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

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

TAL 好未来

Outlines

- **Motivation**
- Method
- Experiments
- Conclusion

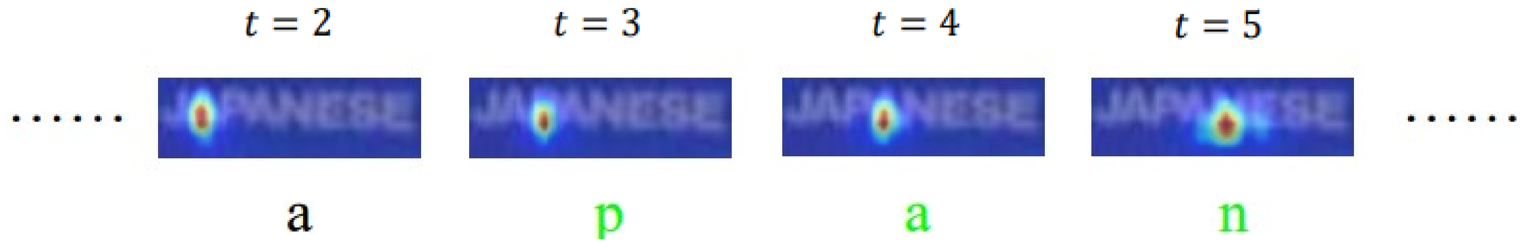
Motivation

- **Attention Mechanism is the Mainstream Method in Scene Text Recognition**
  - ➢ The model predicts corresponding alignments for every characters
  - ➢ Inspired from Neural Machine Translation (NMT) and Image Caption (IC)

- **Existing Methods does not Fully Use the Characteristic of Text Recognition**
  - ➢ Different from NMT and IC, the attention weights in text recognition is concentrated
  - ➢ The attention weights seems like a Gaussian distribution
  - ➢ Existing methods do not modify the attention operation for text recognition specifically

Motivation

- **Existing Attention Operation may Lead to Attention Diffusion**
  - Attention diffusion is the problem that the attention weights are not concentrated
  - The noise features are introduced into the decoding and lead to wrong predictions

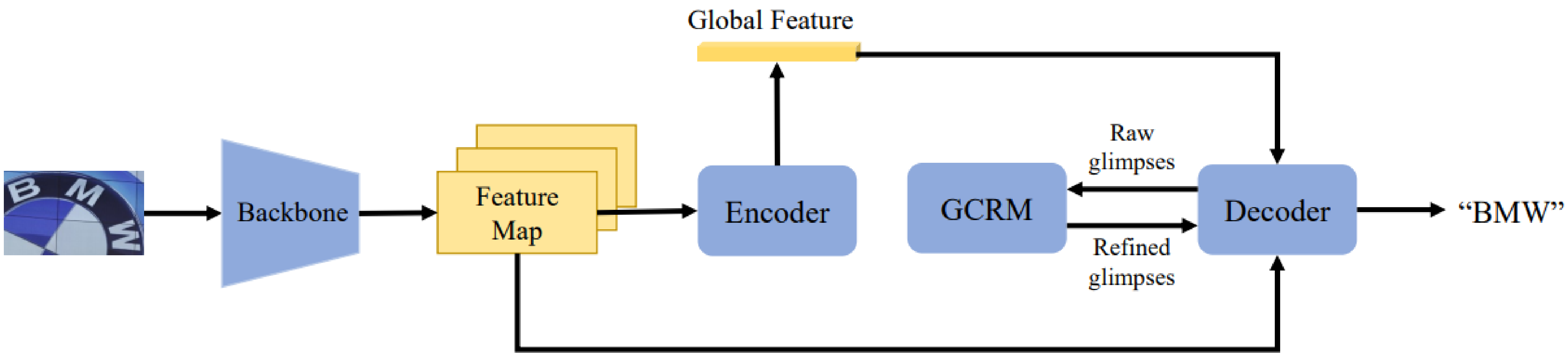


- **Can We Introduce Gaussian Distribution into the Attention Operation?**
  - In this paper, we propose a Gaussian Constrained Refinement Module
  - With refinement the attention weights become more concentrated and accurate

Outlines

- Motivation
- **Method**
- Experiments
- Conclusion

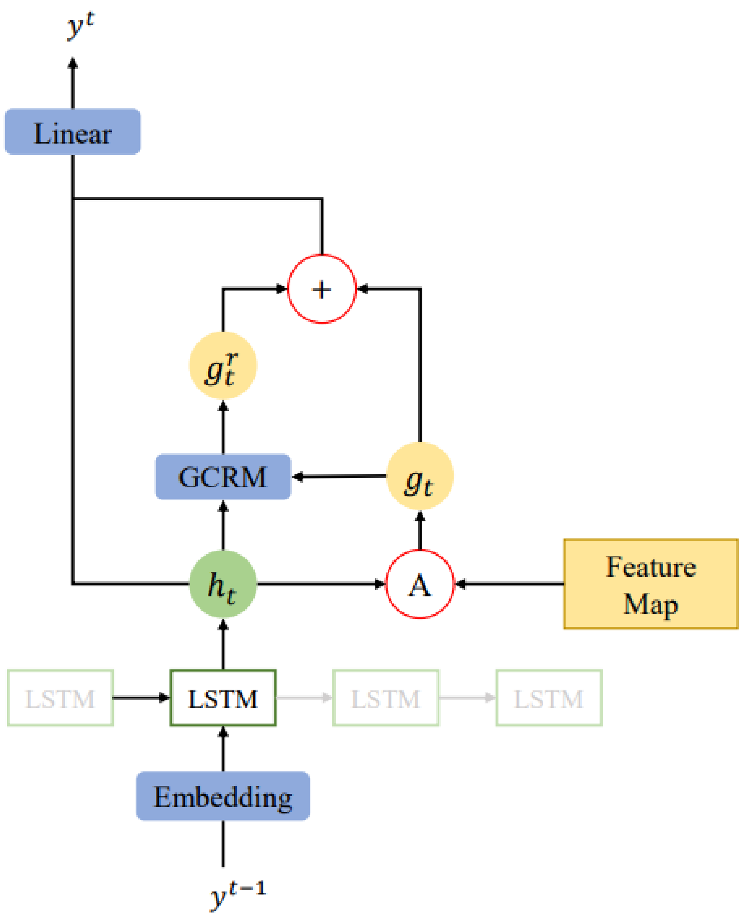
Method: Framework



- Our Gaussian Constrained Attention Network is based on SAR[1]
- The proposed GCRM refines the raw attention weight to make it more concentrated and accurate
- The Backbone adopts a 31-layer ResNet
- The Encoder is a 2-layer LSTM

[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

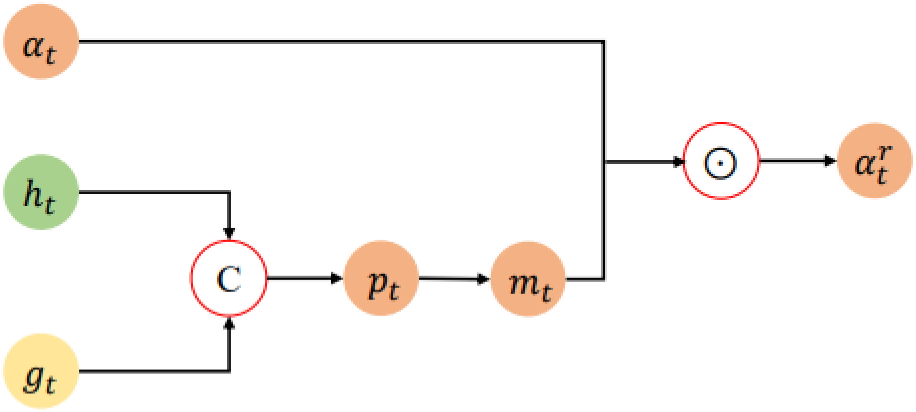
Method: Decoder



- Annotation
  - y^t: the predicted character at time step t
  - h_t: the hidden state of LSTM at time step t
  - A: the attention operation
  - g_t: the glimpse calculated attention weight and feature map
  - $+$: Element-wise sum

- The attention-based decoder is combined with proposed GCRM
- In each iteration, the attention weight is calculated between feature map and the hidden state h_t, then the glimpse g_t is generated
- GCRM tries to refine the raw attention weight with the h_t and g_t, then generates the glimpse g_t^r with the refined attention weight
- The final prediction is predicted based on h_t, g_t and g_t^r

Method: GCRM



- Annotation
  - α_t: the raw attention weight at time step t
  - h_t: the hidden state of LSTM at time step t
  - C: the concatenate operation
  - g_t: the raw glimpse at time step t
  - $+$: element-wise sum
  - p_t: the predicted Gaussian parameters at time step t
  - m_t: the constructed Gaussian mask
  - $\odot$: element-wise multiplication
  - α_t^r:the attention weight after refined at time step t

- GCRM predicts an additional Gaussian mask to refine the raw attention weight α_t
- Pipeline of the GCRM:
  a) The Gaussian parameters p_t are first predicted by hidden state h_t and raw glimpse g_t
  b) The Gaussian mask m_t is constructed with the predicted parameters p_t
  c) The Gaussian mask m_t is applied to refine the raw attention weight α_t with element-wise multiplication

Outlines

- Motivation
- Method
- **Experiments**
- Conclusion

Experiments

Performance

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	<u>94.0</u>	-	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**	<u>80.8</u>	**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**	<u>90.1</u>	93.3	<u>77.1</u>	**81.2**	<u>85.6</u>

- GCAN achieves best performance on 2 datasets and second best performance on 3 datasets
- GCAN is much better than our baseline SAR

Experiments

Ablation Study

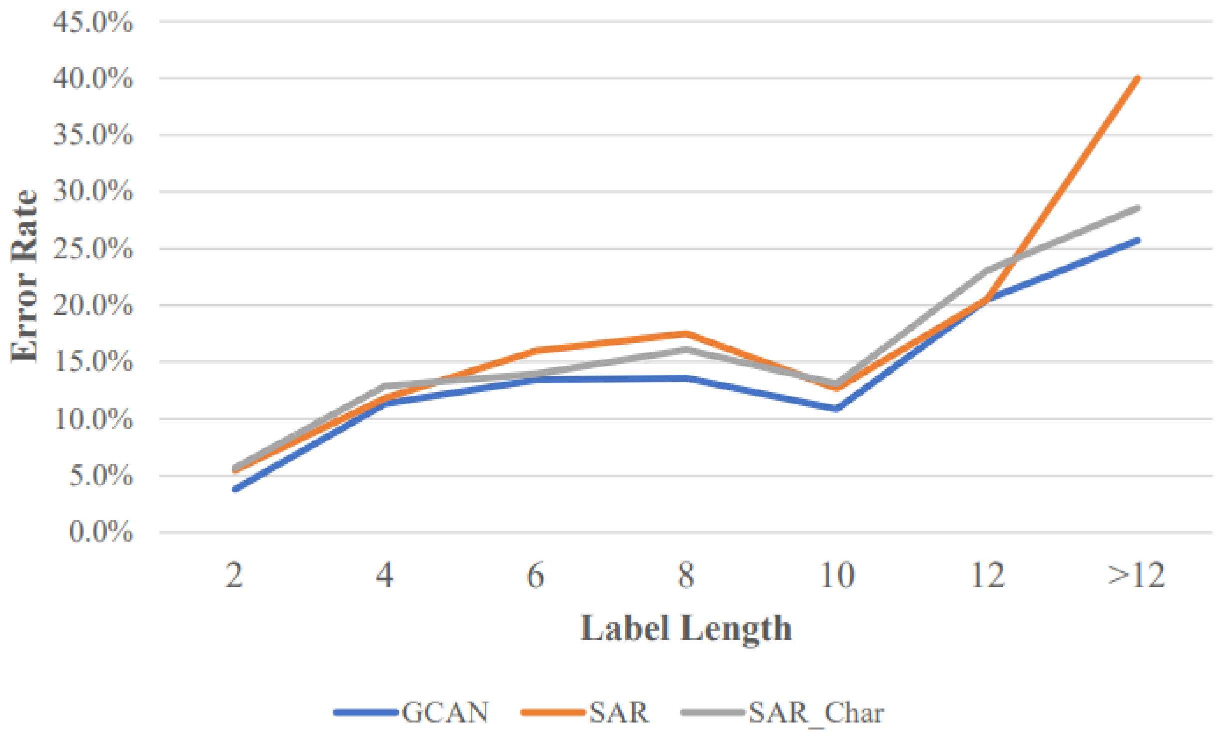
Methods	L_{att}	IIIT5K	SVT	SVTP	IC15
SAR-reproduced		93.0	86.7	77.0	76.1
SAR-reproduced	✓	93.1	87.3	78.8	75.4
with Estimation		93.4	88.1	78.4	75.6
with Estimation	✓	93.8	88.4	78.8	**77.5**
with GCRM		93.6	86.9	79.1	77.0
with GCRM	✓	**94.4**	**90.1**	**81.2**	77.1

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

- GCRM brings significant improvements with or without the character-level supervision
- *Estimation* represents to estimate the Gaussian parameters instead of predicted by GCRM
- GCAN consumes less than **10ms** more during both training and inference compared with SAR

Experiments

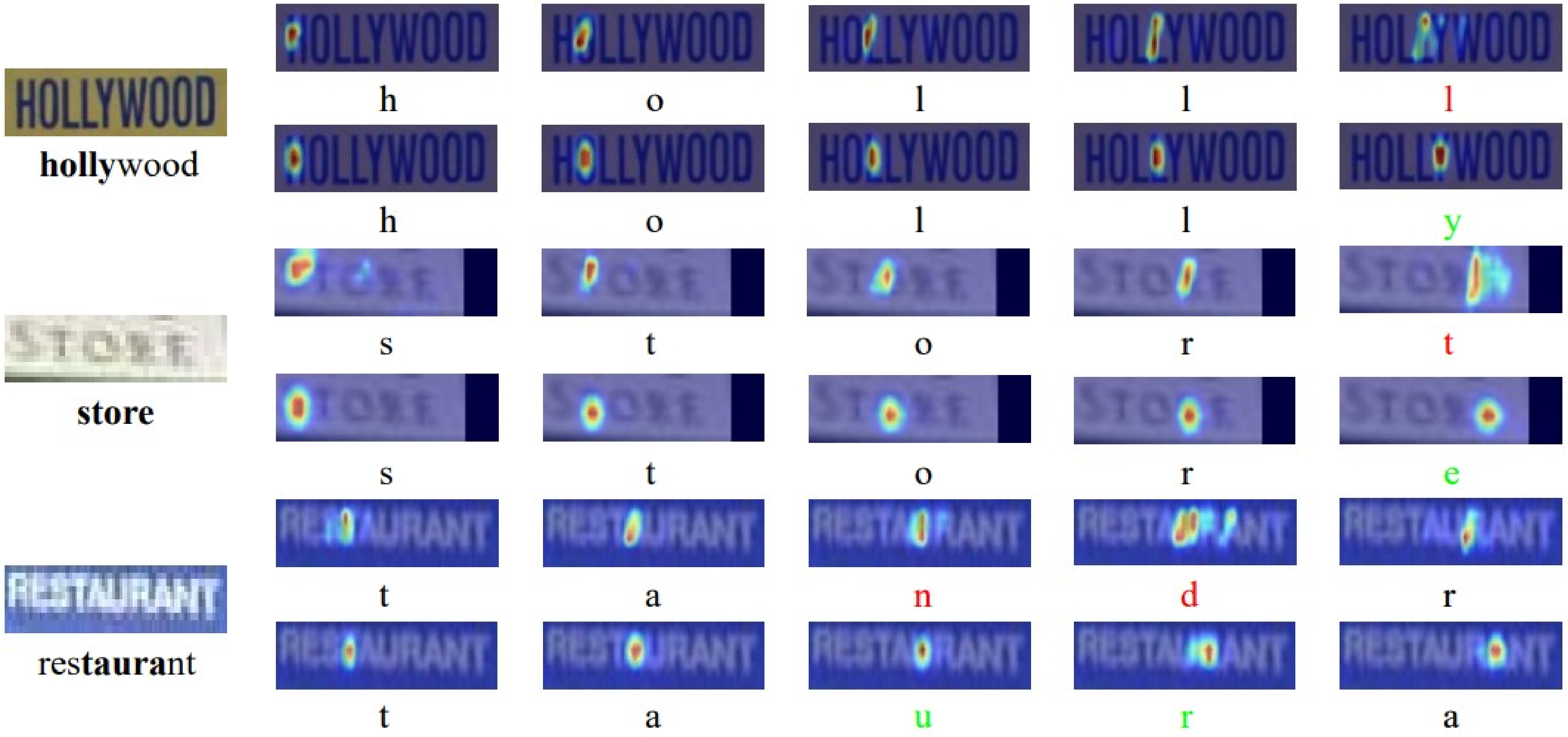
Performance Comparison of Different Text Length



- The error rate on different length of text
- GCAN is more robust with longer text

Experiments

Visualization



- ➢ The first line of each image is the attention weights of SAR, and the second line is from our GCAN
- ➢ With the proposed GCRM, the problem of attention diffusion is alleviated significantly

Outlines

- Motivation
- Method
- Experiments
- **Conclusion**

Conclusion

- In this paper we introduce the problem of attention diffusion
- We propose a novel Gaussian Constrained Refinement Module (GCRM)to deal with attention diffusion
- GCRM is flexible and can be applied into the most existing attention-based methods
- Combining GCRM and SAR, our Gaussian Constrained Attention Network achieves convincing performance
- In future, the specifical attention operation for text recognition is worth exploring

Gaussian Constrained Attention Network for Scene Text Recognition

Thanks for Your Watching !

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

TAL 好未来