DEEP STRUCTURED OUTPUT LEARNING FOR UNCONSTRAINED TEXT RECOGNITION

Max Jaderberg, Karen Simonyan, Andrea Vedaldi, Andrew Zisserman

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

TEXT RECOGNITION

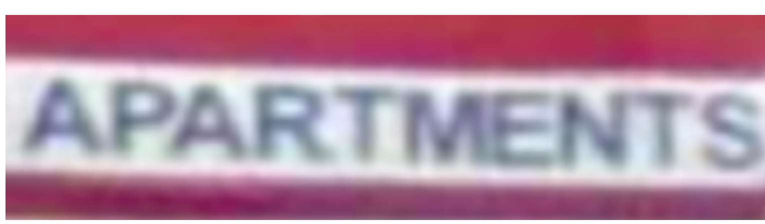
Localized text image as input, character string as output

TEXT RECOGNITION

State of the art — **constrained** text recognition

word classification *[Jaderberg, NIPS DLW 2014]*

static ngram and word language model *[Bissacco, ICCV 2013]*



APARTMENTS

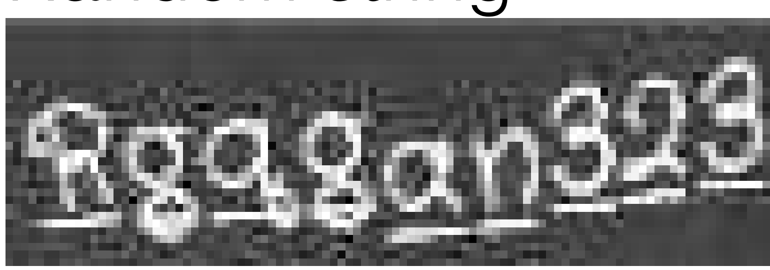
TEXT RECOGNITION

State of the art — **constrained** text recognition

word classification *[Jaderberg, NIPS DLW 2014]*

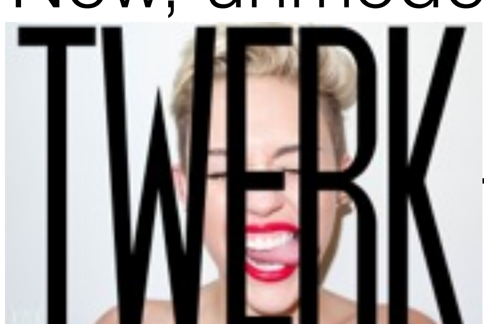
static ngram and word language model *[Bissacco, ICCV 2013]*

Random string



?

New, unmodeled word



?

TEXT RECOGNITION

Unconstrained text recognition

e.g. for house numbers *[Goodfellow, ICLR 2014]*
business names, phone numbers, emails, etc

Random string

RGQGAN323

New, unmodeled word

TWERK

OVERVIEW

- Two models for text recognition *[Jaderberg, NIPS DLW 2014]*
  - Character Sequence Model
  - Bag-of-N-grams Model

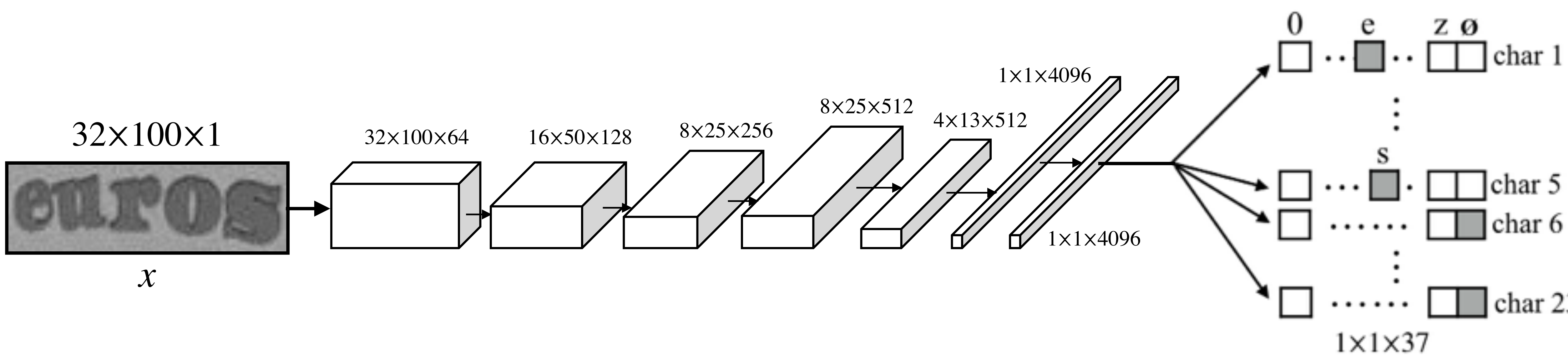
- Joint formulation
  - CRF to construct graph
  - Structured output loss
  - Use back-propagation for joint optimization

- Experiments
  - Generalize to perform zero-shot recognition
  - When constrained recover performance

CHARACTER SEQUENCE MODEL

Deep CNN to encode image.
Per-character decoder.

$w = (c_1, c_2, \ldots, c_N)$



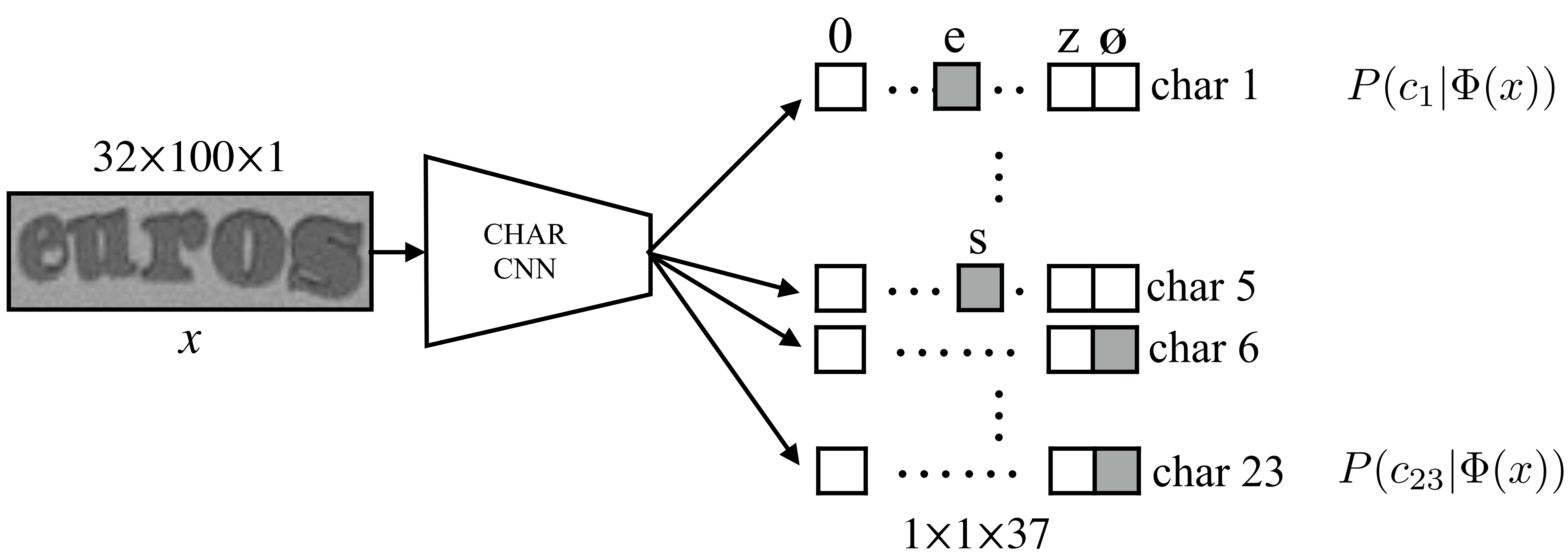
5 convolutional layers, 2 FC layers, ReLU, max-pooling
23 output classifiers for 37 classes (0-9,a-z,null)

Fixed 32x100 input size — distorts aspect ratio

CHARACTER SEQUENCE MODEL

Deep CNN to encode image.
Per-character decoder.

$w = (c_1, c_2, \dots, c_N)$



$$w^* = \arg\max_w P(w|x) = \arg \max_{c_1, c_2, \dots, c_{N_{\max}}} \prod_{i=1}^{N_{\max}} P(c_i|\Phi(x))$$

Bag-of-N-grams Model

Represent string by the character N-grams contained within the string

spires →

s
p
i
r
e

1-grams

sp
pi
ir
re
es

2-grams

spi
pir
ire
res

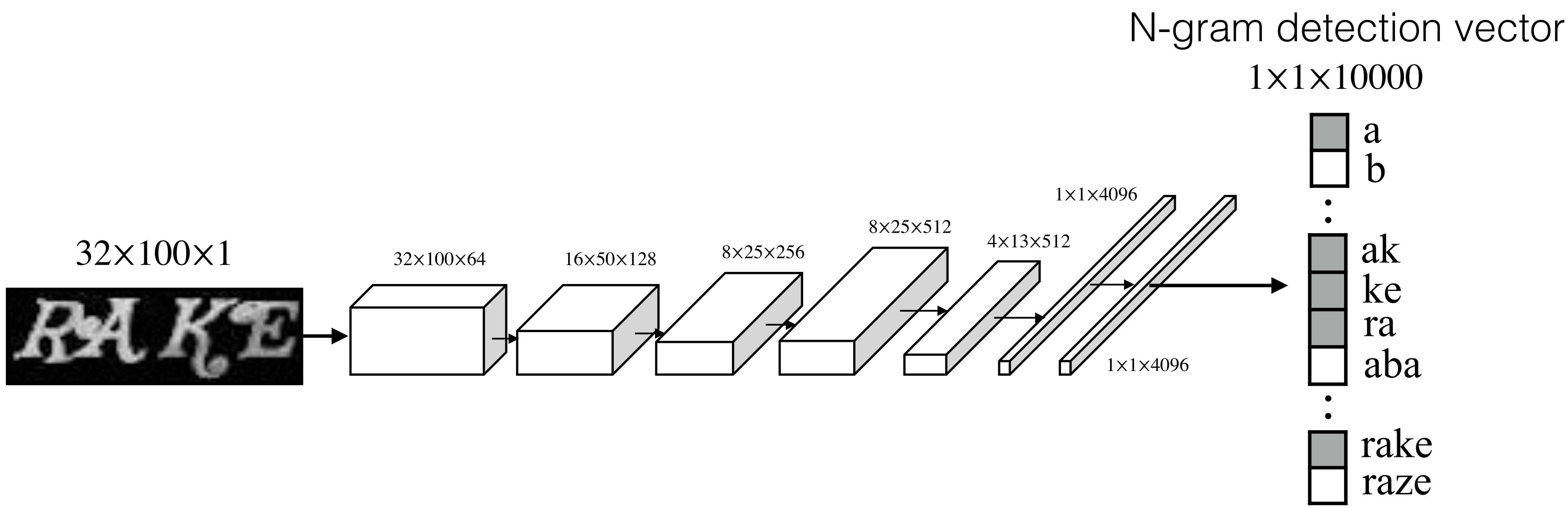
3-grams

spir
pire
ires

4-grams

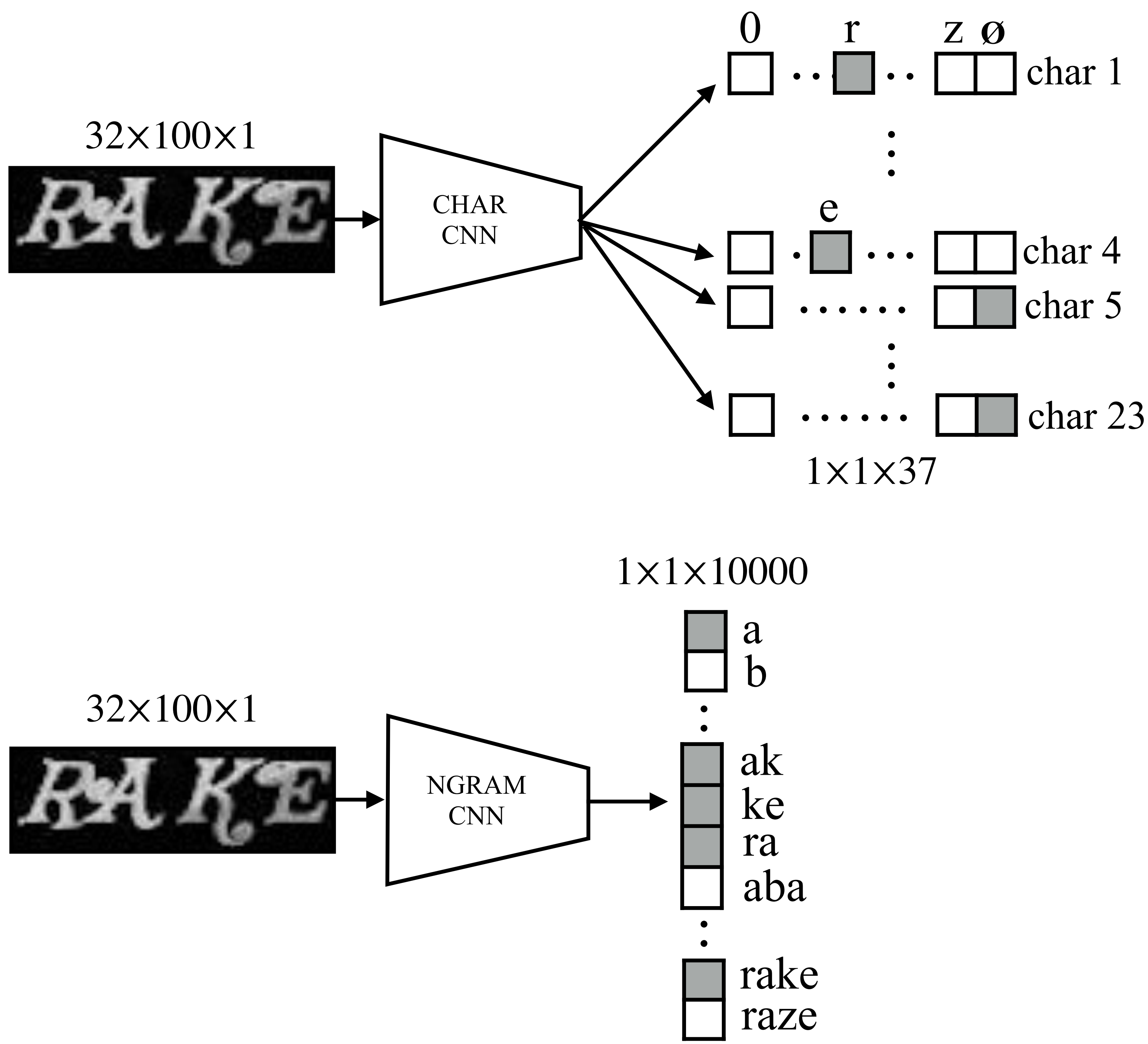
BAG-OF-N-GRAMS MODEL

Deep CNN to encode image.
N-grams detection vector output.
Limited (10k) set of modeled N-grams.



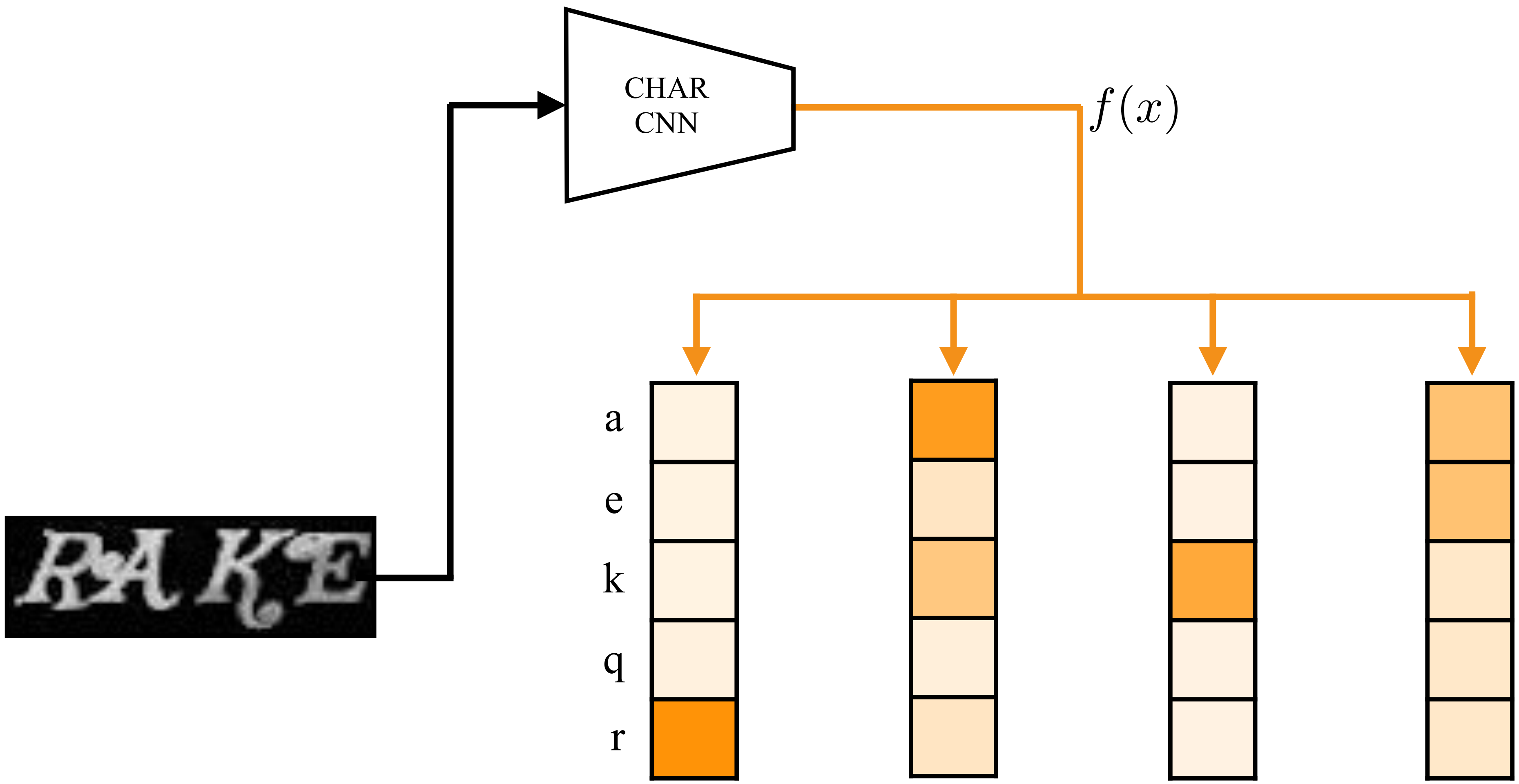
JOINT MODEL

Can we combine these two representations?



Joint Model

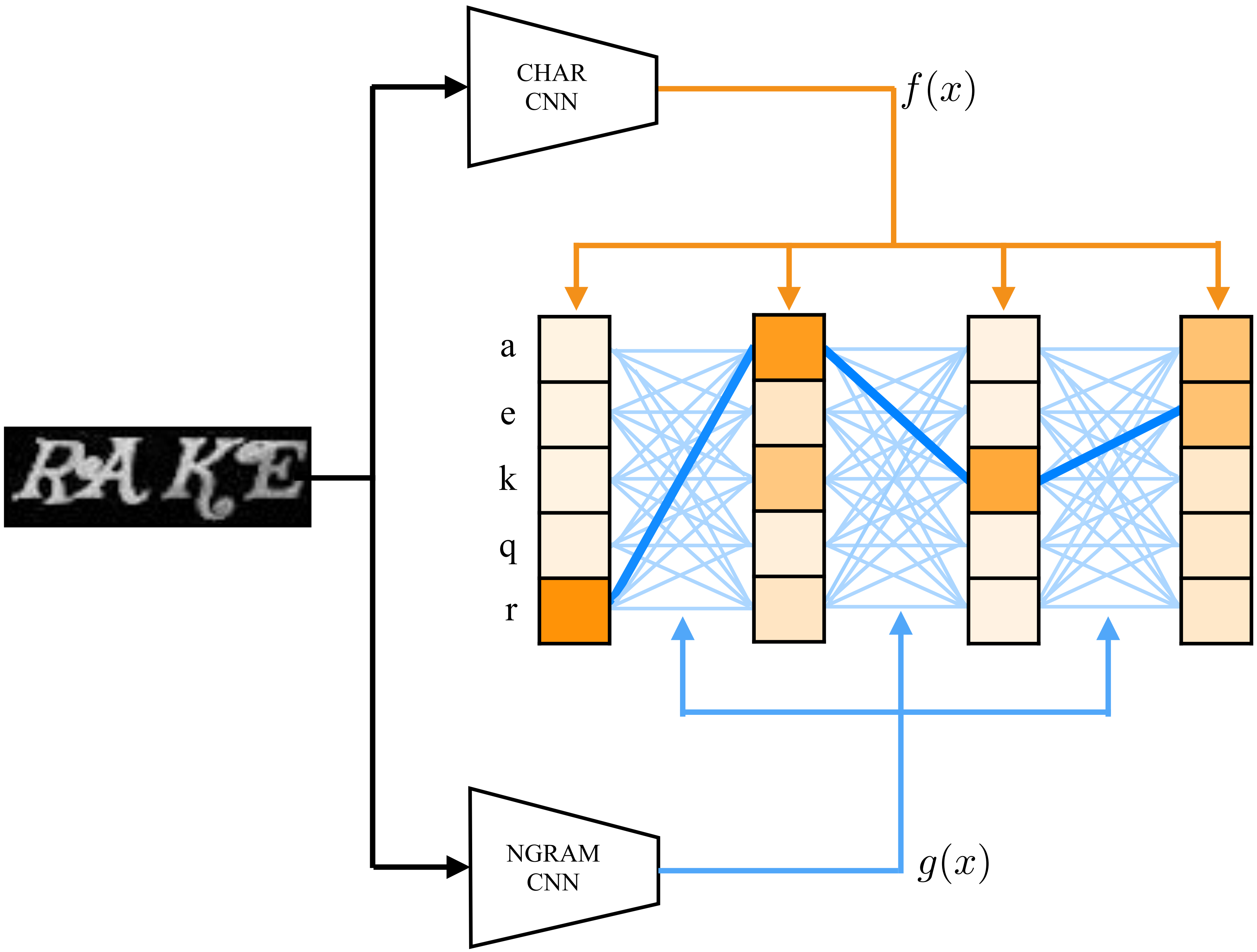
$$S(w, x) = \sum_{i=1}^{N_{\max}} f_{c_i}^i(x)$$



Joint Model

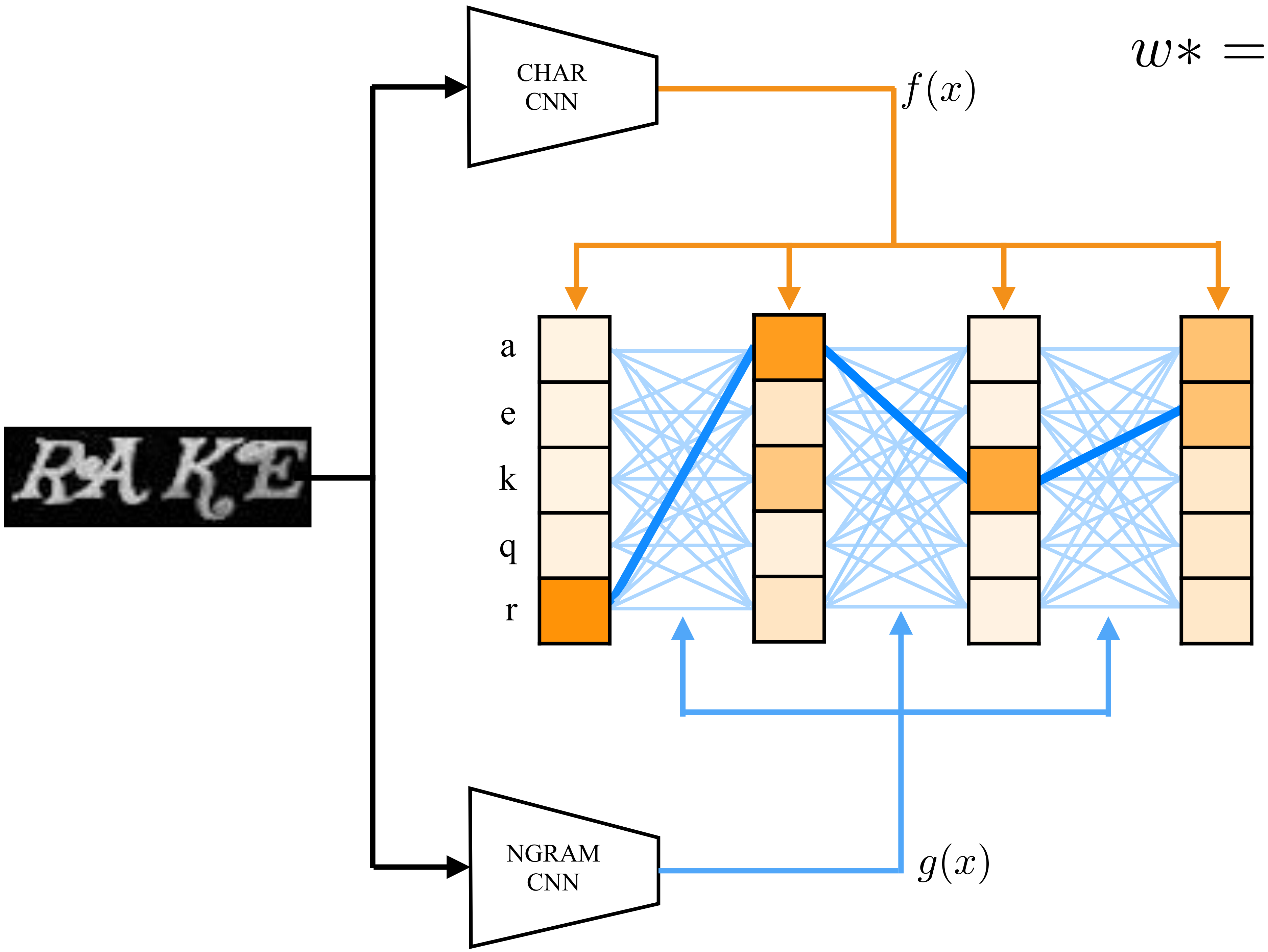
maximum number of chars

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



JOINT MODEL

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



$$w* = \arg\max_{w} S(w,x)$$

beam search

Structured Output Loss

Score of ground-truth word should be greater than or equal to the highest scoring incorrect word + margin.

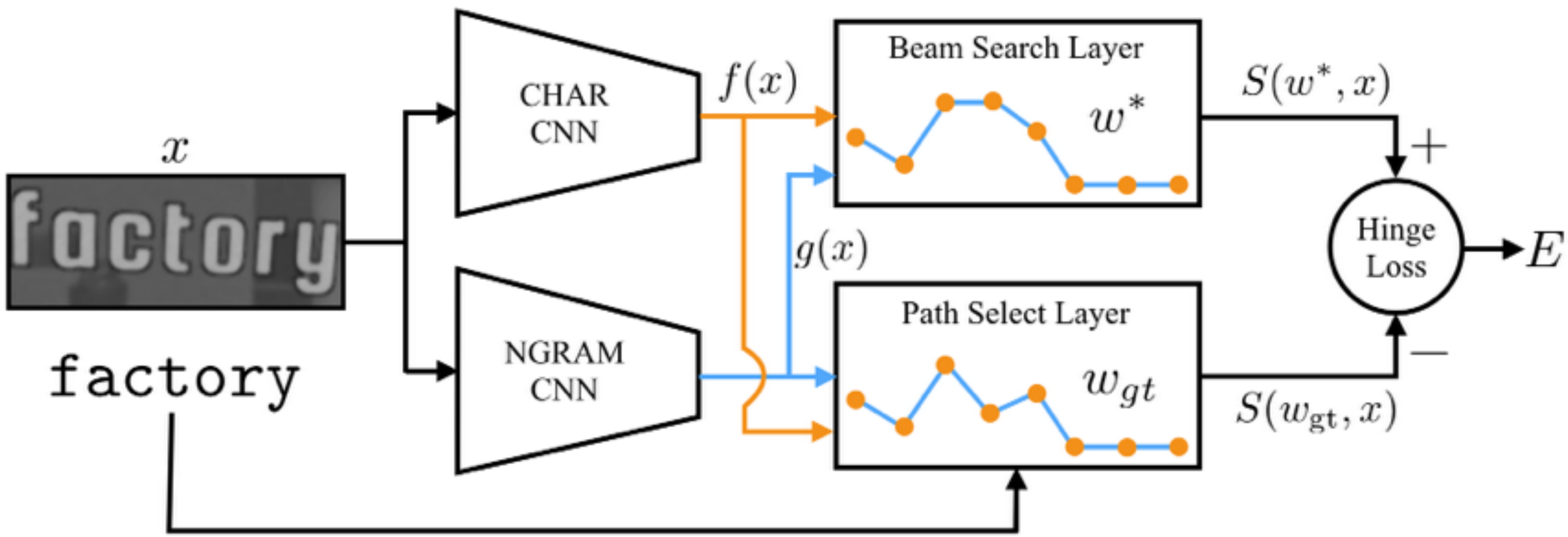
$$S(w_{\mathrm{gt}}, x) \;\geq\; \mu + S(w^*, x)$$

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

Enforcing as soft constraint leads to a hinge loss

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

Structured Output Loss



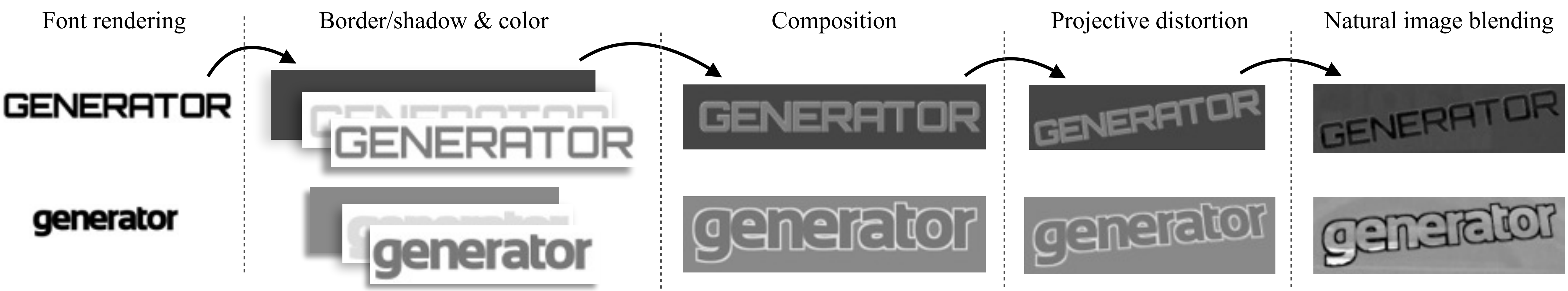
$$S(w_{\mathrm{gt}}, x) \;\geq\; \mu + S(w^*, x)$$

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

EXPERIMENTS

DATASETS

All models trained **purely on synthetic data**
[Jaderberg, NIPS DLW 2014]



Realistic enough to transfer to test on **real-world** images

Datasets

Synth90k

Lexicon of 90k words.

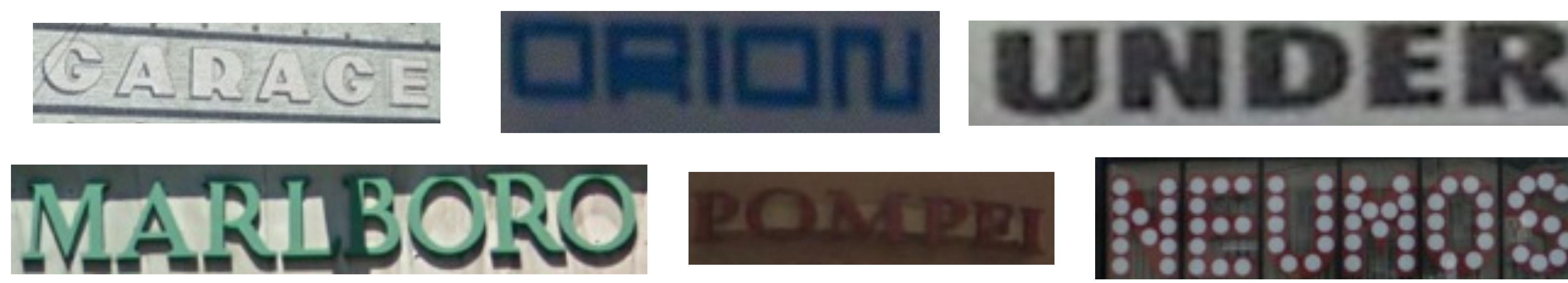
9 million images, training + test splits

Download from [***http://www.robots.ox.ac.uk/~vgg/data/text/***](http://www.robots.ox.ac.uk/~vgg/data/text/)

DATASETS

ICDAR 2003, 2013

Street View Text



IIIT 5k-word

TRAINING

Pre-train CHAR and NGRAM model independently.

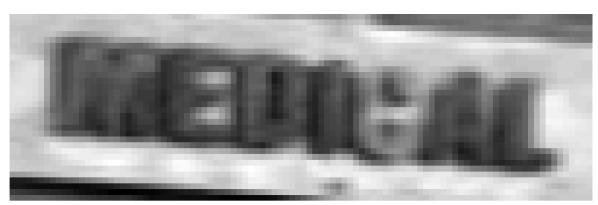
Use them to initialize joint model and continue jointly training.

EXPERIMENTS - JOINT IMPROVEMENT

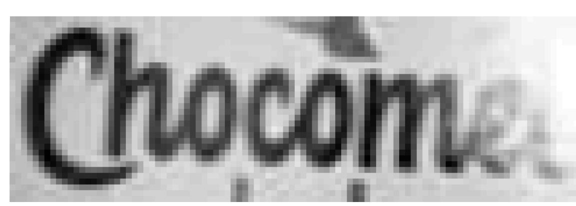
Train Data	Test Data	CHAR	JOINT
Synth90k	Synth90k	87.3	**91.0**
	IC03	85.9	**89.6**
	SVT	68.0	**71.7**
	IC13	79.5	**81.8**

joint model outperforms character sequence model alone

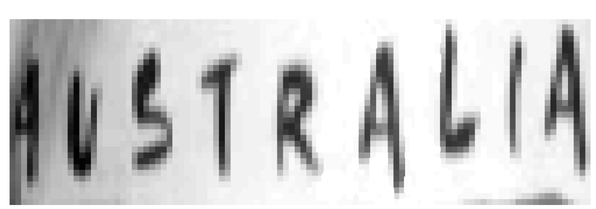
CHAR: grahaws
JOINT: grahams
GT: grahams



CHAR: mediaal
JOINT: medical
GT: medical

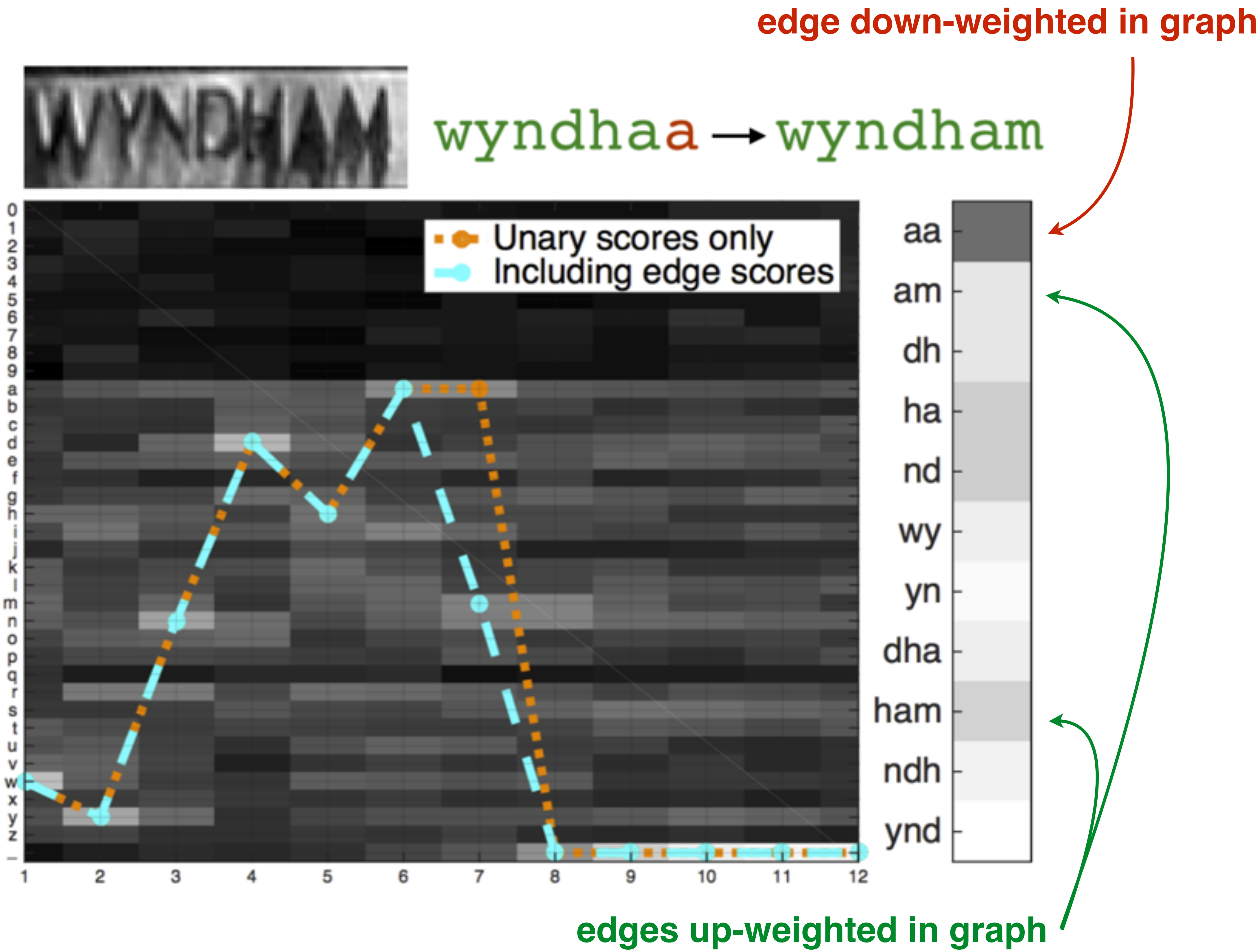


CHAR: chocoma_
JOINT: chocomel
GT: chocomel



CHAR: iustralia
JOINT: australia
GT: australia

Joint Model Corrections



EXPERIMENTS - ZERO-SHOT RECOGNITION

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**

large difference for CHAR model when not trained on test words

joint model recovers performance

Experiments - Comparison

Model Type	Model	No Lexicon		
		IC03	SVT	IC13
Unconstrained	*Baseline (ABBYY)*	-	-	-
Language Constrained	Wang, ICCV '11	-	-	-
	Bissacco, ICCV '13	-	78.0	87.6
	Yao, CVPR '14	-	-	-
	Jaderberg, ECCV '14	-	-	-
	Gordo, arXiv '14	-	-	-
	Jaderberg, NIPSDLW '14	**98.6**	**80.7**	**90.8**
Unconstrained	CHAR	85.9	68.0	79.5
	JOINT	89.6	71.7	81.8

Experiments - Comparison

Model Type	Model	No Lexicon			Fixed Lexicon			
		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

SUMMARY

- Two models for text recognition
- Joint formulation
  - Structured output loss
  - Use back-propagation for joint optimization
- Experiments
  - Joint model improves accuracy on language-based data.
  - Degrades elegantly when not from language (N-gram model doesn’t contribute much)
  - Set benchmark for unconstrained accuracy, competes with purely constrained models.

jaderberg@google.com

UNIVERSITY OF OXFORD

EPSRC
Engineering and Physical Sciences Research Council

Google DeepMind