Scenergy: Scene Text Recognition with Conditional Random Fields

Pranav Goyal
2017078

Surabhi S. Nath
2016271

Motivation

- Scene text recognition has been an important and challenging problem for the computer vision community
- Most existing techniques do not exploit the contextual information
- Utilizing top-down and bottom-up cues can enable better recognition
- Probabilistic graphical models can model the relations across text entities in the image
- Mishra et al. used a probabilistic approach to capture character cues to improve recognition of words in an image
- We extend their idea to sentence-level recognition in street signboard images

Related Works

There have been several techniques proposed for scene text recognition in

Detection Methods	Recognition Methods	Auxiliary Methods
1. Anchor based 2. Region proposal 3. Text instance level 4. Bottom up 5. Multi oriented text [1, 2, 3, 4, 5]	1. CTC based 2. Attention based 3. 2 Stage end to end 4. Probabilistic Models [6, 7, 8, 9]	1. Synthetic Data 2. Semi-supervised 3. Deblurring 4. Contextual Information 5. Adversarial [10, 11, 12, 13, 14]

Aim

- To utilize top down spatial lexicon knowledge and bottom up word level cues in a conditional random field framework and perform scene text recognition on street signboard images
- To compare performance on baseline:
  - Against different evaluation metrics, namely word match accuracy, edit distance and phonetic distance
  - Against different graph constructions, namely, linear, dense and weighted dense conditional random fields

Conditional Random Fields

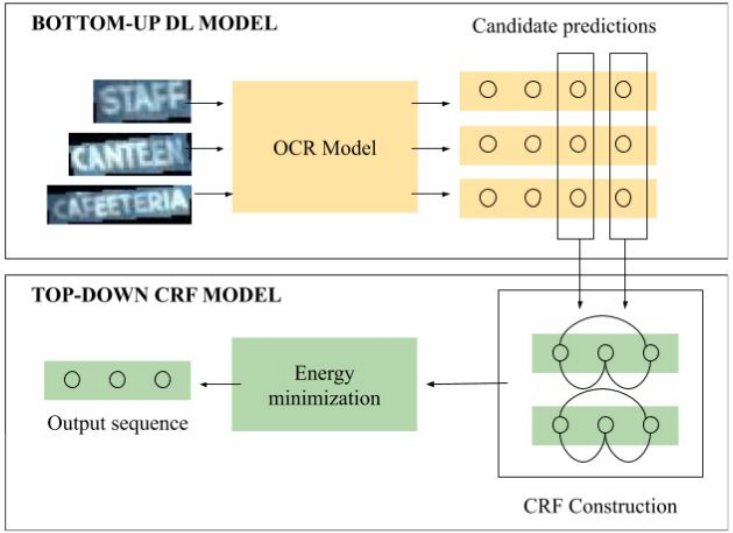
- Conditional random field (CRF) is a type of Markov graphical model
- It is a discriminative, undirected model which does not assume independence
- The output sequence from an input sequence, is attained as follows:

$$\begin{aligned}\hat{y} &= argmax_y p(Y|X) = argmax_y \frac{p(YX)}{p(X)} \\ &= argmax_y p(YX) \\ &= argmax_y \sum_{i=1}^{n} (-E(x_i, y_i) - \sum_j E(y_i, y_j)) \\ &= argmin_y \sum_{i=1}^{n} (E(x_i, y_i) + \sum_j E(y_i, y_j))\end{aligned}$$

where n is the length of the sequence and $E(x_i,y_i)$, $E(y_i,y_j)$ are the unigram, bigram energy terms

Methodology

- The bottom-up DL model and the top-down CRF model word together to output the most probable sequence of word predictions.
- Word crops are obtained from a signboard image and passed into the network



Methodology

1. DL Model

- ASTER, was used as base DL model, which was fine-tuned on our data
- ASTER comprises of 3 stages:
  - Transformation Network for rectification of input image
  - Encoder for feature extraction
  - Decoder for obtaining character combination based on character level probability values
- Instead of the most probable character combination, a set of candidate words along with their confidences are obtained.
- The confidences are converted to energies and passes to the preprocessing unit.

Methodology

2. Preprocessing Unit

- The candidate predictions for each word obtained from the DL model are pruned and reduced to at most 3.
- This is achieved by taking the top 3 predictions with the highest unigram confidence value.
- These candidate predictions are then fed to the CRF Model.

Methodology

3. CRF Construction

- Candidate predictions from the Preprocessing Unit are acquired for every word.
- Combinations of these candidates are formed to obtain potential sequences and the energy for each sequence is calculated.
- The CRF Energy consists of a unigram term, obtained from the DL model, and a bigram term, obtained from the edge connections in the CRF.

$$\sum_{i=1}^{n}(\lambda_u E(x_i, y_i) + \lambda_v \sum_j E(y_i, y_j))$$

- The minimum energy sequence is chosen as the optimal output.

Methodology

- Four different graphical constructions are tested:
  - Pure DL model: λ_b is set to 0.
  - Linear CRF model: The bigram energy term is only considered for adjacent neighbours.
  - Dense CRF model: The bigram term is considered assuming all words y_i are connected to all other words y_j where $i \neq j$.
  - Dense CRF model with positional weighting: The bigram term is considered assuming all words y_i are connected to all other words y_j where $i \neq j$, however, the energies are assigned a weight of $1/d$, where d is the distance between the 2 words in the sequence

Methodology

4. Evaluation Metrics

1. Word Match Accuracy

$$WM = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} II(\hat{y}_{ij}, y_{ij})}{\sum_{i=1}^{m} \sum_{j=1}^{n_i} n_i}$$

2. Sentence level Average Edit Distance

$$AED = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} ED(\hat{y}_{ij}, y_{ij})}{m}$$

3. Sentence level Average Phonetic Distance

$$APD = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} PD(\hat{y}_{ij}, y_{ij})}{m}$$

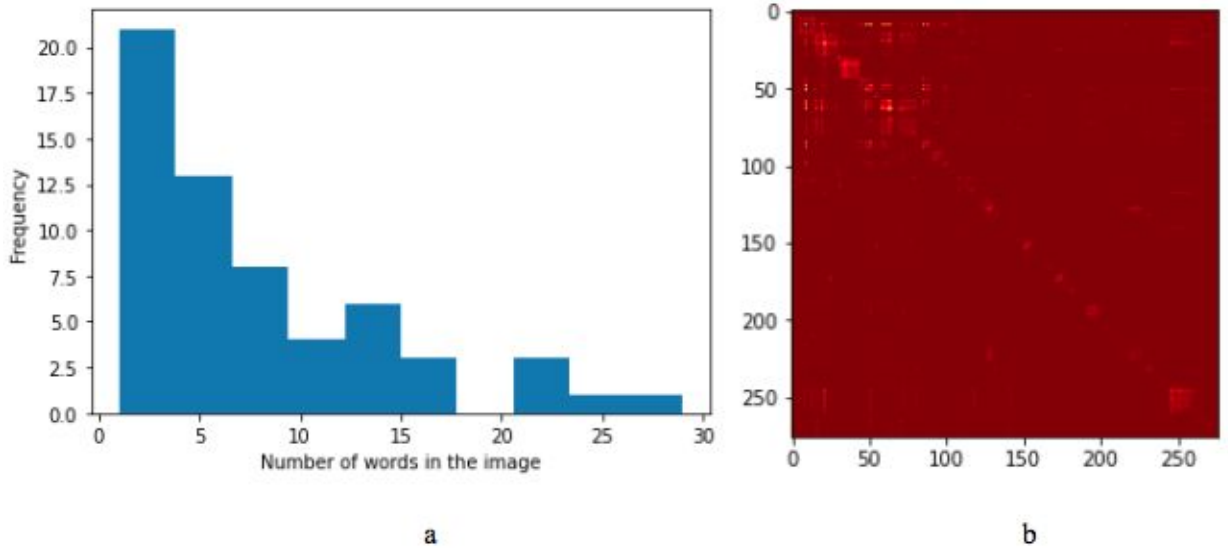
Dataset

- The dataset consists of 1300 street signboard images
- Each signboard is associated with coordinates of words and text.
- We extracted 9044 English word cropped images
- The bigram probabilities represented the probability of co-occurrence 2 words in the same image.
- The bigram value for (a, b) is obtained by calculating the number of signboards with both a and b, divided by the total number of signboard images.

Dataset

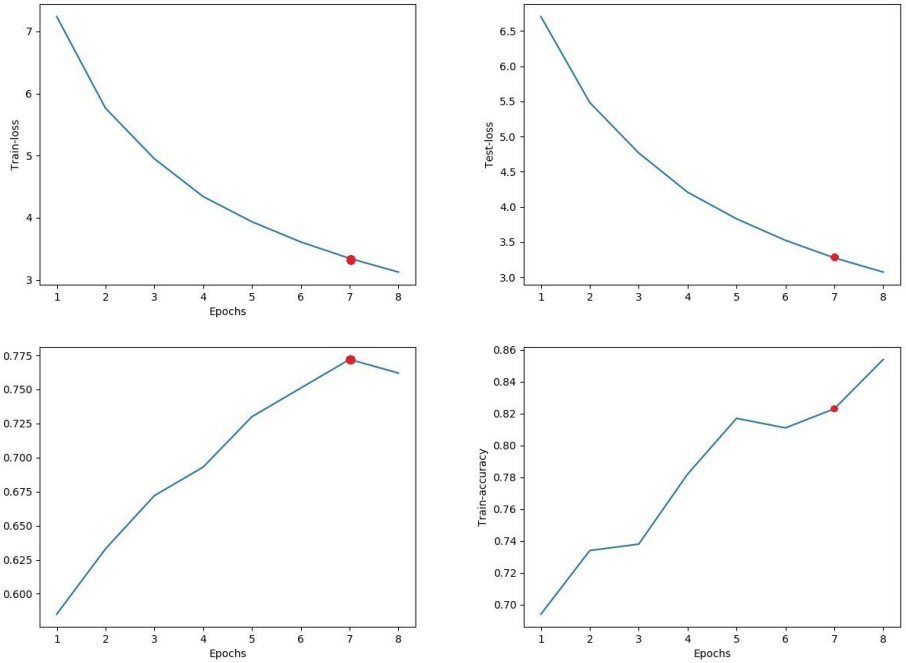
Sample images from dataset with ground truth

Dataset Statistics



Data Statistics for the test set: a) histogram of number of words in an image, b) heatmap of the bigram probabilities

Experiments: Train and Test Plots



Loss and accuracy plots on train and test data

Results

- The CRF top down cues were able to improve performance as compared to the pure DL model on both train and test set.
- On the test set, the linear CRF model performs better than the dense CRF model, while, the the dense CRF model with positional weighting performs better that the linear CRF model.
- On the train set, the performance of dense CRF with and without positional weighting marginally exceeds performance as compared to the linear CRF.
- This highlights the usefulness of lexicon derived bottom up cues and the role of positional effect on word sequence determination.

Results

		Train Set	Test Set
Pure DL Model	Word Match Accuracy	81.88	76.73
	Average Edit Distance	5.33	6.24
	Average Phonetic Distance	2.69	3.22
Linear CRF	Word Match Accuracy	84.07	80.06
	Average Edit Distance	5.19	6.11
	Average Phonetic Distance	2.63	3.19
Dense CRF	Word Match Accuracy	84.70	78.24
	Average Edit Distance	5.14	6.35
	Average Phonetic Distance	2.59	3.35
Dense CRF with Positional Weighting	Word Match Accuracy	84.65	80.36
	Average Edit Distance	5.12	5.94
	Average Phonetic Distance	2.59	3.10

Observations

- Simply choosing the DL model output with highest confidence (lowest energy), may not always be optimal.
- The linear CRF captures local properties, while the dense CRF captures global signboard level properties.
- The dense CRF has an upper hand when the signboard images are repeated, as in the case of different angles and illuminations images of the same signboard.
- Such cases of multiple images of the same signboard are more prevalent in the train set.
- In the test set, the ability to correctly determine local phrases drives performances
- Linear CRF and dense CRF with weighting are able to give more importance to adjacent word contributions hence perform well on the test set.

Conclusion

In conclusion, the work highlights the usefulness of lexicon derived bottom up cues and the role of positional effect on word sequence determination.

References

1. C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for multi-class object layout", ICCV, 2009.
2. J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context", IJCV, 81(1):2–23, 2009.
3. T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to
4. predict where humans look", ICCV, 2009.
5. Uchida S, "Text localization and recognition in images and video", Handbook of document image processing and recognition, Springer, London, pp 843–883, 2014.
6. H. Lin, P. Yang and F. Zhang, "Review of Scene Text Detection and Recognition", Archives of Computational Methods in Engineering, 2020.
7. Liu, Yuliang and Jin, Lianwen, "Deep Matching Prior Network: Toward Tighter Multi-oriented Text Detection", CVPR, 2017.
8. Liao, Minghui, Zhu, Zhen, Shi, Baoguang, Xia, Gui-song and Bai, Xiang, "Rotation-Sensitive Regression for Oriented Scene Text Detection", CVPR, 2018.
9. Wang T, Wu DJ, Coates A, Ng AY, "End-to-end text recognition with convolutional neural networks", International conference on pattern recognition, pp 3304–3308, 2012.
10. Liao, Minghui, Shi, Baoguang, Bai, Xiang, Wang, Xinggang and Liu, Wenyu, "TextBoxes: A Fast Text Detector with a Single Deep Neural Network", AAAI, 2017.
11. H. Lin, P. Yang and F. Zhang, "Review of Scene Text Detection and Recognition", Archives of Computational Methods in Engineering, 2020.
12. Dan, Deng, Haifeng, Liu, Xuelong, Li and Deng, Cai, "PixelLink: Detecting Scene Text via Instance Segmentation", AAAI, 2018.
13. Shi, Baoguang, Bai, Xiang and Belongie, Serge, "Detecting Oriented Text in Natural Images by Linking Segments", CVPR, 2017.
14. Ghosh, Suman K, Valveny, Ernest and Bagdanov, Andrew D, "Visual attention models for scene text recognition", ICDAR, 2017.
15. Busta, Michal, Neumann, Lukas and Matas, Jiri, "Deep TextSpotter: An End-To-End Trainable Scene Text Localization and Recognition Framework", ICCV, 2017.
16. Jerod J Weinman, "Scene Text Recognition Using Similarity and a Lexicon with Sparse Belief Propagation", TPAMI, 2009.

References

17. Jaderberg, Max, Simonyan, Karen, Vedaldi, Andrea and Zisserman, Andrew, "Synthetic data and artificial neural networks for natural scene text recognition", NIPS, 2014.
18. Tian, Shangxuan, Lu, Shijian and Li, Chongshou, "Wetext: Scene text detection under weak supervision", ICCV, 2017.
19. Hradis, Michal, Kotera, Jan, Zemcik, Pavel and Sroubek, Filip, "Convolutional neural networks for direct text deblurring", BMVC, 2015.
20. Zhu, Anna, Gao, Renwu and Uchida, Seiichi, "Could scene context be beneficial for scene text detection?" Pattern Recognition, 2016.
21. Yuan, Xiaoyong, He, Pan and Li, Xiaolin Andy, "Adaptive Adversarial Attack on Scene Text Recognition", CVPR, 2018.
22. Pan YF, Hou X, Liu CL, "A hybrid approach to detect and localize texts in natural scene images", IEEE Transactions on Image Processing, 20:800–813, 2011.
23. Jaderberg M, Simonyan K, Vedaldi A, Zisserman A, "Deep structured output learning for unconstrained text recognition", International conference on learning representations, pp 1–10, 2015.
24. H. Zhang, C. Liu, C. Yang, X. Ding and K. Wang. "An Improved Scene Text Extraction Method Using Conditional Random Field and Optical Character Recognition". International Conference on Document Analysis and Recognition, 2011.