# Synthetic Data & Artificial Neural Networks for Natural Scene Text Recognition

Mark Jaderberg, Karen Simonyan, Andrea Vedaldi, Andrew Zisserman

# OUTLINE

- Objective
- Challenges
- Synthetic Data Engine
- Models
- Experiments and Results
- Discussion and Questions

# Objective

To build a framework for Text Recognition in Natural Images

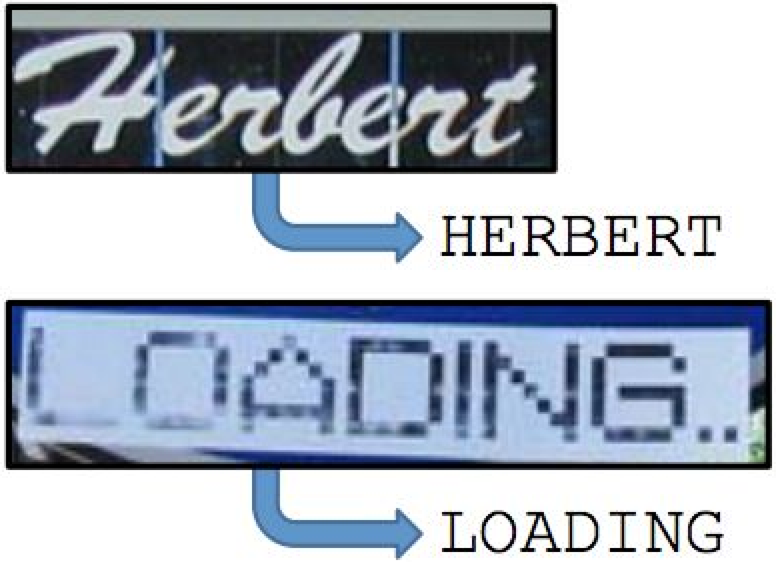



Image Credits: Synthethic Data and Artificial Neural Networks for Natural Scene Text Recognition (Poster)

# Challenges

- Inconsistent lighting, distortions, background noise, variable fonts, orientations etc..
- Existing Scene Text datasets are very small and cover limited vocabulary.

# Synthetic Data Engine

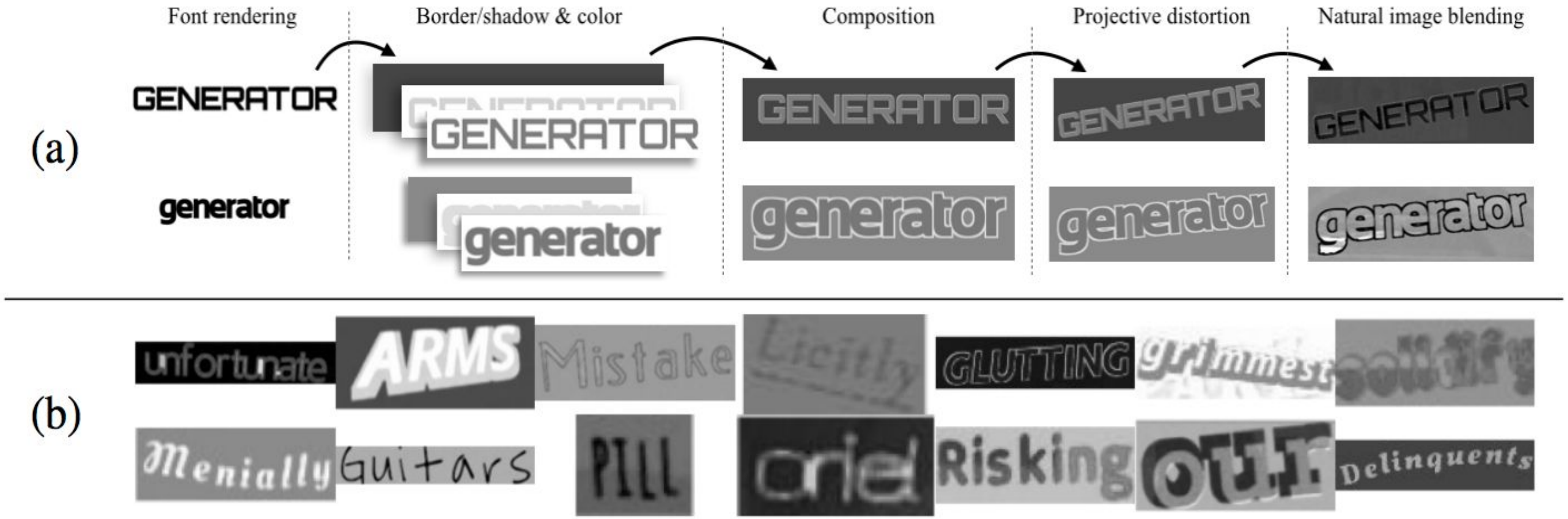



Figure 1: (a) The text generation process after font rendering, creating and coloring the image-layers, applying projective distortions, and after image blending. (b) Some randomly sampled data created by the synthetic text engine.

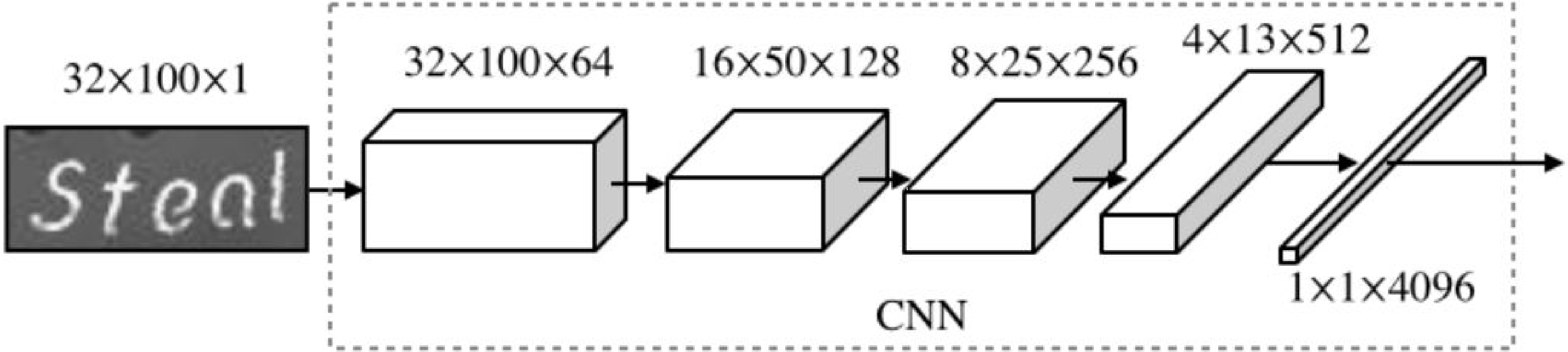
Credits: Synthethic Data and Artificial Neural Networks for Natural Scene Text Recognition

# Models

Authors propose 3 Deep Learning Models:

- Dictionary Encoding
- Character Sequence Encoding
- Bag of NGrams encoding

# Base Architecture

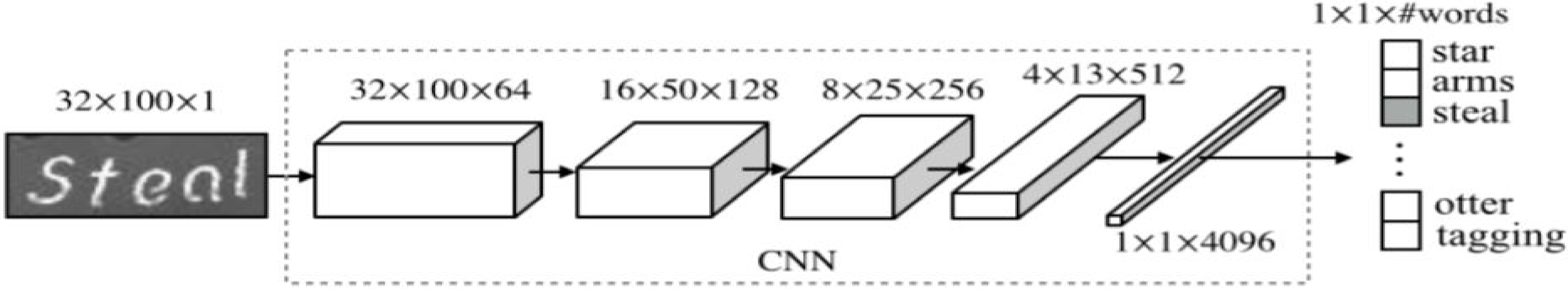


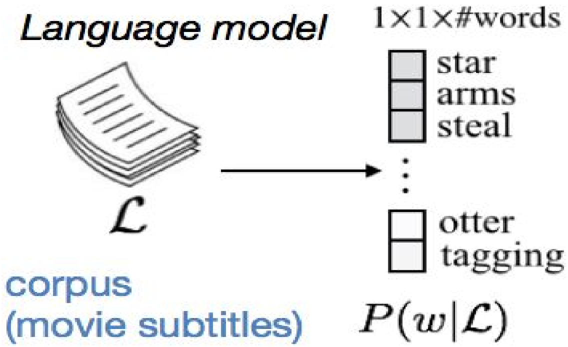

- 2 x 2 Max Pooling after 1st, 2nd and 3rd Convolutional Layer
- SGD for optimization
- Dropout for regularization

Credits: Synthethic Data and Artificial Neural Networks for Natural Scene Text Recognition

# Dictionary Encoding (DICT) [Constrained Language Model]

Multiclass Classification Problem (One class per word *w* in Dictionary *W*)



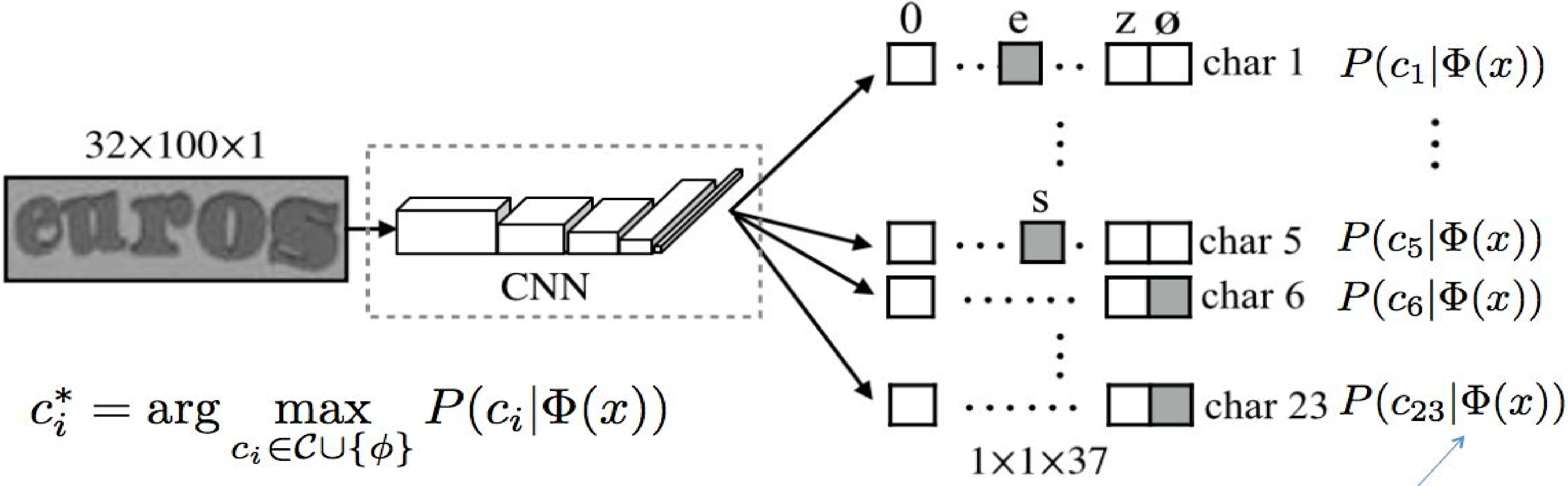

predicted word → $$w^* = \arg\max_{w \in \mathcal{W}} P(w|x)P(w|\mathcal{L})$$

The number of classes can be scaled to **90k classes.** Requires *incremental training* – initialize learning with 5k classes, incrementally increase number of classes as learning progresses.

Slide Credits: Synthethic Data and Artificial Neural Networks for Natural Scene Text Recognition (Poster)

# Character Sequence Encoding (CHAR)

CNN with multiple independent classifiers (one for each character)




$$c_i^* = \arg \max_{c_i \in \mathcal{C} \cup \{\phi\}} P(c_i|\Phi(x))$$

$$w^* = c_1^* c_2^* \ldots c_i^* \quad \forall c_i^* \notin \{\phi\}$$

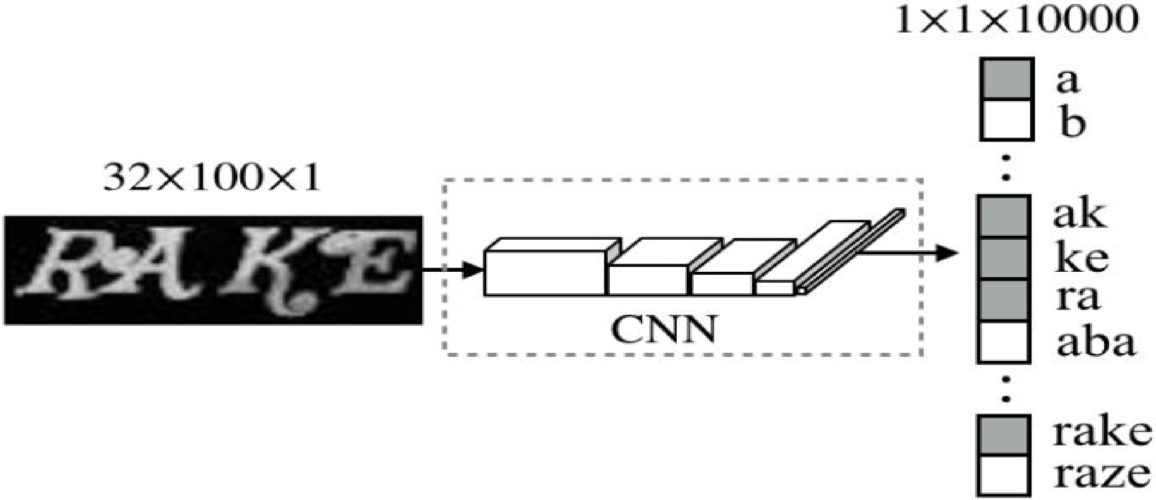
null character

- No language model but need to fix max length of the word.
- Suitable for unconstrained recognition

Slide Credits: Synthethic Data and Artificial Neural Networks for Natural Scene Text Recognition (Poster)

# BAG of N-Grams Encoding (NGRAM)

Represent a word as bag of N-grams.
Eg G(Spires) = { s, p, i, r, e, s, sp, pi, ir, re, es, spi, pir, ire, res }




Visually model 10k common 1, 2, 3, and 4-grams.

10k independent binary classifiers.

Result is N-gram detection vector.

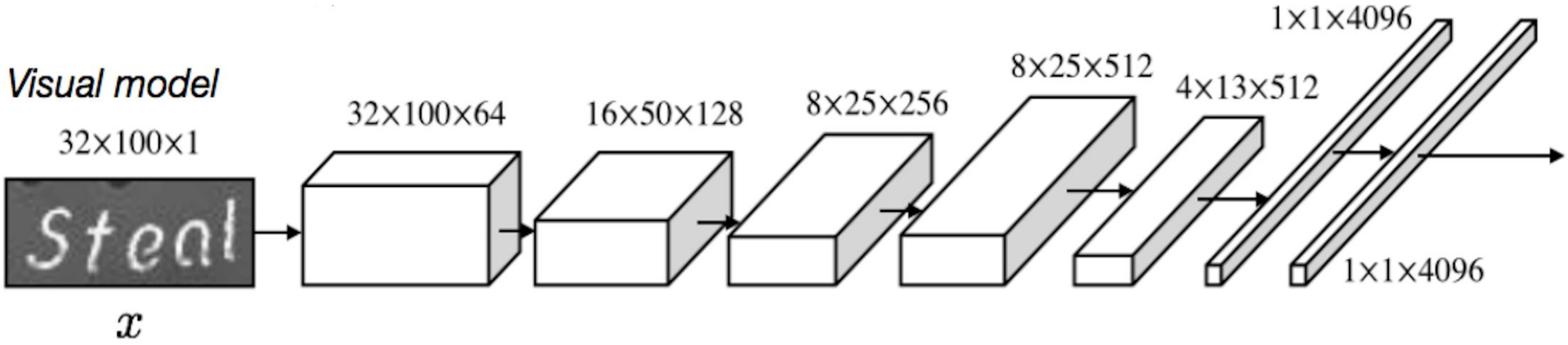
Two ways to recover words:

- Find nearest neighbour of output with ideal outputs of dictionary words.
- Train a linear SVM for each dictionary word, using training data outputs.

Slide Credits: Synthethic Data and Artificial Neural Networks for Natural Scene Text Recognition (Poster)

# +2 Models

- Lack of overfitting on basic models suggests their under-capacity.
- Try larger models to investigate the effect of additional model capacity.




- Extra convolutional layer with 512 filters
- Extra 4096 unit fully connected layer at the end

# Experiments and Results

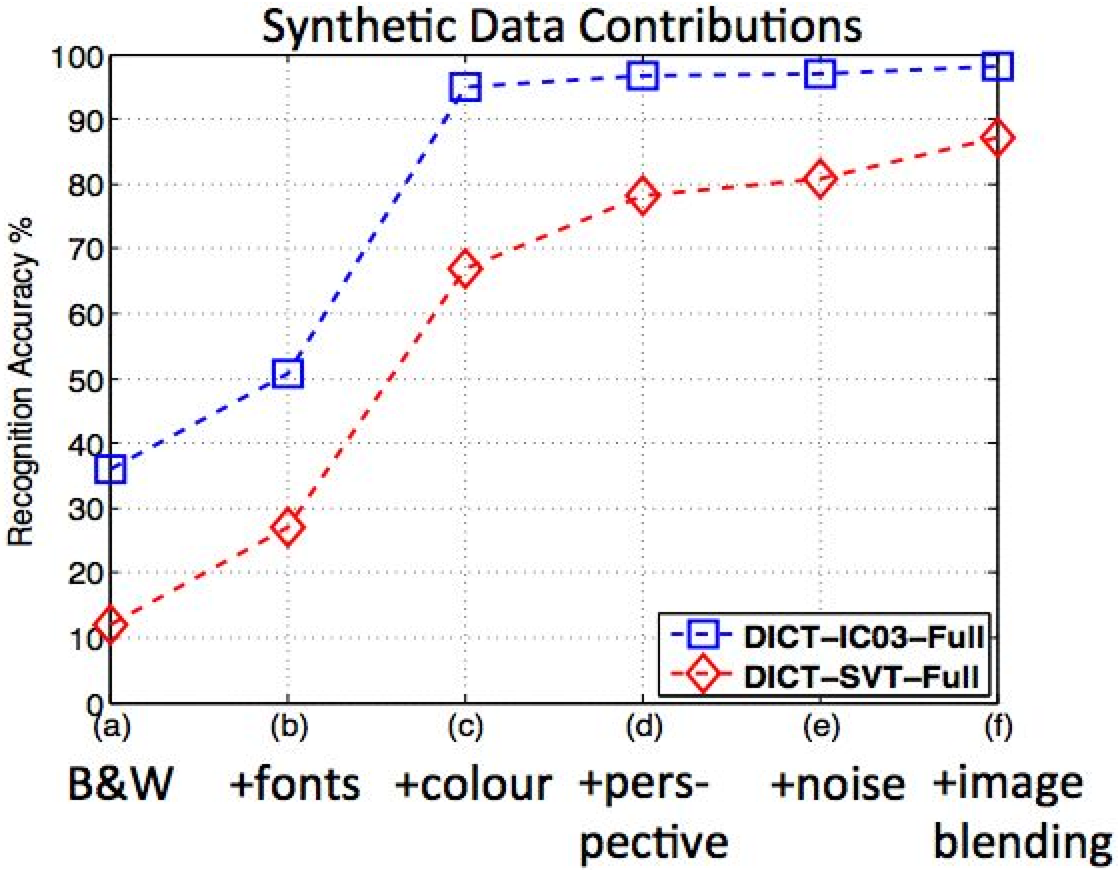



Image Credits: Synthethic Data and Artificial Neural Networks for Natural Scene Text Recognition (Poster)

# Base Models vs +2 Models

| Model | Trained Lexicon | Synth | IC03-50 | IC03 | SVT-50 | SVT | IC13 |
|---|---|---|---|---|---|---|---|
| DICT IC03 FULL | IC03 FULL | 98.7 | 99.2 | 98.1 | - | - | - |
| DICT SVT FULL | SVT FULL | 98.7 | - | - | 96.1 | 87.0 | - |
| DICT 50K | 50K | 93.6 | 99.1 | 92.1 | 93.5 | 78.5 | 92.0 |
| DICT 90K | 90K | 90.3 | 98.4 | 90.0 | 93.7 | 70.0 | 86.3 |
| DICT +2 90K | 90K | 95.2 | 98.7 | 93.1 | 95.4 | 80.7 | 90.8 |
| CHAR | 90K | 71.0 | 94.2 | 77.0 | 87.8 | 56.4 | 68.8 |
| CHAR +2 | 90K | 86.2 | 96.7 | 86.2 | 92.6 | 68.0 | 79.5 |
| NGRAM NN | 90K | 25.1 | 92.2 | - | 84.5 | - | - |
| NGRAM +2 NN | 90K | 27.9 | 94.2 | - | 86.6 | - | - |

# Quality of Synthetic Data

| Model | Trained Lexicon | Synth | IC03-50 | IC03 | SVT-50 | SVT | IC13 |
|---|---|---|---|---|---|---|---|
| DICT IC03 FULL | IC03 FULL | **98.7** | **99.2** | **98.1** | - | - | - |
| DICT SVT FULL | SVT FULL | **98.7** | - | - | **96.1** | **87.0** | - |
| DICT 50K | 50K | 93.6 | 99.1 | 92.1 | 93.5 | 78.5 | 92.0 |
| DICT 90K | 90K | 90.3 | 98.4 | 90.0 | 93.7 | 70.0 | 86.3 |
| DICT +2 90K | 90K | 95.2 | 98.7 | 93.1 | 95.4 | 80.7 | 90.8 |
| CHAR | 90K | 71.0 | 94.2 | 77.0 | 87.8 | 56.4 | 68.8 |
| CHAR +2 | 90K | 86.2 | 96.7 | 86.2 | 92.6 | 68.0 | 79.5 |
| NGRAM NN | 90K | 25.1 | 92.2 | - | 84.5 | - | - |
| NGRAM +2 NN | 90K | 27.9 | 94.2 | - | 86.6 | - | - |

# Effect of Dictionary Size

| Model | Trained Lexicon | Synth | IC03-50 | IC03 | SVT-50 | SVT | IC13 |
|---|---|---|---|---|---|---|---|
| DICT IC03 FULL | IC03 FULL | 98.7 | 99.2 | 98.1 | - | - | - |
| DICT SVT FULL | SVT FULL | 98.7 | - | - | 96.1 | 87.0 | - |
| DICT 50K | 50K | 93.6 | 99.1 | 92.1 | 93.5 | 78.5 | 92.0 |
| DICT 90K | 90K | 90.3 | 98.4 | 90.0 | 93.7 | 70.0 | 86.3 |
| DICT +2 90K | 90K | 95.2 | 98.7 | 93.1 | 95.4 | 80.7 | 90.8 |
| CHAR | 90K | 71.0 | 94.2 | 77.0 | 87.8 | 56.4 | 68.8 |
| CHAR +2 | 90K | 86.2 | 96.7 | 86.2 | 92.6 | 68.0 | 79.5 |
| NGRAM NN | 90K | 25.1 | 92.2 | - | 84.5 | - | - |
| NGRAM +2 NN | 90K | 27.9 | 94.2 | - | 86.6 | - | - |

| Model | IC03-50 | IC03-Full | SVT-50 | SVT | IC13 | IIIT5k-50 | IIIT5k-1k |
|---|---|---|---|---|---|---|---|
| *Baseline (ABBYY)* | 56.0 | 55.0 | 35.0 | - | - | 24.3 | - |
| Wang, ICCV '11 | 76.0 | 62.0 | 57.0 | - | - | - | - |
| Bissacco, ICCV '13 | - | - | 90.4 | 78.0 | 87.6 | - | - |
| Yao, CVPR '14 | 88.5 | 80.3 | 75.9 | - | - | 80.2 | 69.3 |
| Jaderberg, ECCV '14 | 96.2 | 91.5 | 86.1 | - | - | - | - |
| Gordo, arXiv '14 | - | - | 90.7 | - | - | 93.3 | 86.6 |
| DICT-IC03-Full | **99.2** | **98.1** | - | - | - | - | - |
| DICT-SVT-Full | **-** | **-** | **96.1** | **87.0** | - | - | - |
| DICT+2-90k | **98.7** | **98.6** | **95.4** | **80.7** | **90.8** | **97.1** | **92.7** |
| CHAR+2 | **96.7** | **94.0** | **92.6** | 68.0 | 79.5 | **95.5** | 85.4 |
| NGRAM+2-SVM | **96.5** | **94.0** | - | - | - | - | - |

Slide Credits: Synthethic Data and Artificial Neural Networks for Natural Scene Text Recognition (Poster)

# Examples

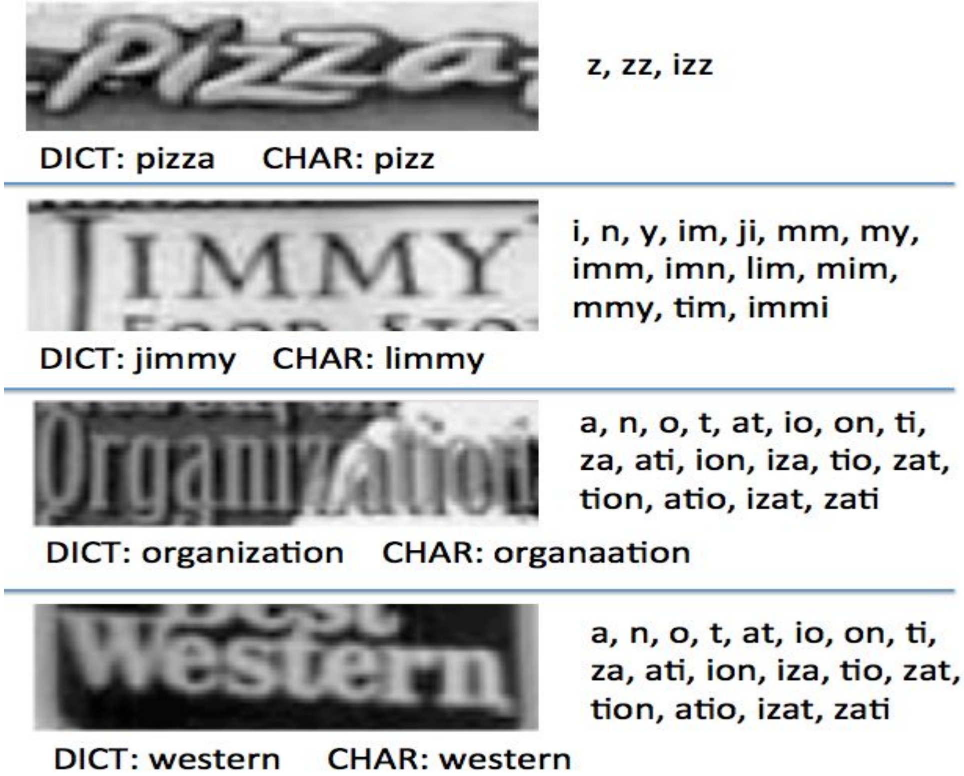




Image Credits: Synthethic Data and Artificial Neural Networks for Natural Scene Text Recognition (Poster)

# Applications

- Image Retrieval
- Self Driving Cars

# Discussion and Questions

- How fair is it to assume knowledge of target lexicon ?
- Has synthetic data been used in any other domains ?
- Can we use RNN models for predicting words character level classification ?
- Are there better ways of mapping Ngrams to words ?
- How are collisions handled in Ngrams model ?
- How diverse does the text synthesis output need to be ?

# References

[1] Synthethic Data and Artificial Neural Networks for Natural Scene Text Recognition

[2] Synthethic Data and Artificial Neural Networks for Natural Scene Text Recognition (Poster)

# Thank You  :)