Short-Term Scientific Mission Grant
- APPLICATION FORM[1] -

Action number: CA20111

Applicant name: Rashid Barket

Details of the STSM

Title: Tree Transformers for Symbolic Mathematics

Start and end date: June 30 – July 13

Detail of the cost in EUROS:

- Transport (upload screen capture): 97.08

- Hotel/day (upload screen capture): 104 x 13 = 1349

- Food/day: 35 x 14 = 490

TOTAL: 1936.08

Screenshots attached below

Goals of the STSM

Purpose and summary of the STSM.

(max.200 word)

We aim to show that in the field of machine learning for mathematics, current approaches using transformers do not take advantage of the structural hierarchy inherent in the data. There have been two approaches to introduce tree transformers, one by proposing a new positional encoding for tree structures [1] and one for altering the attention mechanism to consider the location of a node in a tree [2]. As of now, no research has been published on how tree versions of transformers can benefit mathematics or theorem-proving tasks. In some of my previous work [3], I showed how a tree version of an LSTM was beneficial for classification tasks in symbolic integration. We now propose to investigate whether using tree transformers can improve a variety of other mathematics tasks, such as symbolic integration [4] and symbolic regression [5]. I have experience with tree transformers, and my collaborator has experience in generative AI problems for mathematics. We plan to combine our knowledge together to work on these tasks. The ultimate goal is to show that tree transformers are better suited for symbolic mathematics tasks. We also wish to show how the work done in symbolic mathematics tasks can further develop machine learning in theorem proving.

---

1 This form is part of the application for a grant to visit a host organisation located in a different country than the country of affiliation. It is submitted to the COST Action MC via-e-COST. The Grant Awarding Coordinator coordinates the evaluation on behalf of the Action MC and informs the Grant Holder of the result of the evaluation for issuing the Grant Letter.

Working Plan

Description of the work to be carried out by the applicant.

(max.500 word)

The first step of the project will be to explore symbolic integration, which the corresponding researchers both have expertise in. We wish to see if and how switching from an ordinary transformer to a tree transformer will improve the results and generalisability of the models. We expect to see an increase in accuracy based on some initial testing from a (different) classification task I am currently working on [3]. However, the main findings would be to show how the models can generalise. This can be shown in two possible ways. The first is to train on data from specific data generation methods introduced in previous literature and then test on data that comes from external sources. We can also train on one type of data generator and then test on another. We aim to show that the tree based model can integrate sums of functions but only learn on each individual expression. That is, if int(f) and int(g) are in the dataset, we hypothesise that a tree transformer can predict int(f+g) even though int(f+g) is not in the dataset. The tree structure should help a tree transformer learn this more easily, but this may not necessarily be true about a regular transformer.

We will also look at the problem of symbolic regression [5]. Symbolic regression is an emerging field in machine learning and aims to infer mathematical expressions directly from data to provide interpretable insights into the relationships between variables. There are both transformer approaches and deep reinforcement learning approaches, but both rely on the same type of data. That is, S={x_i, y_i} input/output pairs and a symbolic equation that defines them. Given S, the ML method will predict the equation. Tree transformers will be useful in the decoder part of the architecture, and we will experiment with whether the new tree embedding scheme will aid in the equation discovery process.

These are the two primary tasks we wish to improve with tree transformers. We will first start by trying the tree transformer embedding scheme in place of the original positional encoder for both tasks. If time permits, we will also look at modifying the attention mechanism as well to see if we can get further gains on both tasks. Another potential goal is trying other tasks, such as solving ordinary differential equations and deriving integro-differential equations.

Expected outputs and contribution to the Action MoU objectives and deliverables.

Main expected results and their contribution to the progress towards the Action objectives (https://europroofnet.github.io/objectives/) and deliverables (https://europroofnet.github.io/deliverables/).

Working groups to which this mission contributes:

(max.500 words)

The expected results are implementing a new encoding scheme for symbolic expressions based on tree transformers [1]. As far as we are aware, this has never been done for any symbolic mathematics task before. With the new encoding scheme, we will then show that this helps with various symbolic mathematics tasks, such as symbolic integration and symbolic regression. If we have a positive result, we aim to publish this at a top machine learning conference such as NeurIPS, ICLR, or ICML. We are confident that the novelty of modifying transformers to suit the tree-based nature of symbolic mathematics is a major contribution towards mathematical reasoning in machine learning models. The generated datasets and code will also be available online for anyone to use.

This contributes to Working Group 5 (WG5): Machine Learning in Proofs. ML in symbolic mathematics and ML in theorem proving have many overlaps, and the work done on symbolic integration and regression has applications to theorem proving. Higher order logic formulas can also be represented as unary-binary trees, just like in symbolic mathematics. For instance, suppose we have the formula $\forall P.(P(a) \rightarrow P(b))$. This can be represented as the tree:

```
   ∀P
   |
   →
  / \
P(a)  P(b)
```

Similar to Symbolic Mathematics, ML tasks in theorem proving have only explored transformers using sequence embeddings and have not used the hierarchal information from the formulas in the ML architecture.

This is directly related to Research Coordination Objective 6: Develop the use of artificial intelligence and machine learning techniques on proofs. As symbolic mathematics is related to theorem proving, this also relates to Capacity Building Objective 6: Transfer knowledge in terms of expertise, scientific tools and human resources across the different disciplines and between academia and industry. The author is also sponsored by Maplesoft, a software company that develops the computer algebra system Maple. This creates more collaboration opportunities between academia and industry. Lastly, this is also related to deliverable D8: Detailed technical report on the evaluation of techniques for learning proof search guidance and premise selection in automated theorem provers. This report has emphasised how syntax trees are used in work like tree (recursive) neural networks. In systems like ENIGMA-NG, tree neural networks produce vector embeddings that represent the structure of terms and clauses. These embeddings are then used for tasks like clause selection in saturation-based automated theorem provers. This idea has not yet been explored in transformers, and the work done in this STSM will help progress these objectives and deliverables.

References

[1] Vighnesh Shiv, and Chris Quirk. (2019). Encodings to enable tree-based transformers. Proceedings of the 33rd International Conference on Neural Information Processing Systems. url=https://papers.nips.cc/paper_files/paper/2019/hash/6e0917469214d8fbd8c517dcdc6b8dcf-Abstract.html

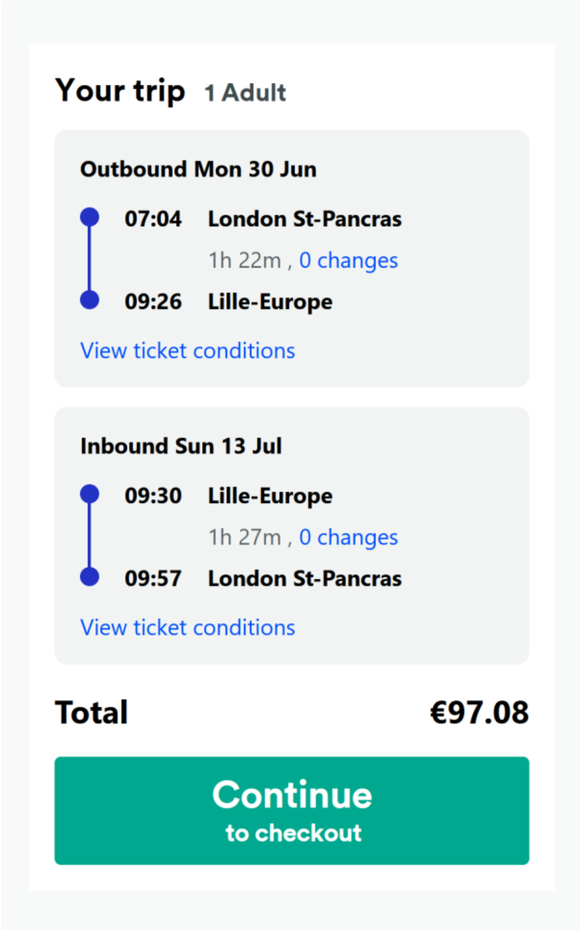
[2] Aushian Wang, Hung-Yi Lee, and Yun-Nung Chen. (2019). Tree Transformer: Integrating Tree Structures into Self-Attention. In Proceedings of the 2019 Conference EMNLP-IJCNLP, pages 1061–1070 doi=10.18653/v1/D19-1098

[3] Barket, R., England, M., Gerhard, J. (2024). Symbolic Integration Algorithm Selection with Machine Learning: LSTMs Vs Tree LSTMs. In: Mathematical Software – ICMS 2024. ICMS 2024. Lecture Notes in Computer Science, vol 14749. Springer, Cham. https://doi.org/10.1007/978-3-031-64529-7_18

[4] Guillaume Lample and Francois Charton. (2020). Deep learning for symbolic mathematics. In Proc. International Conference on Learning Representations (ICLR). doi:10.48550/arxiv.1912.01412\

[5] Makke, N, Chawla, S. (2024). Interpretable scientific discovery with symbolic regression: a review. *Artif Intell Rev* **57**, 2. https://doi.org/10.1007/s10462-023-10622-0

Train ticket return cost screenshot (Hotel cost on next page)



Hotel cost screenshot

Choose your room

Start date
30 Jun

End date
13 Jul

Travellers
1 room, 1 traveller

Prices are lower than typical

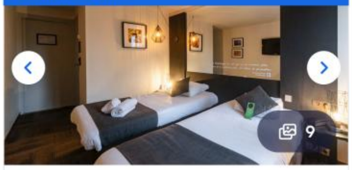
The total price is lower than our estimated range for similar properties on our site (in this destination and others). More details

€1,266
€1,963
€2,938

All rooms | 1 bed | 2 beds | 3+ beds

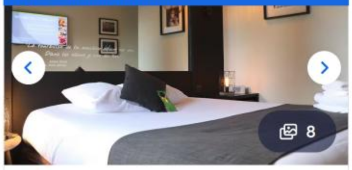
Showing 4 of 4 rooms

Our lowest price



Standard Twin Room
8.0 Very good
3 reviews
17 sq m
Sleeps 3
2 Single Beds
Free WiFi
Collect and Redeem
More details >

Popular among travellers



Standard Room, 1 Double Bed
9.0 Wonderful
2 reviews
17 sq m
Sleeps 3
1 Double Bed
Free WiFi
Collect and Redeem
More details >

Upgrade your stay

Chambre Premium Twin
17 sq m
Sleeps 3
2 Single Beds
Free WiFi
Collect and Redeem

Minibar | Netflix | Iron/ironing board

Standard Room, 1 Double Bed
9.0 Wonderful
2 reviews
17 sq m
Sleeps 3
1 Double Bed
Free WiFi
Collect and Redeem
More details >

Cancellation policy — per stay

More details on all policy options

Non-refundable	+ €0
Fully refundable before 27 Jun Reserve now, pay later	+ €145

Extras — per night

No extras	+ €0
Full Breakfast	+ €38

10% off
~~€1,494~~ €1,349
for 13 nights, 1 room
€104 per night
includes taxes & fees

Reserve