Transcribing Content from Structural Images with Spotlight Mechanism

Yu Yin[1] Zhenya Huang[1] Enhong Chen[1]* Qi Liu[1] Fuzheng Zhang[2] Xie Xing[2] Guoping Hu[3]

[1]Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China,
{yxonic,huangzhy}@mail.ustc.edu.cn, {cheneh,qiliuql}@ustc.edu.cn;
[2]Microsoft Research Asia, {fuzzhang,xing.xie}@microsoft.com; [3]iFLYTEK Research, gphu@iflytek.com

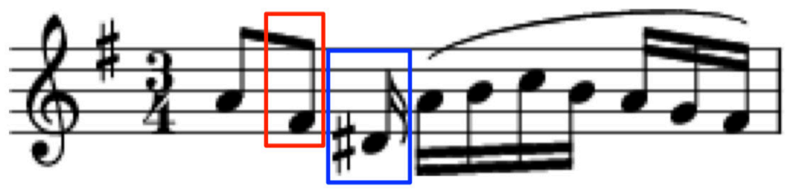
Abstract

Transcribing content from structural images is a challenging task as not only the content objects should be recognized, but the internal structure should also be preserved. In our work, we propose a hierarchical Spotlight Transcribing Network (STN) framework followed by a two-stage "where-to-what" solution. We first decide "where-to-look" through a novel spotlight mechanism to focus on different areas of the original image following its structure. Then, we decide "what-to-write" by developing a GRU based network with the spotlight areas for transcribing the content accordingly.

Introduction

Transcribing Content from Images

- OCR
- Scene text recognition

Straightforward content

(a) Music score example

Structural Images

Previous works ignore large proportion of structural images, where the content objects are **well-formed** in **complex manners**, e.g., music scores (Figure (a)) and formulas (Figure (b)).

Challenges

- Content objects usually follow a fine-grained grammar, and are organized in a complex manner

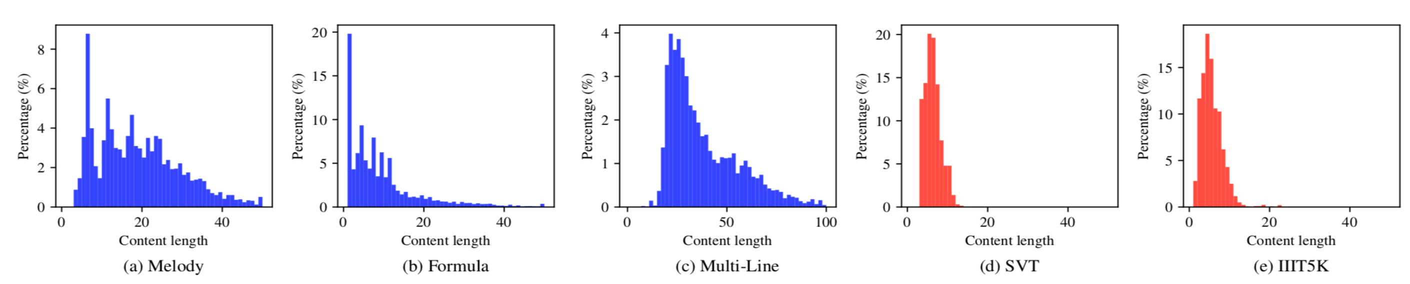
$$f(x) = \frac{\boxed{\sqrt{}}x-1}{x-2}$$

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

(b) Formula example

- Content objects in structural images, even if they just take a small proportion, may carry much semantics
- There exist plenty of similar objects puzzling the transcribing task

Preliminaries



Structural images: printed graphics that organized in a complex structure.

Characteristics:

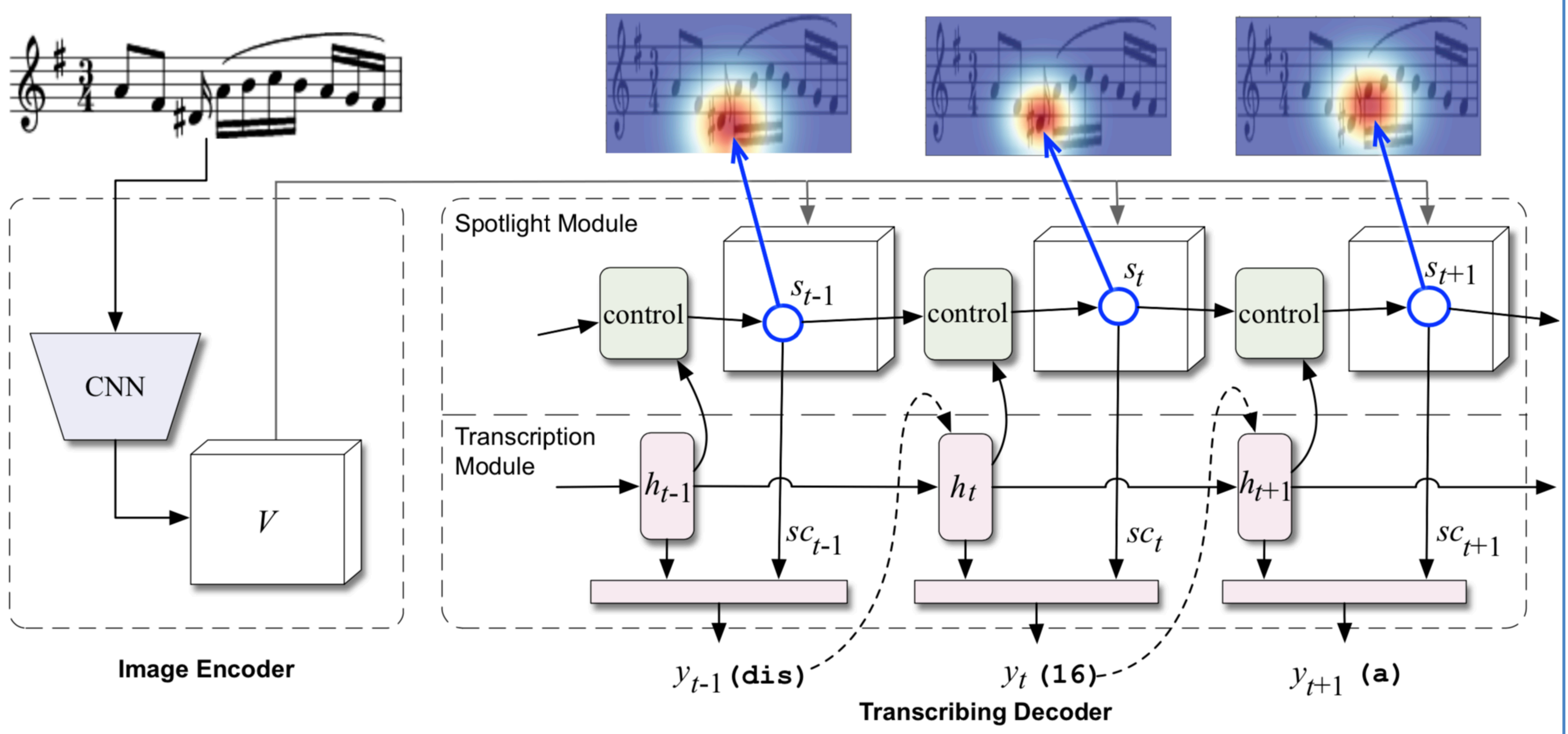
- Much semantics
- Larger output space
- Reversible

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

Definition 3.1. (**Structural Image Transcription Problem**). Given a structural $W \times H$ image x, our goal is to transcribe the content from it as a sequence $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\}$ as close as possible to the source code sequence y, where each $\hat{y}_t$ is the predicted token taking from the specific language corresponding to the image.

Spotlight Transcribing Network (STN)



The overall architecture of Spotlighted Transcribing Network (STN) consists of two main components:

1. **Image encoder**: a CNN based feature extractor;
2. A hierarchical **transcribing decoder**:
   - **Spotlight Module**: find out "*where-to-look*";
   - **Transcription Module**: generates the token sequence.

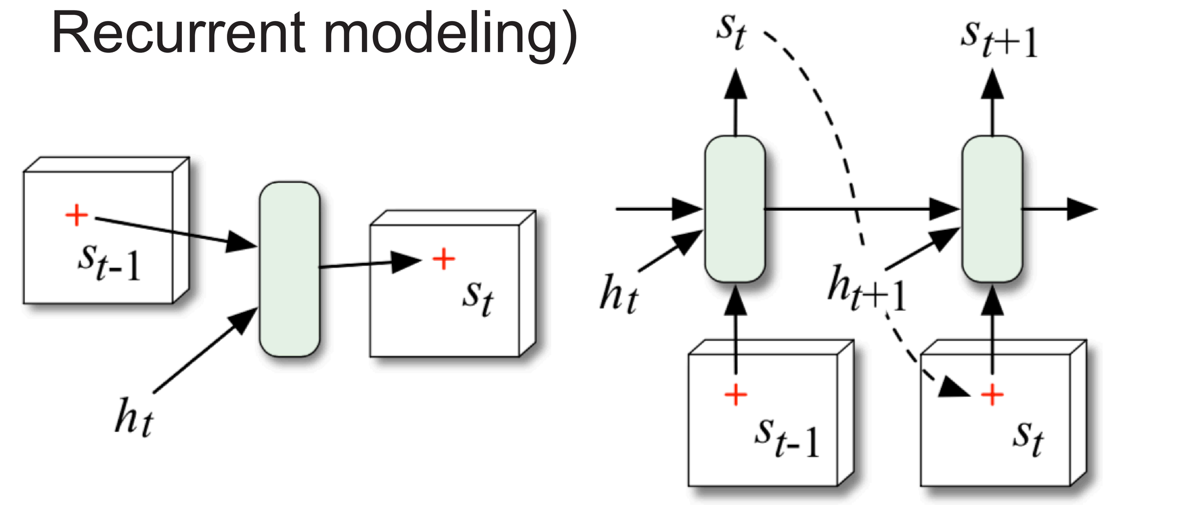
Spotlight Mechanism

Given spotlight handle $s_t = (x_t, y_t, \sigma_t)^{\mathrm{T}}$, assign weights to encoded vectors following *Gassian distribution*.

Spotlight Control

We provide two control modules:

- **Markovian** control module (as in *STNM* with Markov property)
- **Recurrent** control module (as in *STNR* with Recurrent modeling)



(a) Markovian control module. (b) Recurrent control module.

Training and Refining STN

The model is trained by standard backpropagation, with reinforcement learning as a refinement:

- State: the internal states in STN framework;
- Action: token generation;
- Reward: reconstruction similarity between original image and compiled image.

Experiments

Transcribing performance

Outperforms traditional attention based methods.

(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**

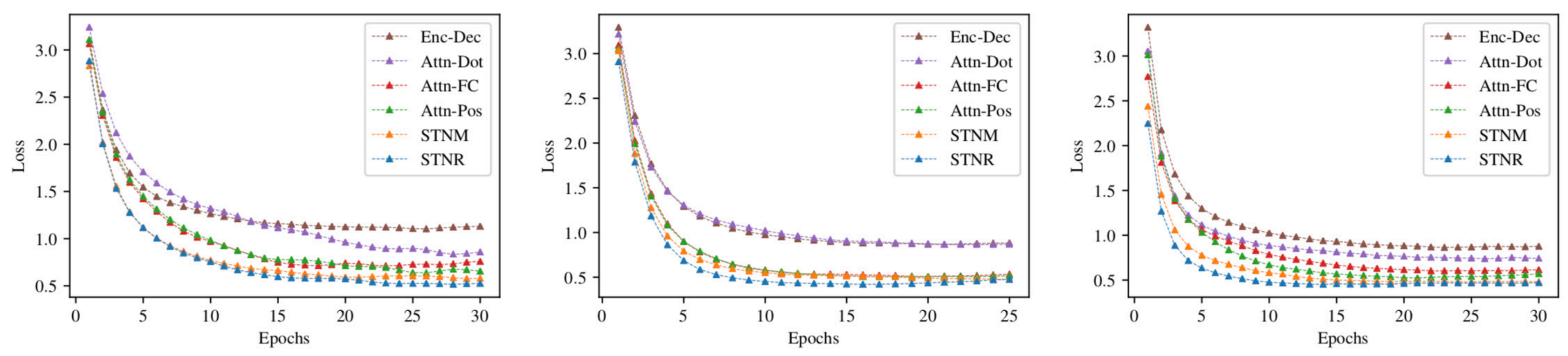
Baseline	Testing set percentage 40%	30%	20%	10%
EncDec	0.405	0.427	0.445	0.451
AttnDot	0.530	0.563	0.600	0.611
AttnFC	0.657	0.701	0.717	0.725
AttnPos	0.716	0.723	0.732	0.741
STNM	0.717	0.726	0.740	0.749
STNR	**0.739**	**0.751**	**0.759**	**0.778**

(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**

Validation loss

Converges faster and achieves lower validation loss.



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

Spotlight visualization

- STNR finds a more reasonable reading path;
- STNR clearly distinguishes similar regions properly.

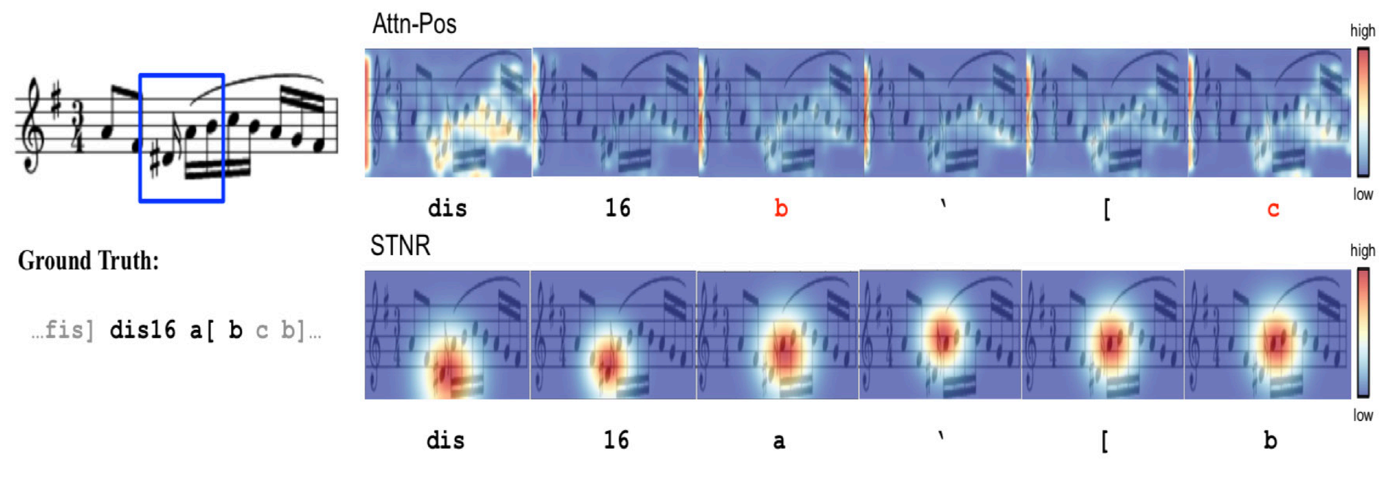


Figure 7: Comparison between attention and spotlight mechanism on Melody dataset.

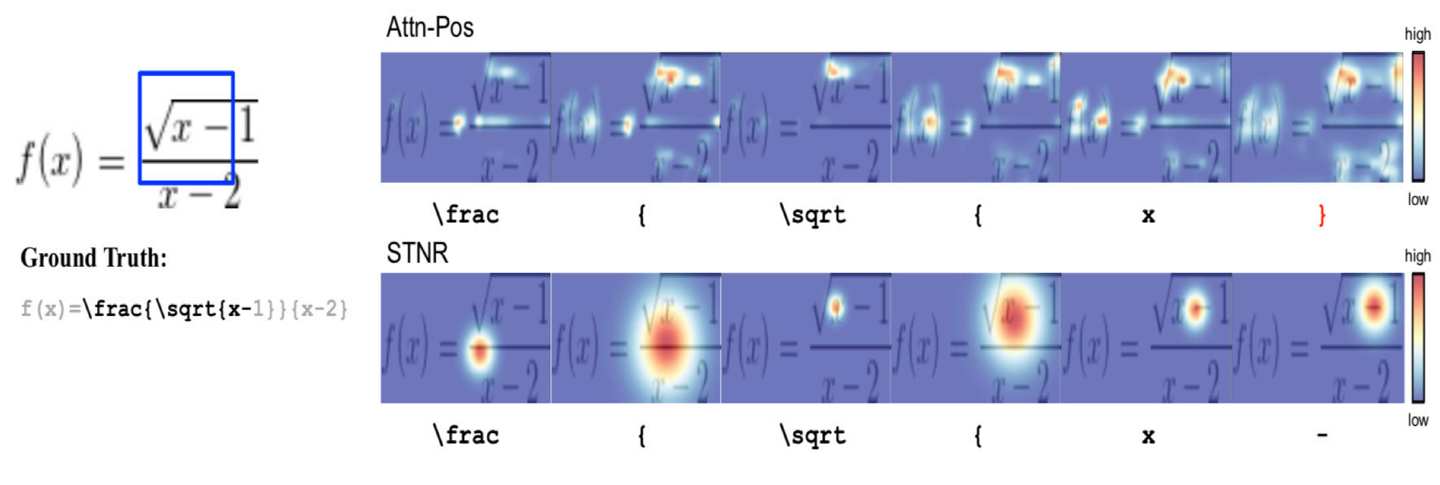


Figure 8: Comparison between attention and spotlight mechanism on Formula dataset.