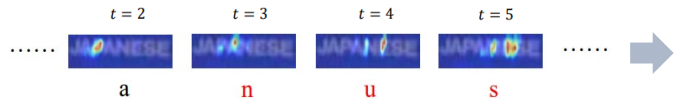
Gaussian Constrained Attention Network for Scene Text Recognition

Zhi Qiao[1], Xugong Qin[1], Yu Zhou[1*], Fei Yang[2], Weiping Wang[1]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]TAL Education Group *Corresponding Author

Motivation

- Attention operation is inspired from Neural Machine Translation and Image Caption
- The attention alignments in text recognition are concentrated and like a **Gaussian distribution**
- Existing methods do not fully use this characteristic and may suffer from the problem of **attention diffusion**

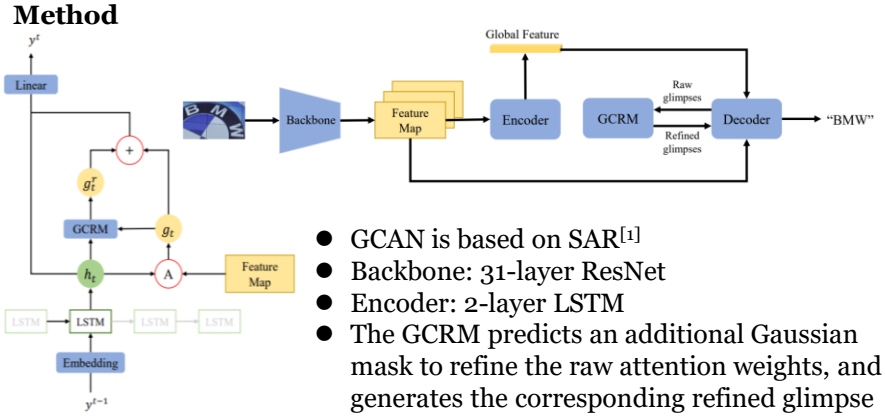


Attention Diffusion is the problem that attention weights are not concentrated, which may bring noise features into the decoding

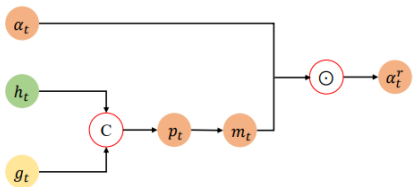
Introduction

- We propose a novel **Gaussian Constrained Refinement Module** (GCRM) to refine the raw attention weights
- We apply GCRM into SAR[1], and propose our **Gaussian Constrained Attention Network** (GCAN)

Method



- GCAN is based on SAR[1]
- Backbone: 31-layer ResNet
- Encoder: 2-layer LSTM
- The GCRM predicts an additional Gaussian mask to refine the raw attention weights, and generates the corresponding refined glimpse



- Pipeline of GCRM:
  - Predicting Gaussian parameters with the hidden state h_t and raw glimpse g_t
  - Constructing the Gaussian Mask with the predicted parameters p_t
  - Applying Gaussian mask m_t to the raw attention weights α_t with multiplication

Experiment

- Ablation Study

Methods	Training	Inference
SAR-reproduced	67.9ms	45.4ms
GCAN	75.5ms	54.3ms

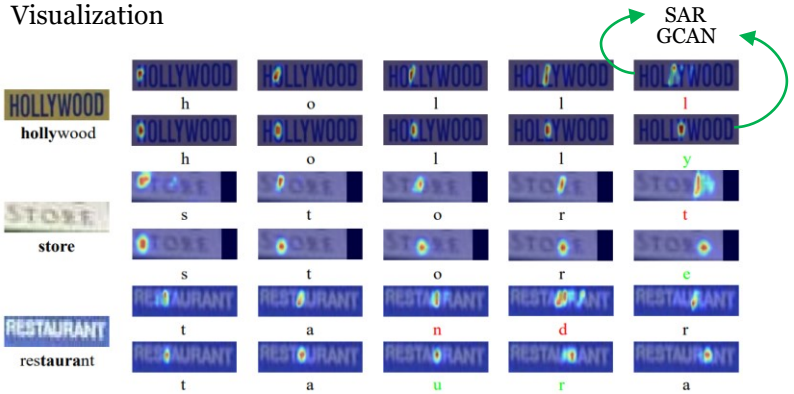
➢ GCAN consumes less than **10ms** more compared with SAR

Methods	L_{att}	IIIT5K	SVT	SVTP	IC15
SAR-reproduced		93.0	86.7	77.0	76.1
SAR-reproduced	✓	93.1	87.3	78.8	75.4
with Estimation		93.4	88.1	78.4	75.6
with Estimation	✓	93.8	88.4	78.8	**77.5**
with GCRM		93.6	86.9	79.1	77.0
with GCRM	✓	**94.4**	**90.1**	**81.2**	77.1

➢ GCRM improves the performance with or without character-supervision

➢ Estimation represents to estimate the Gaussian parameters without GCRM

- Visualization

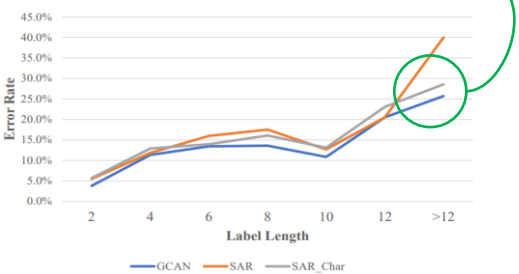


➢ First line of each image is the visualized attention weights of SAR

➢ Second line is the visualized attention weights of GCAN

- Performance about Different Text Length

GCAN is more robust with long text



- Compared with previous methods

Methods	IIIT5K	SVT	IC13	IC15	SVTP	CUTE
Shi *et al.* [13]	81.2	82.7	89.6	-	-	-
Shi *et al.* [50]	81.9	81.9	88.6	-	71.8	59.2
Lee *et al.* [49]	78.4	80.7	90.0	-	-	-
Yang *et al.* [36]	-	-	-	-	75.8	69.3
Cheng *et al.* [37]	87.4	85.9	93.3	70.6	-	-
Cheng *et al.* [52]	87.0	82.8	-	68.2	73.0	76.8
Liu *et al.* [57]	92.0	85.5	91.1	74.2	78.9	-
Bai *et al.* [65]	88.3	87.5	**94.4**	73.9	-	-
Liu *et al.* [66]	87.0	-	92.9	-	-	-
Liu *et al.* [67]	89.4	87.1	94.0	-	73.9	62.5
Shi *et al.* [18]	93.4	89.5	91.8	76.1	78.5	79.5
Liao *et al.* [53]	91.9	86.4	91.5	-	-	79.9
Zhan *et al.* [56]	93.3	**90.2**	91.3	76.9	79.6	83.3
Xie *et al.* [60]	-	-	-	68.9	70.1	82.6
Li *et al.* [26]	91.5	84.5	91.0	69.2	76.4	83.3
Luo *et al.* [25]	91.2	88.3	92.4	74.7	76.1	77.4
Yang *et al.* [54]	**94.4**	88.9	93.9	**78.7**	80.8	**87.5**
Wang *et al.* [58]	94.3	89.2	93.9	74.5	80.0	84.4
SAR-reproduced	93.0	86.7	90.8	76.1	77.0	83.7
GCAN (Ours)	**94.4**	90.1	93.3	77.1	**81.2**	85.6

➢ GCAN achieves best performance on 2 benchmarks

[1] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," in AAAI, 2019, pp. 8610–8617