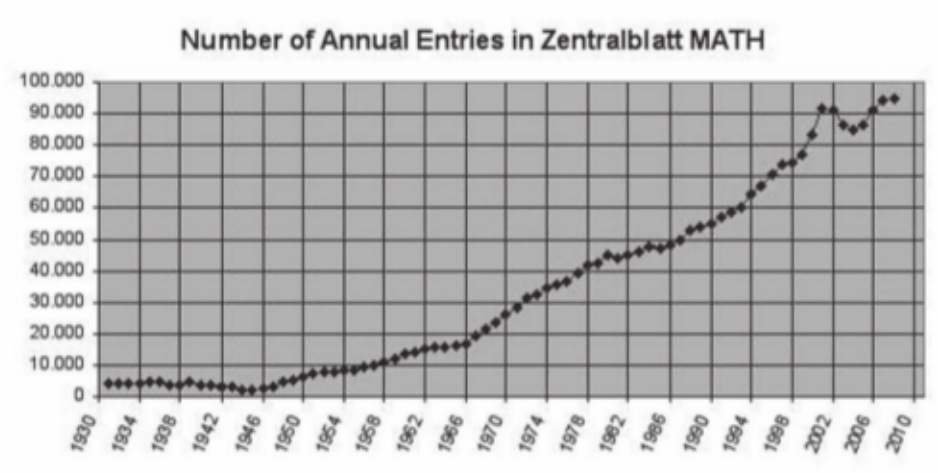

HOL


Metamath


Lean

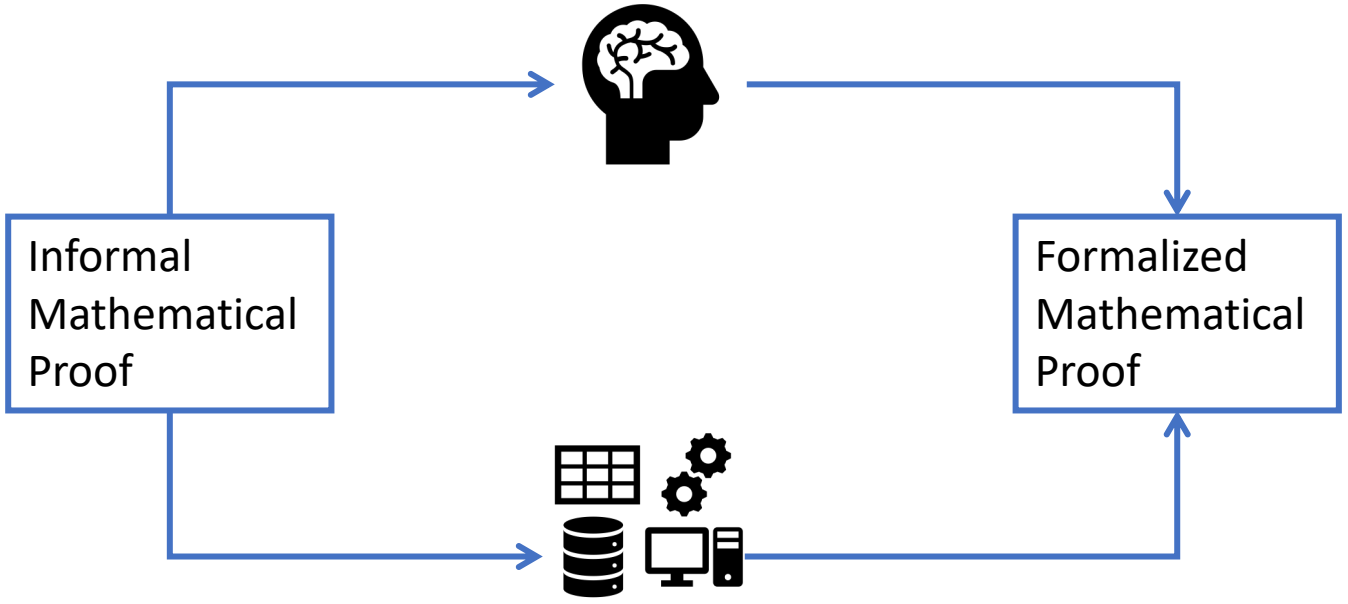
Isabelle

- Mizar contains over 10k definitions and over 50k proofs, yet…



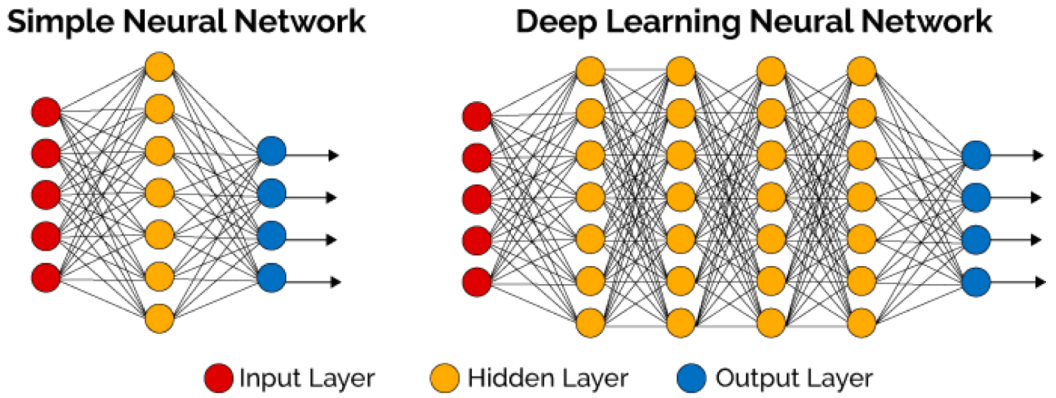
Machine Learning in Auto-formalization

- Function approximation view toward formalization and the prospect of machine learning approach to formalization.



Deep Learning

- Some theoretical results
  - Universal approximation theorem (Cybenko, Hornik), Depth separation theorem (Telgarsky, Shamir), etc
- Algorithmic techniques and novel architecture
  - Backpropagation, SGD, CNN, RNN, etc
- Advance in hardware and software
  - GPU, Tensorflow, etc
- Availability of large dataset
  - ImageNet, IWSLT, etc



Deep Learning in Theorem Proving

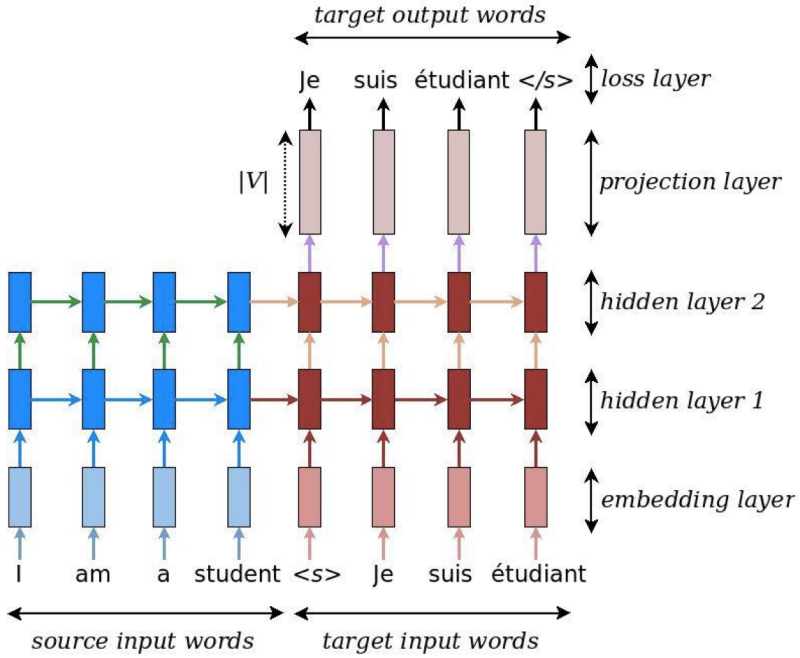
- Applications focus on doing ATP on existing libraries.

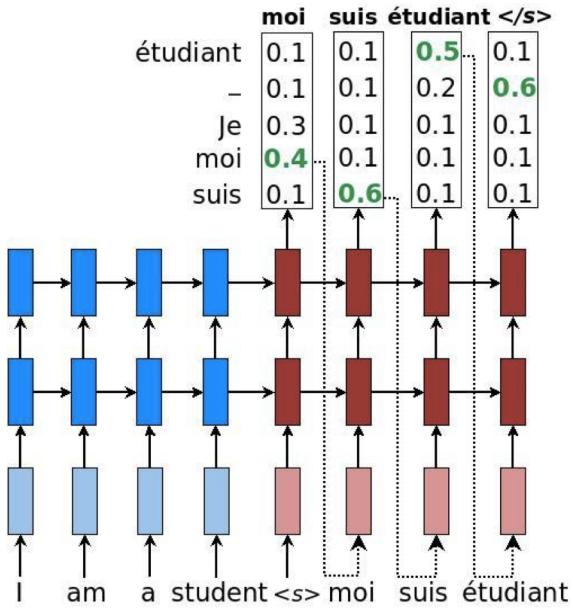
Year	Authors	Architecture	Dataset
Jun, 2016	Alemi et al.	CNN, LSTM/GRU	MMLFOF (Mizar)
Aug, 2016	Whalen	RL, GRU	Metamath
Jan, 2017	Loos et al.	CNN, WaveNet, RecursiveNN	MMLFOF (Mizar)
Mar, 2017	Kaliszyk et al.	CNN, LSTM	HolStep (HOL-Light)
Sep, 2017	Wang et al.	FormulaNet	HolStep (HOL-Light)
May, 2018	Kaliszyk et al.	RL	MMLFOF (Mizar)

- Opportunities of deep learning in formalization.

An Initial Experiment

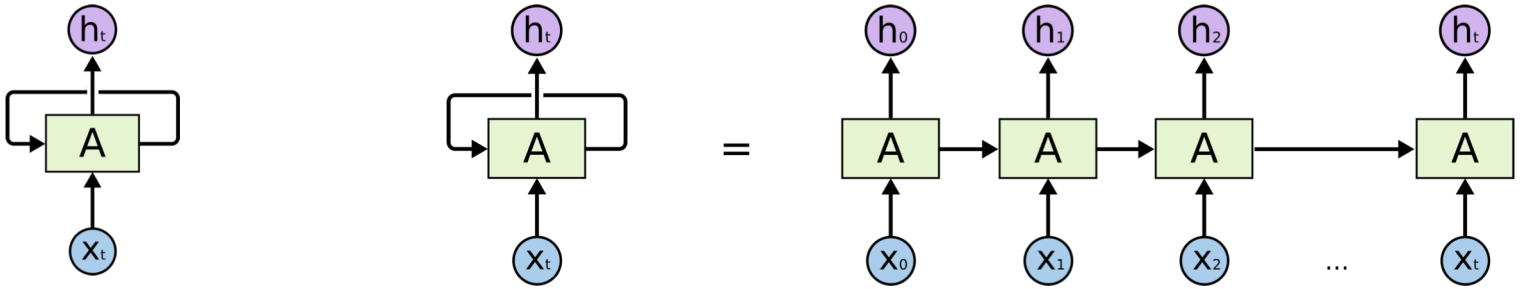
- Visit to Prague in January.
- Neural machine translation (Seq2seq model, Luong 2017).
- Can be considered as a complicated differentiable function.

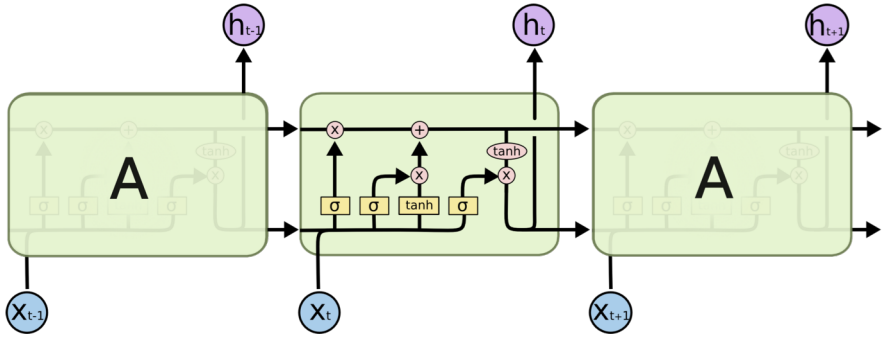


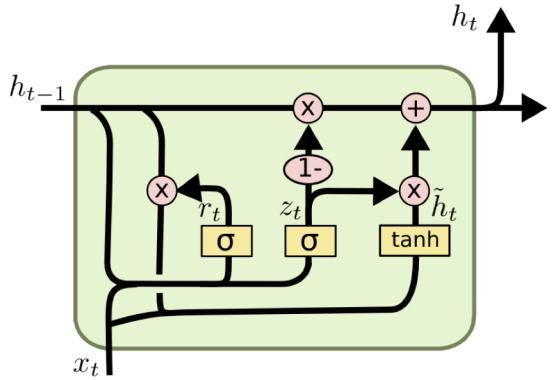


An Initial Experiment

- Recurrent neural network (RNN) and Long short-term memory cell (LSTM)







$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

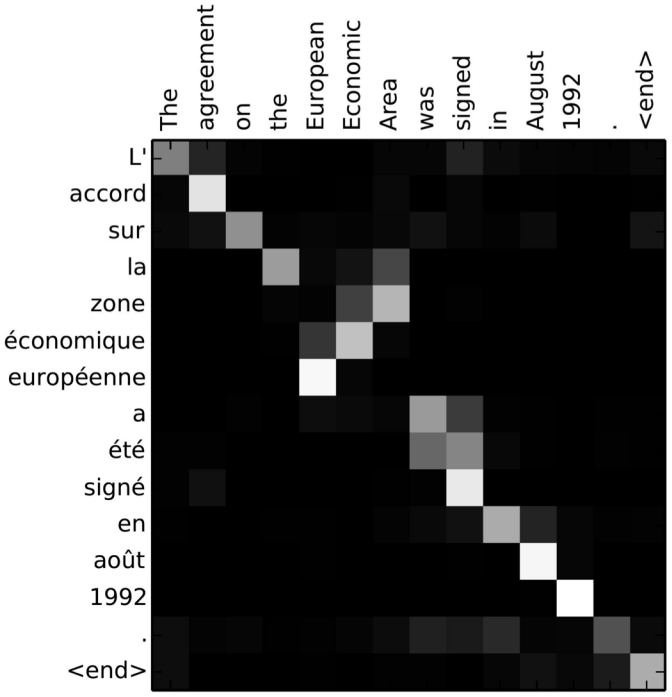
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

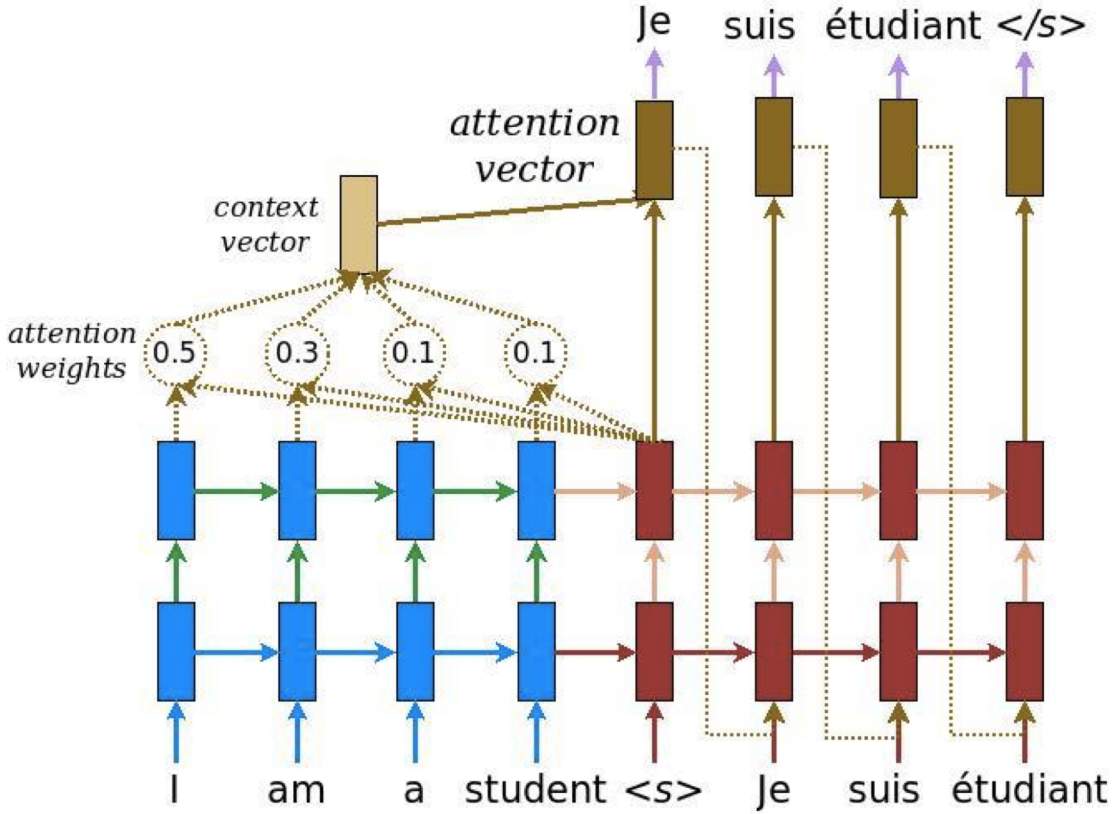
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

An Initial Experiment

- Attention mechanism





$$\alpha_{ts} = \frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'=1}^{S} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)} \quad \text{[Attention weights]} \quad (1)$$

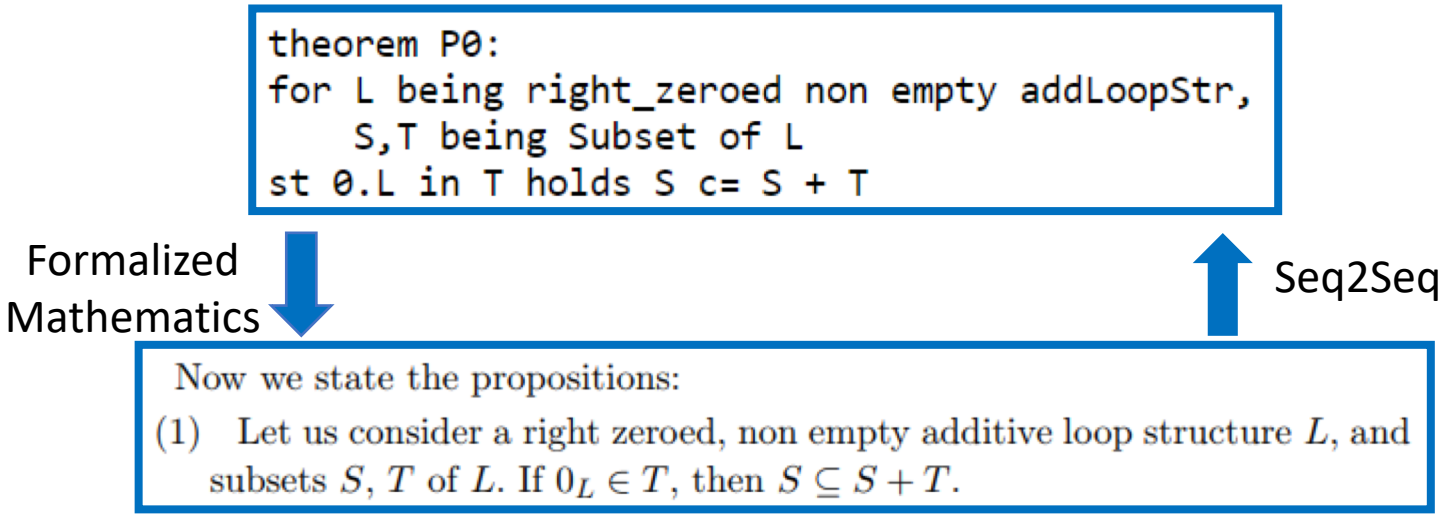
$$\boldsymbol{c}_t = \sum_s \alpha_{ts} \bar{\boldsymbol{h}}_s \quad \text{[Context vector]} \quad (2)$$

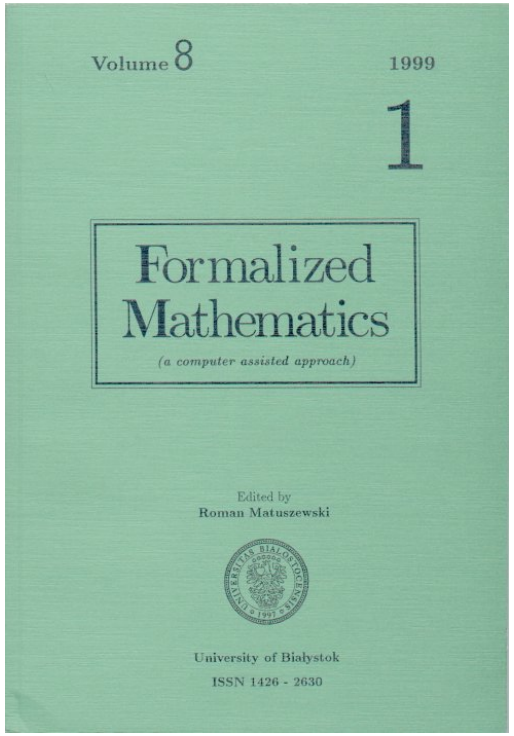
$$\boldsymbol{a}_t = f(\boldsymbol{c}_t, \boldsymbol{h}_t) = \tanh(\boldsymbol{W_c}[\boldsymbol{c}_t; \boldsymbol{h}_t]) \quad \text{[Attention vector]} \quad (3)$$

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \boldsymbol{W} \bar{\boldsymbol{h}}_s & \text{[Luong's multiplicative style]} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_1} \boldsymbol{h}_t + \boldsymbol{W_2} \bar{\boldsymbol{h}}_s\right) & \text{[Bahdanau's additive style]} \end{cases} \quad (4)$$

An Initial Experiment

- Raw data from Grzegorz Bancerek (2017†).
- Formal abstracts of *Formalized mathematics*, which are generated latex from Mizar (v8.0.01_5.6.1169)
- Extract Latex-Mizar statement pairs as training data. Use Latex as source and Mizar as target.





An Initial Experiment

- In total, 53368 theorems (schema) statements were divided by a 10:1 ratio.
- Both Latex and Mizar tokenized to accommodate the framework.

Latex	If $ X \mathrel { = } { \rm the ~ } { { { \rm carrier } ~ { \rm of } ~ { \rm } } } { A _ { 9 } } $ and $ X $ is plane , then $ { A _ { 9 } } $ is an affine plane .
Mizar	X = the carrier of AS & X is being_plane implies AS is AffinPlane ;
Latex	If $ { s _ { 9 } } $ is convergent and $ { s _ { 8 } } $ is a subsequence of $ { s _ { 9 } } $, then $ { s _ { 8 } } $ is convergent .
Mizar	seq is convergent & seq1 is subsequence of seq implies seq1 is convergent ;

An Initial Experiment

- Preliminary result (among the 4851 test statements)

Attention mechanism	Number of identical statements generated	Percentage
No attention	120	2.5%
Bahdanau	165	3.4%
Normed Bahdanau	1267	26.12%
Luong	1375	28.34%
Scaled Luong	1270	26.18%
Any	1782	36.73%

- A good correspondence between Latex and Mizar, probably easy to learn.

An Initial Experiment

- Sample unmatched statements

Attention mechanism	Mizar statement
Correct statement	`for T being Noetherian sup-Semilattice for I being Ideal of T holds ex_sup_of I , T & sup I in I ;`
No attention	`for T being lower-bounded sup-Semilattice for I being Ideal of T holds I is upper-bounded & I is upper-bounded ;`
Bahdanau	`for T being T , T being Ideal of T , I being Element of T holds height T in I ;`
Normed Bahdanau	`for T being Noetherian adj-structured sup-Semilattice for I being Ideal of T holds ex_sup_of I , T & sup I in I ;`
Luong	`for T being Noetherian adj-structured sup-Semilattice for I being Ideal of T holds ex_sup_of I , T & sup I in I ;`
Scaled Luong	`for T being Noetherian sup-Semilattice , I being Ideal of T ex I , sup I st ex_sup_of I , T & sup I in I ;`

An Initial Experiment

- Neural translation w.r.t. number of training steps

Rendered Latex	Suppose s_1 is convergent and s_2 is convergent. Then $lim\ (s_1 + s_2) = lim\ s_1 + lim\ s_2$
Snapshot-1000	x in dom f implies (x * y) * (f | (x | (y | (y | y)))) = (x | (y | (y | (y | y)))) ;
Snapshot-3000	seq is convergent & lim seq = 0c implies seq = seq ;
Snapshot-5000	seq1 is convergent & lim seq2 = lim seq2 implies lim_inf seq1 = lim_inf seq2 ;
Snapshot-7000	seq is convergent & seq9 is convergent implies lim (seq + seq9) = (lim seq) + (lim seq9) ;
Snapshot-9000	seq1 is convergent & lim seq1 = lim seq2 implies (seq1 + seq2) + (lim seq1) = (lim seq1) + (lim seq2) ;
Snapshot-12000	seq1 is convergent & seq2 is convergent implies lim (seq1 + seq2) = (lim seq1) + (lim seq2) ;
Correct	seq1 is convergent & seq2 is convergent implies lim (seq1 + seq2) = (lim seq1) + (lim seq2) ;

Further Experiments

- More data available in April after the work of Naumowicz et al. [T23]
- Not only theorems, but also all the individual proof steps.
- Results are 1,056,478 pairs of Latex– Mizar sentences.

- [T22] André Greiner-Petter, Moritz Schubotz, Howard Cohl and Bela Gipp. *MathTools: An Open API for Convenient MathML Handling*
- [T23] Grzegorz Bancerek, Adam Naumowicz and Josef Urban. *System Description: XSL-based Translator of Mizar to LaTeX*
- [T24] Moritz Schubotz. *VMEXT2: A Visual Wikidata aware Content MathML Editor*
- [T25] Richard Marcus, Michael Kohlhase and Florian Rabe. *Demo: TGView3D for Immersive Theory Graph Exploration*
- [T26] Michael Kohlhase. *Demo: Math Object Identifiers -- Towards Research Data in Mathematics*

Further Experiments

- Division of data

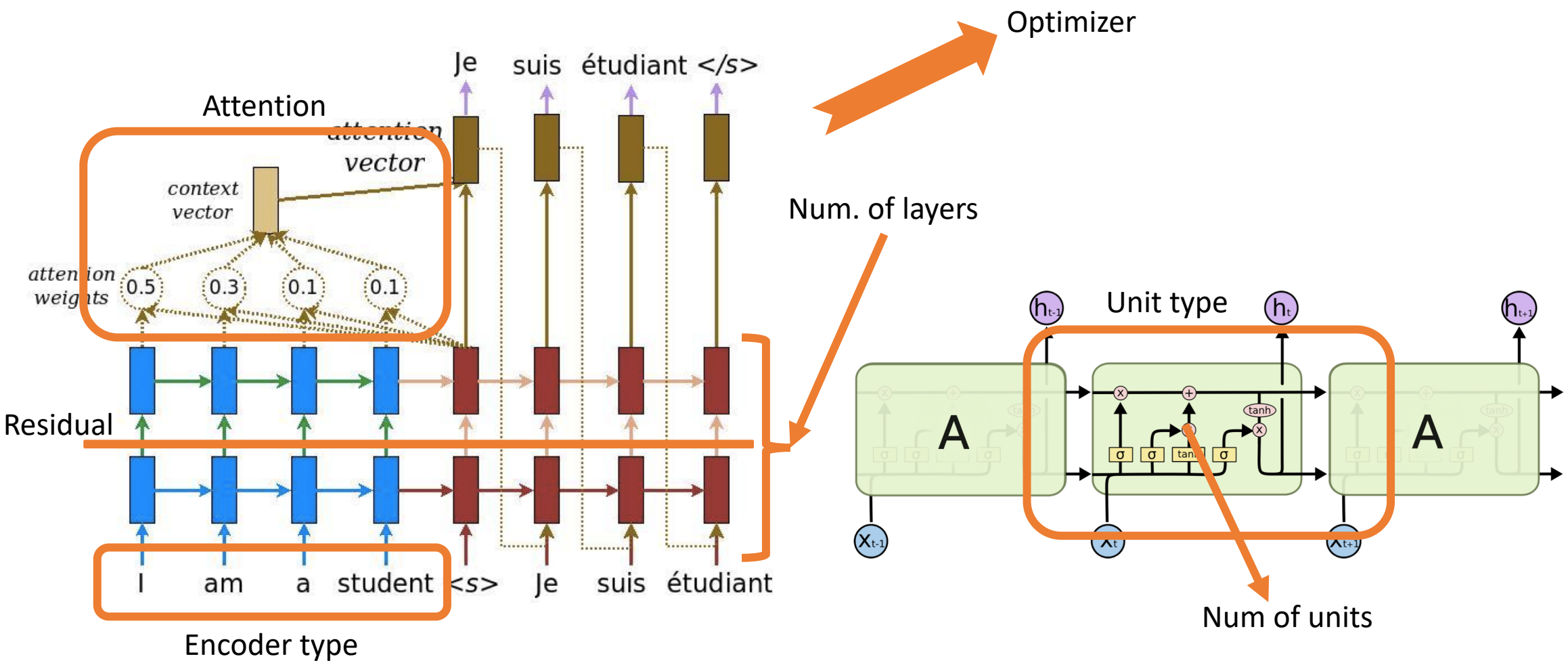
Category	Num of pairs/tokens
Total	1,056,478
Training data	947,231
Validation data (for NMT model selection)	2,000
Testing data (for NMT model selection)	2,000
Inference data	105,247
Unique tokens for Latex	7,820
Unique tokens for Mizar	16,793
Overlap between Training and Inference	57,145

- Overlapping data constitutes 54.3% of the inference set.

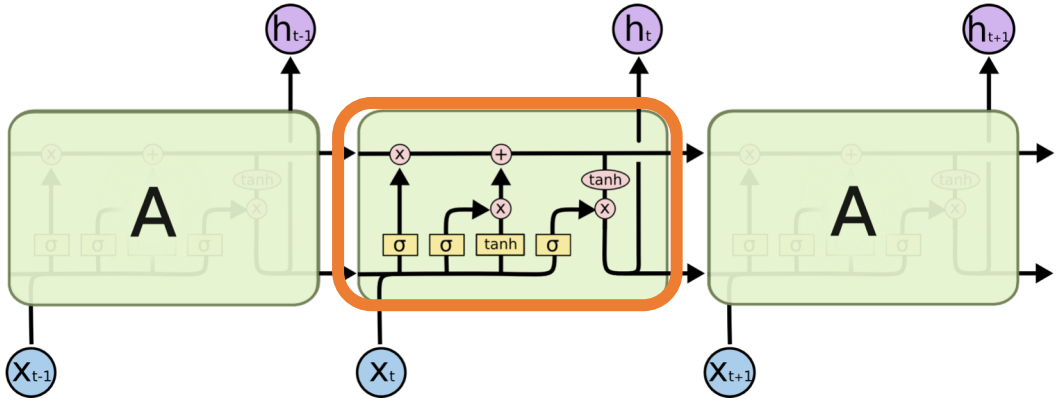
Further Experiments

- Tweaking hyperparameters

Name	Values	Description
Unit type	• LSTM (default) • GRU • Layer-norm LSTM	Type of the memory cell in RNN
Attention	• No attention (default) • (Normed) Bahdanau • (Scaled) Luong	The attention mechanism
Num. of layers	• 2 layers (default) • 3 / 4 / 5 / 6 layers	RNN layers in encoder and decoder
Residual	• False (default) • True	Enables residual layers (to overcome exploding/vanishing gradients)
Optimizer	• SGD (default) • Adam	The gradient-based optimization method
Encoder type	• Unidirectional (default) • Bidirectional	Type of encoding methods for input sentences
Num. of units	• 128 (default) • 256 / 512 / 1024 / 2048	The dimension of parameters in a memory cell


Optimizer
Attention
Je
suis
étudiant </s>
vector
context vector
attention weights
0.5
0.3
0.1
0.1
Num. of layers
Residual
Unit type
A
A
I
am
a
student
<s>
Je
suis
étudiant
Num of units
Encoder type

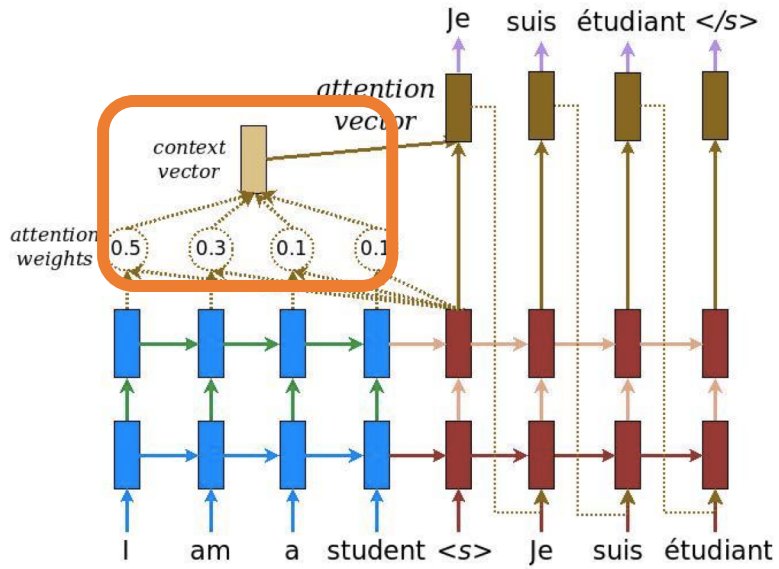
- Memory-cell unit types



Parameter	Final Test Perplexity	Final Test BLEU	Identical Statements (%)	Identical No-overlap (%)
LSTM	**3.06**	**41.1**	**40121 (38.12%)**	**6458 (13.43%)**
GRU	3.39	34.7	37758 (35.88%)	5566 (11.57%)
Layer-norm LSTM	11.35	0.4	11200 (10.64%)	1 (0%)

Table 5. Evaluation on type of memory cell (attention not enabled)

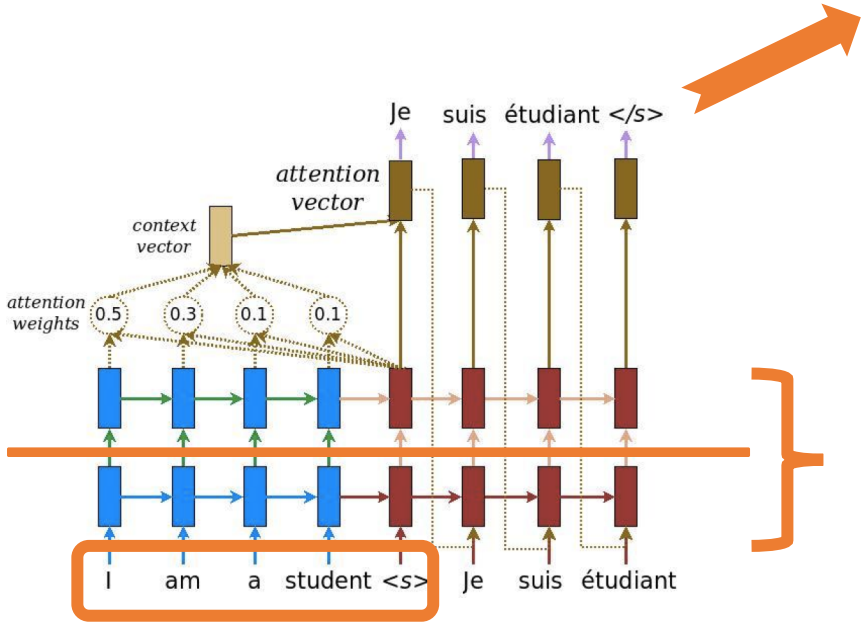
- Attention



Parameter	Final Test Perplexity	Final Test BLEU	Identical Statements (%)	Identical No-overlap (%)
No Attention	3.06	41.1	40121 (38.12%)	6458 (13.43%)
Bahdanau	3	40.9	44218 (42.01%)	8440 (17.55%)
Normed Bahdanau	1.92	63.5	60192 (57.19%)	18057 (37.54%)
Luong	**1.89**	64.8	60151 (57.15%)	18013 (37.45%)
Scaled Luong	2.13	**65**	**60703 (57.68%)**	**18105 (37.64%)**

Table 6. Evaluation on type of attention mechanism (LSTM cell)

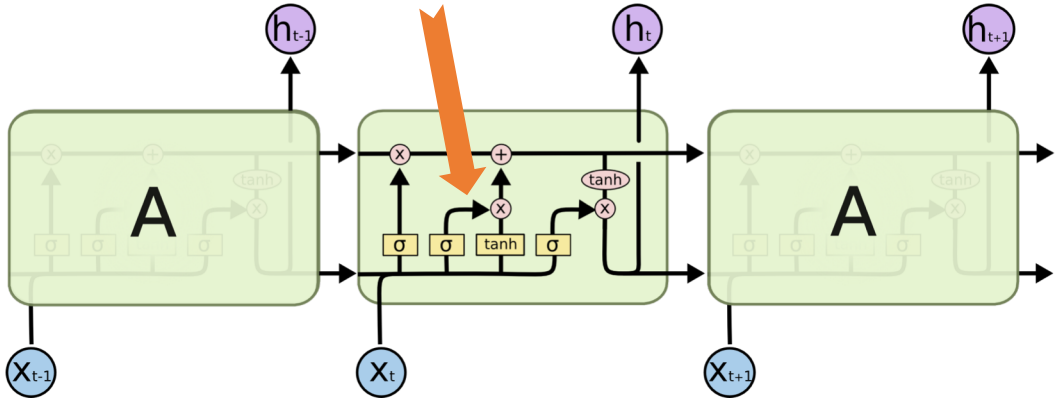
- Residuals, layers, etc.



Parameter	Final Test Perplexity	Final Test BLEU	Identical Statements (%)	Identical No-overlap (%)
2-Layer	3.06	41.1	40121 (38.12%)	6458 (13.43%)
3-Layer	2.10	64.2	57413 (54.55%	16318 (33.92%)
4-Layer	2.39	45.2	49548 (47.08%)	11939 (24.82%)
5-Layer	5.92	12.8	29207 (27.75%)	2698 (5.61%)
6-Layer	4.96	20.5	29361 (27.9%)	2872 (5.97%)
2-Layer Residual	1.92	54.2	57843 (54.96%)	16511 (34.32%)
3-Layer Residual	1.94	62.6	59204 (56.25%)	17396 (36.16%)
4-Layer Residual	1.85	56.1	59773 (56.79%)	17626 (36.64%)
5-Layer Residual	2.01	63.1	59259 (56.30%)	17327 (36.02%)
6-Layer Residual	NaN	0	0 (0%)	0 (0%)
2-Layer Adam	1.78	56.6	61524 (58.46%)	18635 (38.74%)
3-Layer Adam	1.91	60.8	59005 (56.06%)	17213 (35.78%)
4-Layer Adam	1.99	51.8	57479 (54.61%)	16288 (33.86%)
5-Layer Adam	2.16	54.3	54670 (51.94%)	14769 (30.70%)
6-Layer Adam	2.82	37.4	46555 (44.23%)	10196 (21.20%)
2-Layer Adam Res.	1.75	56.1	63242 (60.09%)	19716 (40.97%)
3-Layer Adam Res.	1.70	55.4	64512 (61.30%)	20534 (42.69%)
4-Layer Adam Res.	1.68	57.8	64399 (61.19%)	20353 (42.31%)
5-Layer Adam Res.	1.65	64.3	64722 (61.50%)	20627 (42.88%)
6-Layer Adam Res.	1.66	59.7	65143 (61.90%)	20854 (43.35%)
2-Layer Bidirectional	2.39	**69.5**	63075 (59.93%)	19553 (40.65%)
4-Layer Bidirectional	6.03	63.4	58603 (55.68%)	17222 (35.80%)
6-Layer Bidirectional	2	56.3	57896 (55.01%)	16817 (34.96%)
2-Layer Adam Bi.	1.84	56.9	64918 (61.68%)	20830 (43.30%)
4-Layer Adam Bi.	1.94	58.4	64054 (60.86%)	20310 (42.22%)
6-Layer Adam Bi.	2.15	55.4	60616 (57.59%)	18196 (37.83%)
2-Layer Bi. Res.	2.38	24.1	47531 (45.16%)	11282 (23.45%)
4-Layer Bi. Res.	NaN	0	0 (0%)	0 (0%)
6-Layer Bi. Res.	NaN	0	0 (0%)	0 (0%)
2-Layer Adam Bi. Res.	1.67	62.2	65944 (62.66%)	21342 (44.37%)
4-Layer Adam Bi. Res.	**1.62**	66.5	65992 (62.70%)	21366 (44.42%)
6-Layer Adam Bi. Res.	1.63	58.3	**66237 (62.93%)**	**21404 (44.50%)**

Table 7. Evaluation on various hyperparameters w.r.t. layers

- Unit dimension in cell



Parameter	Final Test Perplexity	Final Test BLEU	Identical Statements (%)	Identical No-overlap (%)	Training Time (hrs.)
128 Units	3.06	41.1	40121 (38.12%)	6458 (13.43%)	1
256 Units	1.59	64.2	63433 (60.27%)	19685 (40.92%)	3
512 Units	1.6	**67.9**	66361 (63.05%)	21506 (44.71%)	5
1024 Units	**1.51**	61.6	**69179 (65.73%)**	**22978 (47.77%)**	11
2048 Units	2.02	60	59637 (56.66%)	16284 (33.85%)	31

Table 8. Evaluation on number of units

- Greedy covers and edit distances

	Identical Statements	0	≤ 1	≤ 2	≤ 3
Best Model - 1024 Units	69179 (total) 22978 (no-overlap)	65.73% 47.77%	74.58% 59.91%	86.07% 70.26%	88.73% 74.33%
Top-5 Greedy Cover - 1024 Units - 4-Layer Bi. Res. - 512 Units - 6-Layer Adam Bi. Res. - 2048 Units	78411 (total) 28708 (no-overlap)	74.50% 59.68%	82.07% 70.85%	87.27% 78.84%	89.06% 81.76%
Top-10 Greedy Cover - 1024 Units - 4-Layer Bi. Res. - 512 Units - 6-Layer Adam Bi. Res. - 2048 Units - 2-Layer Adam Bi. Res. - 256 Units - 5-Layer Adam Res. - 6-Layer Adam Res. - 2-Layer Bi. Res.	80922 (total) 30426 (no-overlap)	76.89% 63.25%	83.91% 73.74%	88.60% 81.07%	90.24% 83.68%
Union of All 39 Models	83321 (total) 32083 (no-overlap)	79.17% 66.70%	85.57% 76.39%	89.73% 82.88%	91.25% 85.30%

Table 9. Coverage w.r.t. a set of models and edit distances

- Translating from Mizar back to Latex

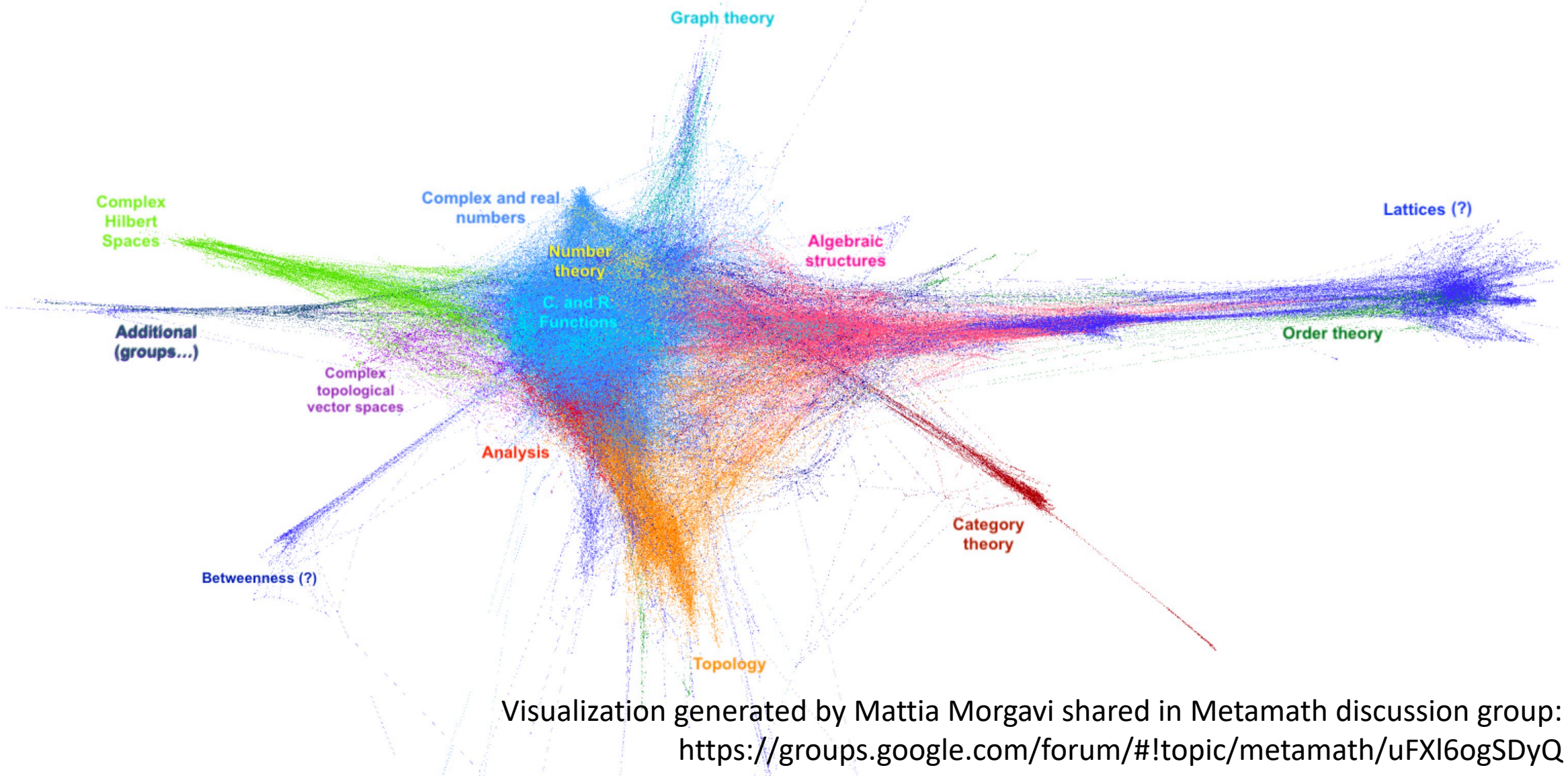
Parameter	Final Test Perplexity	Final Test BLEU	Identical Statements	Percentage
512 Units Bidirectional Scaled Luong	2.91	57	54320	51.61%

Table 10. Evaluation on number of units

Discussion

- Formalization using deep learning is a promising direction.
- Deep learning and AI, open to further development.
- Understanding mathematical statements versus general natural language understanding.
- Implication of achieving auto-formalization.
- Lots of challenges await us.

Thanks



Visualization generated by Mattia Morgavi shared in Metamath discussion group:
https://groups.google.com/forum/#!topic/metamath/uFXl6ogSDyQ