Hunter TEXNet | CSCI 350 Project Proposal

Alex Taradachuck Ralph “Blake” Venté

March 2020

1 Project Title

Our title is *Hunter TEXNet* following the tradition of LeCun’s *LeNet* Neural Network architecture (LeCun et al. 1998, p. 7) but our investigation will not be limited to these models.

2 Description

2.1 Previous Work

The prospect of accurately transcribing mathematical expression into a markup representation is enticing because it opens the doors for bringing new life to old mathematical texts or those for which the source code is unavailable.

A Harvard project called *What you get is what you see* (WYGIWYS) by Deng, Kanervisto, and Rush 2016 details strategies for machine translation of mathematical notation using an attention-based encoder-decoder neural model. Notably, the researchers were interested in the influence of markup alone on the efficacy of a model — without the user providing explicit information about underlying grammars (Deng, Kanervisto, and Rush 2016, p. 1).

2.2 Objectives

Our goal is to build on previous work with a system that exploits the recursive nature of mathematical language and that has high invariance to the typeface used. We hope to achieve this by combining an Optical Character Recognition (OCR) model, that operates on atomic tokens, while an outer model parses the input and segments the image into these tokens. Our motivation behind this line of reasoning is the view of Mathematics as a Chomsky Normal Form language Miller and Viola 1998. We explain this more in our Synthetic Example.

Viewing this as two separate problems means we can isolate learning the grammar from character recognition complexity. This idea isn’t new: one essential component of WYGIWYS provides spatial understanding which is factored into later layers of their Neural Network (ANN) (Deng, Kanervisto, and Rush 2016, p. 4), but the final model does not explicitly separate grammar from characters. We intend on de-coupling the concerns to open up doors for grammar model reuse. Then, different typefaces may be learned separately by the OCR model.

As mentioned in Deng, Kanervisto, and Rush 2016, the model is sensitive to what we choose as an “atomic token.” We will build on their findings. Also, they mention that the model is vulnerable to ambiguities in LATEX itself — that different inputs may result in the same output. We will also use a normalized form of LATEX to mitigate this issue. Another similarity is that we will generate `png` images and do pre-processing using `imagemagick`’s `convert` utility.

2.2.1 Innovations

Our research differs because we will employ models to learn the underlying grammar of mathematical notation and character recognition explicitly and independently.

We want to use this specialization to segment the image and dispatch an OCR model on specific subsets of the image. This may allow for us to translate expressions of arbitrary length, complexity, and typeface with no reduction in performance.

By partitioning this task, we also hope to reduce dependency on Neural Networks. Tentatively, we're investigating using an ANN with Long Short Term Memory architecture (LSTM) for grammar because of prior research: LSTM models are able to derive meaning in context, which is important for grammar recognition (Deng, Kanervisto, and Rush 2016, p. 1). Then, we explore Support Vector Machines with kernel functions for OCR for a more compact model with comparable performance because of their history with OCR tasks.

2.2.2 Synthetic Example

When considering the following formula, we wouldn't want our model to be confused — is the 0 stacked on top of the 1 or are they on separate lines?

$$f = \begin{cases} 0, & a \leq b \\ 1, & c \leq d \end{cases}$$

To us it's intuitive to break up the task of understanding this expression into two distinct steps: first, recognize that we are in some form of a nested environment, handle it (the outer { in this case), then recursively parse the next internal portion (in this case, the first line of the nested structure).

Instead of taking the expression in as a whole, we hope to build a system that pays attention to specifically the portions of the text that are "outer" before recursively translating the inner function. As this is both of our first experiences with OCR or machine translation, we hope to learn much more about ANN's and other models capable of parsing recursive structures. Hopefully then, we be able to translate this intuitive reasoning into a model that uses this assumption of recursive structure for performance benefits.

3 Tasks and Roles

Of course distribution of work at this stage is tentative and subject to any new findings. Blake will complete the first 3, and Alex will complete the second 3.

1. Mine a greater number of examples from real world text, expanding the corpus by doubling the number of images compared to the Deng database.
   (a) Find the examples that the Harvard model cannot accommodate.
   (b) Find real-world data to augment the Harvard database.
2. Create model to learn the spatial grammar of arbitrary markup.
   (a) Recursively search the image, assigning semantic markup for every child node in the search.
   (b) Recognize "base cases", that is, segment the image into atomic tokens for dispatch to the OCR recognizer.
3. Create a high-performance OCR engine that recognizes characters.
   (a) Begin with the standard Computer Modern typeface in LaTeX (and all of its mathematical fonts).
   (b) Investigate the difference in performance between a single OCR engine responsible for the token translation of all typefaces we test compared to individual OCR models for each typeface.

4 Topics

4.1 Machine Learning

1. Optical Character Recognition (OCR)

2. Image Segmentation / Document Recognition
3. Artificial Neural Network (ANN)
   (a) Natural Language Processing (NLP)
   (b) Long Short Term Memory in ANN's
   (c) Recurrence and Convolution in ANN's
4. Support Vector Machines (SVM)

5 Deliverables

5.1 Required Objectives

At the submission deadline, we will have the following prepared:

1. all data generated and normalized, building on the work of the Harvard team;
2. all source code containing our finished models and documentation;
3. research paper outlining the intricacies of our models and their performance on generated examples;
4. a live demo of the model; and
5. a 2 minute video.

5.2 Stretch Goals

Minimal interactive web front-end where a user will be able to upload an image of a mathematical expression and receive the LaTeX code associated with it. This would also serve as the platform for one of our demos.

6 Evaluation

We will evaluate our models with a confusion matrix, Hamming distance, and statistic where appropriate. Similarly to the Harvard team, we use the perplexity metric, common to machine translation tasks (Jelinek et al. 1977, p. 1). We will pay special attention to evaluating how well our model handles nested structures compared to the models in Wygiwys.

References

[Che+17] Yu Cheng et al. "A Survey of Model Compression and Acceleration for Deep Neural Networks. CoRR abs/1710.09282 (2017)". In: *arXiv preprint arXiv:1710.09282* (2017).

[DKR16] Yuntian Deng, Anssi Kanervisto, and Alexander M Rush. "What you get is what you see: A visual markup decompiler". In: *arXiv preprint arXiv:1609.04938* 10 (2016), pp. 32–37.

[Jel+77] Fred Jelinek et al. "Perplexity—a measure of the difficulty of speech recognition tasks". In: *The Journal of the Acoustical Society of America* 62.S1 (1977), S63–S63.

[LeC+98] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[MV98] Erik G Miller and Paul A Viola. "Ambiguity and constraint in mathematical expression recognition". In: *AAAI/IAAI*. 1998, pp. 784–791.

[Reu+19] Christian Reul et al. "OCR4all—An Open-Source Tool Providing a (Semi-) Automatic OCR Workflow for Historical Printings". In: *Applied Sciences* 9.22 (2019), p. 4853.

[SS05] Alan Sexton and Volker Sorge. "Database-driven mathematical character recognition". In: *International Workshop on Graphics Recognition*. Springer. 2005, pp. 218–230.

[Suz+03] Masakazu Suzuki et al. “INFTY: an integrated OCR system for mathematical documents”. In: *Proceedings of the 2003 ACM symposium on Document engineering*. 2003, pp. 95–104.