SCATTER: Selective Context Attentional Scene Text Recognizer
Supplementary Materials

Ron Litman*, Oron Anschel*, Shahar Tsiper, Roee Litman, Shai Mazor and R. Manmatha
Amazon Web Services
{litmanr, oronans, tsiper, rlit, smazor, manmatha}@amazon.com

1. Regular Vs Irregular Text

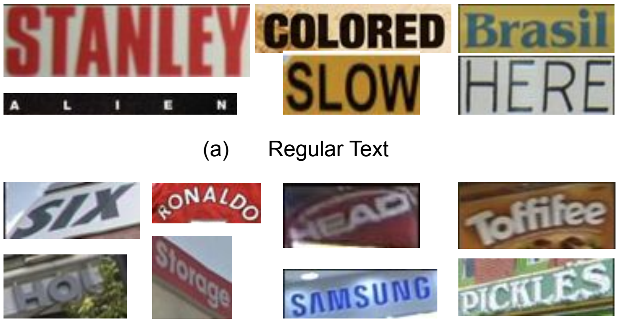
Recent works distinguish between two types of scene-text datasets: *Irregular text* where the text may be arbitrarily shaped (e.g. curved text), and *regular text* where the sequence of characters is nearly horizontally aligned. In Fig. 1 we bring examples that demonstrate the main differences between these two types.

2. Network Pruning - Compute Constraint

Fig. 2 in the main manuscript, shows accuracy levels for all intermediate decoders on several different stacking arrangements for training (e.g., using 1, 3, and 5 blocks). Table 1 shows the exact results of Fig. 2, with the additional results of all stacking arrangements for training, from a single block up to five blocks. The results demonstrate that in general, it is favorable to train a deep network (with more blocks) and then prune, compared to training with a shallow architecture in the first place. For example, if the target architecture inference time should include only 2 BiLSTM layers (similar to [1]). Training a 5-block SCATTER and pruning to a single block (11th row in Table 1) achieves **+0.4%** pp and **+1.3%** pp on regular and irregular text respectively, compared to training a single block (first row of the table) in the first place.

3. Examples of Intermediate Predictions

Following the discussion in Sec. 5.3. of the paper, we provide additional examples of intermediate predictions for both regular and irregular text in Table 2. Table 2 shows that in some cases (the first two rows for each text type) the earlier decoders fail to predict the word in the image, while the final decoder is correct. In other cases (the last two rows for each text type) at least one of the intermediate decoders predicts the correct text, however, the final decoder fails to do so. The described phenomena suggests that one could develop selection, voting or ensemble technique to improve results by choosing the correct prediction out of the available selective-decoders outputs.



(a) Regular Text

(a) Irregular Text

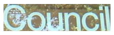
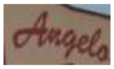
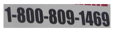
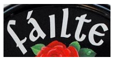
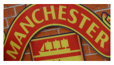
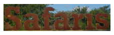
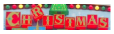
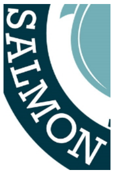
Figure 1: Examples of regular (IIIT5k, SVT, IC03, IC13) and irregular (IC15, SVTP, CUTE) real-world datasets.

Table 1: Average test accuracy at intermediate decoding stages of the network, compared across different training network depths. * Regular Text and Irregular Text columns are weighted (by size) average results on the regular and irregular datasets respectively.

Training Blocks	**N Blocks After Pruning**	**N LSTM Layers After Pruning**	**Regular Text***	**Irregular Text***
1	1	2	93.2	82.7
2	1	2	93.2	82.6
2	2	4	93.6	83.0
3	1	2	93.8	83.2
3	2	4	93.9	83.2
3	3	6	93.7	83.4
4	1	2	93.4	83.5
4	2	4	93.4	83.9
4	3	6	93.6	83.4
4	4	8	93.7	83.5
5	1	2	93.6	84.0
5	2	4	93.7	83.7
5	3	6	93.8	83.6
5	4	8	94.0	84.1
5	5	10	94.0	83.7

*Authors contribute equally.

Table 2: Examples of intermediate decoders predictions on eight different images, from both regular and irregular text datasets. The presented results in the table, suggests that a selection, voting or ensemble technique could be use to improve results

Text Type	Test Image	Intermediate Decoder				Final Decoder	Ground Truth
		1	2	3	4		
Reguler		oouncil	oouncil	council	council	council	council
		angels	angels	angels	angelo	angelo	angelo
		18008091469	18008091469	18008091469	18008091469	180080914669	18008091469
		lailte	failte	failte	lailte	lailte	failte
Irreguler		annchester	winchester	wanchester	banchester	manchester	manchester
		shanthant	safaris	safaris	safaric	safaris	safaris
		ch__stmas	christmas	christinas	christwas	christwas	christmas
		balmon	salmon	salmon	balmon	balmon	salmon

4. Stable Training of a Deep BiLSTM Encoder

Table 3: The effect of the number of BiLSTM layers used on recognition accuracy. Only by using SCATTER we are able to add BiLSTM layer to improve results. Regular Text and Irregular Text columns are weighted (by size) average results on the regular and irregular datasets. We refer to our re-trained model using the code of Baek et al. 2019 as Baseline.

# Blocks	LSTM Layers	Regular Text	Irregular Text
Baseline	1	92.5	79.0
Baseline	2	92.7	79.1
Baseline	3	92.6	78.7
Baseline	4	92.4	78.6
1	2	93.2	82.7
2	4	93.6	83.0
3	6	93.7	83.4
4	8	93.7	83.5
5	10	94.0	83.7

Previous papers used only a 2-layer BiLSTM encoder. In [2] the authors report a decrease in accuracy while increasing the number of layers in the BiLSTM encoder. In Table 3 we show results of a reproduction of the experiment reported in [2], training a baseline architecture with an increasing number of BiLSTM layers in the encoder. We observe a similar phenomena to [2] – a reduction in accuracy when using more than two BiLSTM layers in the baseline architecture. However, the bottom rows of the table demonstrate that SCATTER allows stacking of more BiLSTM layers, which ultimately leads to an increase in final performance.

References

[1] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. *arXiv preprint arXiv:1904.01906*, 2019. 1

[2] Ling-Qun Zuo, Hong-Mei Sun, Qi-Chao Mao, Rong Qi, and Rui-Sheng Jia. Natural scene text recognition based on encoder-decoder framework. *IEEE Access*, 7:62616–62623, 2019. 2