

Mathematical Formula Recognition

Ayush Kumar Shah (as1211@rit.edu)

Advisor : **Richard Zanibbi**

Research Potential Assessment

13th May 2021

R·I·T

B. THOMAS GOLISANO
College of **COMPUTING AND**
INFORMATION SCIENCES



Document and Pattern Recognition Lab

Mathematical Formula Recognition



Math formulas are a key element in technical documents.

have the following probability over rule r given the previous count table $\mathbb{V}$:

$$P(r_i = v | N) = \frac{N_{R_i}(r_i) + \alpha P_{R_i}^0(r_i)}{v_i + \alpha} \quad (2)$$

here in the case of DP, $\mathbb{V} = \{\mathbb{V}_i\}$, where $\mathbb{V}_i$ is the index for the category of rule counts.

We use the *top-down* sampling algorithm of Chung et al. (2014) which samples cut and edge variables from top down and one at a time. For each node v , we denote the composed rule type that we get when we set the cut of node v to $\mathbf{0}$ as r_1 and the two split rule types that we get when we set the cut to 1 as r_2, r_3 . We sample the cut value z_v of the current node according to the posterior probability:

$$P(z_v = c | N) = \begin{cases} \frac{P_{R_i}(r_1) + \alpha P_{R_i}^0(r_1)}{P_{R_i}(r_1) + \alpha P_{R_i}^0(r_1) + P_{R_i}(r_2) + \alpha P_{R_i}^0(r_2) + P_{R_i}(r_3) + \alpha P_{R_i}^0(r_3)} & \text{if } c = \mathbf{0} \\ \frac{P_{R_i}(r_2) + \alpha P_{R_i}^0(r_2)}{P_{R_i}(r_1) + \alpha P_{R_i}^0(r_1) + P_{R_i}(r_2) + \alpha P_{R_i}^0(r_2) + P_{R_i}(r_3) + \alpha P_{R_i}^0(r_3)} & \text{otherwise} \end{cases} \quad (3)$$

where the posterior probability $P(z_v | N)$ is according to a DP, and $\mathbb{V}, N$ are tables of counts. In the case of DP, $\mathbb{V}, N$ differ only in the rule counts of r_1 where $\mathbb{V}_{R_i}(r_1) = N_{R_i}(r_1) + 1$.

As for edge variables $\mathbb{E}_i$, we refer to the set of composed rules turned on below v_i including the composed rule fragments having v_i as an internal or root node as $\{r_{v_1}, \dots, r_{v_n}\}$. We have the following posterior probability over the edge variable $\mathbb{E}_i$:

$$P(\mathbb{E}_i = c | N) \propto \prod_{j=1}^m P(r_{v_j} | N^{j-1}) \prod_{v_2, v_3 \in \text{children}(v_i)} \text{deg}(c) \quad (4)$$

where $\text{deg}(v)$ is the number of incoming edges for node v , $\text{in}(v)$ is the set of nodes in all subtrees under v , and $\{r_{v_j}\}$ is the tree specified when we set $r_{v_j} = \mathbf{0}$. $\mathbb{V}^{j-1}$ is $\mathbb{V}^{j-2}$ in $\mathbb{V}^{j-1}$ are tables of counts where $N^{j-1} = N$, $N_{R_i}^j(r_1) = N_{R_i}^{j-1}(r_1) + 1$ in the case of DP.

After we have sampled one SHRIG derivation from the forest, we still need to keep track of the place where each nonterminal edge attaches. As we have maintained the graph fragment it represents in each node of the forest, we can retrieve the attachment nodes of each hyperedge in the r.h.s. by tracing at which graph nodes two fragments fuse with each other. We perform this rule extraction procedure from top-down and maintain the order of attachment nodes of each r.h.s. nonterminal edge. When we further rewrite a nonterminal edge, we need to make sure that it keeps the order of the attachment nodes in its parent rule.

As for the unaligned words, we just insert all the omitted unaligned words in the composition procedure. We also add additional rules including the surrounding 2 unaligned words context to make sure there are terminals on the string side.

3.3 Phrase-to-Graph-Fragment Alignment Extraction

Aside from the rules sampled using the MCMC algorithm, we also extract a phrase-to-graph-fragment alignment table from the fragment decomposition forest. This step can be considered as a mapping of larger phrases made of multiple identified spans (plus unaligned words) to a larger fragments made of multiple concept fragments (plus the way they connect using unaligned edges).

Our extraction happens along with the forest construction procedure. In line 1 of Algorithm 1 we extract one rule for each smallest phrase-fragment pairs before and after the unaligned edges are attached. We also extract one rule for each newly constructed node after line 11 if the fragment side of the node is single-rooted.² We do not extract rules after line 3 because it usually introduces additional noise of meaningful concepts which are unrecognized in the concept identification stage.

4 Decoding

4.1 Concept identification

During the decoding stage, first we need to identify meaningful spans in the sentence and map them to graph fragments on the graph side. Then we use SHRIG rules to parse each sentence from bottom up and left to right, which is similar to constituent parsing. The recall of the concept identification stage from Flanagan et al. (2014) is 0.79, which means 21.9% of the meaningful concepts are already lost at the beginning of the next stage.

Our strategy is to use lemma and POS tags information after the concept identification stage, we use it to recall some meaningful concepts. We find that, except for some special function words, most nouns, verbs and adjectives should be aligned. We use the lemma information to retrieve unaligned words whose morphological form does not appear in our training data. We also use

²Here we will also look at the surrounding 2 unaligned words to fix partial alignment and not be introduced by meaningful unaligned words

2.3 Joint Inference Framework

We extend expression (1) to facilitate joint inference on mention heads and coreference as follows:

$$\begin{aligned} \arg \max_{\mathbf{y}} \sum_{u \in \mathcal{U}, v \in \mathcal{M}} f_{u,v} y_{u,v} + \sum_{m \in \mathcal{M}} g_m y_m, \\ \text{s.t. } \sum_{u \in \mathcal{U}} y_{u,v} \leq 1 \quad \forall v \in \mathcal{M}, \\ \sum_{u \in \mathcal{U}} y_{u,v} \leq y_v \quad \forall v \in \mathcal{M}, \\ y_{u,v} \in \{0, 1\}, \quad y_m \in \{0, 1\} \quad \forall u, v, m \in \mathcal{M}. \end{aligned}$$

Here, $\mathcal{M}$ is the set of all mention head candidates. y_m is the decision variable for mention head candidate m . $y_m = 1$ if and only if the mention head m is chosen. To consider coreference decisions and mention head decisions together, we add the constraint $\sum_{u \in \mathcal{U}} y_{u,v} \leq y_v$, which ensures that if a candidate mention head v is not chosen, then it will not have coreference links with other mention heads.

2.4 Joint Learning Framework

To support joint learning of the parameters $\mathbf{w}_1$ and $\mathbf{w}_2$ described above, we define a joint training objective function $\mathcal{C}(\mathbf{w}_1, \mathbf{w}_2)$ for mention head detection and coreference, which uses a max-margin approach to learn both weight vectors. Suppose we have a collection of documents $\mathcal{D}$, and we generate $\mathcal{M}_d$ mention head candidates for each document d ($d \in \mathcal{D}$). We use an indicator function $\delta(u, m)$ to represent whether mention heads u, m are in the same coreference cluster based on gold annotations ($\delta(u, m) = 1$ iff they are in the same cluster). Similarly, $\Omega(m)$ is an indicator function representing whether mention head m is valid in the gold annotations.

For simplicity, we first define

$$\begin{aligned} u' &= \arg \max_{u \in \mathcal{M}} \mathbf{w}_1^T \phi(u, m) - \delta(u, m), \\ u'' &= \arg \max_{u \in \mathcal{M}, \delta(u, m)=1} \mathbf{w}_2^T \phi(u, m) \Omega(m). \end{aligned}$$

We then minimize the following joint training objective function $\mathcal{C}(\mathbf{w}_1, \mathbf{w}_2)$.

$$\begin{aligned} \mathcal{C}(\mathbf{w}_1, \mathbf{w}_2) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \frac{1}{n_d} \sum_m (C_{coref,m}(\mathbf{w}_2) \\ + C_{local,m}(\mathbf{w}_1) + C_{trans,m}(\mathbf{w}_1)) + R(\mathbf{w}_1, \mathbf{w}_2). \end{aligned}$$

$\mathcal{C}(\mathbf{w}_1, \mathbf{w}_2)$ is composed of four parts. The first part is the loss function for coreference, where we have

$$\begin{aligned} C_{coref,m}(\mathbf{w}_2) = -\mathbf{w}_2^T \phi(u', m) \Omega(m) \\ + (\mathbf{w}_2^T \phi(u', m) - \delta(u', m)) (\Omega(m) \vee \Omega(u')). \end{aligned}$$

It is similar to the loss function for a latent left-linking coreference model³. As the second component, we have the quadratic loss for the mention head detection model,

$$C_{local,m}(\mathbf{w}_1) = \frac{1}{2} (\mathbf{w}_1^T \phi(m) - \Omega(m))^2.$$

Using the third component, we further maximize the margin between valid and invalid mention head candidates when they are selected as the best-left-link mention heads for any valid mention head. It can be represented as

$$C_{trans,m}(\mathbf{w}_1) = \frac{1}{2} (\mathbf{w}_1^T \phi(u') - \Omega(u'))^2 \Omega(m).$$

The last part is the regularization term

$$R(\mathbf{w}_1, \mathbf{w}_2) = \frac{\lambda_1}{2} \|\mathbf{w}_1\|^2 + \frac{\lambda_2}{2} \|\mathbf{w}_2\|^2.$$

2.5 Stochastic Subgradient Descent for Joint Learning

For joint learning, we choose stochastic subgradient descent (SGD) approach to facilitate performing SGD on a per mention head basis. Next, we describe the weight update algorithm by defining the subgradients.

The partial subgradient w.r.t. mention head m for the head weight vector $\mathbf{w}_1$ is given by

$$\begin{aligned} \nabla_{\mathbf{w}_1, m} \mathcal{C}(\mathbf{w}_1, \mathbf{w}_2) = \\ \frac{1}{|\mathcal{D}| n_d} (\nabla C_{local,m}(\mathbf{w}_1) + \nabla C_{trans,m}(\mathbf{w}_1)) + \lambda_1 \mathbf{w}_1, \end{aligned} \quad (2)$$

where

$$\begin{aligned} \nabla C_{local,m}(\mathbf{w}_1) &= (\mathbf{w}_1^T \phi(m) - \Omega(m)) \phi(m), \\ \nabla C_{trans,m}(\mathbf{w}_1) &= (\mathbf{w}_1^T \phi(u') - \Omega(u')) \phi(u') \Omega(m). \end{aligned}$$

The partial subgradient w.r.t. mention head m for the coreference weight vector $\mathbf{w}_2$ is given by

$$\begin{aligned} \nabla_{\mathbf{w}_2, m} \mathcal{C}(\mathbf{w}_1, \mathbf{w}_2) = \lambda_2 \mathbf{w}_2 + \\ \begin{cases} \phi(u', m) - \phi(u'', m) & \text{if } \Omega(m) = 1, \\ \phi(u', m) & \text{if } \Omega(m) = 0 \text{ and } \Omega(u') = 1, \\ 0 & \text{if } \Omega(m) = 0 \text{ and } \Omega(u') = 0. \end{cases} \end{aligned} \quad (3)$$

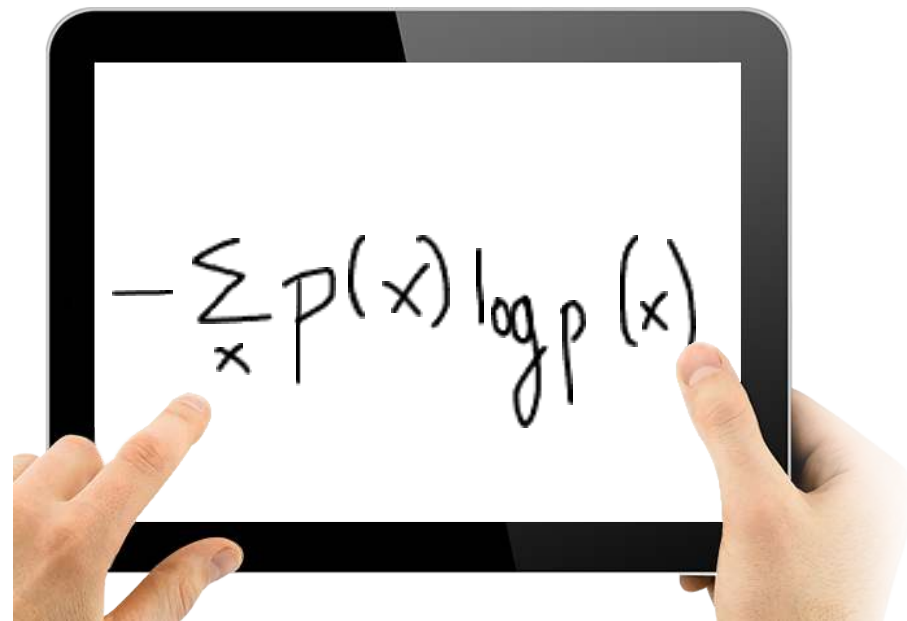
Here λ_1 and λ_2 are regularization coefficients which are tuned on the development set. To learn the mention head detection model, we consider two different parts of the gradient in expression (2). $\nabla C_{local,m}(\mathbf{w}_1)$ is exactly the local gradient of mention head m while we add $\nabla C_{trans,m}(\mathbf{w}_1)$ to represent

³More details can be found in Chang et al. (2013). The difference here is that we also consider the validity of mention heads using $\Omega(u)$.

Motivation: Information Retrieval

$$-\sum_x p(x) \log p(x)$$

What is the name
of this formula?



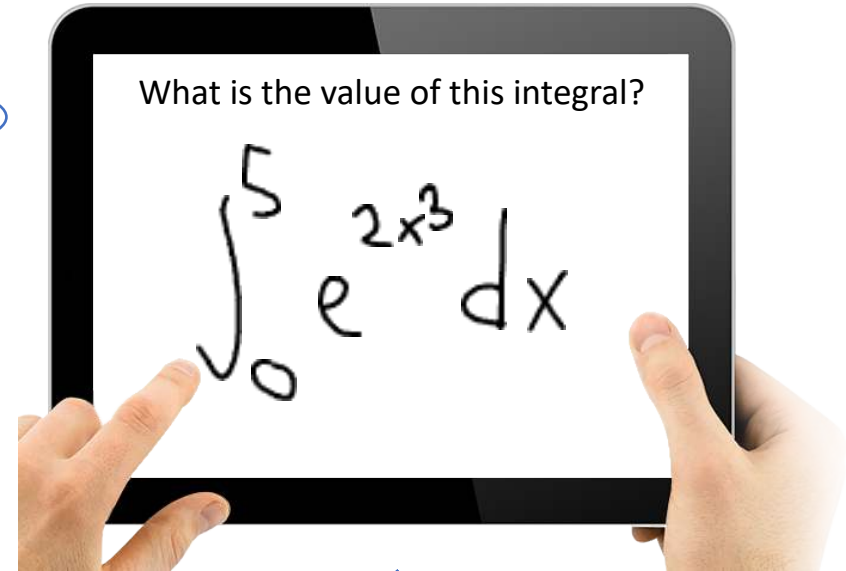
Let's find documents
describing this formula.

Motivation: Inf. Retrieval - Question Answering

What is the value of
this integral?

$$\int_0^5 e^{2x^3} dx$$

How is this
computed?

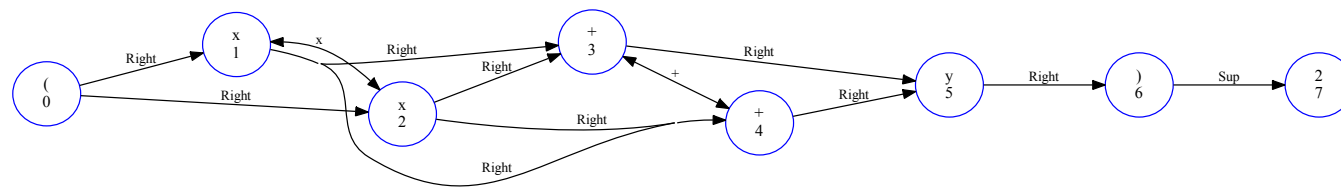


How big is the value of
this integral?

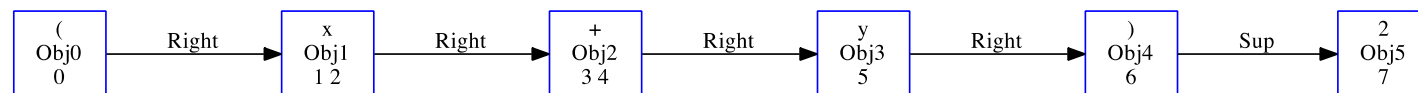
Inputs and Graphs Describing Appearance

$(x + y)^2$

Offline / Online handwritten strokes



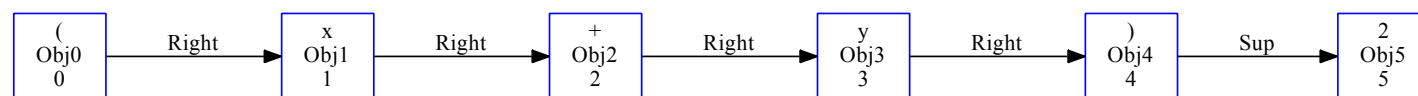
Label Graph (LG) Over Handwritten Strokes



Symbol Layout Tree (SLT) over Symbols

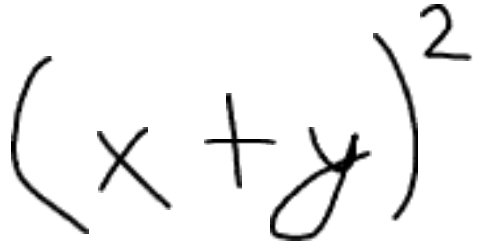
$(x + y)^2$

Typeset formula image

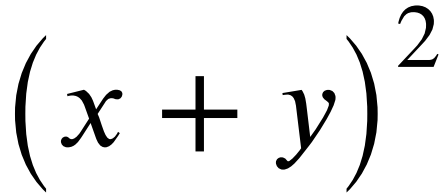


Symbol Layout Tree over Connected Components (= Symbols)

Inputs and SLT (Appearance) Encodings

A handwritten mathematical expression $(x+y)^2$ in black ink, showing natural cursive strokes.

Offline / Online handwritten strokes

A high-quality typeset image of the mathematical expression $(x+y)^2$ in a standard serif font.

Typeset formula image



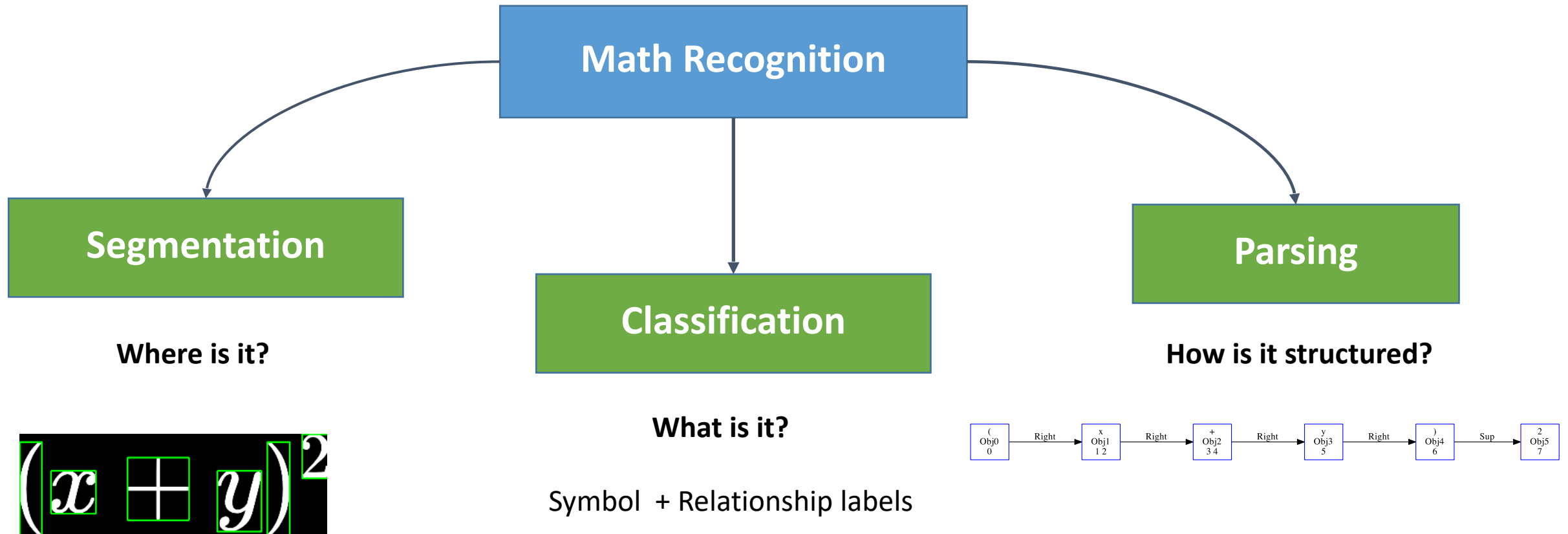
```
<mrow>
  <mo xml:id="0:">{</mo>
  <mrow>
    <mi xml:id="1:2:">x</mi>
    <mrow>
      <mo xml:id="3:4:">+</mo>
      <mrow>
        <mi xml:id="5:">y</mi>
        <msup>
          <mo xml:id="6:">)</mo>
          <mn xml:id="7:">2</mn>
        </msup>
      </mrow>
    </mrow>
  </mrow>
</mrow>
```

MathML Output

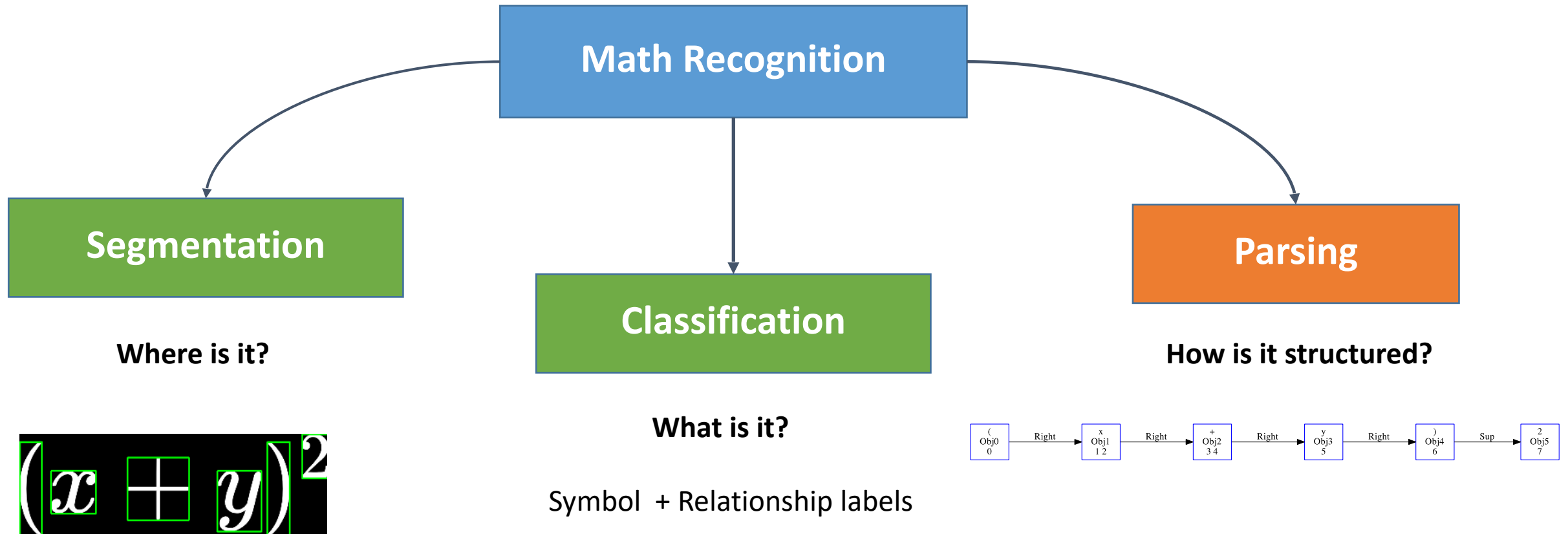
$$(x + y)^{2}$$

LaTeX Output

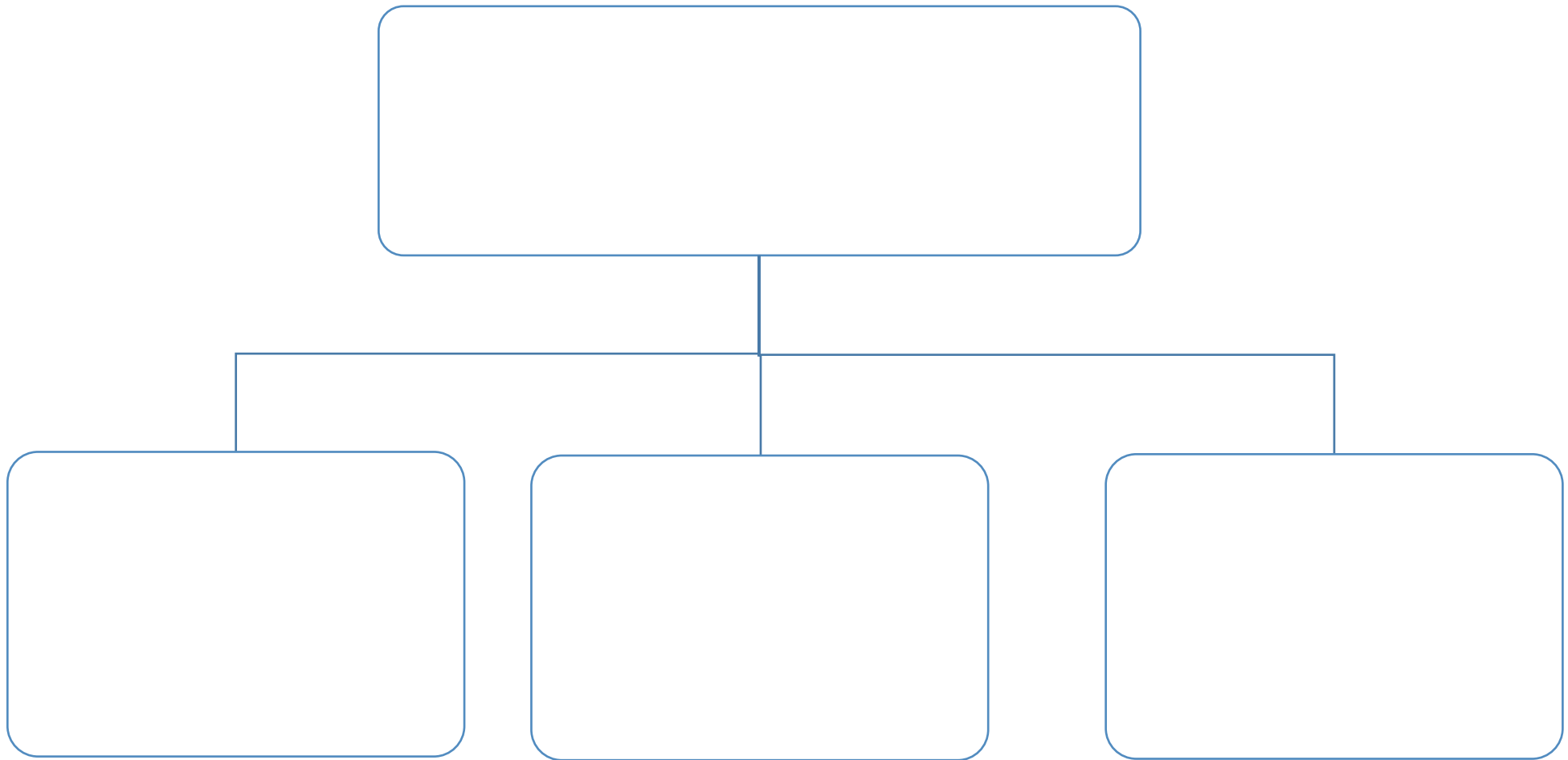
The Problem

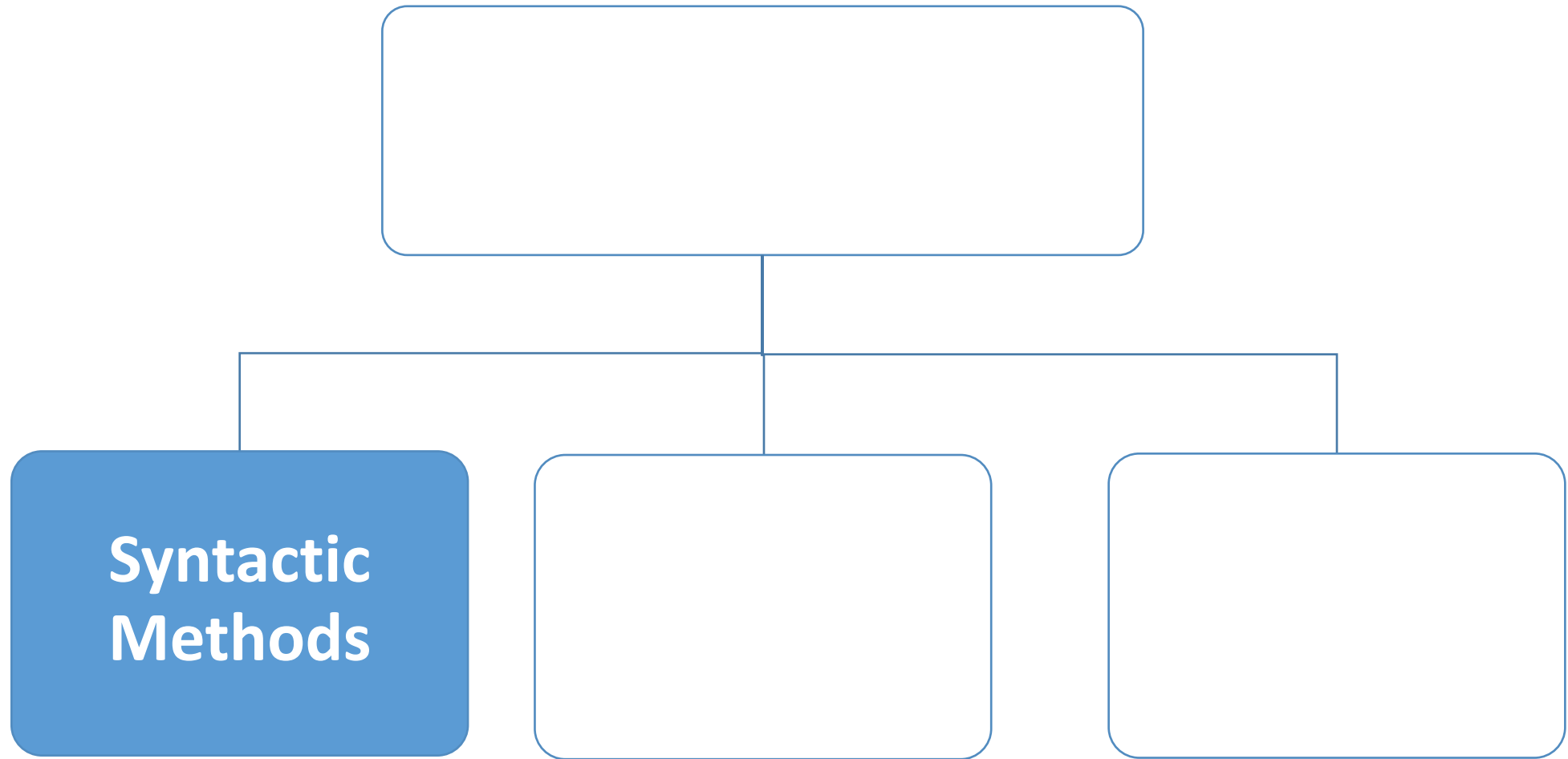


The Problem



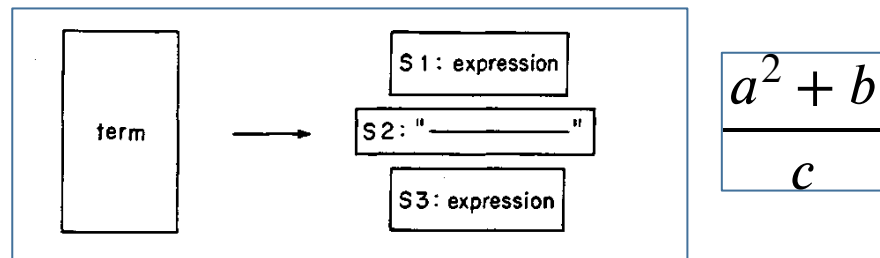
LITERATURE REVIEW: Math Notation Parsing



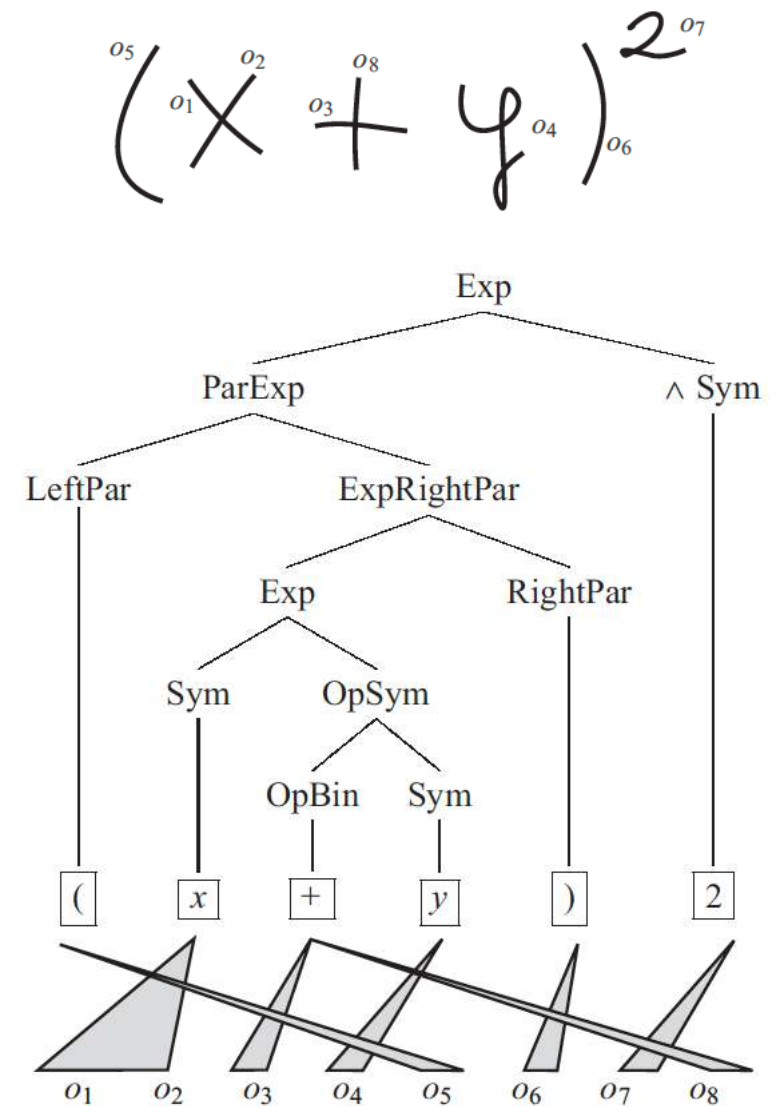


Syntactic Methods

- Anderson's system (1967)
- Grammar rules
 - Production rules, CFG, SCFG (1D, 2D)
- Top-down and bottom-up parsing algorithms
 - e.g., CYK algorithm (Dynamic programming) for SCFG



[1] Example replacement rule (2D)



[2] Parse tree of expression $(x + y)^2$

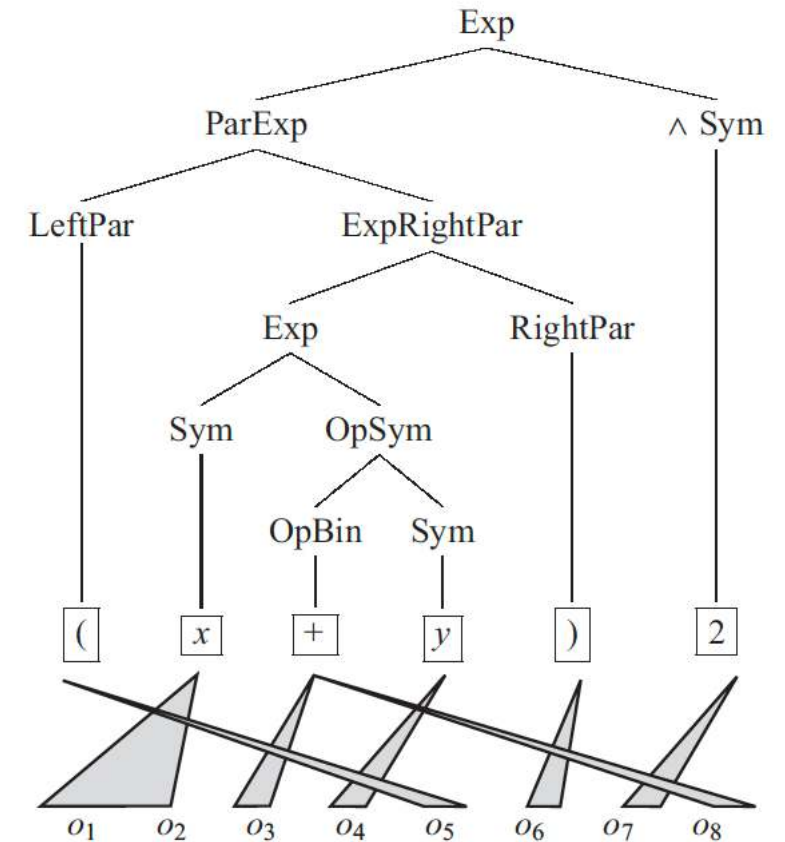
[1] R. H. Anderson, "Syntax-directed recognition of hand-printed two-dimensional mathematics," in Symposium on Interactive Systems for Experimental Applied Mathematics: Proceedings of the Association for Computing Machinery Inc. Symposium, New York, NY, USA, Aug. 1967, pp. 436–459. doi: 10.1145/2402536.2402585.

[2] F. Álvaro, J.-A. Sánchez, and J.-M. Benedí, "An integrated grammar-based approach for mathematical expression recognition," Pattern Recognition, vol. 51, pp. 135–147, Mar. 2016, doi: 10.1016/j.patcog.2015.09.013.

Syntactic Methods

Limitations

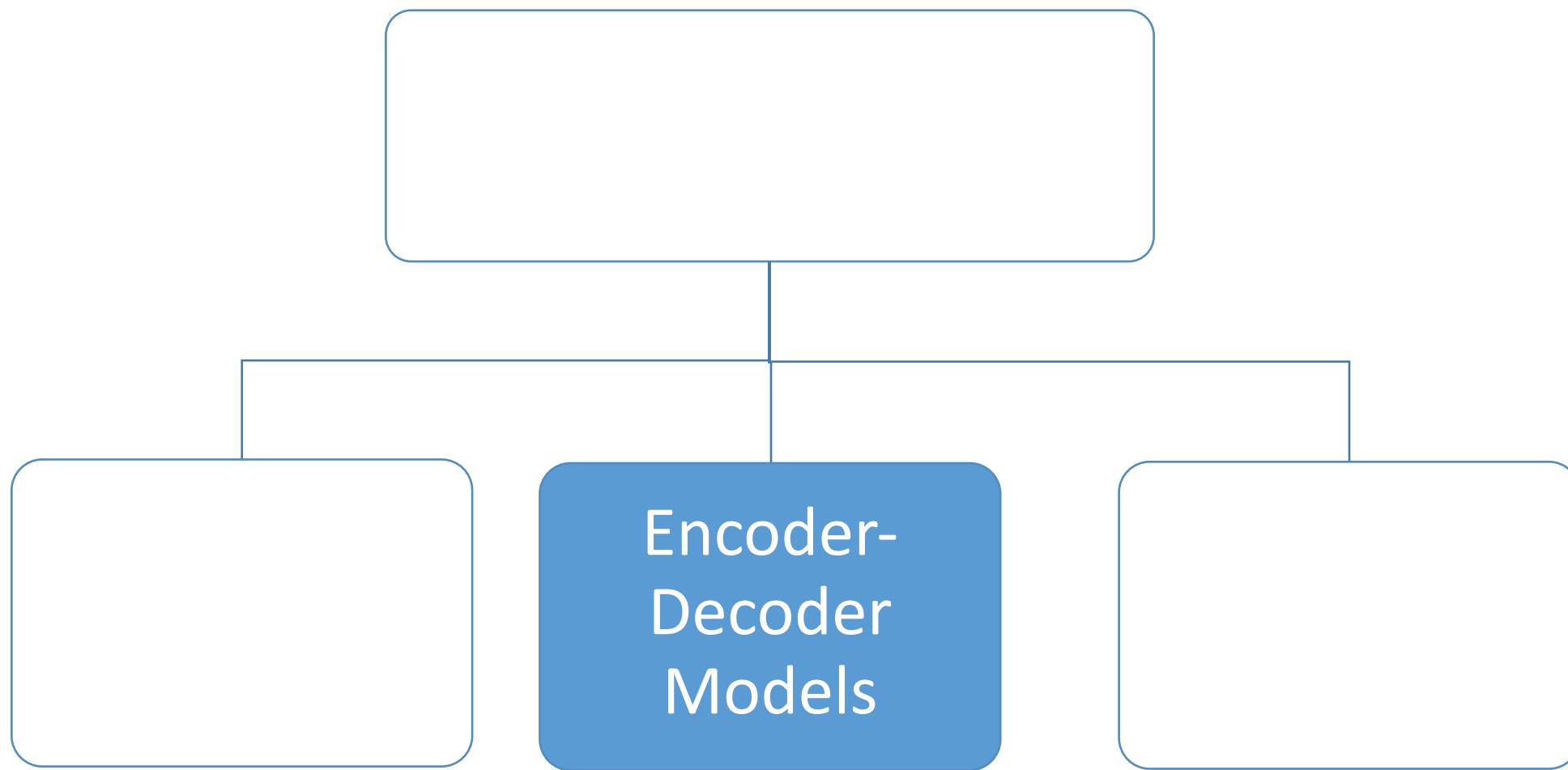
- Computationally expensive
- Cannot pinpoint errors
- Difficult to define a universal grammar (manual grammar construction)
- Need to handle parse failures, e.g., incomplete formulas may be unsyntactic (“a +”)



[2] Parse tree of expression $(x + y)^2$

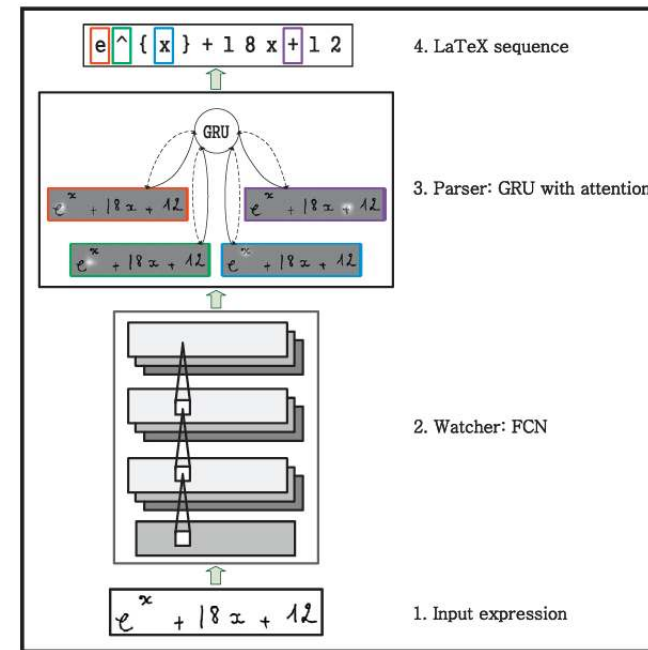
[1] R. H. Anderson, “Syntax-directed recognition of hand-printed two-dimensional mathematics,” in Symposium on Interactive Systems for Experimental Applied Mathematics: Proceedings of the Association for Computing Machinery Inc. Symposium, New York, NY, USA, Aug. 1967, pp. 436–459. doi: 10.1145/2402536.2402585.

[2] F. Álvaro, J.-A. Sánchez, and J.-M. Benedí, “An integrated grammar-based approach for mathematical expression recognition,” Pattern Recognition, vol. 51, pp. 135–147, Mar. 2016, doi: 10.1016/j.patcog.2015.09.013.

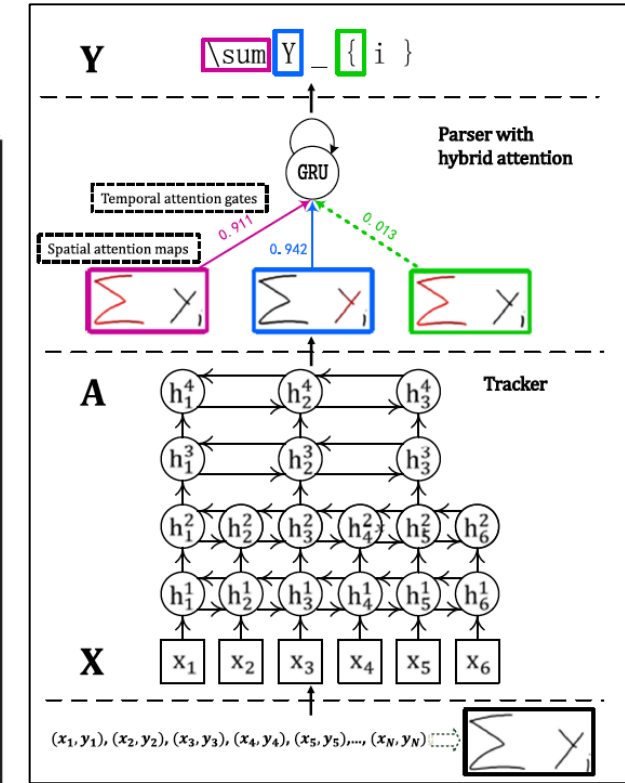


Encoder-Decoder Models

- Motivated by automatic image captioning
- 2D Image/Strokes -> 1D LaTeX
- **Feature extraction:**
 - CNN (offline features)
 - RNN (online features)
- **Encoder:** CNN or RNN
- **Decoder with attention:** RNN



[4] WAP Architecture



[5] TAP Architecture

[3] K. Xu et al., "Show, attend and tell: neural image caption generation with visual attention," in Proceedings of the 32nd International Conference on Machine Learning - Volume 37, Lille, France, Jul. 2015, pp. 2048–2057.

[4] J. Zhang et al., "Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition," Pattern Recognition, vol. 71, pp. 196–206, Nov. 2017, doi: 10.1016/j.patcog.2017.06.017.

[5] J. Zhang, J. Du, and L. Dai, "Track, Attend, and Parse (TAP): An End-to-End Framework for Online Handwritten Mathematical Expression Recognition," IEEE Transactions on Multimedia, vol. 21, no. 1, pp. 221–233, Jan. 2019, doi: 10.1109/TMM.2018.2844689.

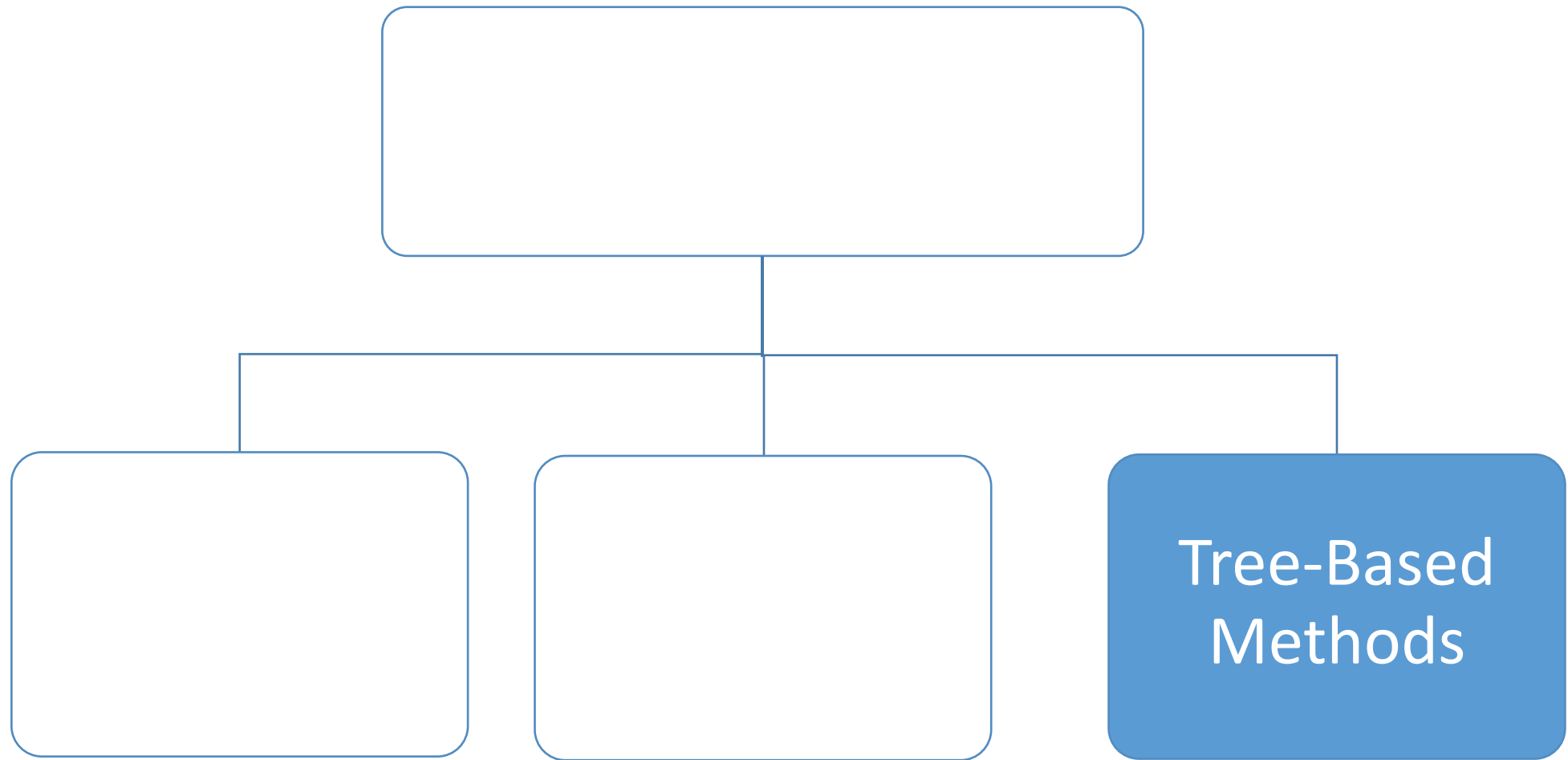
Encoder-decoder models

Advantages

- Data driven i.e., expression grammar not required
- State-of-the-art results for benchmarks (e.g., CROHME)

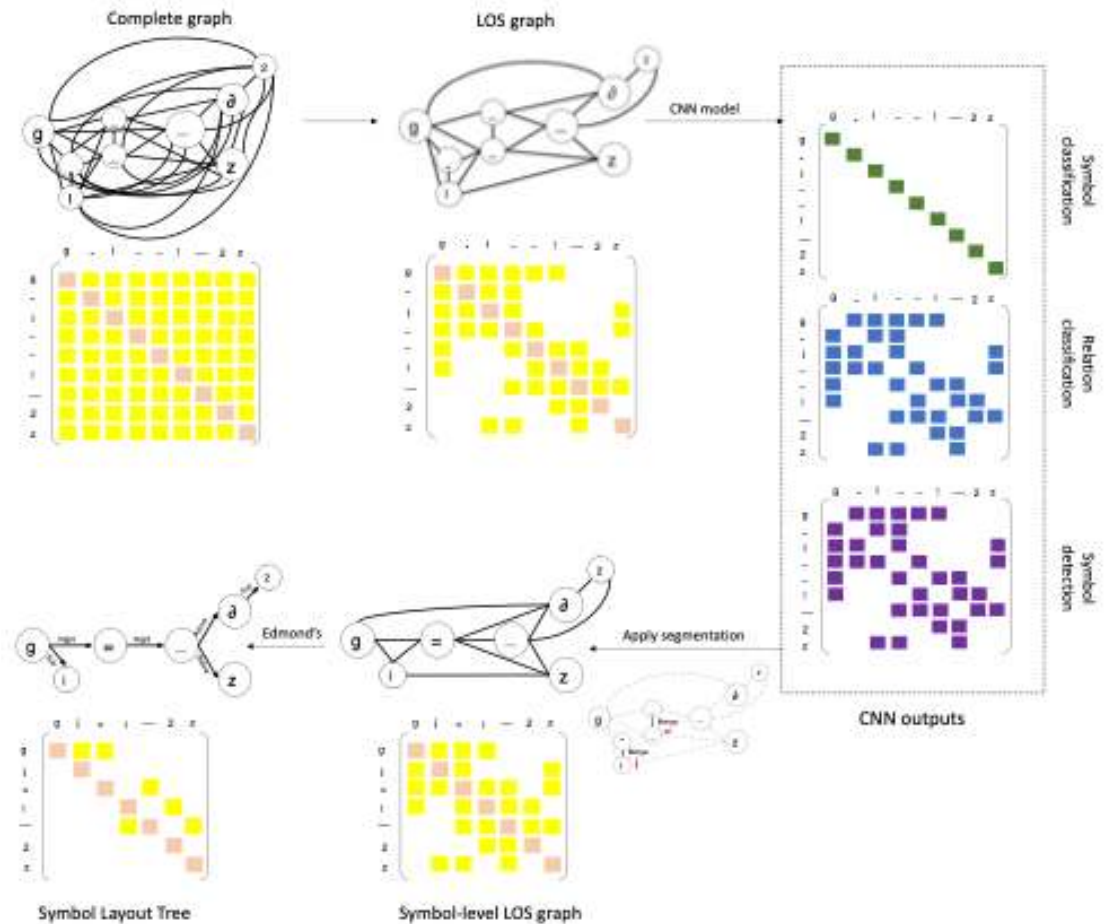
Limitations

- Weak interpretability
- Error diagnosis and system tuning are challenging



Tree-Based Methods

- Graph-search problem (MST extraction) [6]
- QD-GGA [7]
 - Graph over input primitives (strokes or connected components)
 - Score / Probability distributions
 - Line-Of-Sight (LOS) Graphs
 - Alternative edge filtering possible
 - Edmond's arborescence algorithm [8] used to extract MST (SLT)



[7] QD-GGA Parsing

[6] Y. Eto and M. Suzuki, "Mathematical formula recognition using virtual link network," in Proceedings of Sixth International Conference on Document Analysis and Recognition, Sep. 2001, pp. 762–767. doi: 10.1109/ICDAR.2001.953891.

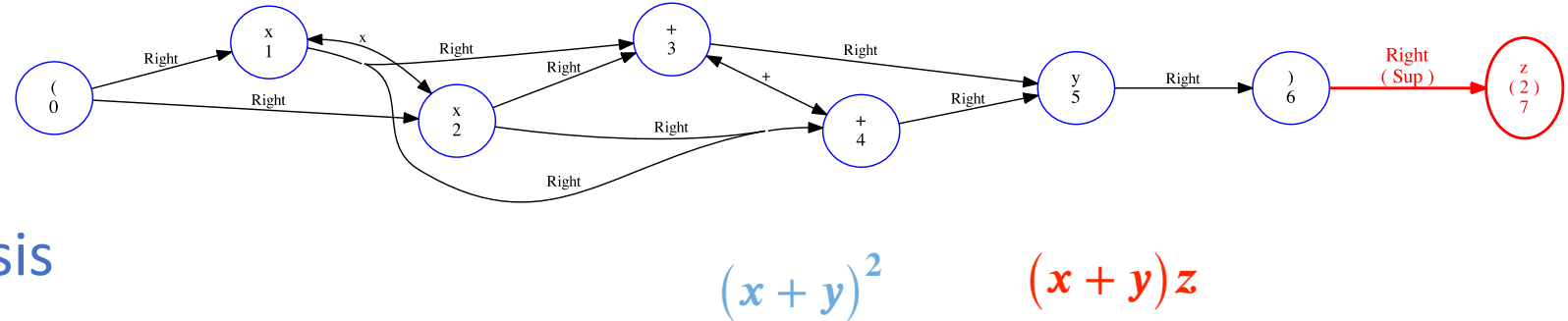
[7] Mahdavi, M.; Sun, L.; Zanibbi, R. Visual Parsing with Query-Driven Global Graph Attention (QD-GGA): Preliminary Results for Handwritten Math Formula Recognition. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); IEEE; pp 2429–2438. <https://doi.org/10.1109/CVPRW50498.2020.00293>.

[8] J. Edmonds, "Optimum branchings," J. RES. NATL. BUR. STAN. SECT. B. MATH. MATH. PHYS., vol. 71B, no. 4, p. 233, Oct. 1967, doi: 10.6028/jres.071B.032.

Tree-Based Methods

Advantages

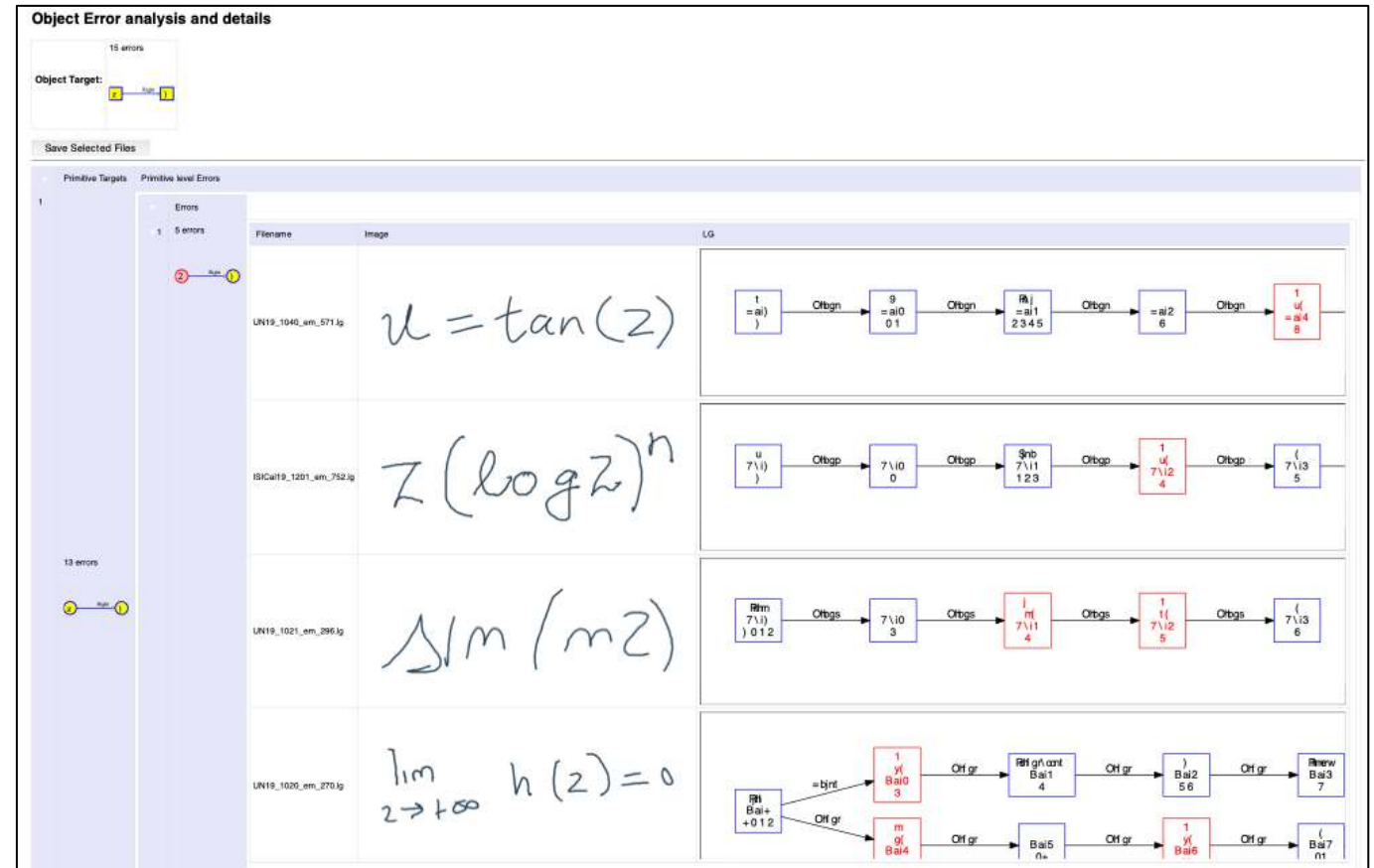
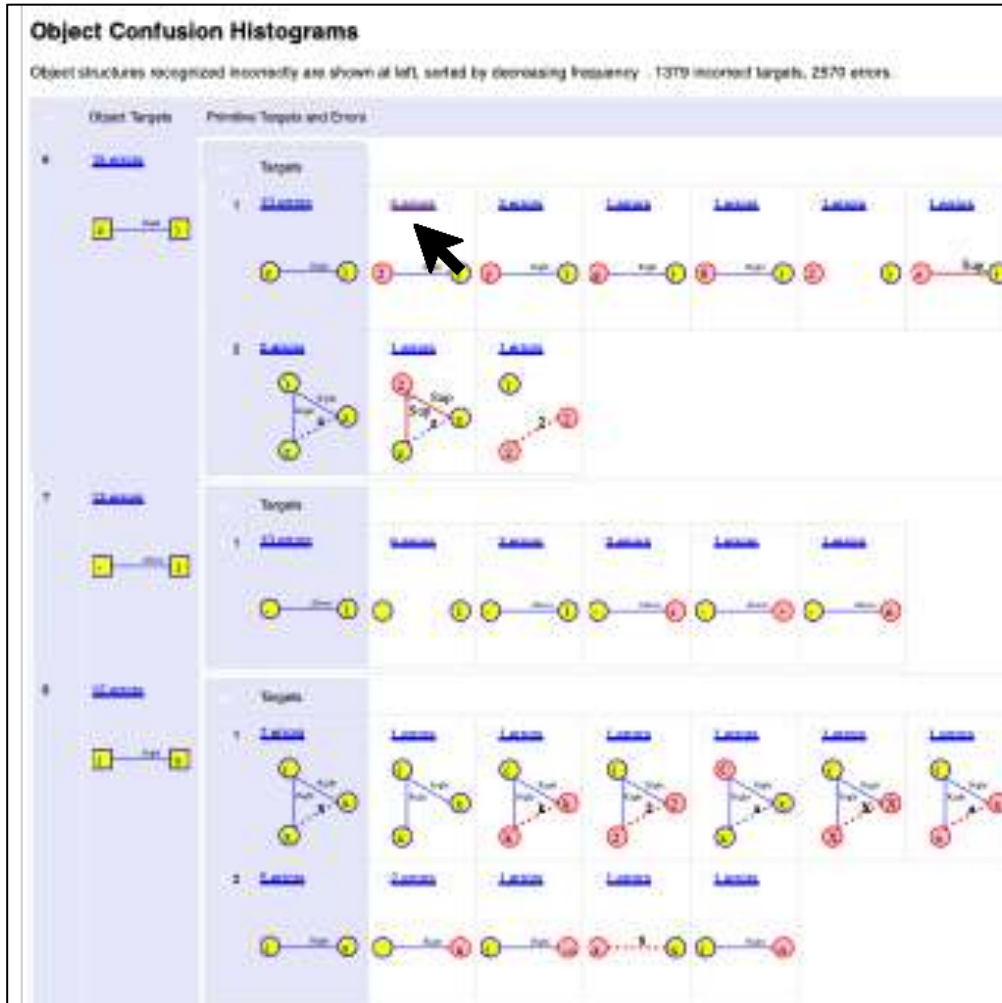
- High interpretability
- Simple/automatic error analysis
- Relatively simple parameter tuning
- Data driven i.e.; expression grammars not required



Limitations

- Lower recognition rates with current models than encoder-decoder
- Currently utilize language models based only on visual features
- (QD-GGA) Loss of spatial information from average pooling of CNN features

PREPARATION FOR EXPERIMENTS



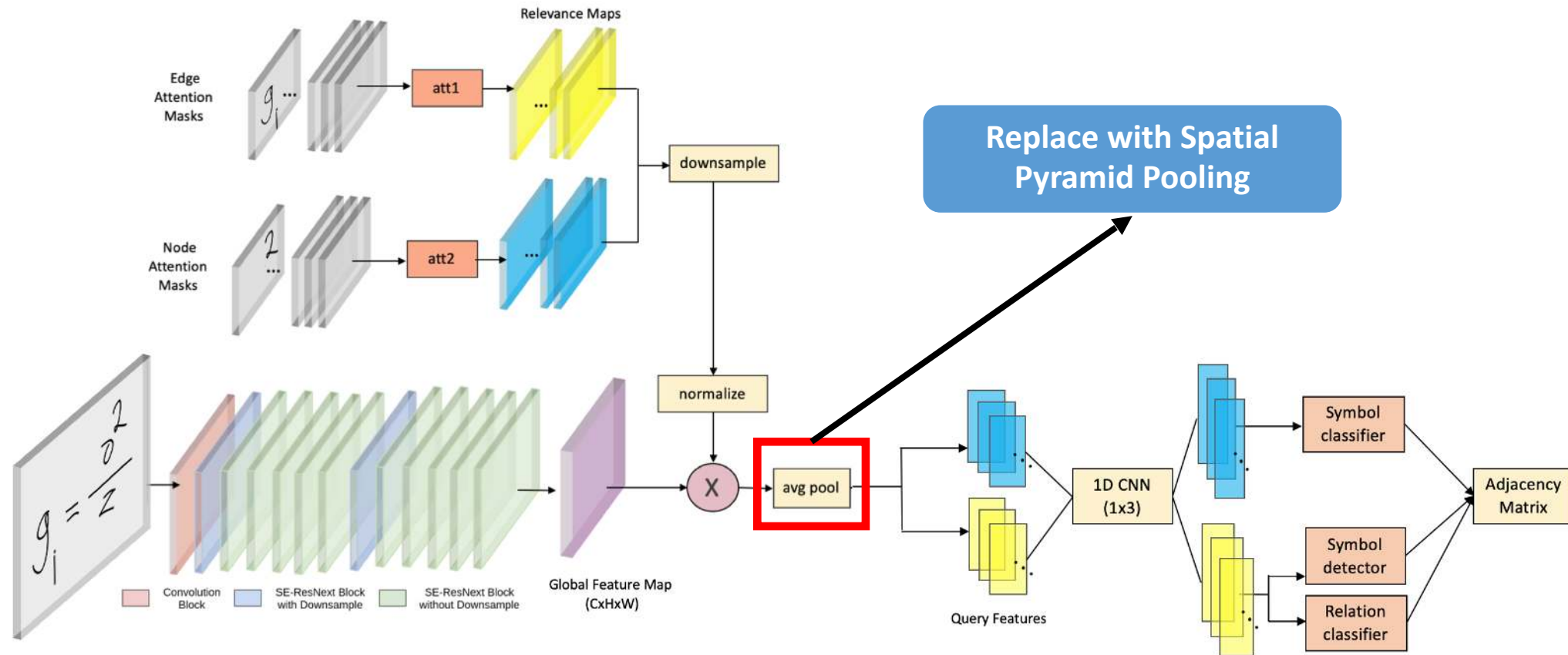
[9] Error analysis

[9] Shah, A. K.; Dey, A.; Zanibbi, R. A Math Formula Extraction and Evaluation Framework for PDF Documents. In 2021 International Conference on Document Analysis and Recognition (ICDAR), to appear; 2021.

[12] H. Mouchère, R. Zanibbi, U. Garain, and C. Viard-Gaudin, "Advancing the state of the art for handwritten math recognition: the CROHME competitions, 2011–2014," *IJDAR*, vol. 19, no. 2, pp. 173–189, Jun. 2016, doi: [10.1007/s10032-016-0263-5](https://doi.org/10.1007/s10032-016-0263-5).

POOLING EXPERIMENTS on QD-GGA

Spatial Pyramid Pooling



[7] QD-GGA Architecture

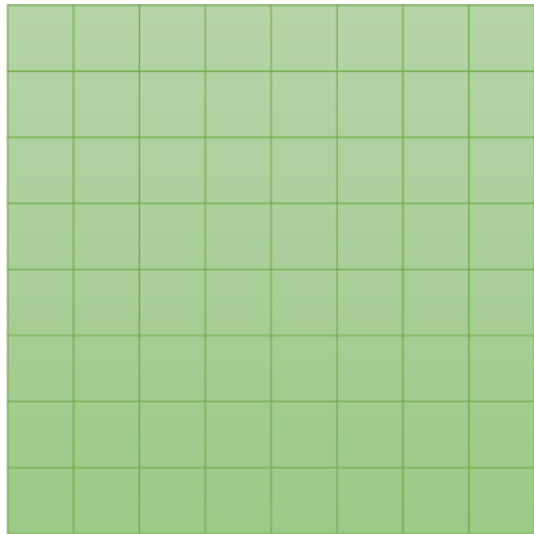
[7] Mahdavi, M.; Sun, L.; Zanibbi, R. Visual Parsing with Query-Driven Global Graph Attention (QD-GGA): Preliminary Results for Handwritten Math Formula Recognition. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); IEEE; pp 2429–2438. <https://doi.org/10.1109/CVPRW50498.2020.00293>.

[10] He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In Computer Vision – ECCV 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2014; pp 346–361. https://doi.org/10.1007/978-3-319-10578-9_23.

Spatial Pyramid Pooling

$m = n = 8$
 $d = 512$

Level 1: 1 x 1

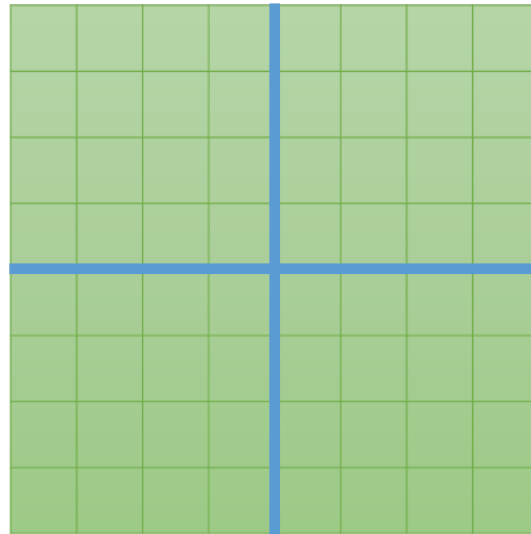


$m \times n \times 512$



1 x 512

Level 2: 2 x 2

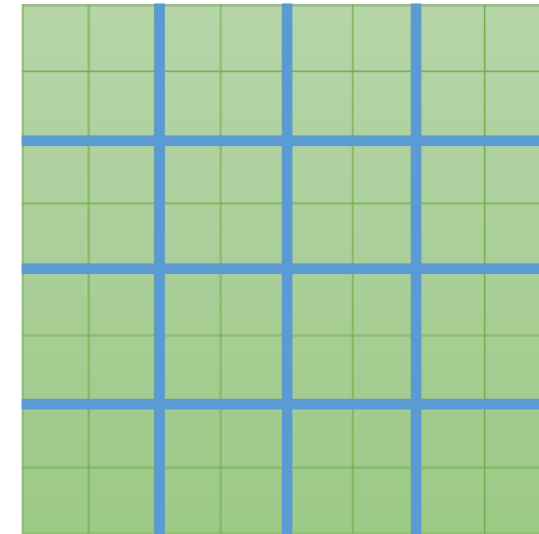


$m \times n \times 512$



4 x 512

Level 3: 4 x 4



$m \times n \times 512$



16 x 512



21 x 512

Fixed-length output

Spatial Pyramid Pooling

Why SPP?

- Captures more spatial information
- Handwritten formulas - variable scales and aspect ratios
- Robust to object deformation
- Reduces overfitting and increases scale invariance

SPP EXPERIMENT

EXPERIMENTAL CONFIGURATIONS FOR SPATIAL PYRAMID POOLING AND BASELINE CONDITIONS

Experiment Config	Levels				No. of bins	Pooling	Output dimension
	1	2	3	4			
Avg (Baseline)	1 x 1				1	Avg.	512-d (512 x 1)
Max	1 x 1				1	Max.	512-d (512 x 1)
SPP_Avg_50	1 x 1	2 x 2	3 x 3	6 x 6	50	Avg.	25600-d (512 x 50)
SPP_Max_50	1 x 1	2 x 2	3 x 3	6 x 6	50	Max.	25600-d (512 x 50)
SPP_Avg_30	1 x 1	2 x 2	3 x 3	4 x 4	30	Avg.	15360-d (512 x 30)
SPP_Max_30	1 x 1	2 x 2	3 x 3	4 x 4	30	Max.	15360-d (512 x 30)

Depth = 512

SPP EXPERIMENT: Results

Dataset	No. of expressions
CROHME 2019 Training set	9993
CROHME 2019 Test set	1199
CROHME 2016 Test set	1147
CROHME 2014 Test set	986

BENCHMARK EXPERIMENT RESULTS ON CROHME 2019 TEST SET

Systems	Symbols		Relationships		Formulas	
	Detection	Det. + Class	Detection	Det. + Class	Structure	Str. + Class
USTC-iFLYTEK	-	-	-	-	91.49	80.73
QD-GGA	97.30	89.09	92.19	90.76	65.55	38.81
Exp: Avg	96.94	88.63	91.48	89.96	64.71	35.12

BENCHMARK RESULTS ON CROHME 2014 and 2016 TEST SETS

Systems	2014	2016
	Str. + Class	Str. + Class
SON	37.63	-
IM2TEX	38.74	-
WAP	46.55	44.55
TAP	61.6	57.02
MAN	54.05	50.56
QD-GGA	32.04	32.84
Graph-to-Graph	54.46	52.05

Conclusion and Future Work

- Research direction towards improving tree-based methods
- Spatial pyramid pooling experiment designed to capture more spatial context
- Future works:
 - Adding global self-attention and context in the model inexpensively
 - Embedding the output adjacency matrix as features
 - Modifying QD-GGA for parsing chemical diagrams

Thank You

This material is based on upon work supported by the Alfred P. Sloan Foundation under Grant No. G-2017-9827 and the National Science Foundation (USA) under Grant No. IIS-1717997.

Thanks to R. Joshi, P. Mali, P. Kukkadapu, A. Keller, M. Mahdavi, L. Sun and J. Diehl. Jian Wu for their contribution to the MathSeer extraction pipeline and QD-GGA.



Alfred P. Sloan
FOUNDATION



Document and Pattern Recognition Lab

For Q&A

Challenges

- Math notations have large symbol vocabulary than ordinary texts.
- Grouping math symbols is not trivial.
- Symbols have different roles in different contexts.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right)$$

312 and 3^{12}

sin, cos, log, ln

$$\sqrt{xy}$$

$$\sum_{i=1}^{10} x_i^2$$

4._□2

$\vec{a}$._ $\vec{b}$

$$x^2 = 2x - \square_1$$

$$\frac{x}{y}$$

Research Agenda

- Problems with current approaches
 - Syntactic approaches are computationally expensive.
 - Encoder-decoder approaches have weak interpretability and error diagnosis.
 - Tree-based methods do not have SOTA recognition rates and current methods lack spatial context.
- Why tree-based?
 - Easy interpretability and error analysis.
 - Natural way of representation.

Research Agenda: Open questions

1. **What are the alternative strategies for pooling so that the model can capture enough spatial context in the inputs?**
2. What strategies can be used for embedding adjacency matrix output probabilities as features during feature extraction?
3. How can we add global self-attention and context to the tree-based methods inexpensively?

Syntactic Methods

- Replacement rules updated to well defined CFG, SCFG (common in NLP) -> production rules defined on the variables and list of terminals
- Toyota et al. -> (CFG) specialized for mathematical formula, called formula description grammar

Year	Systems	Authors	PDF Info	Results for	Rec. Rate
1967	Syntax-directed	Anderson et al.			
2006	1D-CFG	Toyota et al.		INFTY CDB-1	
2009	Linear	Baker et al.	✓		
2011	2D-SCFG	Alvaro et al.		INFTY CDB-1	
2015	Syntax	Ayeb et al.			
2016		Le and Nakagawa		CROHME 2013/14	32.34, 32.86
2017		Le and Nakagawa		CROHME 2014	37.53
2017	Symbol-dominance	Zhang et al.	✓	IM2LATEX-100K, PDFME-10K	
2019	SON	Le et al.		CROHME 2014	37.63

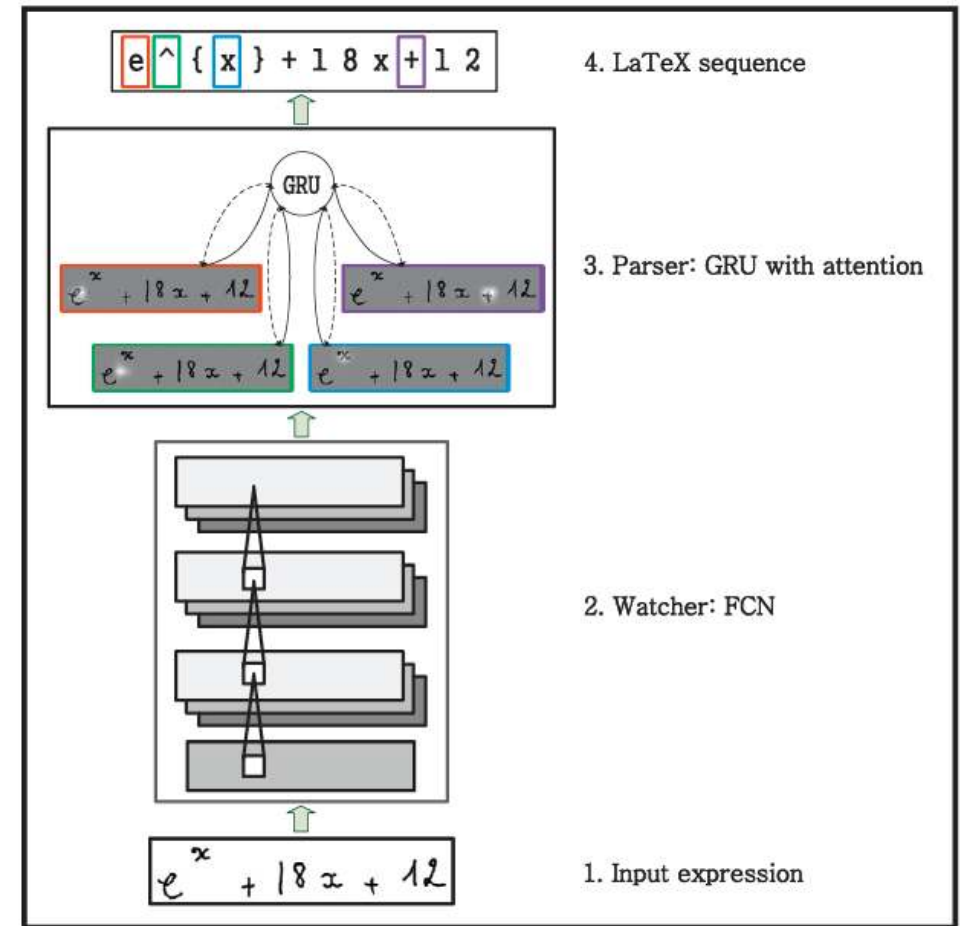
Encoder-decoder models

- Recently common for sequence prediction
- Motivated from image captioning
- Image - > LaTeX
CNN – image features
RNN – decoder (LaTeX)

Year	Systems	Authors	Results for	Rec. Rate
2015		Nguyen et al.	CROHME 2013/14	
2016		Nguyen et al.	CROHME	
2017	IM2TEX	Deng et al.	CROHME 2013/14, IM2LATEX-100K	33.53, 38.74, 77.46
2017		Le and Nakagawa	CROHME 2014	35.19
2017	GRU	Zhang et al.	CROHME 2014	52.43
2017	WAP	Zhang et al.	CROHME 2014/16	46.55, 44.55
2018		Zhang et al.	CROHME 2014/16	52.8, 50.1
2019		Le et al.	CROHME 2014/16	48.78, 45.60
2019		Hong et al.	CROHME 2016	53.35
2019	TAP	Zhang et al.	CROHME 2014/16*	61.16*, 57.02*
2019	MAN	Wang et al.	CROHME 2014/16	54.05, 50.56
2020		Phong et al.	Marmot	

Encoder-decoder models

- Recently common for sequence prediction.
- Motivated from image captioning [3].
- **2D Image** input - > 1D LaTeX output.
- CNN – **offline** features, encoder.
RNN – decoder.
- Attention – focus on relevant regions.
 - Coverage based attention.



[3] WAP Architecture

[3] K. Xu et al., "Show, attend and tell: neural image caption generation with visual attention," in Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, Lille, France, Jul. 2015, pp. 2048–2057.

[4] J. Zhang et al., "Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition," Pattern Recognition, vol. 71, pp. 196–206, Nov. 2017, doi: 10.1016/j.patcog.2017.06.017.

Encoder-decoder models

Coverage-based attention

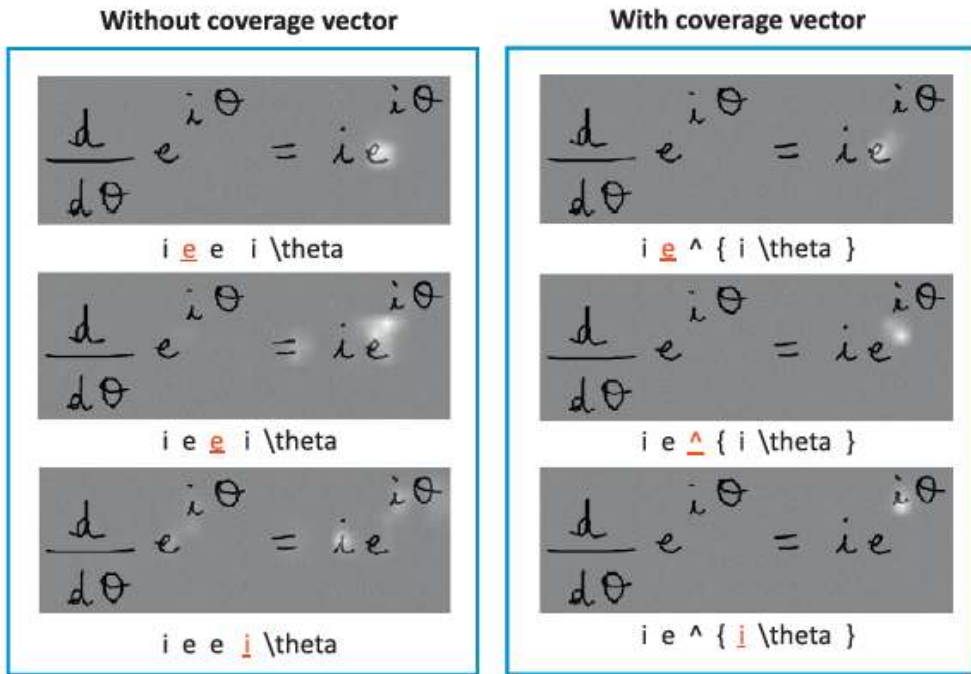
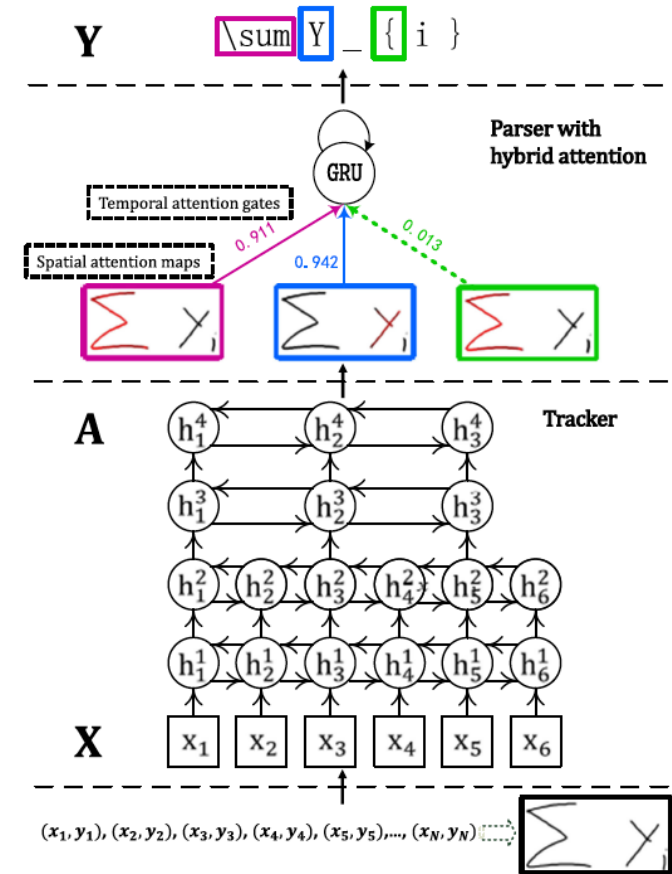


Fig. 4. Examples of attention with and without the coverage vector. The recognized LaTeX sequences of the right side of the equation are printed below each image (the white areas in the images indicate the attended regions, and the underlined text in the LaTeX sequences indicates the corresponding words).

Encoder-decoder models

- Recently common for sequence prediction.
- Motivated from image captioning [3].
- **2D Strokes** input - > 1D LaTeX output.
- RNN – **online** features, encoder, decoder.
- Attention – focus on relevant regions.
 - Guided Hybrid Attention
 - coverage-based spatial attention,
 - temporal attention



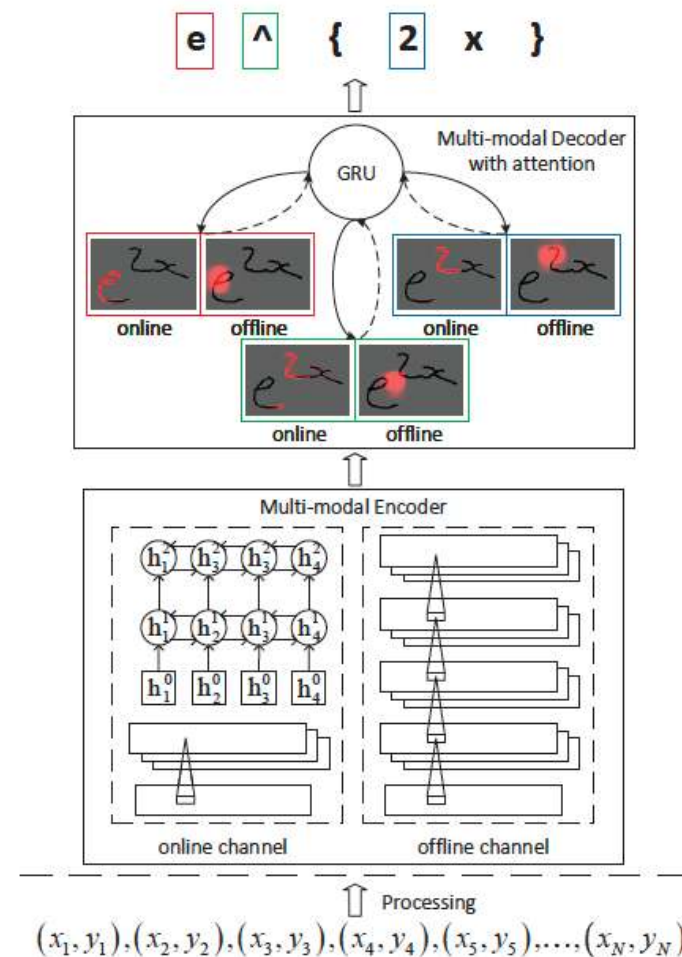
[5] TAP Architecture

[3] K. Xu et al., "Show, attend and tell: neural image caption generation with visual attention," in Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, Lille, France, Jul. 2015, pp. 2048–2057.

[5] J. Zhang, J. Du, and L. Dai, "Track, Attend, and Parse (TAP): An End-to-End Framework for Online Handwritten Mathematical Expression Recognition," IEEE Transactions on Multimedia, vol. 21, no. 1, pp. 221–233, Jan. 2019, doi: 10.1109/TMM.2018.2844689.

Encoder-decoder models

- Recently common for sequence prediction.
- Motivated from image captioning.
- **Strokes / Image** input - > LaTeX output.
- **Online + Offline** features.
- CNN + RNN – encoder.
- RNN – decoder.
- Attention - focus on relevant regions.

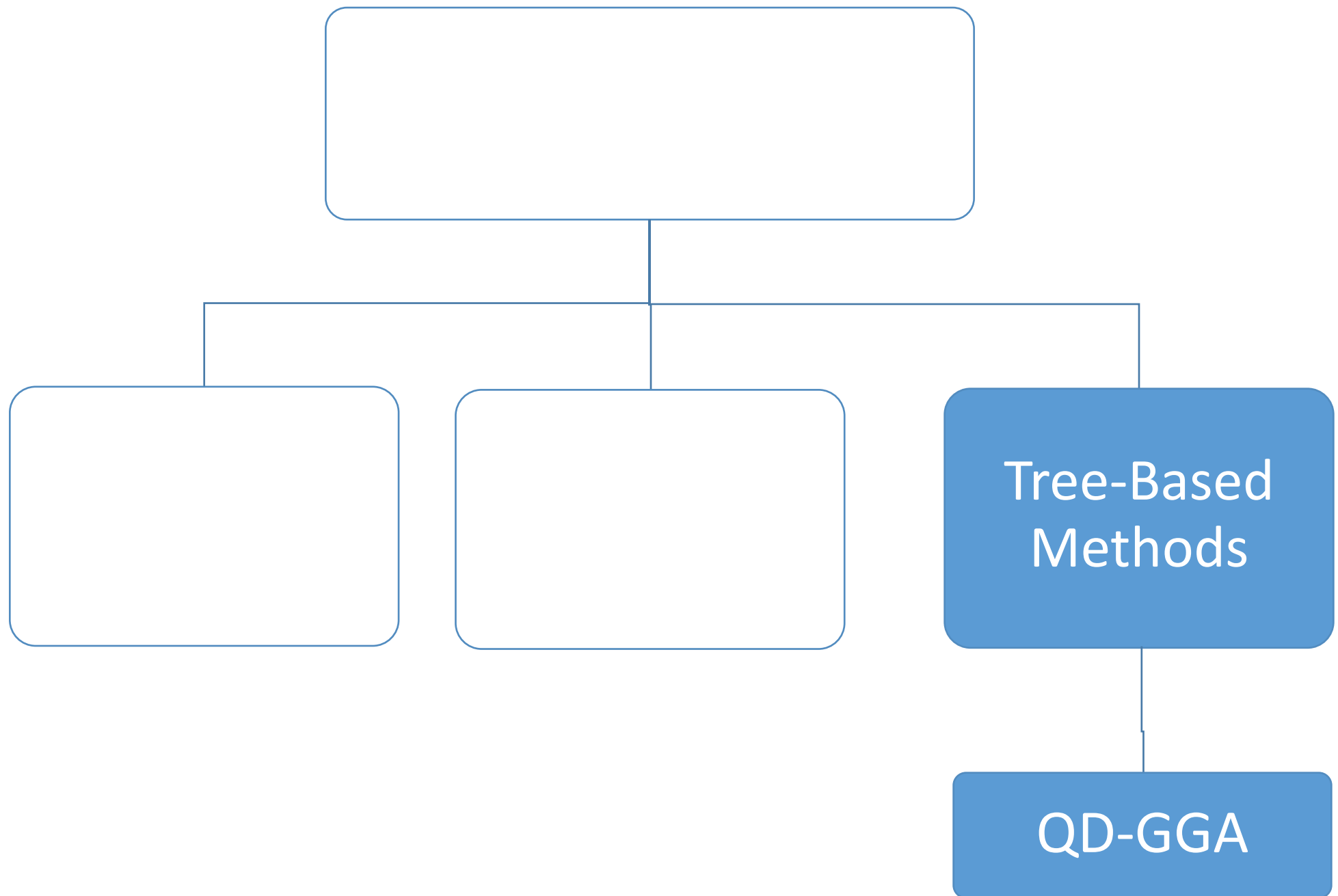


[5] MAN Architecture

Tree-based Methods

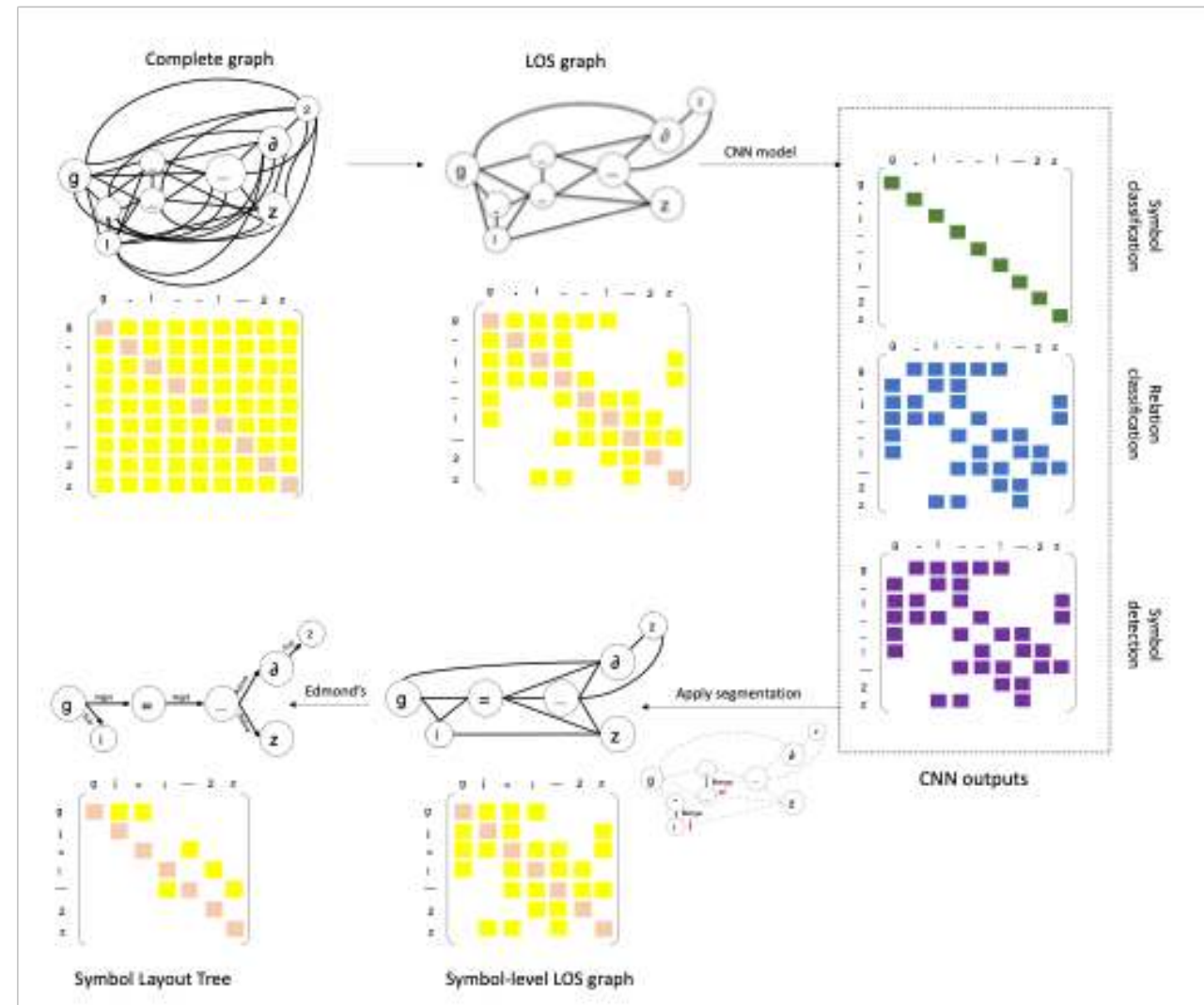
- Graph-search problem (MST extraction)
- Score / Probability distributions
- Graph over input primitives

Year	Systems	Authors	PDF Info	Results for	Rec. Rate
2001		Eto et al.			
2003	INFTY	Suzuki et al.			
2016	INFTY (Updated)	Suzuki et al.	✓		
2016		Hu et al.		CROHME 2014	26.88
2017		Zhang et al.		CROHME 2014	21.59
2019	LPGA	Mahdavi et al.		INFTYCDB-2	
2019	Text2Math	Zou and Lu			
2020	Graph2Tree	Li et al.			
2020	QD-GGA	Mahdavi et al.	✓	CROHME 2014/16/19	32.04, 32.84, 40.65
2021	Graph-to-Graph	Wu et al.		CROHME 2014/16	54.46, 52.05



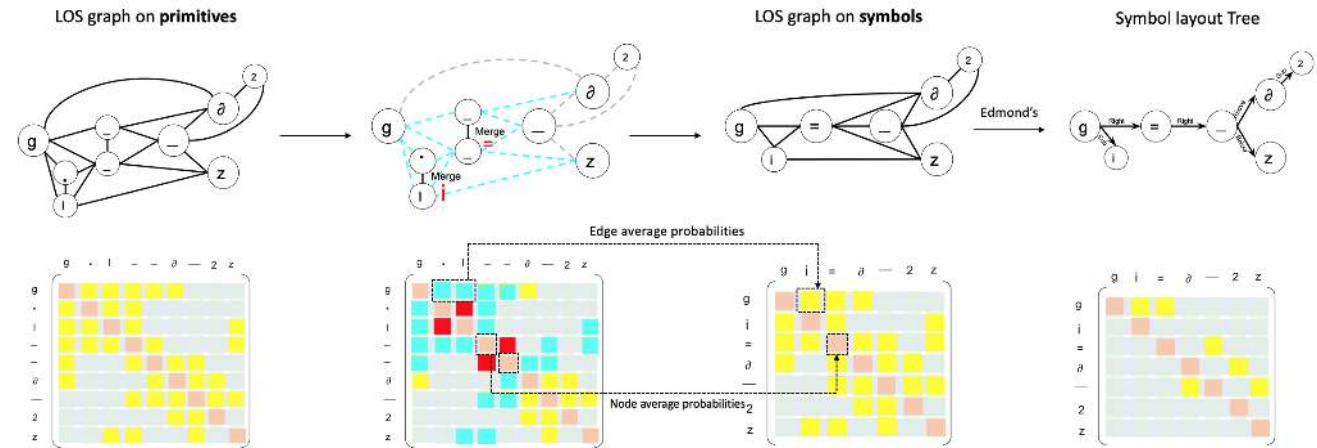
Overview of QD-GGA

- Graph-based MTL framework
- End-to-end training
- **Input:** LOS graph over input primitives
- **Output:** Symbol Layout Tree



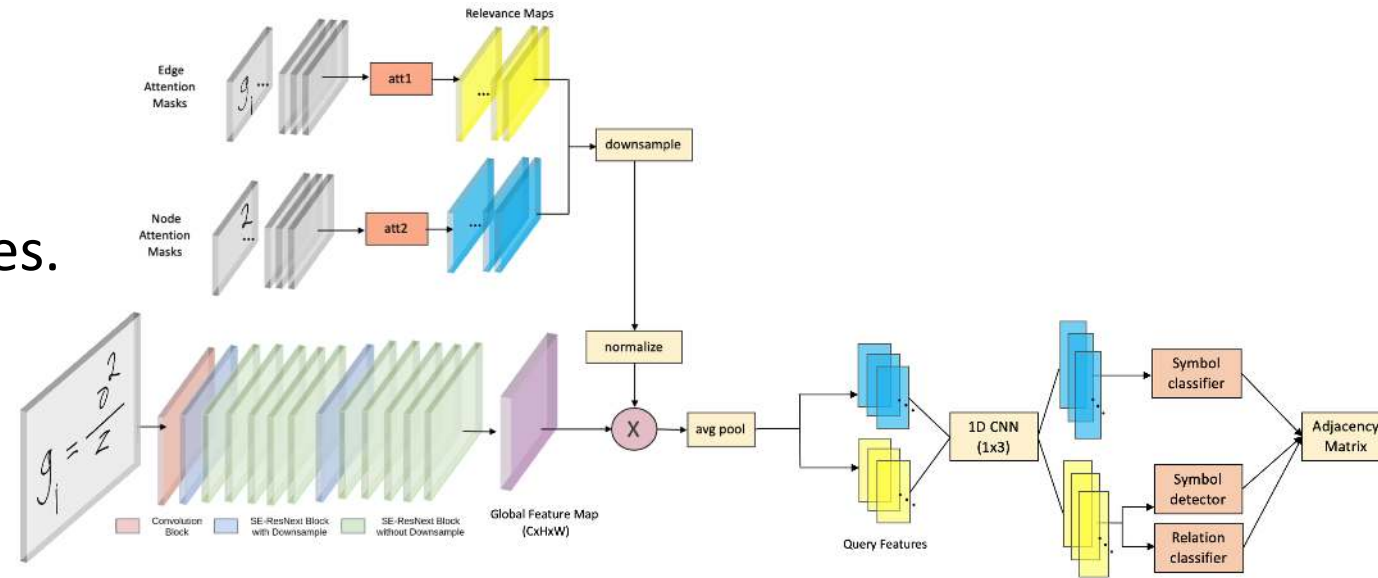
Overview of QD-GGA

- LOS graph on strokes converted to LOS graph on symbols
- Max Spanning Tree extraction from LOS graph
- Edmond's algorithm



Tree-based Methods

- QD-GGA [7]
 - MTL end-to-end framework.
 - **Input:** LOS graph over input primitives.
 - **Output:** Symbol Layout Tree.
 - Edmond's algorithm [8].
- Architecture
 - 2 Attention modules
 - Joint loss
 - 3 adjacency matrices output



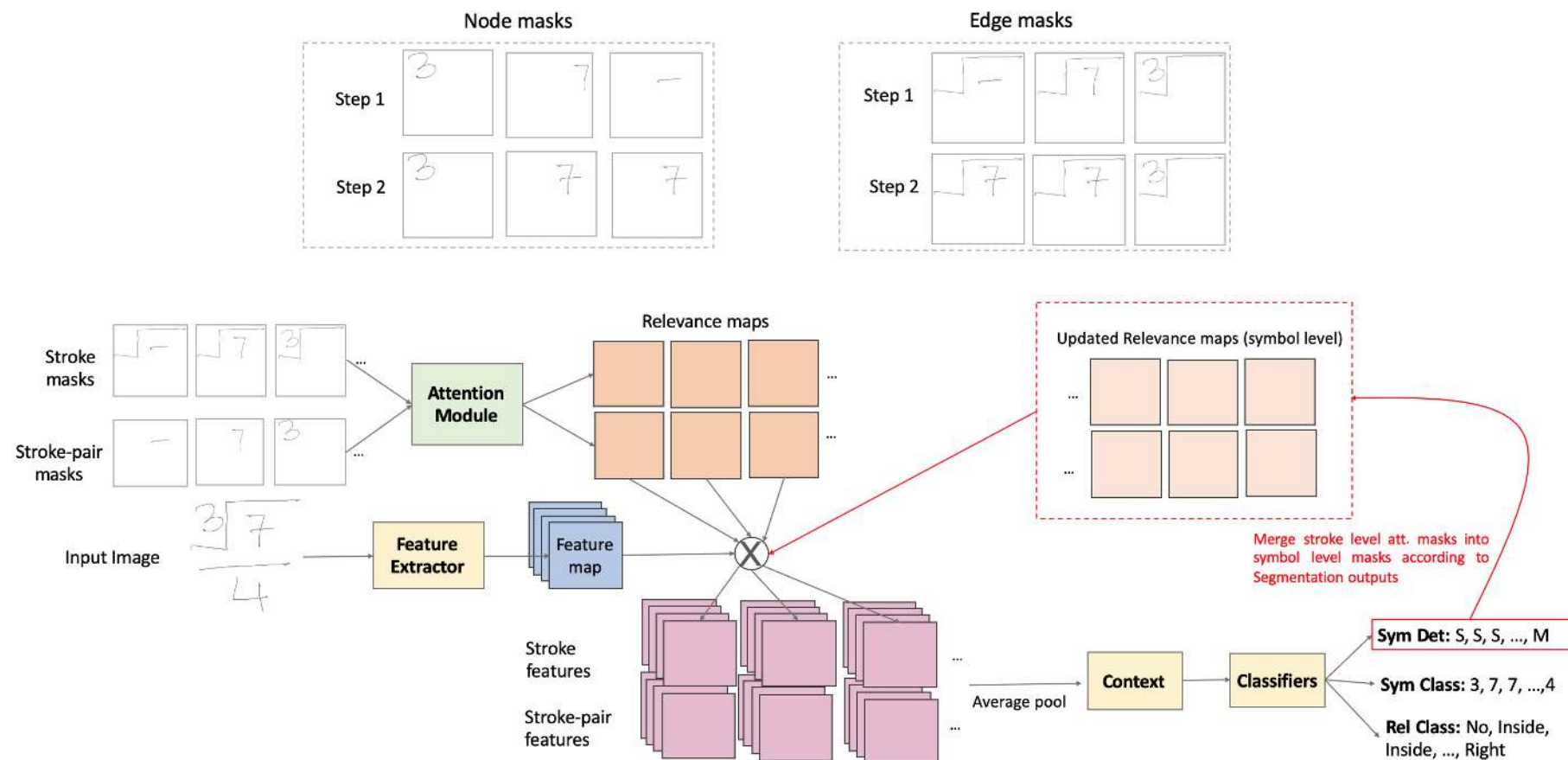
[7] QD-GGA Architecture

[6] Y. Eto and M. Suzuki, "Mathematical formula recognition using virtual link network," in Proceedings of Sixth International Conference on Document Analysis and Recognition, Sep. 2001, pp. 762–767. doi: 10.1109/ICDAR.2001.953891.

[7] Mahdavi, M.; Sun, L.; Zanibbi, R. Visual Parsing with Query-Driven Global Graph Attention (QD-GGA): Preliminary Results for Handwritten Math Formula Recognition. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); IEEE; pp 2429–2438. <https://doi.org/10.1109/CVPRW50498.2020.00293>.

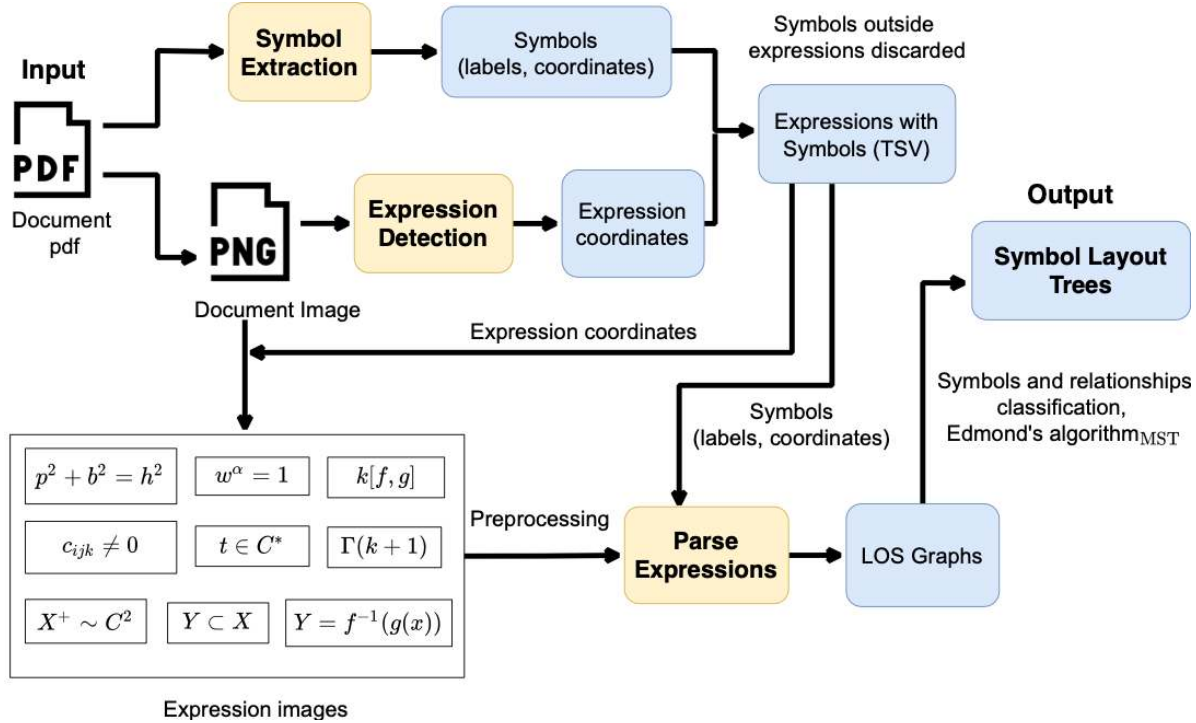
[8] J. Edmonds, "Optimum branchings," J. RES. NATL. BUR. STAN. SECT. B. MATH. MATH. PHYS., vol. 71B, no. 4, p. 233, Oct. 1967, doi: 10.6028/jres.071B.032.

Overview of QD-GGA



PREPARATION FOR EXPERIMENTS

PREPARATION FOR EXPERIMENTS



One can also think of $\zeta(T, k)$ as the zeta function of a nonsingular curve C over $\mathbb{F}_q$ whose field of functions is k . For example, let $C/\mathbb{F}_q$ be a plane curve given by an equation

$$(33) \quad f(X_1, X_2, X_3) = 0$$

where f is nonsingular and homogeneous of some degree and has coefficients in $\mathbb{F}_q$. For each $n \geq 1$ let N_n be the number of projective solutions to (33) in $\mathbb{F}_{q^n}$. The zeta function of the field of functions k of C is the same as the zeta function of the curve C over $\mathbb{F}_q$ which is defined as

$$(34) \quad \zeta(T, C/\mathbb{F}_q) = \exp \left(\sum_{n=1}^{\infty} \frac{N_n T^n}{n} \right).$$

This geometric point of view is very powerful. For example, the Riemann-Roch Theorem on the curve C plays the role of the Poisson summation formula [SCH] and shows that

$$(35) \quad \zeta(T, C/\mathbb{F}_q) = \frac{P(T, C/\mathbb{F}_q)}{(1-T)(1-qT)}$$

where $P \in \mathbb{Z}[T]$ is of degree $2g$, g being the genus of the curve C . It also gives the functional equation $P(T) = q^g T^{2g} P(1/qT)$. The Riemann-Hypothesis for these zeta functions, which was put forth and tested in many examples by Artin, asserts that all the zeroes lie on $|T| = 1/\sqrt{q}$. This was proven by Weil. By now there are several different proofs: Weil [WE3], [WE4], elementary proofs by Stepanov [ST] and Bombieri [BO], and proofs by Deligne [DE] which have the advantage of applying much more generally. One reason for being able to proceed in the function

One can also think of $\zeta(T, k)$ as the zeta function of a nonsingular curve C over $\mathbb{F}_q$ whose field of functions is k . For example, let $C/\mathbb{F}_q$ be a plane curve given by an equation

$$(33) \quad f(X_1, X_2, X_3) = 0$$

where f is nonsingular and homogeneous of some degree and has coefficients in $\mathbb{F}_q$. For each $n \geq 1$ let N_n be the number of projective solutions to (33) in $\mathbb{F}_{q^n}$. The zeta function of the field of functions k of C is the same as the zeta function of the curve C over $\mathbb{F}_q$ which is defined as

$$(34) \quad \zeta(T, C/\mathbb{F}_q) = \exp \left(\sum_{n=1}^{\infty} \frac{N_n T^n}{n} \right).$$

This geometric point of view is very powerful. For example, the Riemann-Roch Theorem on the curve C plays the role of the Poisson summation formula [SCH] and shows that

$$(35) \quad \zeta(T, C/\mathbb{F}_q) = \frac{P(T, C/\mathbb{F}_q)}{(1-T)(1-qT)}$$

where $P \in \mathbb{Z}[T]$ is of degree $2g$, g being the genus of the curve C . It also gives the functional equation $P(T) = q^g T^{2g} P(1/qT)$. The Riemann-Hypothesis for these zeta functions, which was put forth and tested in many examples by Artin, asserts that all the zeroes lie on $|T| = 1/\sqrt{q}$. This was proven by Weil. By now there are several different proofs: Weil [WE3], [WE4], elementary proofs by Stepanov [ST] and Bombieri [BO], and proofs by Deligne [DE] which have the advantage of applying much more generally. One reason for being able to proceed in the function

One can also think of $\zeta(T, k)$ as the zeta function of a nonsingular curve C over $\mathbb{F}_q$ whose field of functions is k . For example, let $C/\mathbb{F}_q$ be a plane curve given by an equation

$$(33) \quad f(X_1, X_2, X_3) = 0$$

where f is nonsingular and homogeneous of some degree and has coefficients in $\mathbb{F}_q$. For each $n \geq 1$ let N_n be the number of projective solutions to (33) in $\mathbb{F}_{q^n}$. The zeta function of the field of functions k of C is the same as the zeta function of the curve C over $\mathbb{F}_q$ which is defined as

$$(34) \quad \zeta(T, C/\mathbb{F}_q) = \exp \left(\sum_{n=1}^{\infty} \frac{N_n T^n}{n} \right).$$

This geometric point of view is very powerful. For example, the Riemann-Roch Theorem on the curve C plays the role of the Poisson summation formula [SCH] and shows that

$$(35) \quad \zeta(T, C/\mathbb{F}_q) = \frac{P(T, C/\mathbb{F}_q)}{(1-T)(1-qT)}$$

where $P \in \mathbb{Z}[T]$ is of degree $2g$, g being the genus of the curve C . It also gives the functional equation $P(T) = q^g T^{2g} P(1/qT)$. The Riemann-Hypothesis for these zeta functions, which was put forth and tested in many examples by Artin, asserts that all the zeroes lie on $|T| = 1/\sqrt{q}$. This was proven by Weil. By now there are several different proofs: Weil [WE3], [WE4], elementary proofs by Stepanov [ST] and Bombieri [BO], and proofs by Deligne [DE] which have the advantage of applying much more generally. One reason for being able to proceed in the function

One can also think of $\zeta(T, k)$ as the zeta function of a nonsingular curve C over $\mathbb{F}_q$ whose field of functions is k . For example, let $C/\mathbb{F}_q$ be a plane curve given by an equation

$$(33) \quad f(X_1, X_2, X_3) = 0$$

where f is nonsingular and homogeneous of some degree and has coefficients in $\mathbb{F}_q$. For each $n \geq 1$ let N_n be the number of projective solutions to (33) in $\mathbb{F}_{q^n}$. The zeta function of the field of functions k of C is the same as the zeta function of the curve C over $\mathbb{F}_q$ which is defined as

$$(34) \quad \zeta(T, C/\mathbb{F}_q) = \exp \left(\sum_{n=1}^{\infty} \frac{N_n T^n}{n} \right).$$

This geometric point of view is very powerful. For example, the Riemann-Roch Theorem on the curve C plays the role of the Poisson summation formula [SCH] and shows that

$$(35) \quad \zeta(T, C/\mathbb{F}_q) = \frac{P(T, C/\mathbb{F}_q)}{(1-T)(1-qT)}$$

where $P \in \mathbb{Z}[T]$ is of degree $2g$, g being the genus of the curve C . It also gives the functional equation $P(T) = q^g T^{2g} P(1/qT)$. The Riemann-Hypothesis for these zeta functions, which was put forth and tested in many examples by Artin, asserts that all the zeroes lie on $|T| = 1/\sqrt{q}$. This was proven by Weil. By now there are several different proofs: Weil [WE3], [WE4], elementary proofs by Stepanov [ST] and Bombieri [BO], and proofs by Deligne [DE] which have the advantage of applying much more generally. One reason for being able to proceed in the function

[9] A Math Formula Extraction Framework

PREPARATION FOR EXPERIMENTS

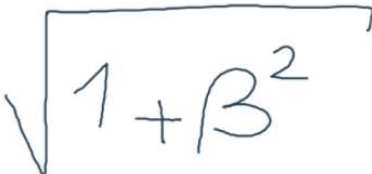
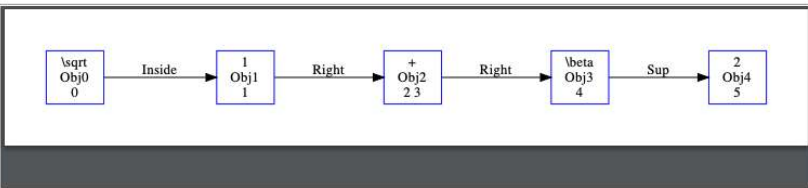
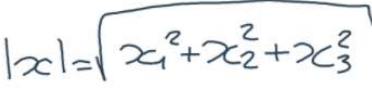
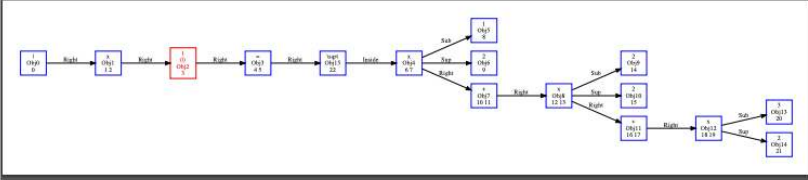
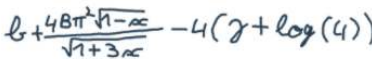
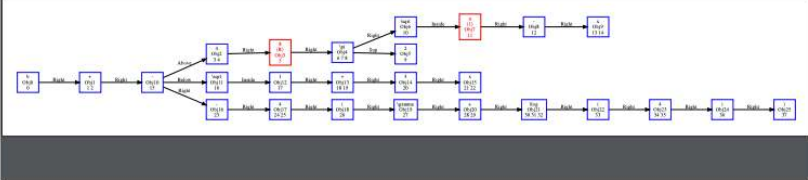
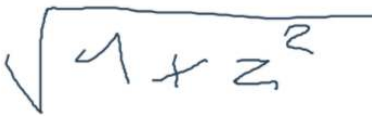
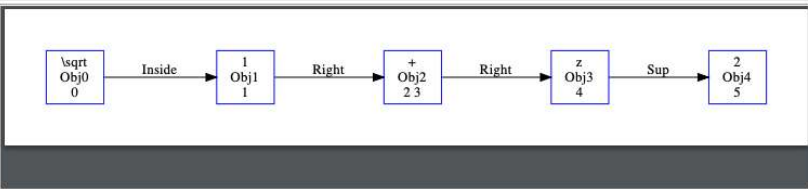
MathSeer Pipeline Results Visualization

Pdf name: K15-1002
 Page: 3

Expression name	Expression image	MathML Output
K15-1002_P0P001		j
K15-1002_P0P002		$g_n = w_1^T \varphi(m)$
K15-1002_P0P003		$\varphi(m)$
K15-1002_P0P004		i
K15-1002_P0P005		j
K15-1002_P0P006		i

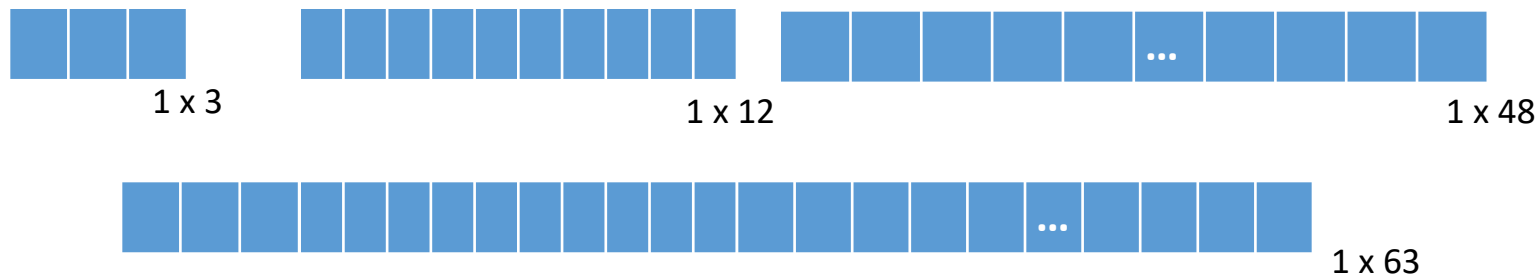
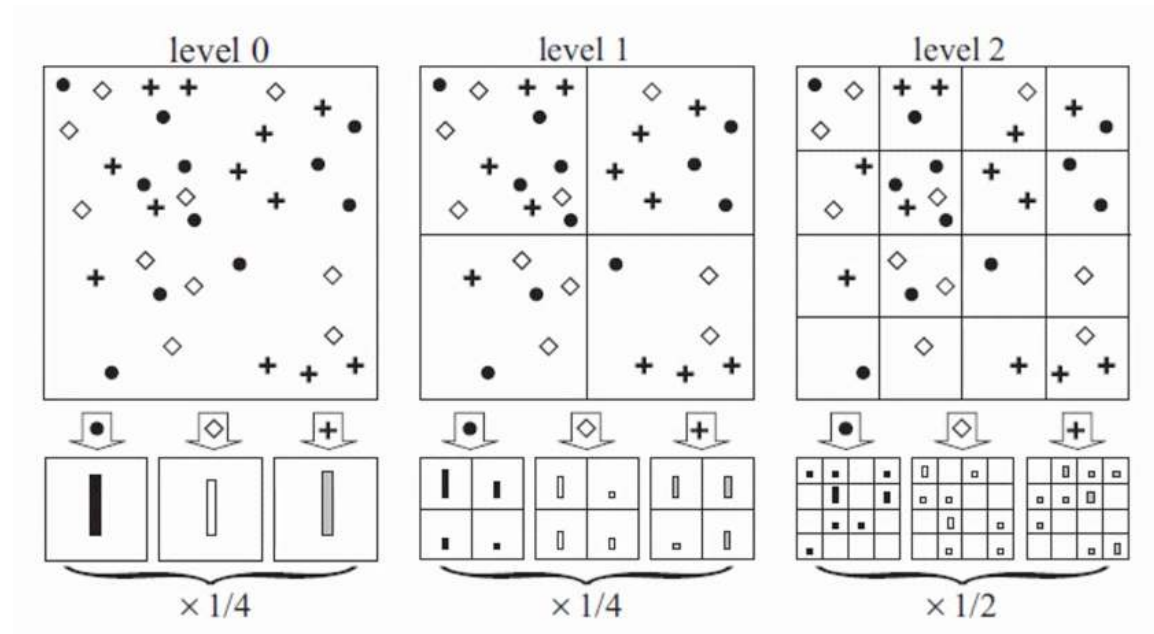
[9] Results visualization

PREPARATION FOR EXPERIMENTS

Primitive Targets		Primitive level Errors		
1	Errors 1 23 errors	Filename	Image	LG
		UN19wb_1112_em_1059.lg		
		UN19_1033_em_478.lg		
		UN19_1027_em_385.lg		
		UN19_1026_em_369.lg		

Details of specific error instance

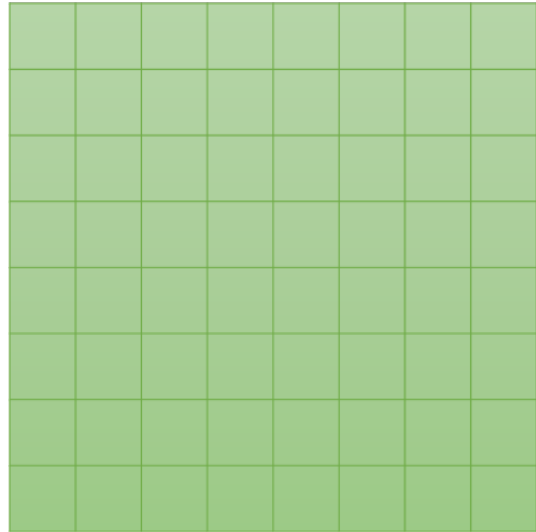
Spatial Pyramid Matching (SPM)



Spatial Pyramid Pooling

$m = n = 8$
 $d = 512$

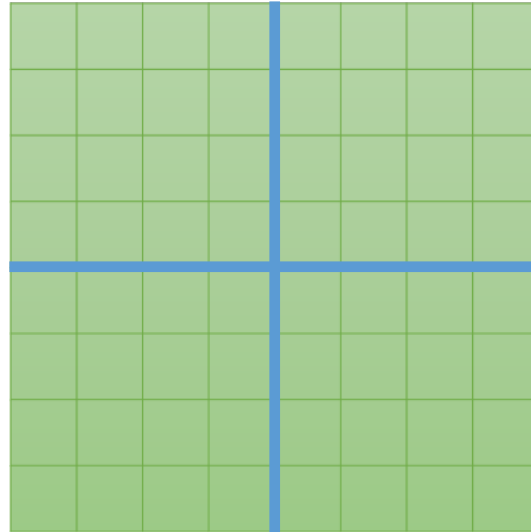
Level 1: 1 x 1



$m \times n \times d$

Size = 8 x 8
Stride = 8 x 8

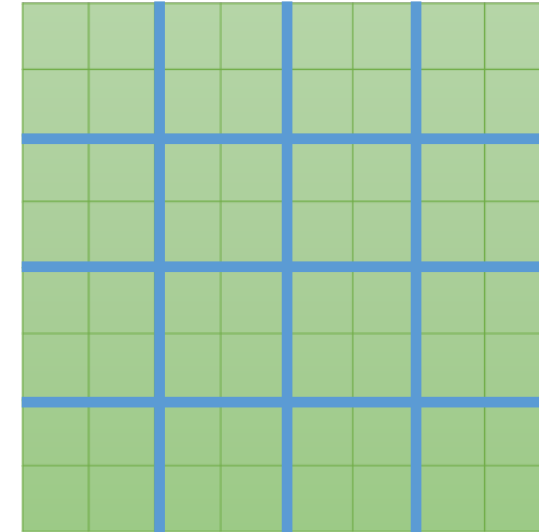
Level 2: 2 x 2



$m \times n \times d$

Size = 4 x 4
Stride = 4 x 4

Level 3: 4 x 4



$m \times n \times d$

Size = 2 x 2
Stride = 2 x 2

$$\text{Size} = \left\lceil \frac{m}{k} \right\rceil \times \left\lceil \frac{n}{k} \right\rceil$$

$$\text{Stride} = \left\lceil \frac{m}{k} \right\rceil \times \left\lceil \frac{n}{k} \right\rceil$$

Dataset and Evaluation

- Dataset
 - CROHME 2019 (9993 expressions) dataset for training.
 - CROHME 2014, 2016, 2019 test sets for evaluation.
 - Expanding the CROHME datasets by decomposing each HME in the dataset to sub-HMEs. [11]
- Evaluation metrics using LgEval tool [12]
 - Formula recognition rates (unlabeled and labeled graph),
 - Detection and classification F-scores for symbols and relationships.

[11] A. D. Le, B. Indurkha, and M. Nakagawa, "Pattern generation strategies for improving recognition of Handwritten Mathematical Expressions," *Pattern Recognition Letters*, vol. 128, pp. 255–262, Dec. 2019, doi: [10.1016/j.patrec.2019.09.002](https://doi.org/10.1016/j.patrec.2019.09.002).

[12] H. Mouchère, R. Zanibbi, U. Garain, and C. Viard-Gaudin, "Advancing the state of the art for handwritten math recognition: the CROHME competitions, 2011–2014," *IJDAR*, vol. 19, no. 2, pp. 173–189, Jun. 2016, doi: [10.1007/s10032-016-0263-5](https://doi.org/10.1007/s10032-016-0263-5).

SPP EXPERIMENT: Preliminary Results

RESULTS ON CROHME 2019 TEST SET (1199 EXPRESSIONS)

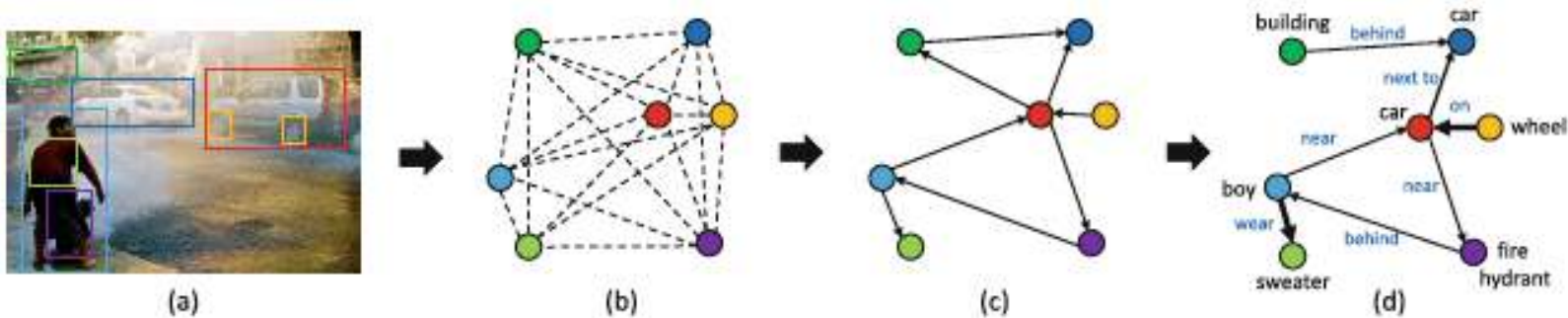
Experiment Configuration		Symbols		Relationships		Formulas	
	Epochs	Detection	Det. + Class	Detection	Det. + Class	Structure	Str. + Class
Avg (Baseline)	99	96.94	88.63	91.48	89.96	64.71	35.12
Max							
SPP_Avg_50							
SPP_Max_50							
SPP_Avg_30							
SPP_Max_30							

Global attention

- Sequential attention on query features
 - How much of other neighbors should contribute for each target query?

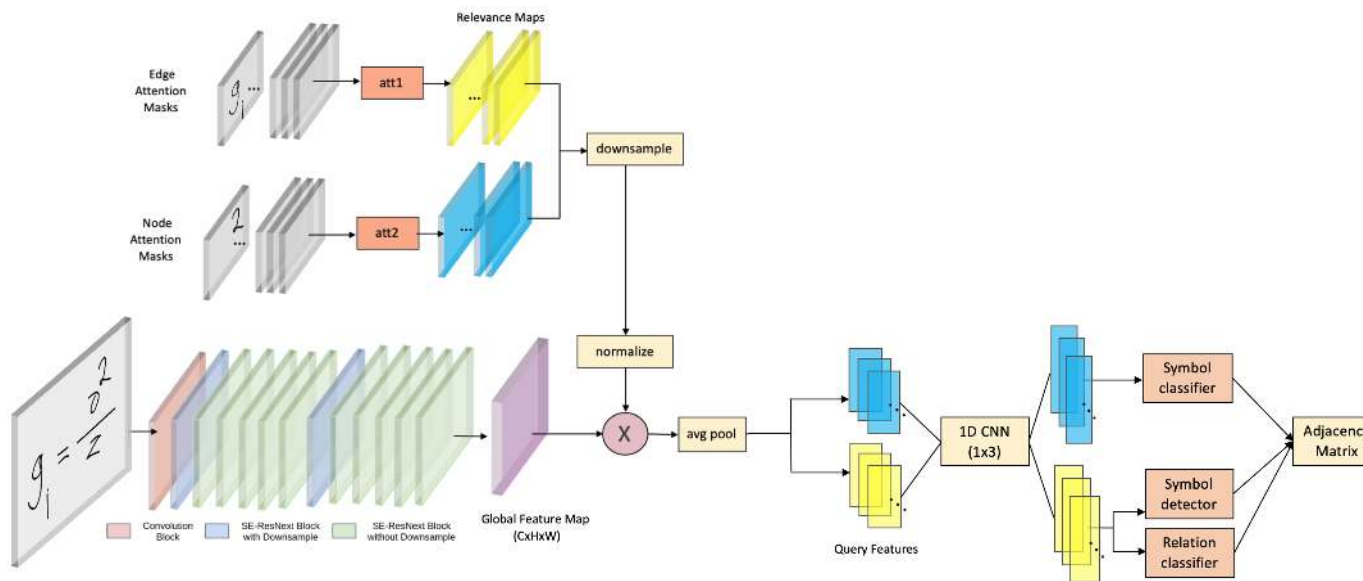
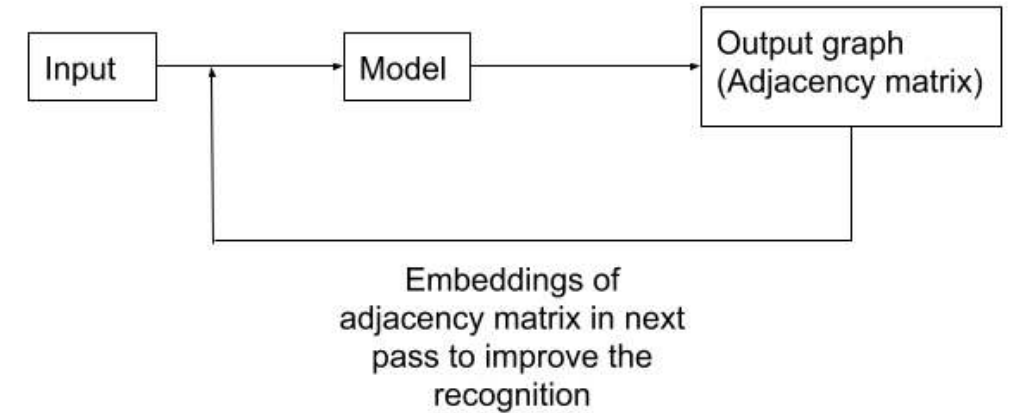
Global attention

- Graph sequential learning module instead of 1-by-3 CNNs



Embedding adjacency matrix as features

- include adjacency matrix probabilities in feature extraction



[7] QD-GGA Architecture