UNIVERSITY OF OXFORD

EPSRC
Engineering and Physical Sciences Research Council

Deep Structured Output Learning for Unconstrained Text Recognition

Max Jaderberg, Karen Simonyan, Andrea Vedaldi, Andrew Zisserman

Visual Geometry Group, Department of Engineering Science, University of Oxford, UK

Google DeepMind

1. Overview

Text recognition in natural scene images. Allow predictions not constrained to dictionary or by static language model.

Contributions

- Combine two complementary text recognition **CNN models with a CRF** in a joint model.
- Formulate the **structured output loss** and use to jointly train the combined model.
- A model able to perform **zero-shot recognition**, and achieving state-of-the-art results in constrained and unconstrained scenarios.

2. Datasets

Synth90k
9 million images covering 90k words, training/test splits defined.
Download: www.robots.ox.ac.uk/~vgg/data/text/

SynthRand
9 million images, 1-10 character random words

ICDAR 2003, ICDAR 2013

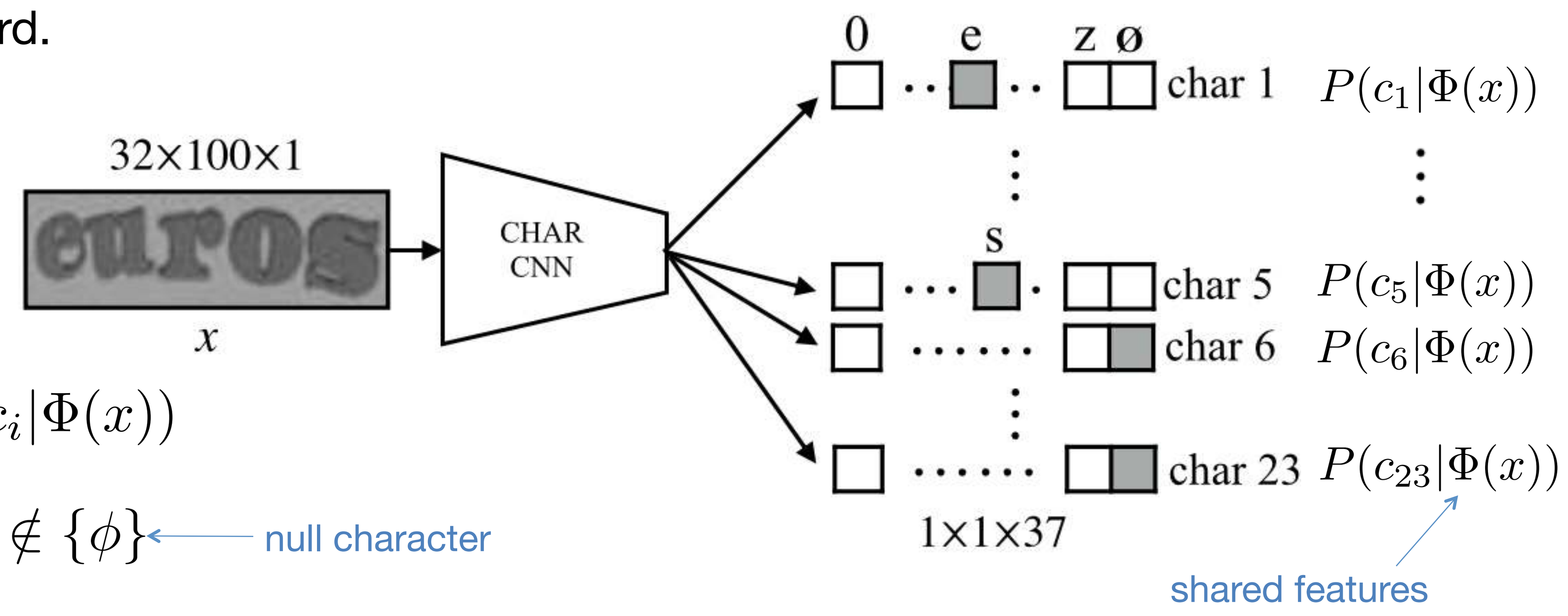
SVT

IIIT5k

3. Text Recognition Models

Character Sequence Encoding (CHAR)

Single CNN with multiple independent classifiers. Each classifier predicts the character at each position of the word.



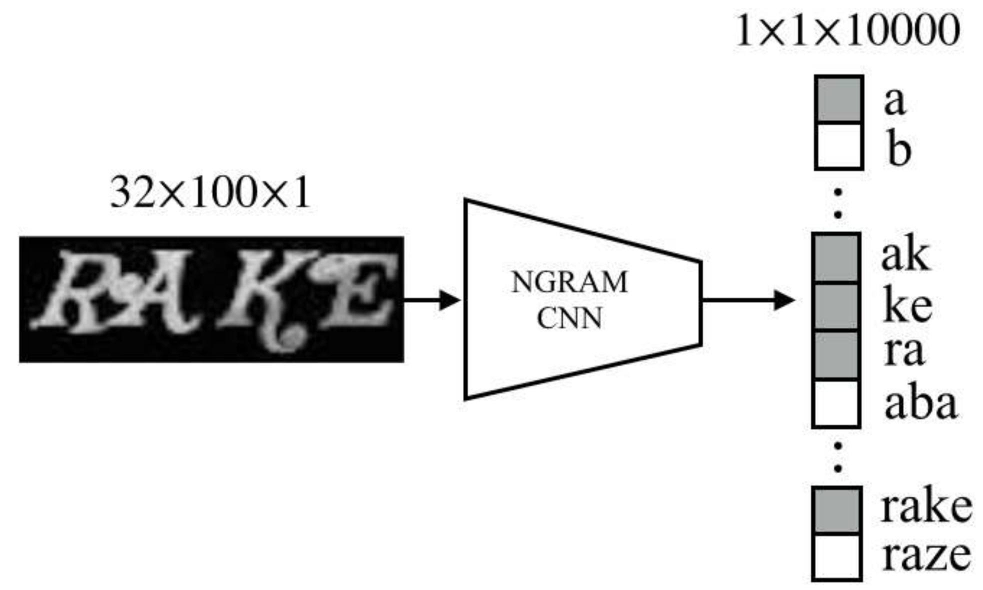
$$c_i^* = \arg \max_{c_i \in \mathcal{C} \cup \{\phi\}} P(c_i|\Phi(x))$$

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

null character

Bag of N-grams Encoding (NGRAM)

Represent a string as a bag-of-N-grams.
E.g. $G(\text{spires}) = \{\text{s, p, i, r, e, s, sp, pi, ir, re, es, spi, pir, ire, res, spire, pires}\}$



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

10k independent binary classifiers.

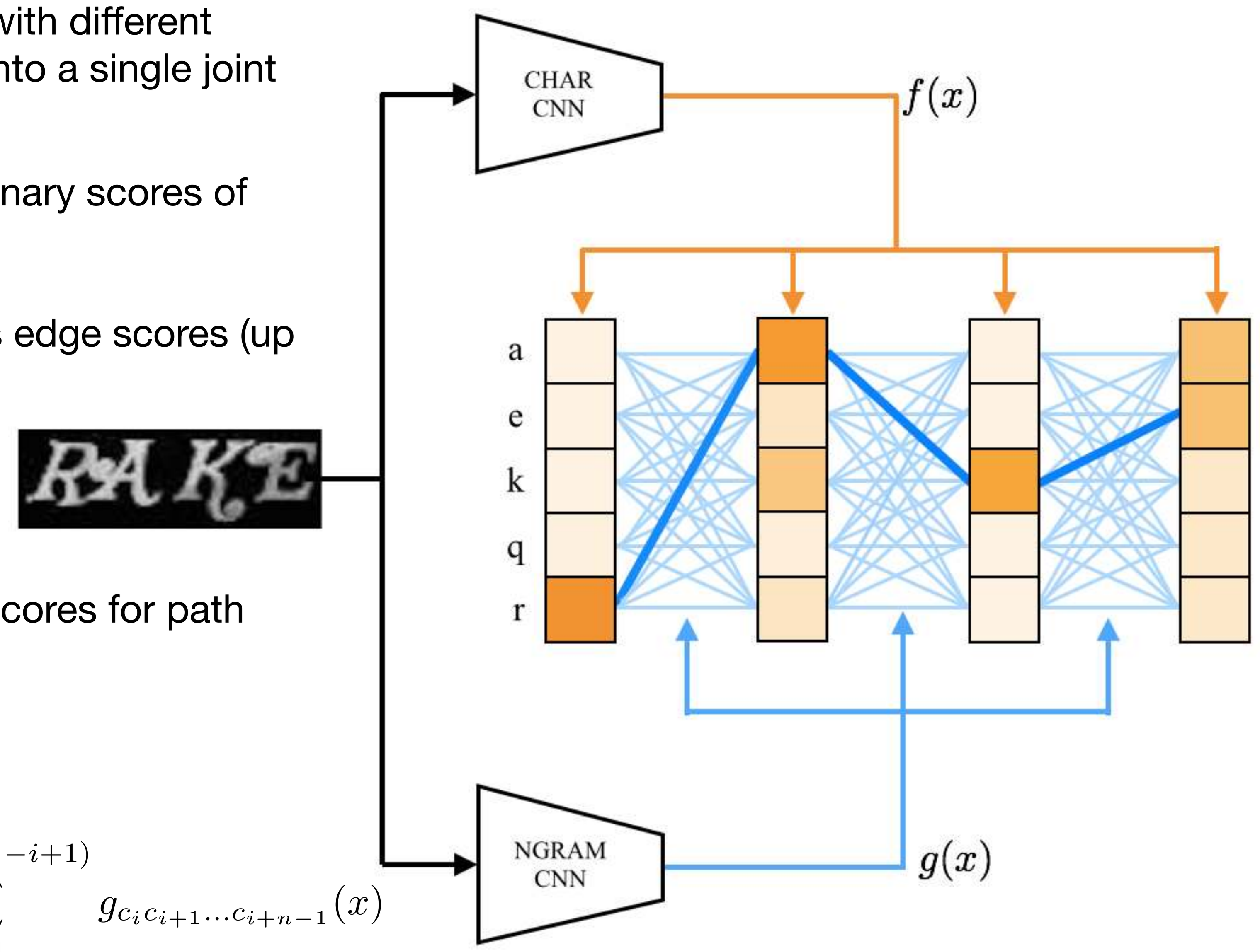
Result is N-gram detection vector.

4. Joint Model & Structured Output Loss

Combine two models with different word representations into a single joint model.

CHAR model defines unary scores of nodes in graph.

NGRAM model defines edge scores (up to 4th order).



Word score is sum of scores for path through graph.

$$S(w,x) = \sum_{i=1}^{N_{\max}} f_{c_i}^i(x) + \sum_{i=1}^{|w|} \sum_{n=1}^{\min(N,|w|-i+1)} g_{c_i c_{i+1} \dots c_{i+n-1}}(x)$$

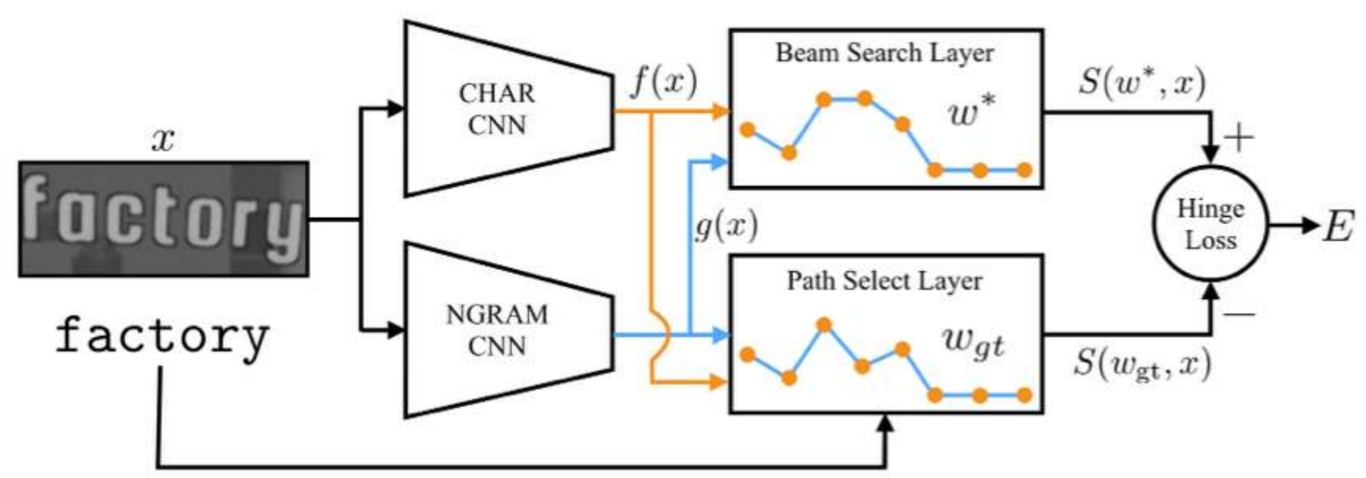
Train with **structured output loss**

$$S(w_{gt}, x) \geq \mu + S(w^*, x)$$

$$\text{where } S(w^*, x) = \max_{w \neq w_{gt}} S(w, x)$$

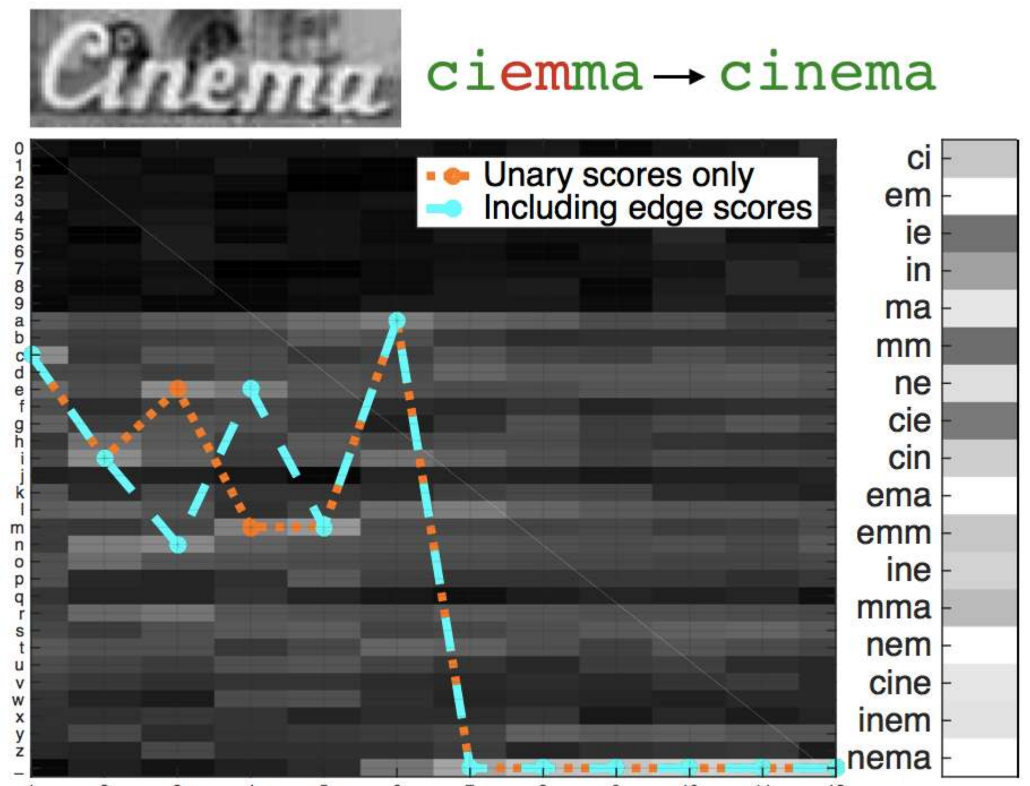
which leads to hinge loss

$$\max_{w \neq w_{\mathrm{gt},i}} \max(0, \mu + S(w,x) - S(w_{\mathrm{gt},i}, x_i))$$



Find max with Beam Search.
Gradients back propagated to networks.

5. Experiments



Train Data	Test Data	CHAR	JOINT
Synth90k	Synth90k	87.3	**91.0**
	Synth72k-90k	87.3	-
	Synth45k-90k	87.3	-
	IC03	85.9	**89.6**
	SVT	68.0	**71.7**
	IC13	79.5	**81.8**
Synth1-72k	Synth72k-90k	82.4	**89.7**
Synth1-45k	Synth45k-90k	80.3	**89.1**
SynthRand	SynthRand	**80.7**	79.5

CHAR: chocoma_ JOINT: chocomel GT: chocomel

CHAR: mediaal JOINT: medical GT: medical

CHAR: iustralia JOINT: australia GT: australia

CHAR: rgqgan323 JOINT: rgqgan323 GT: rgqgan323

		No Lexicon			Fixed Lexicon			
Model Type	Model	IC03	SVT	IC13	IC03-Full	SVT-50	IIIT5k-50	IIIT5k-1k
Unconstrained	*Baseline (ABBYY)*	-	-	-	55.0	35.0	24.3	-
Language Constrained	Wang, ICCV '11	-	-	-	62.0	57.0	-	-
	Bissacco, ICCV '13	-	78.0	87.6	-	90.4	-	-
	Yao, CVPR '14	-	-	-	80.3	75.9	80.2	69.3
	Jaderberg, ECCV '14	-	-	-	91.5	86.1	-	-
	Gordo, arXiv '14	-	-	-	-	90.7	93.3	86.6
	Jaderberg, NIPSDLW '14	**98.6**	**80.7**	**90.8**	**98.6**	**95.4**	**97.1**	**92.7**
Unconstrained	CHAR	85.9	68.0	79.5	96.7	93.5	95.0	89.3
	JOINT	89.6	71.7	81.8	97.0	93.2	95.5	89.6