Building an Auto-formalization Infrastructure from Mathematical Literature through Deep Learning – Project Description

Qingxiang Wang (Shawn)

University of Innsbruck &

Czech Technical University in Prague

March 2018

Overview

- Why Auto-formalization?
- Machine Learning in Auto-formalization
- Deep Learning
- Deep Learning in Theorem Proving
- An Initial Experiment
- 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In 2017, the 16-year old paper was withdrawn: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

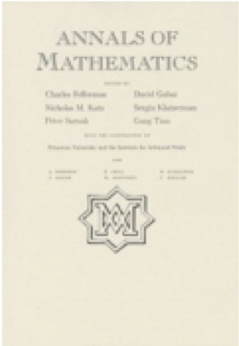
(Submitted on 15 Jul 2017)

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

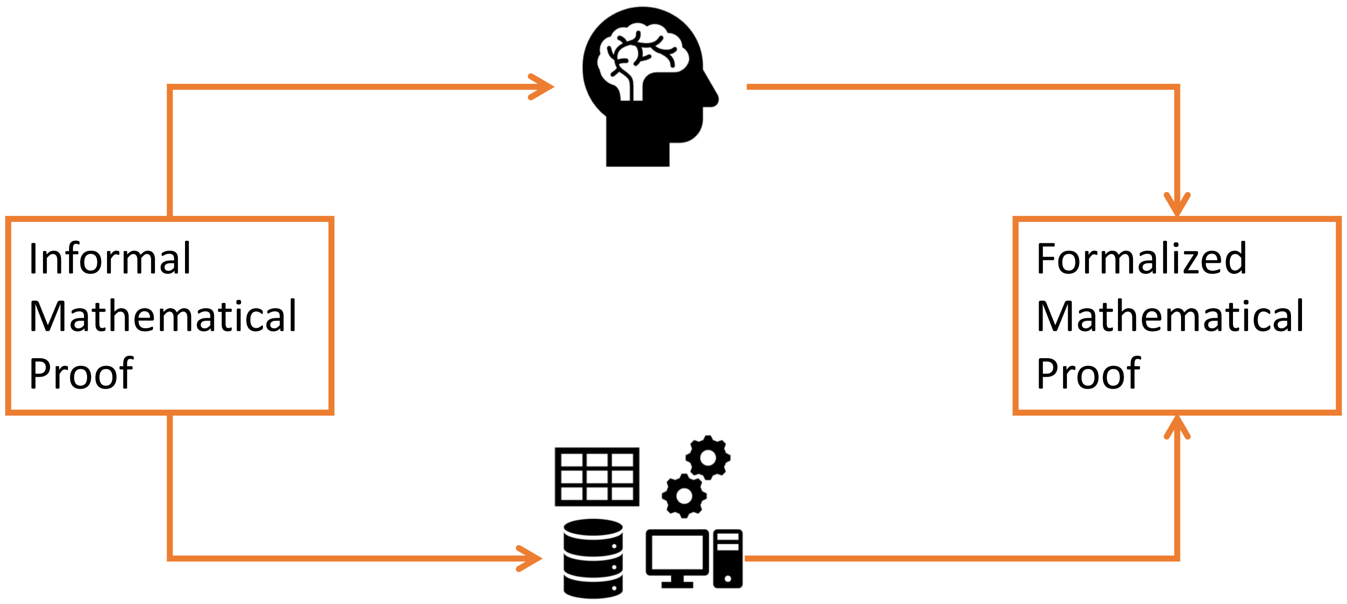
- Formalized libraries.

Coq Mizar HOL Metamath Lean Isabelle

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

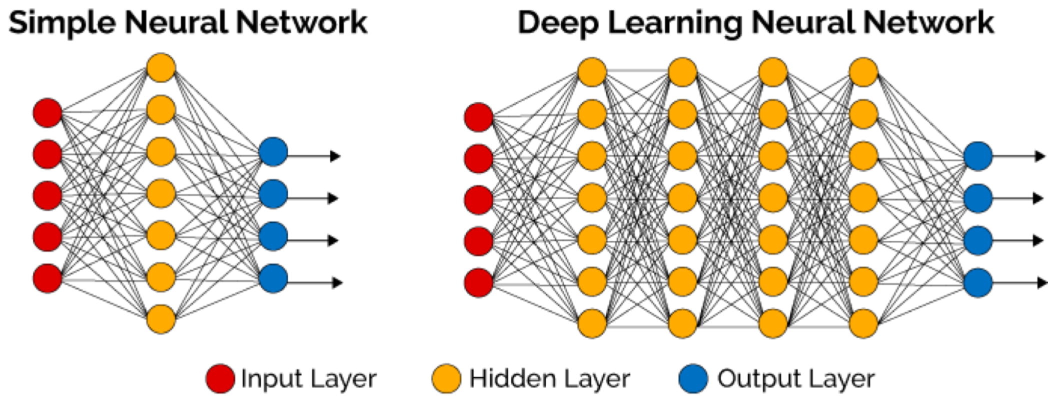
Machine Learning in Auto-formalization

- 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

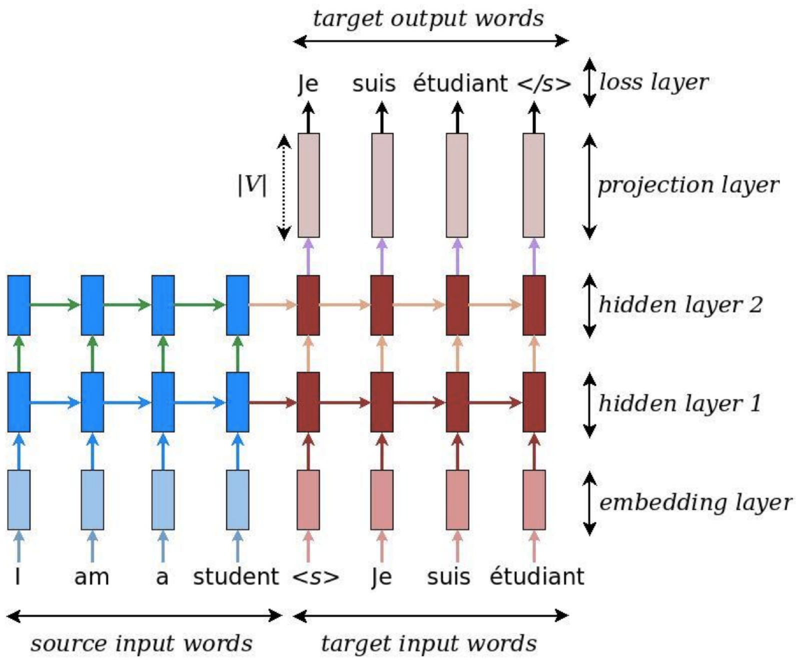
- Applications focus on doing ATP on existing libraries.

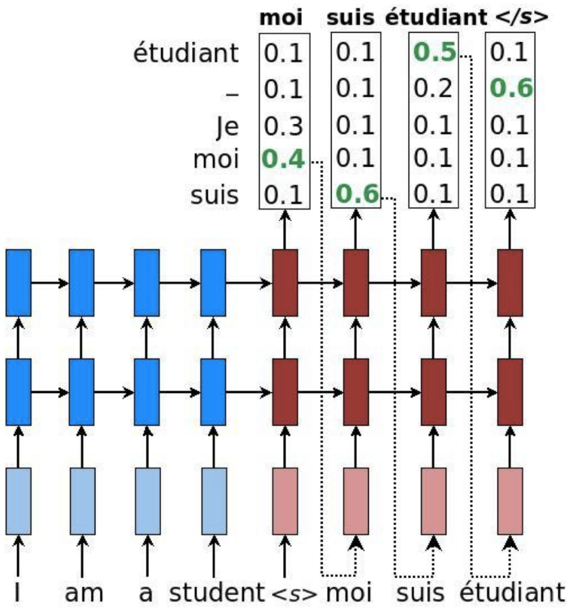
Year	Authors	Architecture	Dataset	Performance
Jun, 2016	Alemi et al.	CNN, LSTM/GRU	MMLFOF (Mizar)	80.9%
Aug, 2016	Whalen	RL, GRU	Metamath	14%
Jan, 2017	Loos et al.	CNN, WaveNet, RecursiveNN	MMLFOF (Mizar)	81.5%
Mar, 2017	Kaliszyk et al.	CNN, LSTM	HolStep (HOL-Light)	83%
Sep, 2017	Wang et al.	FormulaNet	HolStep (HOL-Light)	90.3%

- Opportunities of deep learning in formalization.

An Initial Experiment

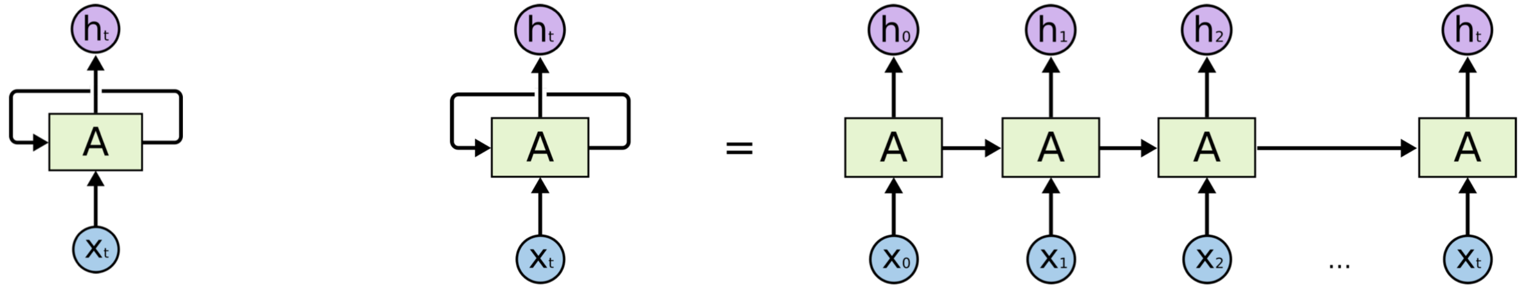
- Visit to Prague in January.
- Neural machine translation (Seq2seq model, Luong 2017).
- Can be considered as a complicated differentiable function.

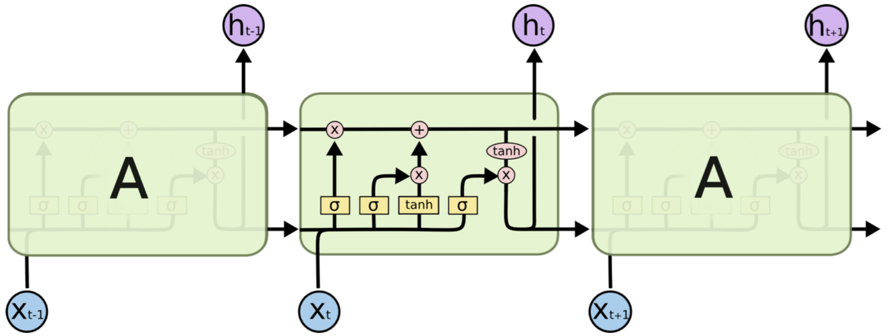


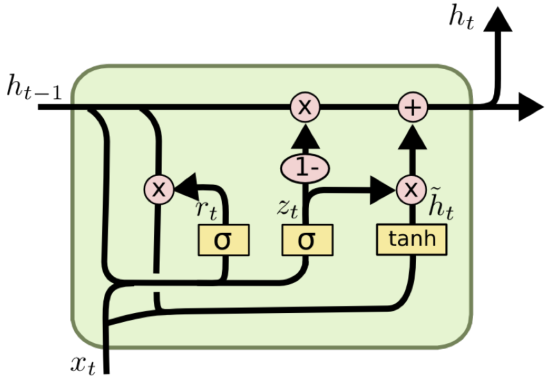


An Initial Experiment

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







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

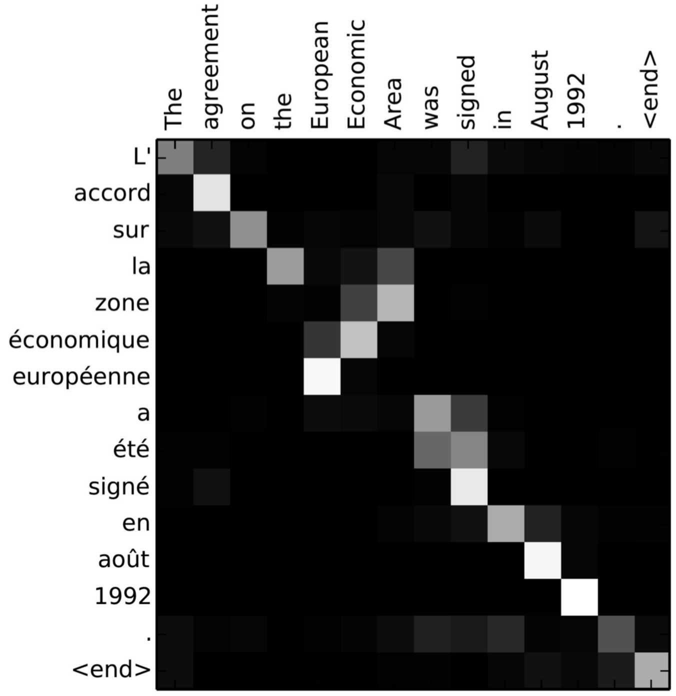
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

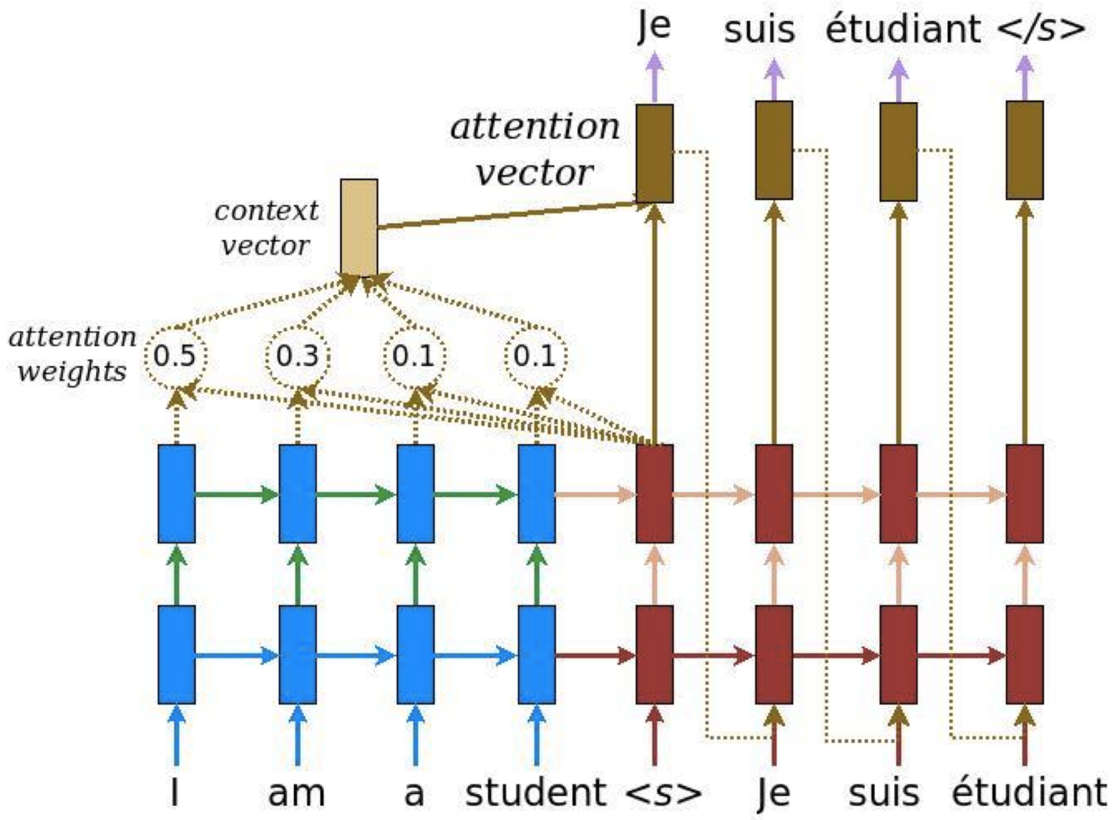
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

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

An Initial Experiment

- Attention mechanism





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

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

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

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

An Initial Experiment

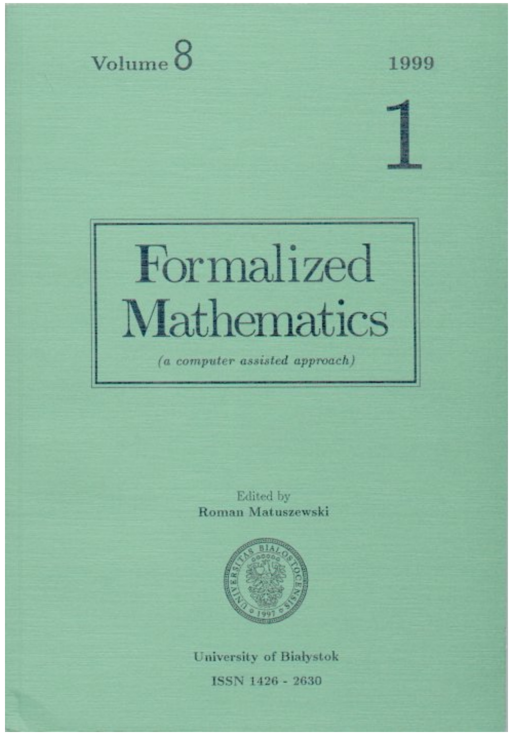
- Raw data from Grzegorz Bancerek (2017†).
- Formal abstracts of *Formalized mathematics*, which are generated latex from Mizar (v8.0.01_5.6.1169)
- Extract Latex-Mizar statement pairs as training data. Use Latex as source and Mizar as target.

```
theorem P0:
for L being right_zeroed non empty addLoopStr,
    S,T being Subset of L
st 0.L in T holds S c= S + T
```

Formalized Mathematics ↓ ↑ Seq2Seq

Now we state the propositions:

(1) Let us consider a right zeroed, non empty additive loop structure L, and subsets S, T of L. If $0_L \in T$, then $S \subseteq S + T$.



An Initial Experiment

- In total, 53368 theorems (schema) statements were divided by 10:1 into:
  - Training set: 48517 statements
  - Test set: 4851 statements
- 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 ;`

- Further exploration in finding parsable statement, or hopefully generating syntactically correct statement.

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

...Ta mathemata [sic] are the things in so far as we take cognizance of them as what we already know them to be in advance, the body of the bodily, the plant-like of the plants, the animal-like of the animals, the thing-ness of the things, and so on. This genuine learning is therefore an extremely peculiar taking, a taking where one who takes only takes what one basically already gets...

Martin Heidegger, Modern Science, Metaphysics and Mathematics