Transcribing Content from Structural Images with Spotlight Mechanism

Yu Yin, Zhenya Huang, Enhong Chen*, Qi Liu, Fuzheng Zhang, Xin Xie, Guoping Hu

Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China; Microsoft Research Asia; iFLYTEK Research

Reporter: Yu Yin

Date: 2018/8/21

Outline

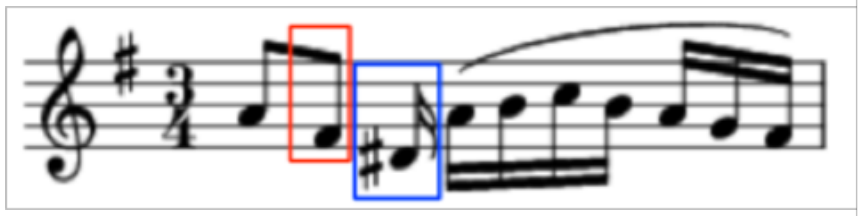
- Introduction
- Preliminaries
- Spotlight Transcribing Network
- Experiments
- Conclusion

Outline

- **Introduction**
- Preliminaries
- Spotlight Transcribing Network
- Experiments
- Conclusion

Image Transcription

- Definition: recognizing semantic information in images into comprehensible forms (e.g., text)
- **tldr: from image to text**
- Examples:



$$f(x) = \frac{\sqrt{x} - 1}{x - 2}$$

Related Research

Related topics:

- Optical Character Recognition (OCR)
- Scene Text Recognition

Related methods:

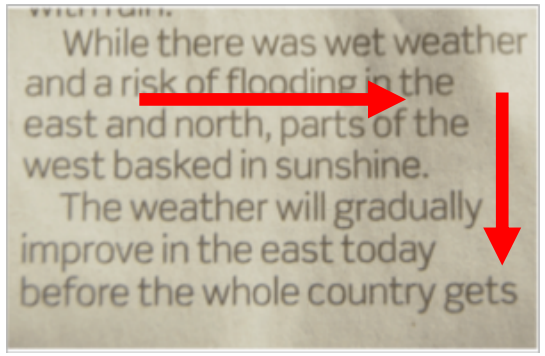
- Rule-based / feature-based methods
- Encoder-decoder architecture

- Mainly focus on images with straightforward content

Structural images?

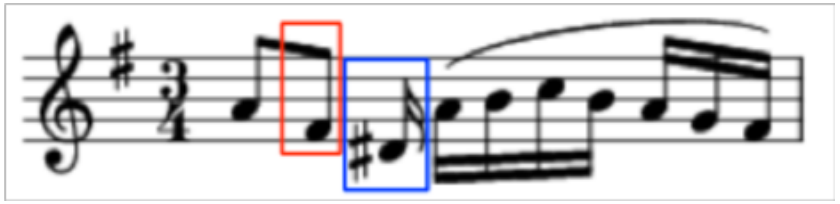
Structural Images

- Natural images / printed documents:



Simple structure

- **Structural images**:

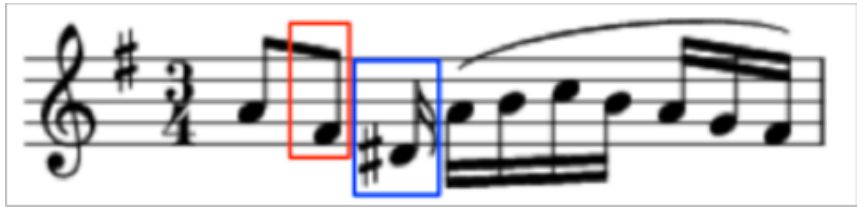


$$f(x) = \frac{\sqrt{x-1}}{x-2}$$

Complex structure

Structural Image Transcription

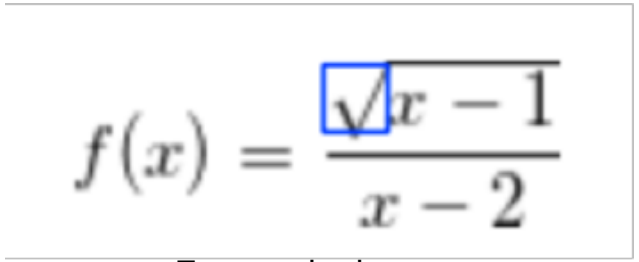
- The problem: structural image transcription
  - From: structural images
  - To: source code in the specific language that fully defines the image (i.e. able to generate back the image)



Music score

```
a8[ fis] dis16 a[ b c ...
```

LilyPond code



$$f(x) = \frac{\sqrt{x-1}}{x-2}$$

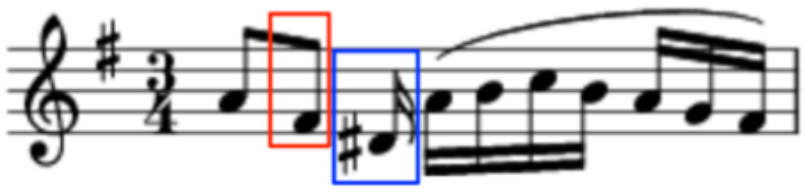
Formula image

```
f(x)=\frac{\sqrt{x-1}}{x-2}
```

TeX code

Challenges

- Unique characteristics:
  - Objects organized in a **complex manner**
  - **Much** semantics in **small** area
  - **Similar** objects



a8[fis] dis16 a[b c ...

(a) Music score example

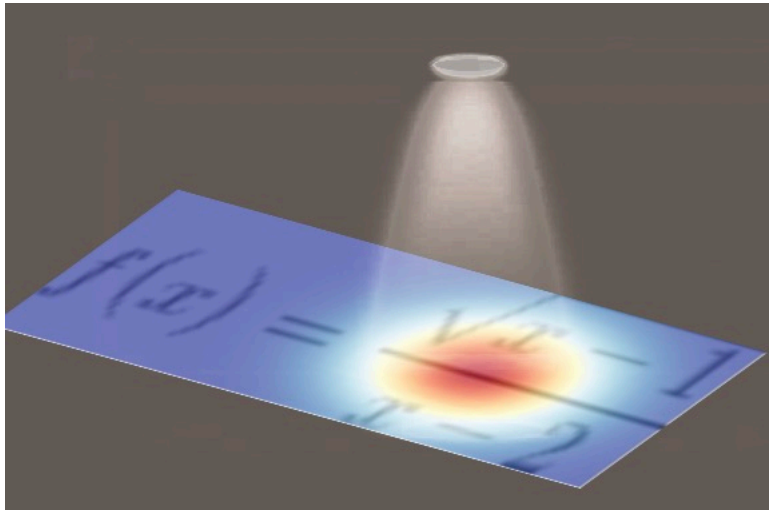
$$f(x) = \frac{\sqrt{x-1}}{x-2}$$

f(x)=\frac{\sqrt{x-1}}{x-2}

(b) Formula example

Our Approach

- Spotlight mechanism:
  - Mimic human visual attention more intuitively
  - Focus on one spot, write down the content, then focus on another spot
  - Simulate this process by shedding a **spotlight** and modeling its movement



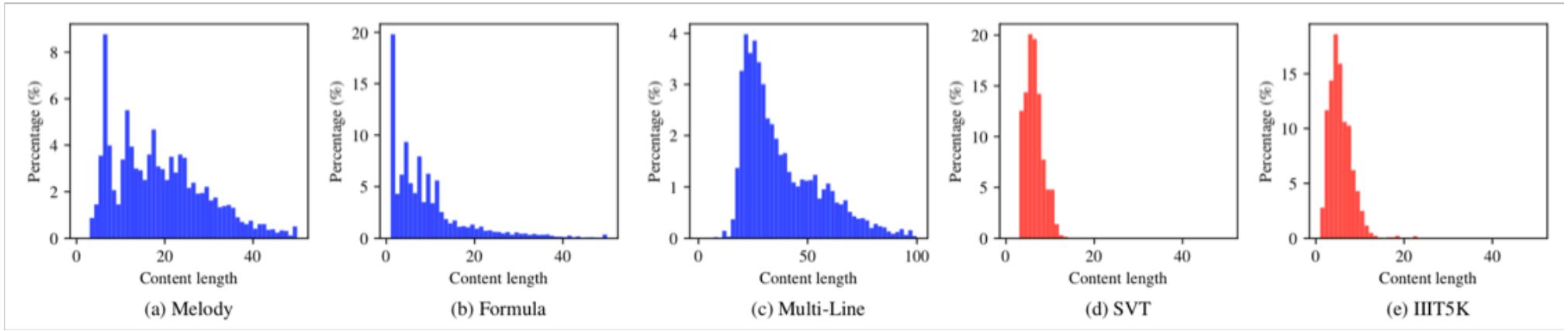
Outline

- Introduction
- **Preliminaries**
- Spotlight Transcribing Network
- Experiments
- Conclusion

Data Description

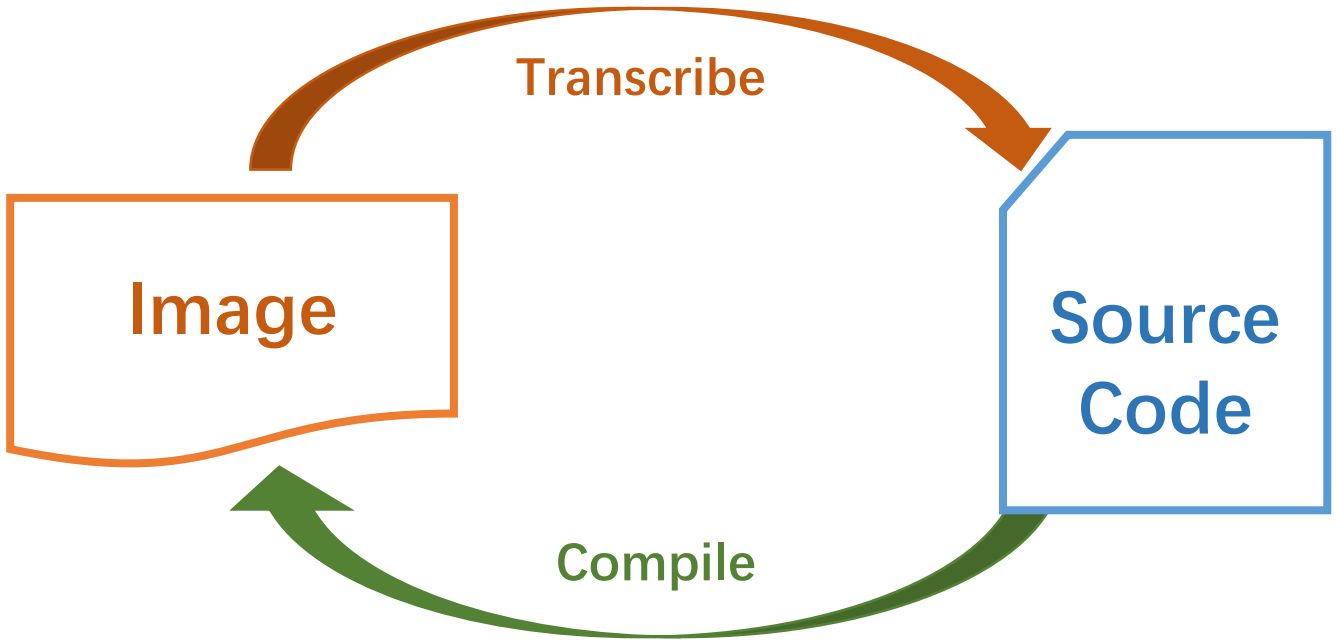
- Datasets: Melody, Formula, Multi-Line
- Compare with scene text recognition: SVT, IIIT5K
- Three main differences:
  - More information to preserve
  - Larger output space
  - Complete transcribing: reversible

Dataset	Image count	Token space	Token count	Avg. tokens per image	Avg. image pixels
Melody	4208	70	82,834	19.7	15,602.7
Formula	61649	127	607,061	9.7	1,190.7
Multi-Line	4595	127	182,112	39.8	9,016.6
SVT	618	26	3,796	5.9	12,733.5
IIIT5K	3000	36	15,269	5.0	11,682.0



Problem Definition

- Input: $W \times H$ gray-scale image x
- Output: token sequence $y = \{y_1, \ldots, y_t\}$ representing source code
- Note: the process is **reversible**



Outline

- Introduction
- Preliminaries
- **Spotlight Transcribing Network**
- Experiments
- Conclusion

Spotlight Transcribing Network

At a glance:

- **Encoder-decoder** architecture
- Two-stage "where-to-what" solution:
  - *Spotlight* module for "**where-to-look**" problem
  - GRU based transcribing module for "**what-to-write**" problem
- Two implementations:
  - **STNM** *with Markov property*
  - **STNR** *with Recurrent modeling*
- Refinement through **reinforcement learning**

Spotlight vs. Attention

Both highlighting different parts when decoding by assigning different weights to feature vectors, but:

- ➢Attention: has to learn to focus
- ✓Spotlight: focus by design

- ➢Attention weights are assigned based on "what" but not "where"
- ✓Spotlight directly finds a reading path that <u>follows the structure</u>

- ➢Attention and transcription are modeled together
- ✓Separate modules for spotlight and transcription

STN Framework

- Spotlight Transcribing Network (STN):
  - Image encoder: CNN based feature extractor
  - Transcribing decoder: **Spotlight Module** and **Transcription Module**

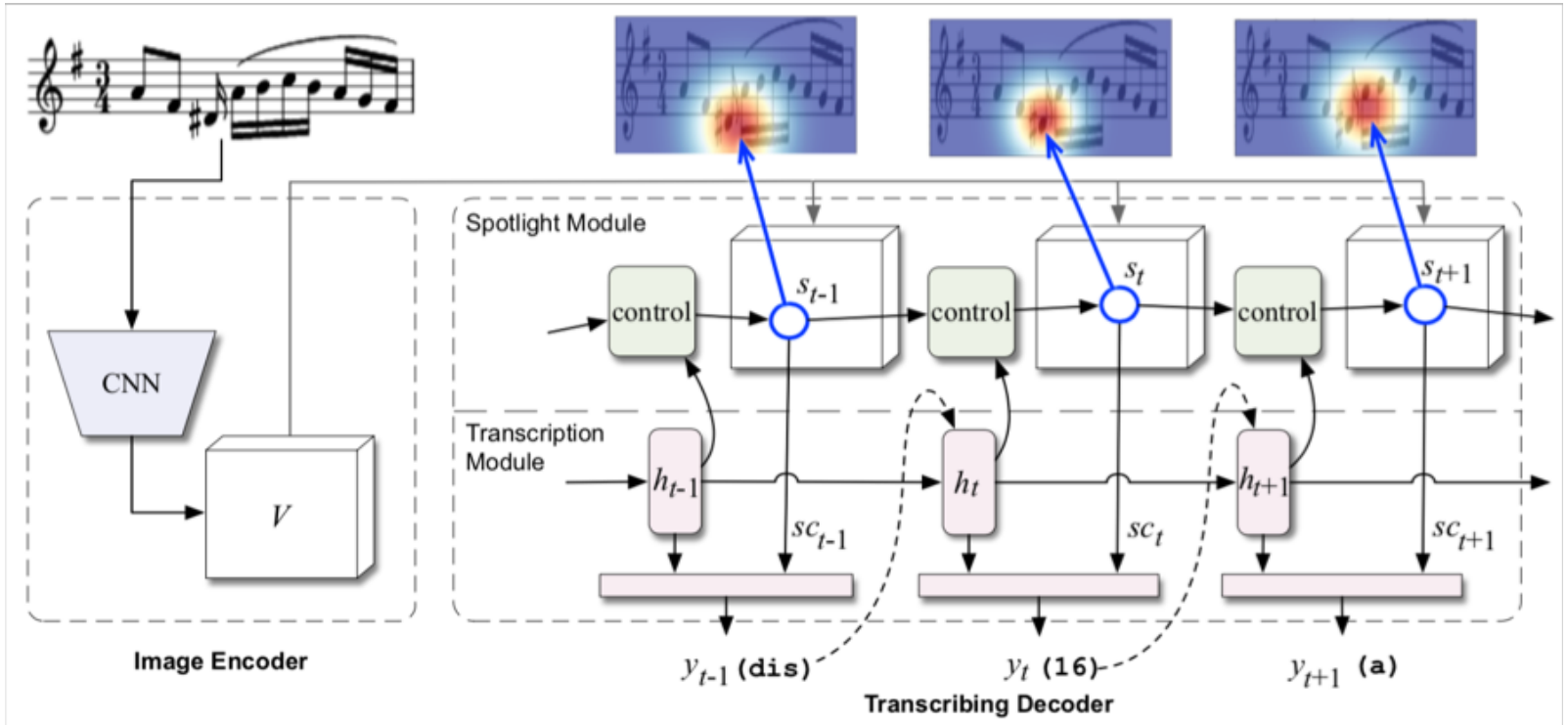


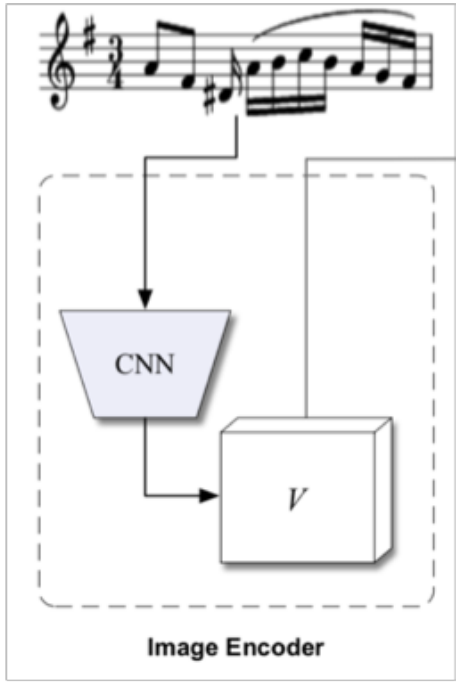
Image Encoder

- Image encoder: feature extraction

Feature vectors

$$V = f(x; \theta_f),$$

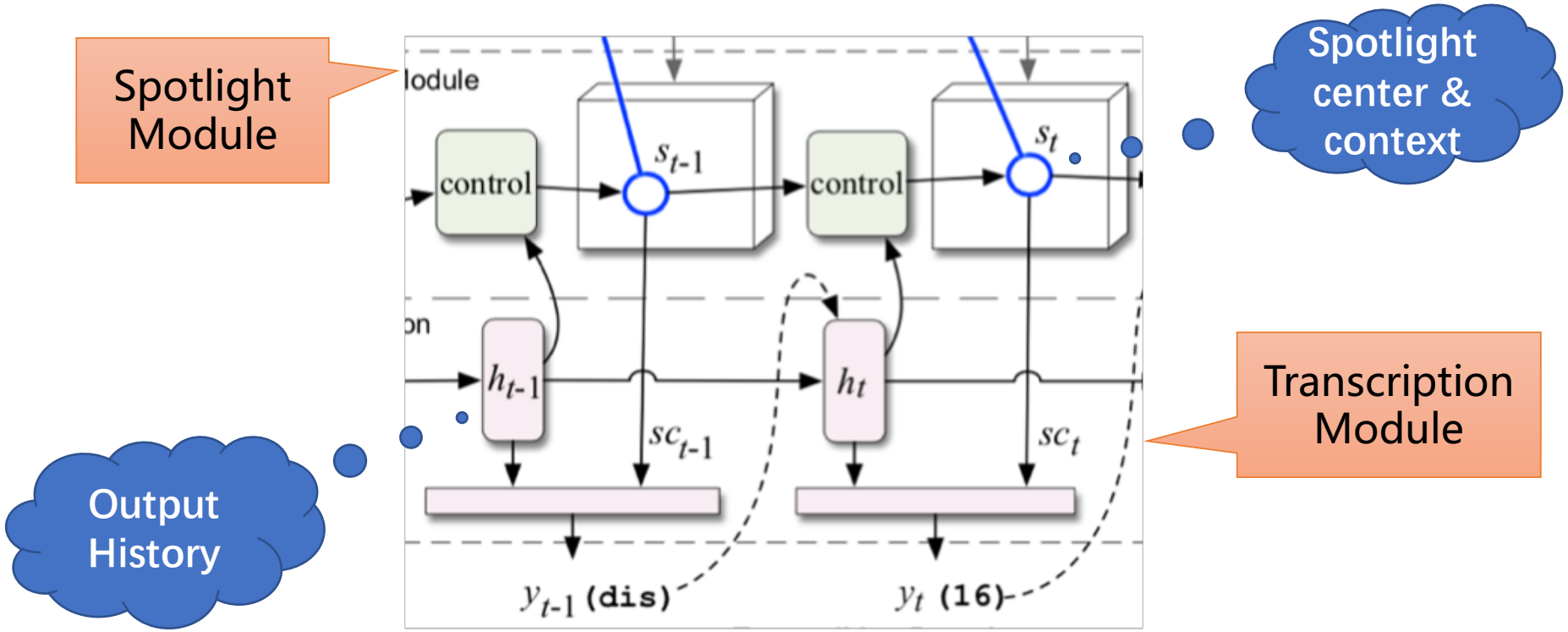
Image



  - CNN: capture high-level semantic information
  - Following ResNet
    - ReLU activation
    - Residual connections
    - Batch normalization
  - Result: $V^{(i,j)}$, each vector at (i,j) represents local semantic information

Transcribing Decoder

- Transcribing decoder:
  - Generating one token at a time: $P(y_t|y_1,\ldots,y_{t-1},V) = \mathrm{Softmax}(d(h_t \oplus sc_t \oplus s_t;\theta_d)),$



Spotlight Mechanism

- Goal: **focus** on certain area when generating, get spotlight context $\underline{sc_t}$

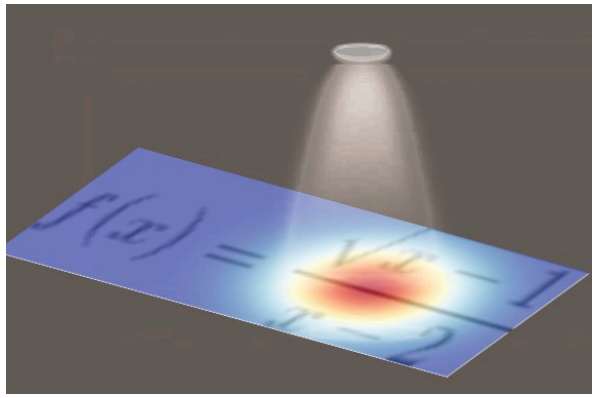
- Approach: shed a spotlight around a spotlight center
  - Spotlight center: *Spotlight handle* (s_t)
    - representing the spotlight: center position and width

$$s_t = (x_t, y_t, \sigma_t)^{\mathrm{T}}$$

  - Assign feature weights following Gaussian distribution:

$$\alpha_t^{(i,j)} \sim \mathcal{N}((i,j)^{\mathrm{T}} | \mu_t, \Sigma_t),$$

$$\mu_t = (x_t, y_t)^T \quad \Sigma_t = \begin{bmatrix} \sigma_t & 0 \\ 0 & \sigma_t \end{bmatrix}.$$

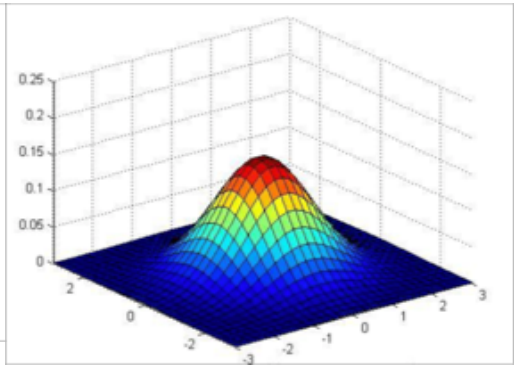


Spotlight Mechanism (cont.)

- Weight calculation:
  - Make it differentiable:

$$\alpha_t^{(i,j)} = \mathrm{Softmax}(b_t) = \frac{\exp(b_t^{(i,j)})}{\sum_{u=1}^{W'} \sum_{v=1}^{H'} \exp(b_t^{(u,v)})},$$

$$b_t^{(i,j)} = -\frac{(i - x_t)^2 + (j - y_t)^2}{\sigma_t^2},$$

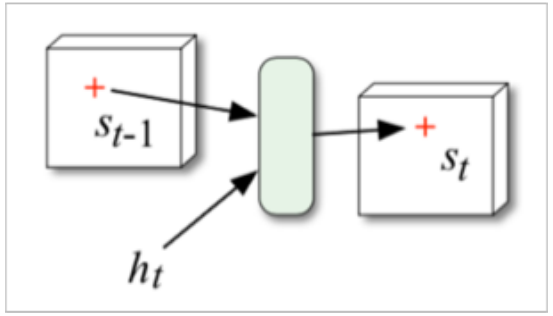


  - Make it parallel:

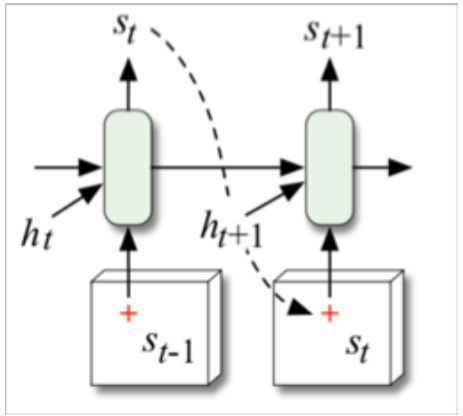
$$\left[\left(\underbrace{\begin{bmatrix} 1 & 2 & \cdots & W' \\ 1 & 2 & \cdots & W' \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 2 & \cdots & W' \end{bmatrix}}_{I} - \underbrace{\begin{bmatrix} x_t & x_t & \cdots & x_t \\ x_t & x_t & \cdots & x_t \\ \cdots & \cdots & \cdots & \cdots \\ x_t & x_t & \cdots & x_t \end{bmatrix}}_{X_t}\right)^2 + \left(\underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2 & 2 & \cdots & 2 \\ \cdots & \cdots & \cdots & \cdots \\ H' & H' & \cdots & H' \end{bmatrix}}_{J} - \underbrace{\begin{bmatrix} y_t & y_t & \cdots & y_t \\ y_t & y_t & \cdots & y_t \\ \cdots & \cdots & \cdots & \cdots \\ y_t & y_t & \cdots & y_t \end{bmatrix}}_{Y_t}\right)^2\right] / \sigma_t^2$$

Spotlight Control

- Controlling spotlight: moving spotlight handle s_t
  - STNM: s_t only depends on s_{t-1}

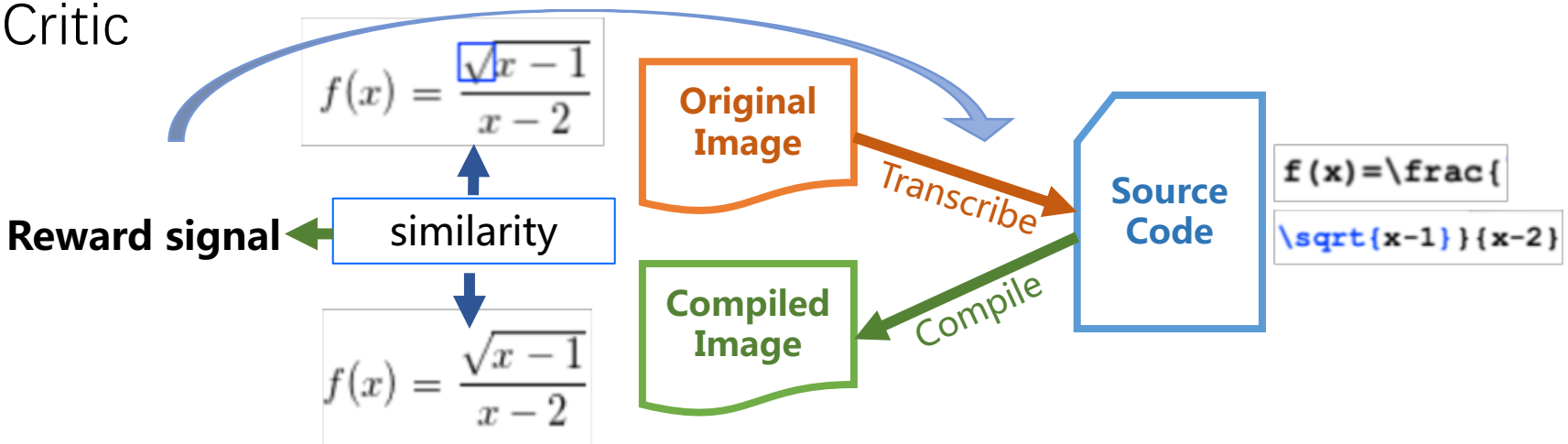


  - STNR: s_t are modeled as a sequence by another GRU



Training and Refining STN

- Training: standard backpropagation
- Refine **spotlight control** for a better reading path
- Reinforcement learning setup:
  - State: output history, spotlight context and handle: $h_t \oplus sc_t \oplus s_t$
  - Action: output generation at each step
  - Reward: *reconstruction similarity* (with grammar constraints)
  - Algorithm: Actor-Critic



Outline

- Introduction
- Preliminaries
- Spotlight Transcribing Network
- **Experiments**
- Conclusion

Experiments

- Data description:

Dataset	Image count	Token space	Token count	Avg. tokens per image	Avg. image pixels
Melody	4208	70	82,834	19.7	15,602.7
Formula	61649	127	607,061	9.7	1,190.7
Multi-Line	4595	127	182,112	39.8	9,016.6

  - Melody and Formula are real-world structural image datasets
  - Multi-line specifically for demonstrating complex structure

- Data partitioned into 60%/40%, 70%/30%, 80%/20%, 90%/10% as training/testing sets
- Available on my website

Experiments

- Baselines:
  - **Enc-Dec**: plain encoder-decoder model
  - **Attn-Dot**: encoder-decoder architecture with attention mechanism that calculate weights by inner production
  - **Attn-FC**: soft visual attention
  - **Attn-Pos**: attention with position embedding

- Metrics: transcription accuracy

Experimental Results

- Accuracy:
  - STN has better performance
  - STN is able to better capture structural information
  - STN better fits images with even more complex structure

(a) **Melody**

Baseline	Testing set percentage			
	40%	30%	20%	10%
EncDec	0.266	0.272	0.277	0.282
AttnDot	0.524	0.548	0.580	0.617
AttnFC	0.683	0.710	0.730	0.756
AttnPos	0.725	0.736	0.741	0.758
STNM	0.729	0.733	0.749	0.759
STNR	**0.738**	**0.748**	**0.758**	**0.767**

(b) **Formula**

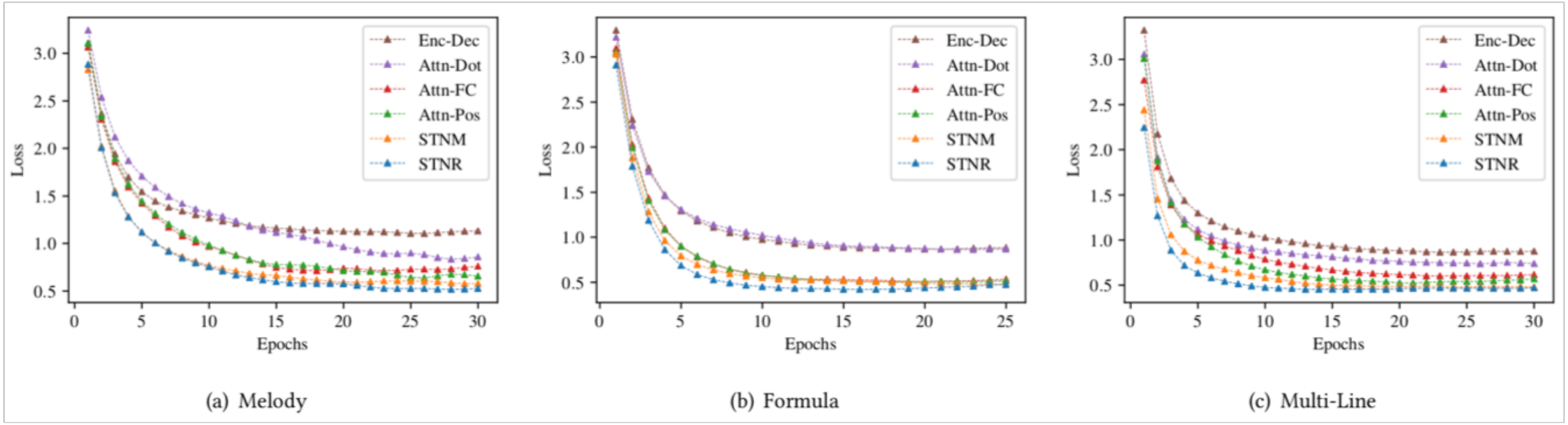
Baseline	Testing set percentage			
	40%	30%	20%	10%
EncDec	0.405	0.427	0.445	0.451
AttnDot	0.530	0.563	0.600	0.611
AttnFC	0.657	0.701	0.717	0.725
AttnPos	0.716	0.723	0.732	0.741
STNM	0.717	0.726	0.740	0.749
STNR	**0.739**	**0.751**	**0.759**	**0.778**

(c) **Multi-Line**

Baseline	Testing set percentage			
	40%	30%	20%	10%
EncDec	0.218	0.227	0.251	0.267
AttnDot	0.334	0.447	0.554	0.599
AttnFC	0.614	0.642	0.686	0.707
AttnPos	0.624	0.652	0.698	0.720
STNM	0.674	0.705	0.731	0.734
STNR	**0.712**	**0.736**	**0.754**	**0.760**

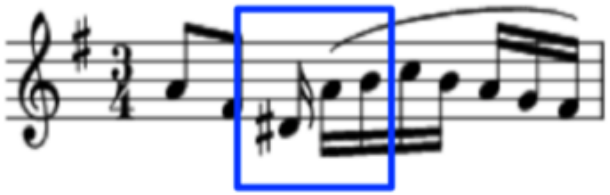
Experimental Results

- Loss:
  - Better performance
  - Fast convergence



(a) Melody (b) Formula (c) Multi-Line

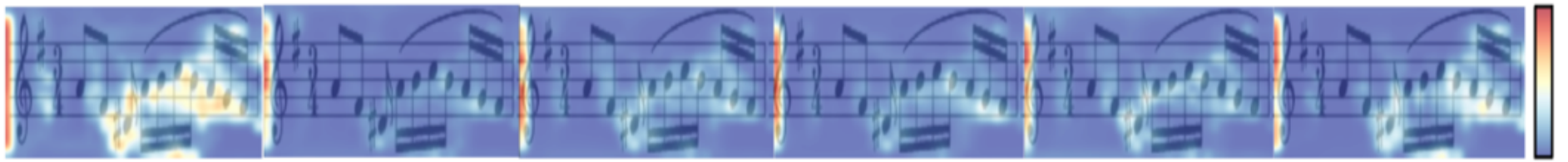
Experimental Results



Ground Truth:

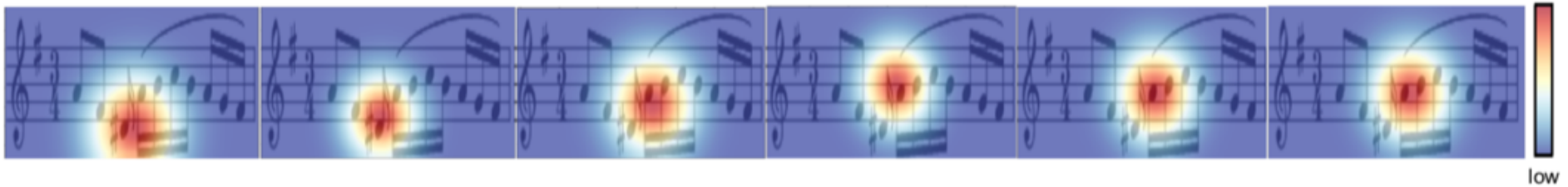
...fis] **dis16 a[b** c b]...

Attn-Pos



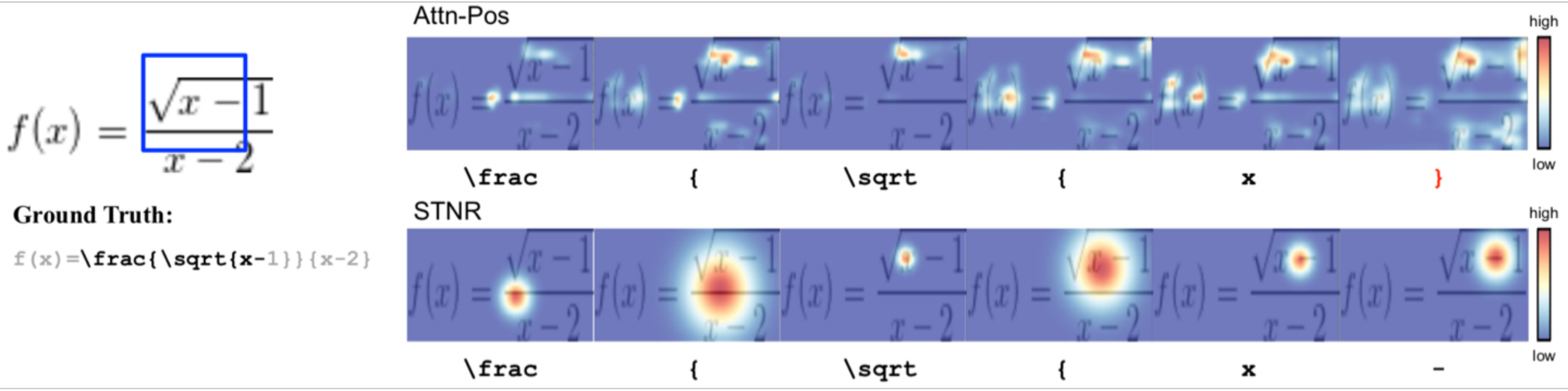
dis 16 b ` [c

STNR



dis 16 a ` [b

Experimental Results



Outline

- Introduction
- Preliminaries
- Spotlight Transcribing Network
- Experiments
- **Conclusion**

Conclusion

- Conclusion
  - Better performance
  - Fast convergence
  - Reasonable reading path: more explainable!

- Future directions:
  - Prior knowledge of the image or the specific language
  - Apply to tasks such as scene text recognition and image captioning
  - Further decouple reading and writing process

Thank you for listening!