Microsoft

Achieving human parity performance in pattern recognition and language understanding by machines

Dr. JC Mao

Corporate Vice President

Artificial Intelligence and Research Group

My First ICPR Paper

https://dblp.uni-trier.de/db/conf/icpr/icpr1988

Multiresolution Rotation Invariant Simultaneous Auto Regressive
Model for Texture Analysis

Wan Jiaruo Mao Jianchang Wang Cheng Dao

Dept. of Electronic Science and Technology
East China Normal University
Shanghai, P.R.China

Abstract

This paper presents two new models called multivariate Rotation-Invariant SAR(RISAR) model and Multiresolution SAR/RISAR(MRSAR/MRRISAR) model.The information included in MRSAR/MRRISAR model is more complete and more independent.Many experiments on image classification using RISAR model and MRRISAR model indicate that these models have many perfect properties, such as, very strong classification power, wide area in which it can be used,very good property of rotation invariance,and high speed for re-

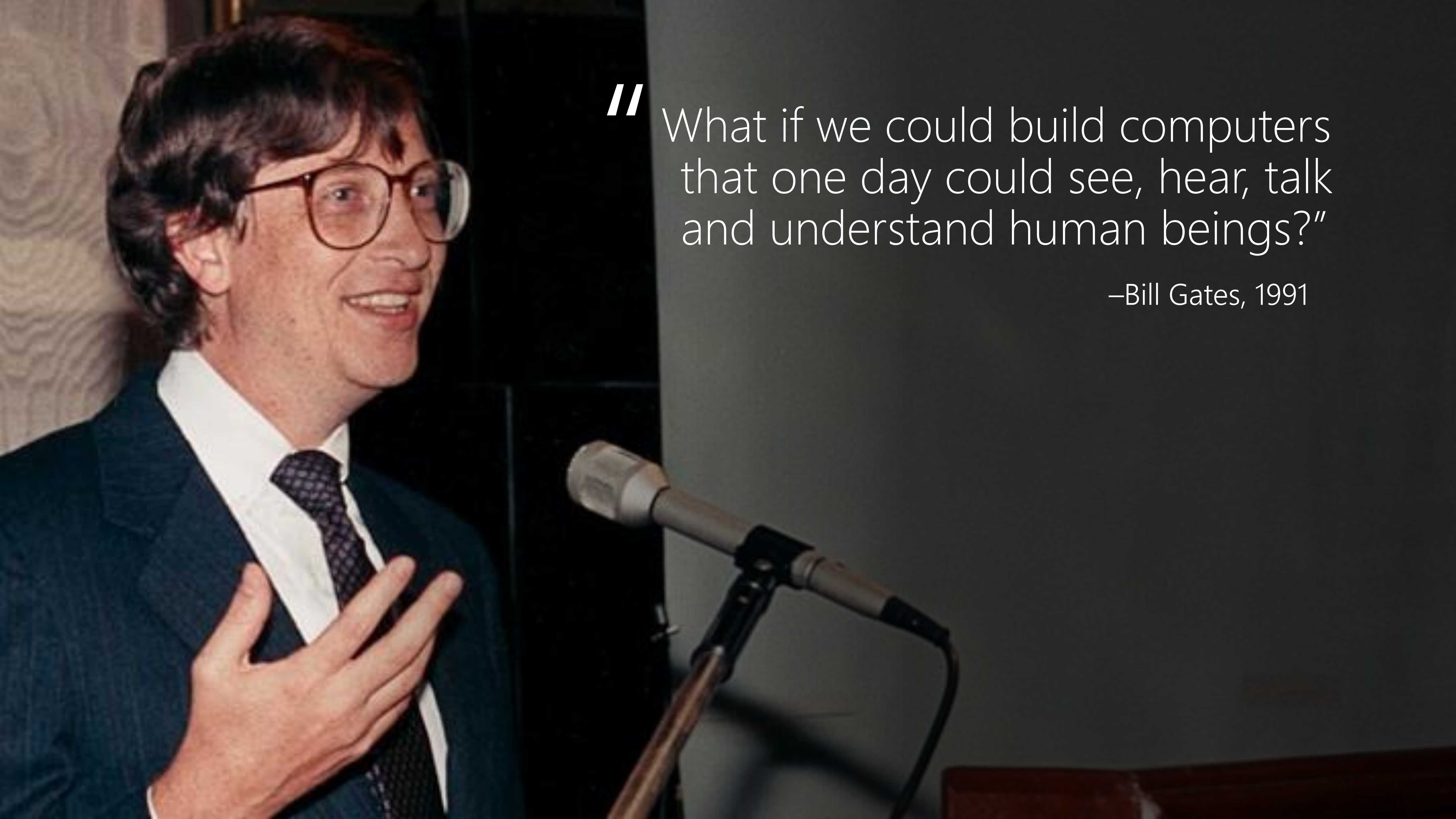
be the set of intensity values of a NXN digitized image which includes only one kind of texture. SAR model for textured image can be expressed as

$$y(s)=\alpha_0+\sum_{r\in D}\alpha(r)*y(s+r)+\varepsilon(s) \qquad (1)$$

where $r=r1+jr2$. α_0,$\{\alpha(r)|r\in D\}$ are model parameters called regressive coefficients. $\varepsilon(s)$ is Gaussian white noice. $y(s+r)$, $r\in D$ are kernel elements of the eqn.(1), and D is the neighbor set of $y(s)$.

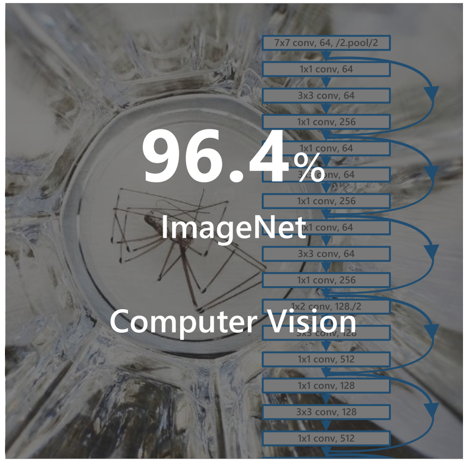
If we put the kernel elements in vector form, two dimensional SAR model could be

No word-processor. Prepared with an ASCII text editor and hand-writing

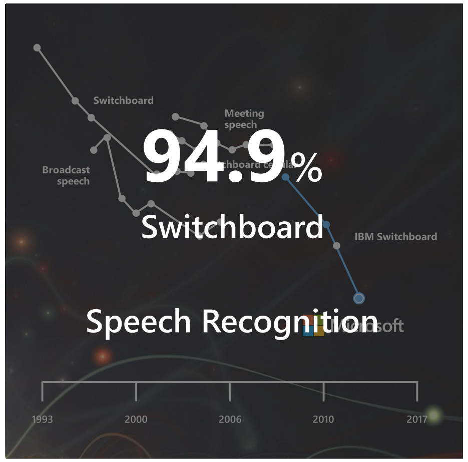

"What if we could build computers that one day could see, hear, talk and understand human beings?"
–Bill Gates, 1991

Microsoft Breakthroughs in AI

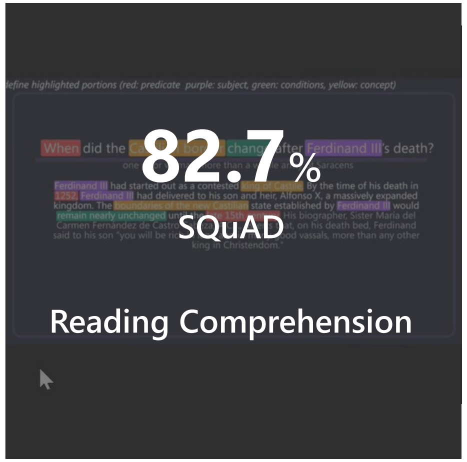
Human Parity Performance



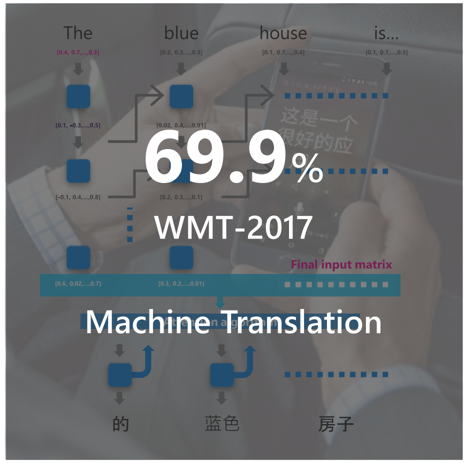
2015



2017



Jan 2018



March 2018

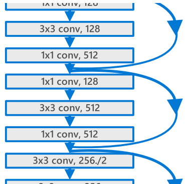
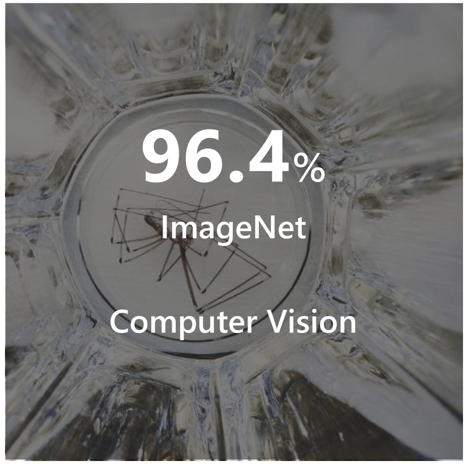


Image Recognition / Classification



2015

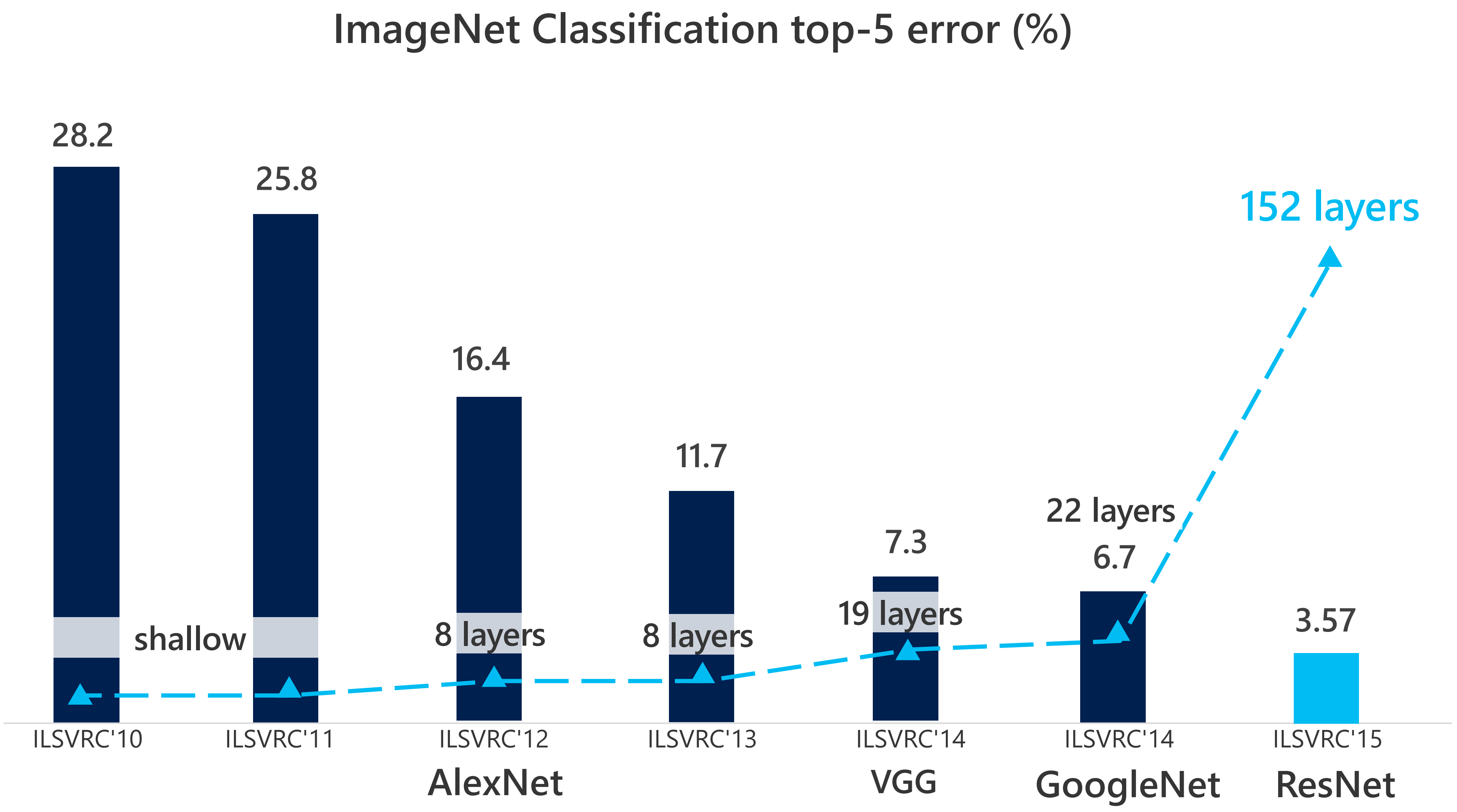
ImageNet
2012 Dataset

- 1000 classes
- 1.2M training images
- 50K validation images
- 100K test images (unpublished)

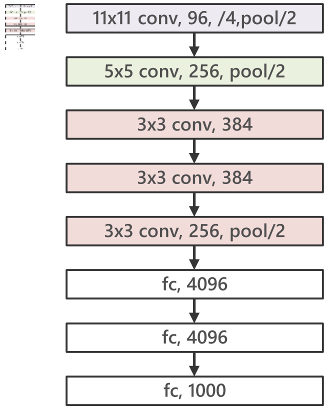
Official measurement
Top-5 error rate

Human performance
Top-5 Error: 5.1%

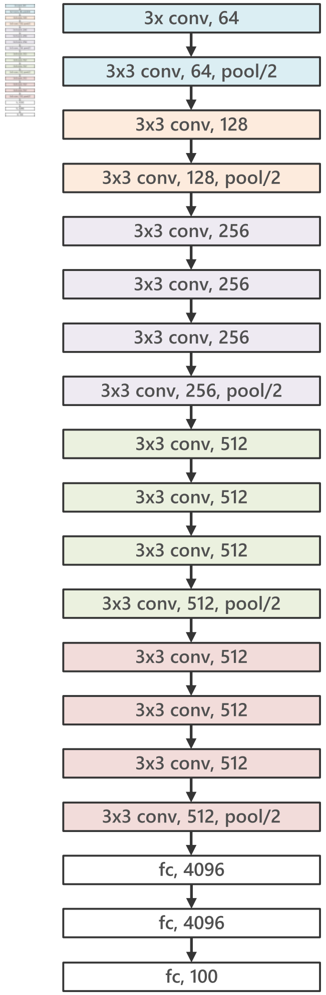
Revolution of Depth



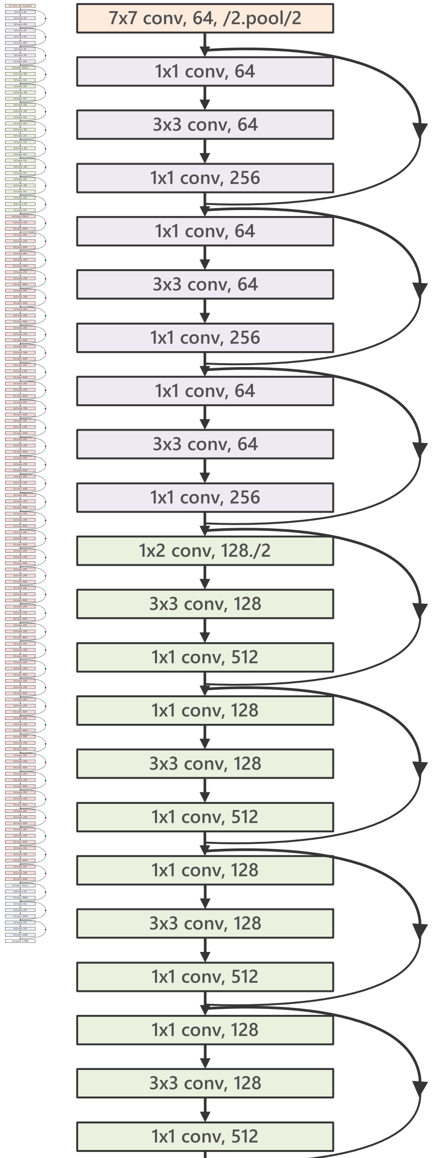
152 Layers ResNet



AlexNet, 8 layers
(ImageNet 2012)



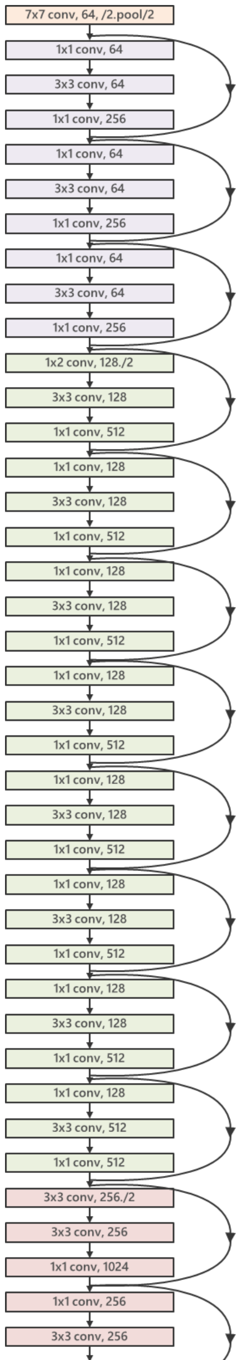
VGG, 19 layers
(ImageNet 2014)



ResNet, 152 Layers
ImageNet 2015)

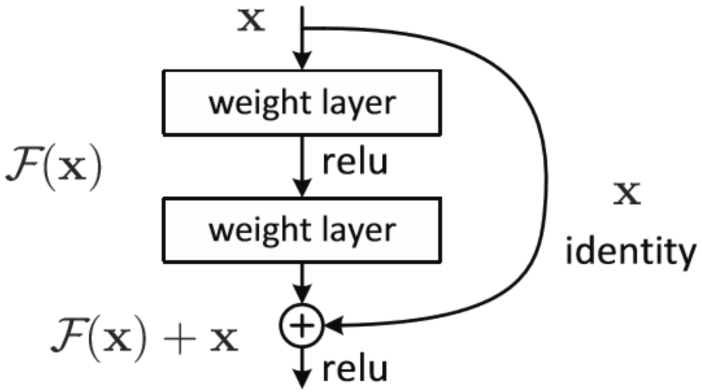
[He et al., Deep Residual Network for Image Recognition, CVPR'2016]

ResNet - Deep Residual Learning



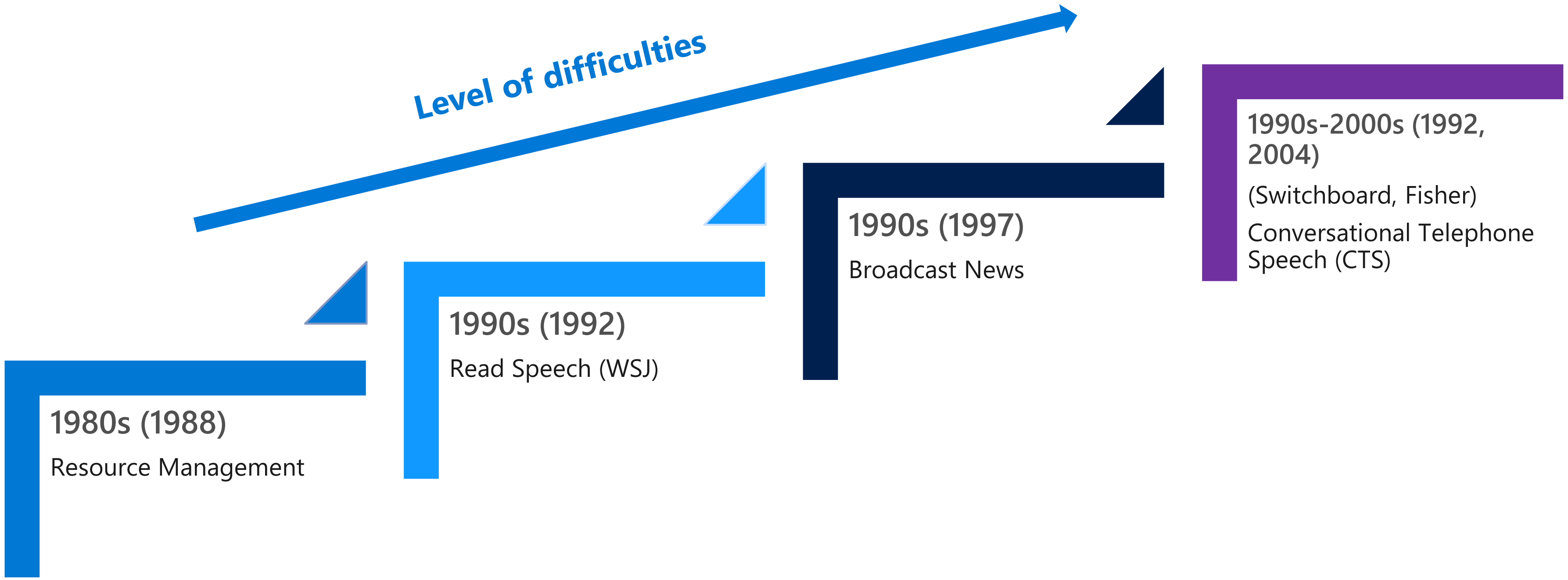
Key enablers

- Residual learning to train a very deep neural network



- The shortcut overcomes the vanishing or exploding gradient
- Use the residual learning unit as a building block to construct a very deep NN
- Ensemble of 6 ResNETs with different depths further improved accuracy

Speech Recognition
(Conversational Telephone Speech)



CTS Data Sets

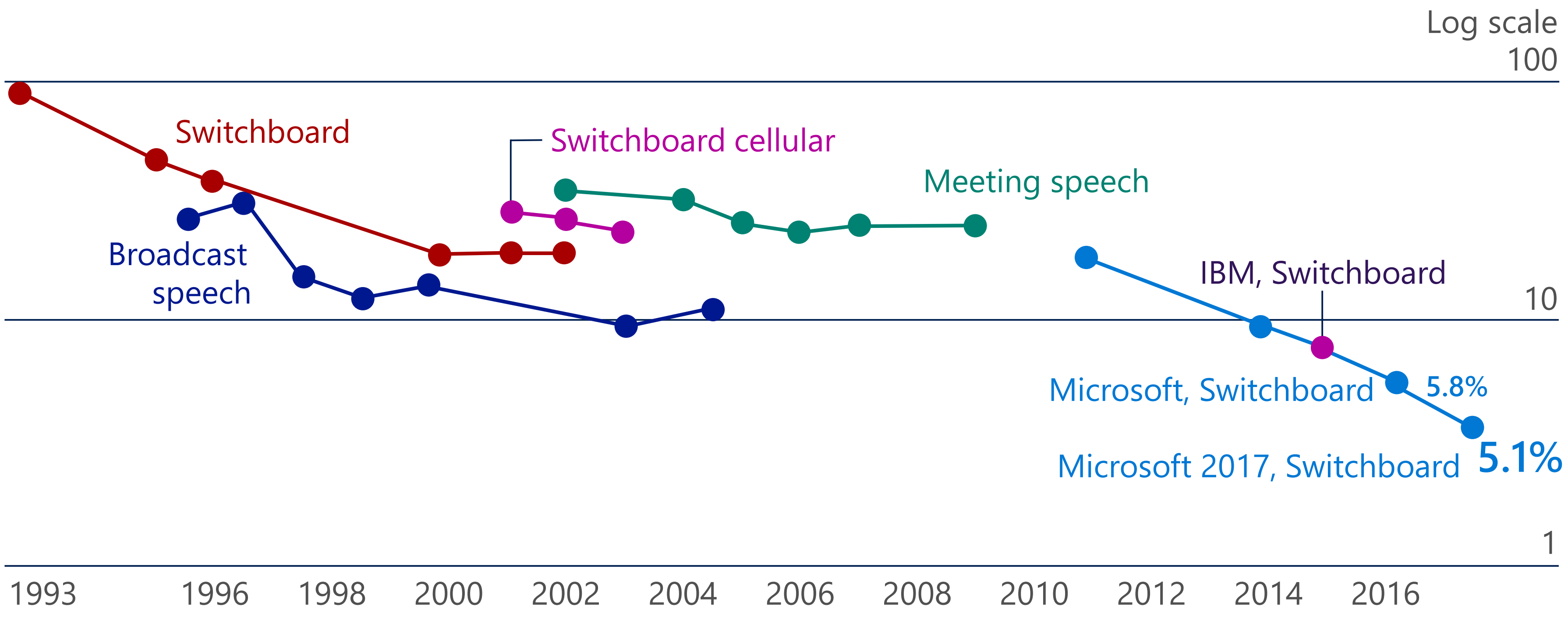
- Acoustic model training data: Switchboard and Fisher) corpora
- Language model training data: LDC corpora 97S62, 2004S13, 2005S13, 2004S11, 2004S09
- Test: NIST 2000 CTS test set

Official measurement
Word Error Rate (WER)
Accuracy = 1 - WER

Human performance
WER: 5.9%
Accuracy: 94.1%

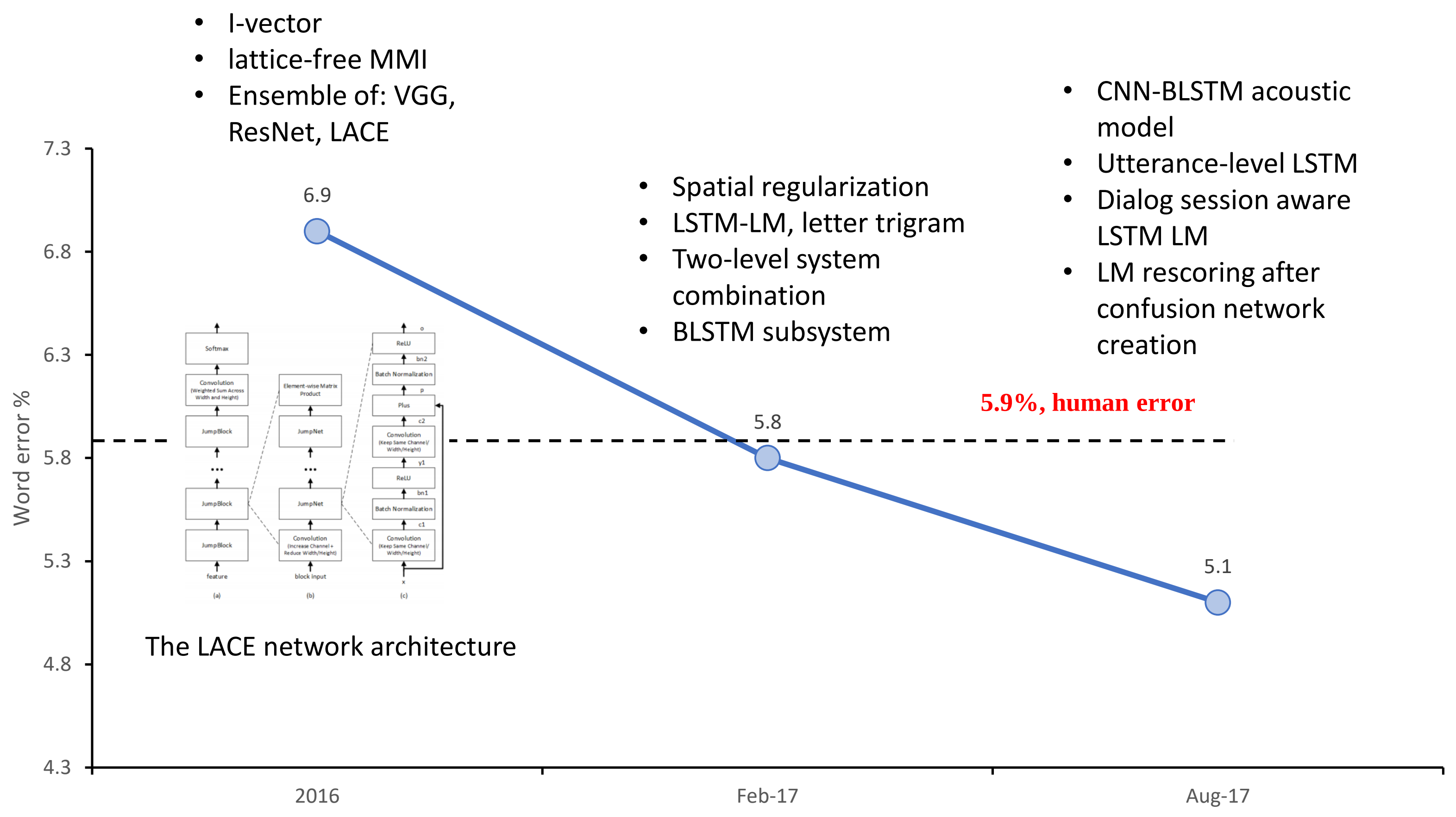
Journey to Human-parity Performance

Speech-recognition word-error rate, selected benchmarks



Source: Microsoft; research papers

Key Enablers



Key to the breakthroughs

- Careful engineering and optimization of Deep Neural Nets (CNN, RNN, LSTM)
- Acoustic modeling using DNNs captures broad context with temporal invariance and frequency-invariance
- Language modeling: RNNs perform better than N-gram
- Ensembles improve robustness

Reading Comprehension

Read a document (passage) and then answer questions about it

Passage (*P*) + Question (*Q*) Answer (*A*)

P Tesla later approached Morgan to ask for more funds to build a more powerful transmitter. **When asked where all the money had gone, Tesla responded by saying that he was affected by the Panic of 1901**, which he (Morgan) had caused. Morgan was shocked by the reminder of his part in the stock market crash and by Tesla's breach of contract by asking for more funds.

Q On what did Tesla blame for the loss of the initial money?

A Panic of 1901

SQuAD

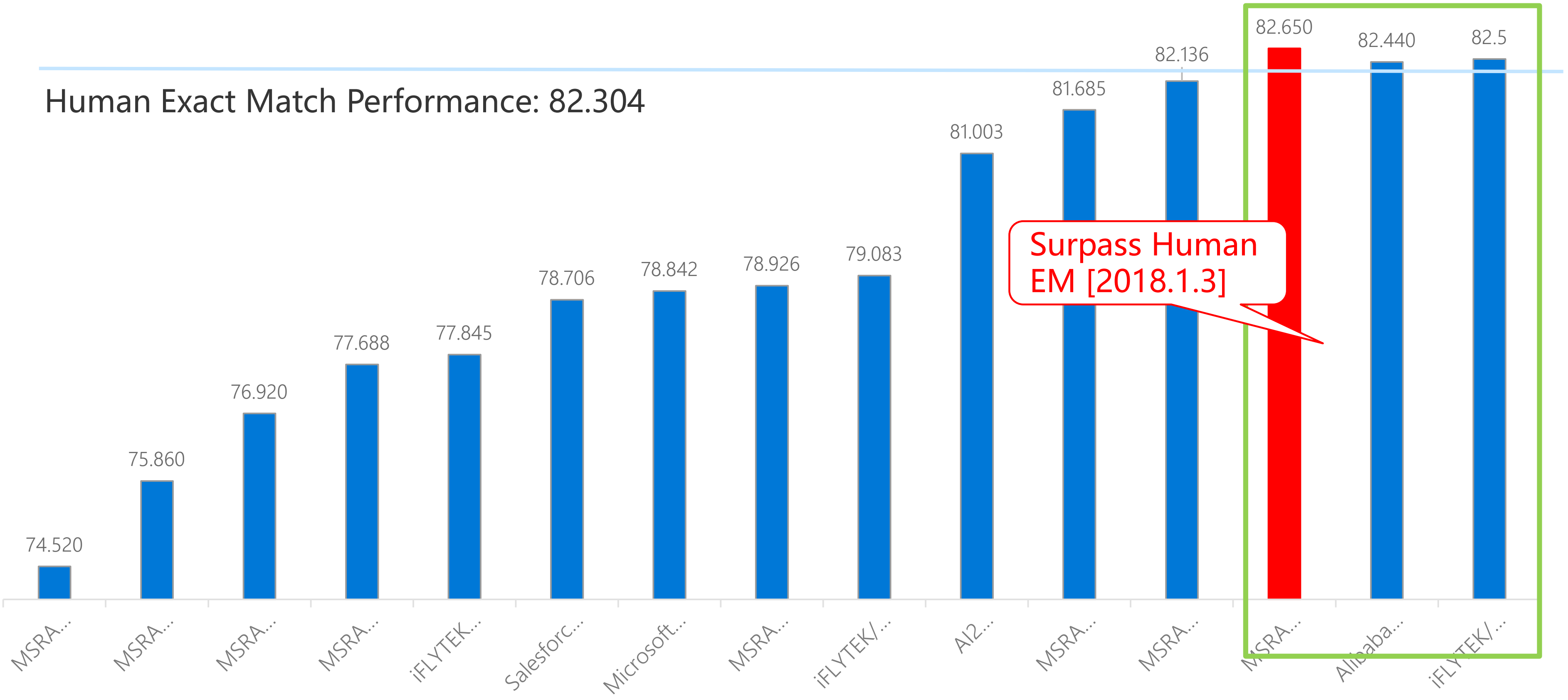
Stanford Question Answering Dataset

- 87,599 training samples
- 10,570 development set
- >10K test samples (unpublished)

Official measurement
Exact Match Accuracy & F1

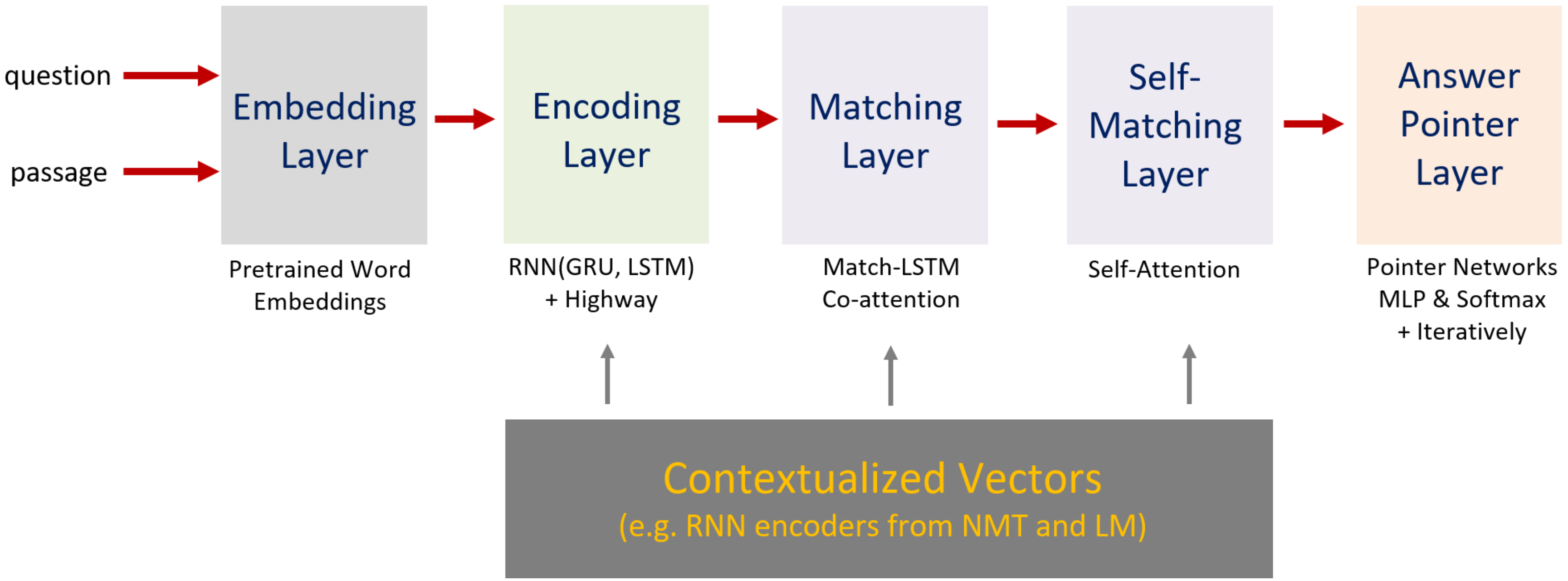
Human performance
Exact Match: 82.3%

Journey to Human-parity Performance on SQuAD 1.1



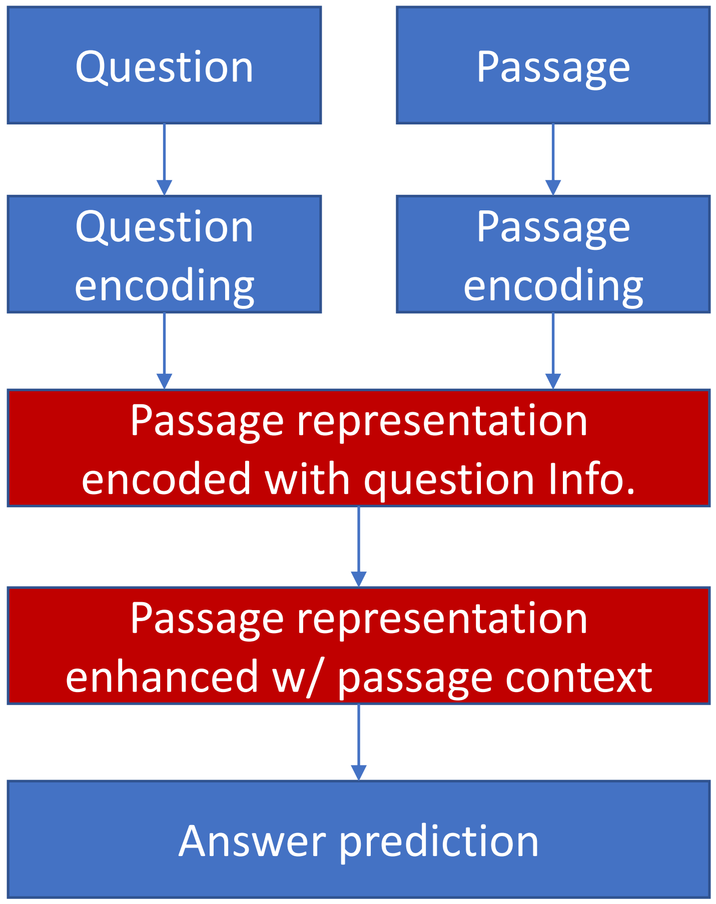
Best System EM Scores on SQuAD Machine Reading Comprehension Dataset (Dec. 6, 2016-Jan. 26, 2018)

R-Net for Reading Comprehension



Key Enablers

- The end to end framework progressively encoding question and passage context into a refined passage representation
- An additional gate to the attention-based RNN, to account for different importance of a passage word to the question
- A novel self-matching mechanism, to encode the context information in the final passage representation



Machine Translation (MT)

Sampled from WMT2017 Chinese-English task

Source input	有 线索 人士 请 拨打 旧金山 警察局 举报 电话 4 15- 575 - 44 44 。
NMT output	For clues, call the San Francisco Police Department at 415-575 - 4444.
Human reference	Anyone with information is asked to call the SFPD Tip Line at 415-575-4444 .

Source input	他 的 职业 生涯 如 过 山 车 一般 。
NMT output	It has been a rollercoaster ride .
Human reference	His career is like a roller coaster.

Source input	霍夫 施泰特尔 表示 : " 这 将 由 检察官 来 确定 " 。
NMT output	That 's what the prosecutor must determine , " said Hofstetter .
Human reference	Mr Hoff Steitel said: "It will be up to the prosecutors to determine.

Machine Translation (MT) Journey

1980s
Traditional MT

1990s
Statistical MT

2010s
Deep Neural MT

2018
Microsoft AI is the first to reach Human Parity on Chinese to English news translation

Microsoft's Human-parity Machine Translation System

- WMT newstest2017, Chinese→English
- Compare with translations by human experts
- Score translations by system and by human experts w.r.t. the source sentences
- Reach human-parity performance

#	Ave%	Ave z	System
1	69.0	0.237	Combo-6
	68.5	0.220	Reference-HT
	68.9	0.216	Combo-5
	68.6	0.211	Combo-6
2	67.3	0.141	Reference-PE
3	62.3	-0.094	Sogou
	62.1	-0.115	Reference-WMT
4	56.0	-0.398	Online-A-1710
	54.1	-0.468	Online-B-1710

Table 4: Human Evaluation Results

Microsoft MT system

Earlier Best performer

Human Experts

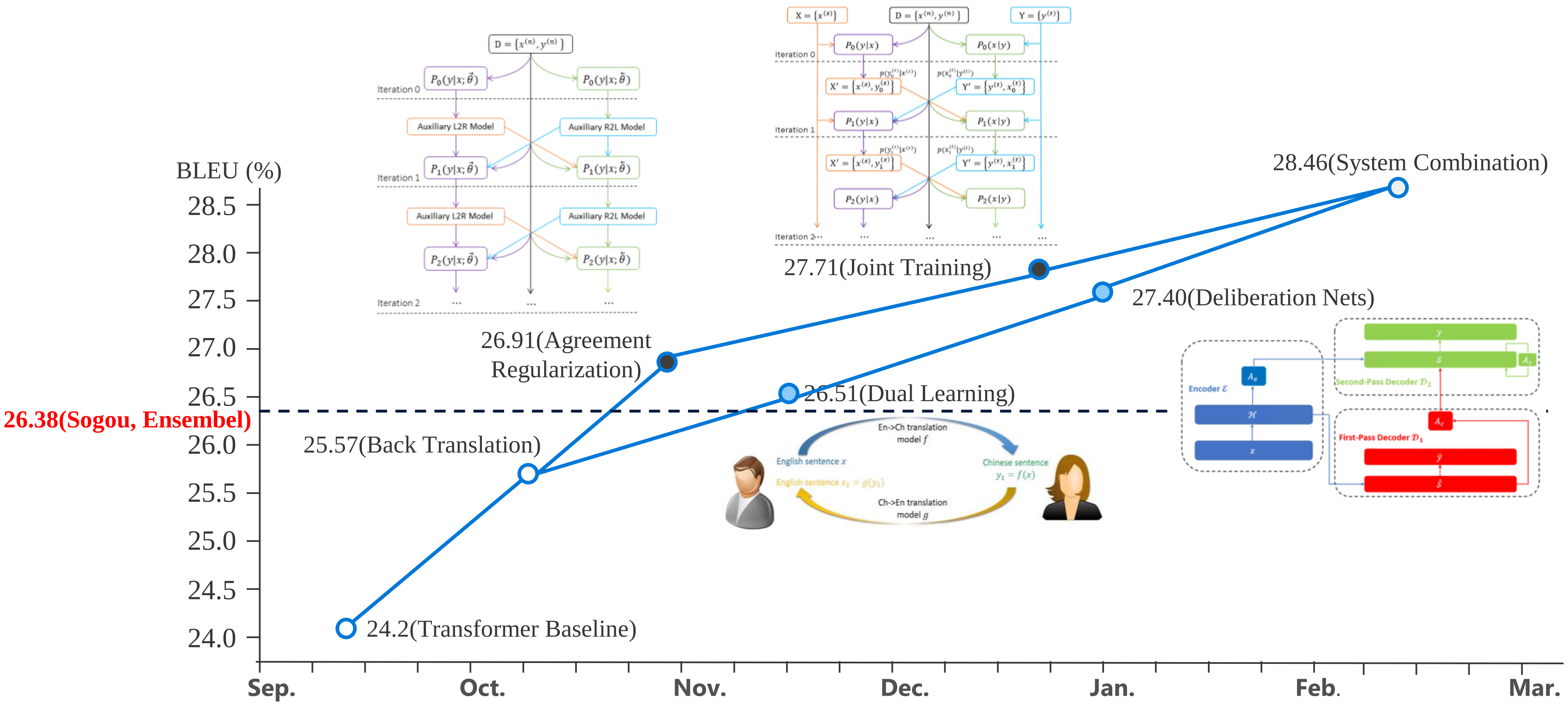
Crowd-sourcing translation

Microsoft product

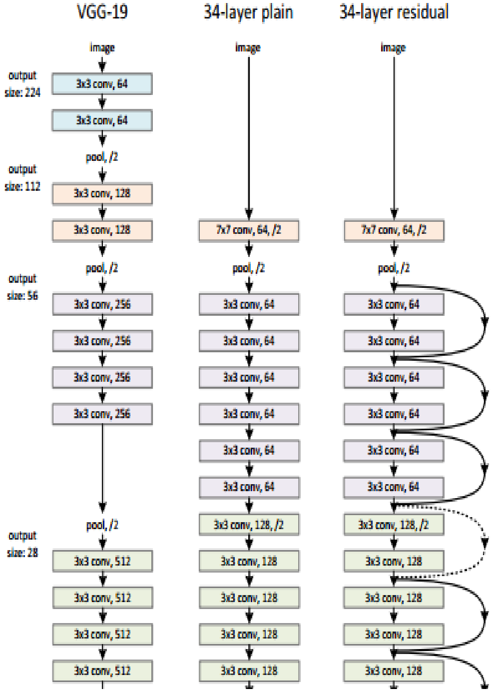
Google product

Achieving Human Parity on Automatic Chinese to English News Translation, Hany Hassan et al, https://arxiv.org/pdf/1803.05567.pdf

Key Technology Enablers



Key Technology Enablers to Deep Learning



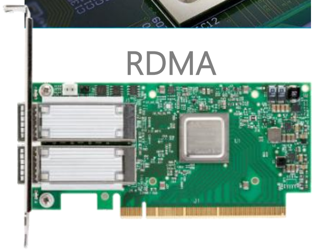
TensorFlow

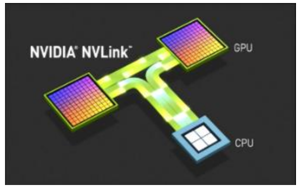
Microsoft CNTK

Caffe2

dmlc mxnet

PYTORCH



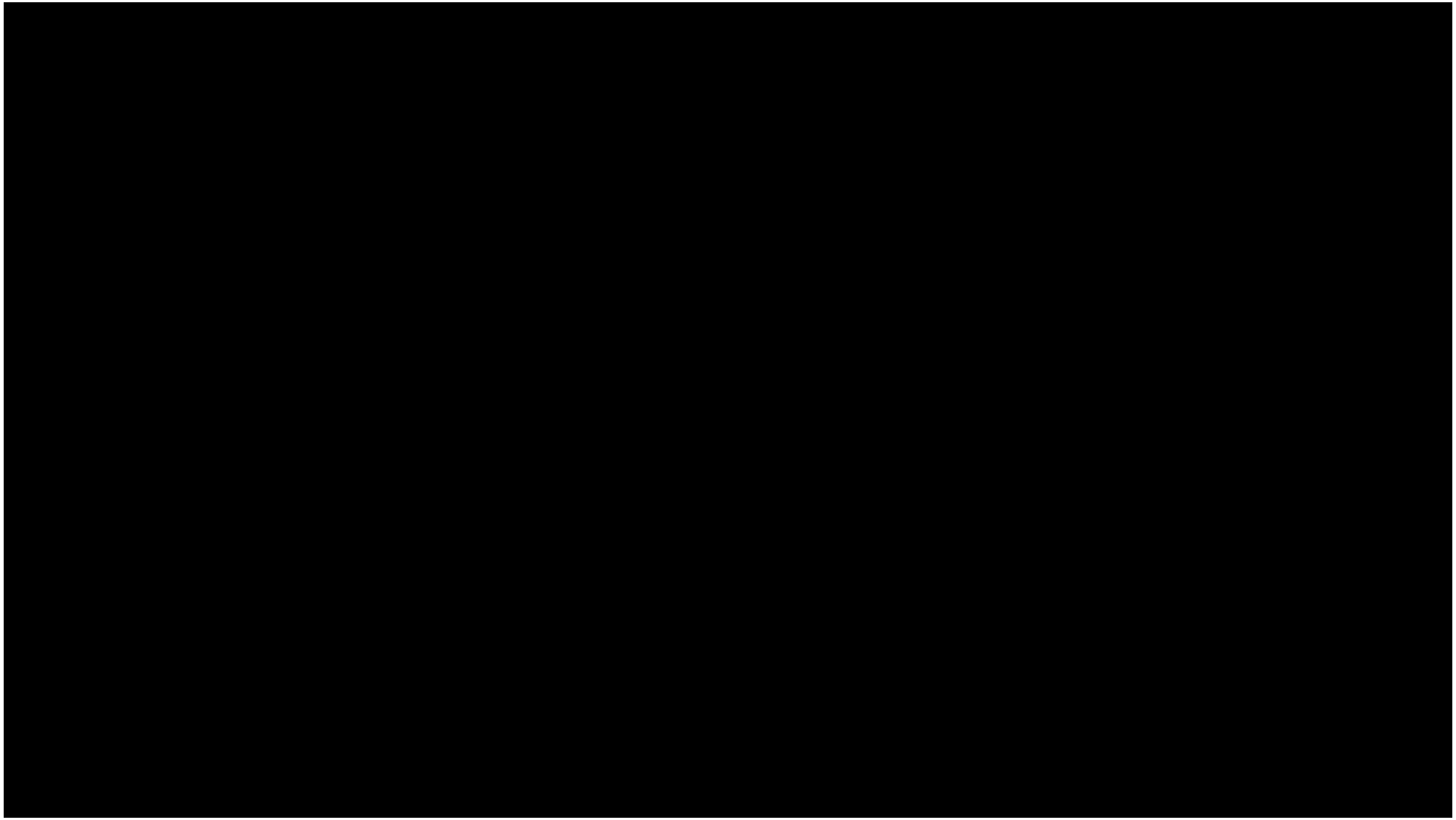


Big Data

Algorithms and Frameworks

Computing Power

Modern Meeting Demo



What is OCR (Optical Character Recognition)?

Twenty Years of Document Image Analysis in PAMI

George Nagy, *Senior Member, IEEE*

Abstract—The contributions to document image analysis of 99 papers published in the *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* are clustered, summarized, interpolated, interpreted, and tactfully evaluated.

Index Terms—Document image analysis, image processing, OCR, character recognition, forms processing, graphics recognition.

1 PAMI AND DIA

INSTEAD of attempting to survey the entire field of document image analysis (DIA), we review only results reported in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Parochial as this may seem, it gives a sharp, well-defined cross-section of the evolution of DIA research. The 99 relevant papers that were found (less than five percent of all articles) are contemplated from a perspective bolstered by lively students, eclectic reading, participation in conferences, and discussions with knowledgeable and opinionated colleagues.

This section considers the role played by *PAMI* in relation to other sources of published information and current commercial practice, differentiates document image analysis from allied disciplines, and describes its major constituents. It should be skipped by old hands with their own cognitive map of the field. The next five sections summarize the DIA tasks addressed in *PAMI* in the last two decades. Only *PAMI* papers are cited, but a short bibliography provides additional entry points to the literature. The conclusion is the author's classification of the domain into *problems solved* and *problems remaining*.

1.1 PAMI vs. Other Sources

It would appear that 99 articles among the several thousand published about DIA in the past 20 years can represent at best a fraction of the state of the art. However, *PAMI* covers much more ground than this ratio would indicate. Before *PAMI*'s birth in 1979, character recognition research appeared mainly in *IEEE Transactions on Computers (TC)* (*IEEE Transaction on Electronic Computers (EC)* until 1968) and in the *Proceedings of the IEEE*, in addition to occasional specialized conferences and workshops. (One of the earliest, the 1966 IEEE Pattern Recognition Workshop in Puerto Rico, resulted in the *IEEE Computer Group's*[1] pattern recognition database.) The *Journal of Pattern Recognition* has regularly published articles on character recognition since its inception in 1971, as has *Pattern Recognition Letters*. Relevant articles appear occasionally in the *IEEE Transactions on Information Theory, Systems, Man, and Cybernetics, Neural Networks,* and *Image Processing*, as well as in a dozen commercially published journals of artificial intelligence, pattern recognition, computer vision, and image processing. The best source of current trade news is the monthly, *Imaging and Document Solutions*. In 1998, Elsevier launched the *International Journal of Document Analysis and Recognition* with the goal of capturing the fractionated DIA and OCR literature. One indication of the dispersal of this literature is that few of the citations in *PAMI* articles reference *PAMI*.

Since 1973, the biennial International Conference on Pattern Recognition (ICPR) has been a steady source of ideas. It has been supplemented since 1991 by the International Conference on Document Analysis and Recognition (ICDAR). Worthwhile contributions have appeared at the annual SPIE Document Recognition and Retrieval (DR&R) symposia in San Jose, and at the peregrinating biennial Document Analysis Systems (DAS) and Structural and Syntactic Pattern Recognition (SSPR) workshops, each of which attracts about 100 participants. During its five-year life span, the Symposium on Document Analysis and Information Retrieval (SDAIR) fostered interaction between DIA and IR specialists. It also featured the results of a large-scale in-house evaluation of commercial OCR technology. Several countries have instituted national conferences on OCR or DIA. The articles found in *PAMI* reflect the international constituency of document analysis. We are often reminded that English is blessed with one of the simplest scripts in the world.

1.2 PAMI vs. Current Practice

We consider DIA and OCR as essentially engineering disciplines, although a case can be made for a more fundamental role. Published work (not only in *PAMI*) has been moderately successful in anticipating emerging applications. The many papers on hand-printed and handwritten character recognition (cf. article by Plamondon and Srihari in this issue) probably did contribute to current products, but some of the print recognition methods explored by researchers risk lagging the capabilities of give-away shrink-wrapped page readers. The research emphasis in

1. Predecessor of the Computer Society.

• *The author is with the Rensselaer Polytechnic Institute, Troy, NY 12180. E-mail: nagy@ecse.rpi.edu.*

Manuscript received 10 Aug. 1999; accepted 12 Oct. 1999.
Recommended for acceptance by K. Bowyer.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 110406.

TABLE 3
A Document Taxonomy

Type	Example	DIA Task	Ancillary data
plain text (narrative or descriptive)	Moby Dick, Gettysburg Address	extract correct word order	English lexicon
newspaper, magazine	NY Times, Vogue	separate and reassemble articles; pointers to illustrations	publication-specific format
scholarly & technical text	IEEE-PAMI, Dr. Dobbs Journal	index: author, title, page; pointers to refs, figs, tables, footnotes, equations	abbreviations, acronyms, units
formal text	program listing, chess, bridge, recipe	extract executable, or compilable, form	program, chess, bridge syntax
letter, envelope	information request, complaint, recommendation	extract routing info; index: sender, date, subject	directories
directory	telephone directory, street index	extract name-attribute pairs	previous edition
structured list	organization chart, table of contents, catalog	recover hierarchy; cross-references	previous edition
business form	order, invoice, subscription, survey, IRS-1040	link field content to dbms; convert to SGML or XML format;	formatted data, dbms, workflow system, lexicons
engineering drawing	assembly or part drawing; isometric view	convert to CAD format	part lists, drawing standards
schematic diagram	circuits, utility maps	extract net list or convert to CAD format	P-SPICE, manhole inventory
map	topographic quad, street map, road map	convert to GIS format	gazetteer, other maps, GIS
music score	Moonlight Sonata	recover MIDI representation	music syntax
table	airline schedules, stock quotes	construct formal model: headers ↔ entries	airline and stock abbreviations, previous edition

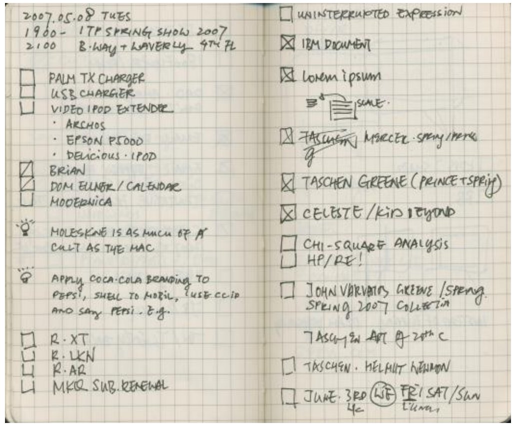
compression methods simply to avoid disk access during page analysis. Run-length coding (RLC) and Freeman chain codes were used early on. Methods that came along later include reduced terminal sequences of context-free grammars [43], coding on hexagonal meshes [94], production rules for subblocks [58], and filtered contours [10]. The July 1980 special edition of the *Proceedings of the IEEE* on digital encoding of graphics contains many excellent surveys, mostly targeted at facsimile. For lossless, bilevel page compression, JBIG is gradually replacing CCITT-G3 and G4. The major remaining application of character encoding is font libraries.

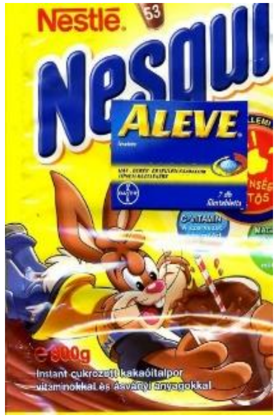
2.2 Binarization

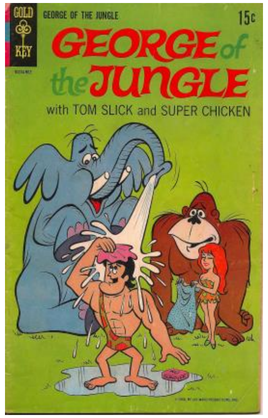
Most early document scanners had hardware reflectance thresholds, but current scanners typically produce 8-bit gray-scale (or color) output. Researchers from the University of Oslo and Michigan State University conducted a sustained, thorough comparison and evaluation of published adaptive binarization methods (including their own) on hydrographic charts [83], [84], [85], [86]. Niblack's method, based on a threshold set below the mean gray-level of a 15×15 window by a fixed fraction (0.2) of the standard deviation of the gray-levels, gave the best results on their maps. (A small modification is necessary when it is evident that the entire window is covered by a large foreground blot). They recommended postprocessing with the method of Yanowitz and Bruckstein, which iteratively creates a threshold surface that is essentially a low-pass-filtered version of the reflectance map. They also reported that character segmentation and recognition did not necessarily benefit from direct gray-scale processing as opposed to adaptive binarization [86].

Textured backgrounds are particularly difficult to handle. Liu and Srihari [53] provide a solution for postal address readers. It requires: 1) preliminary binarization

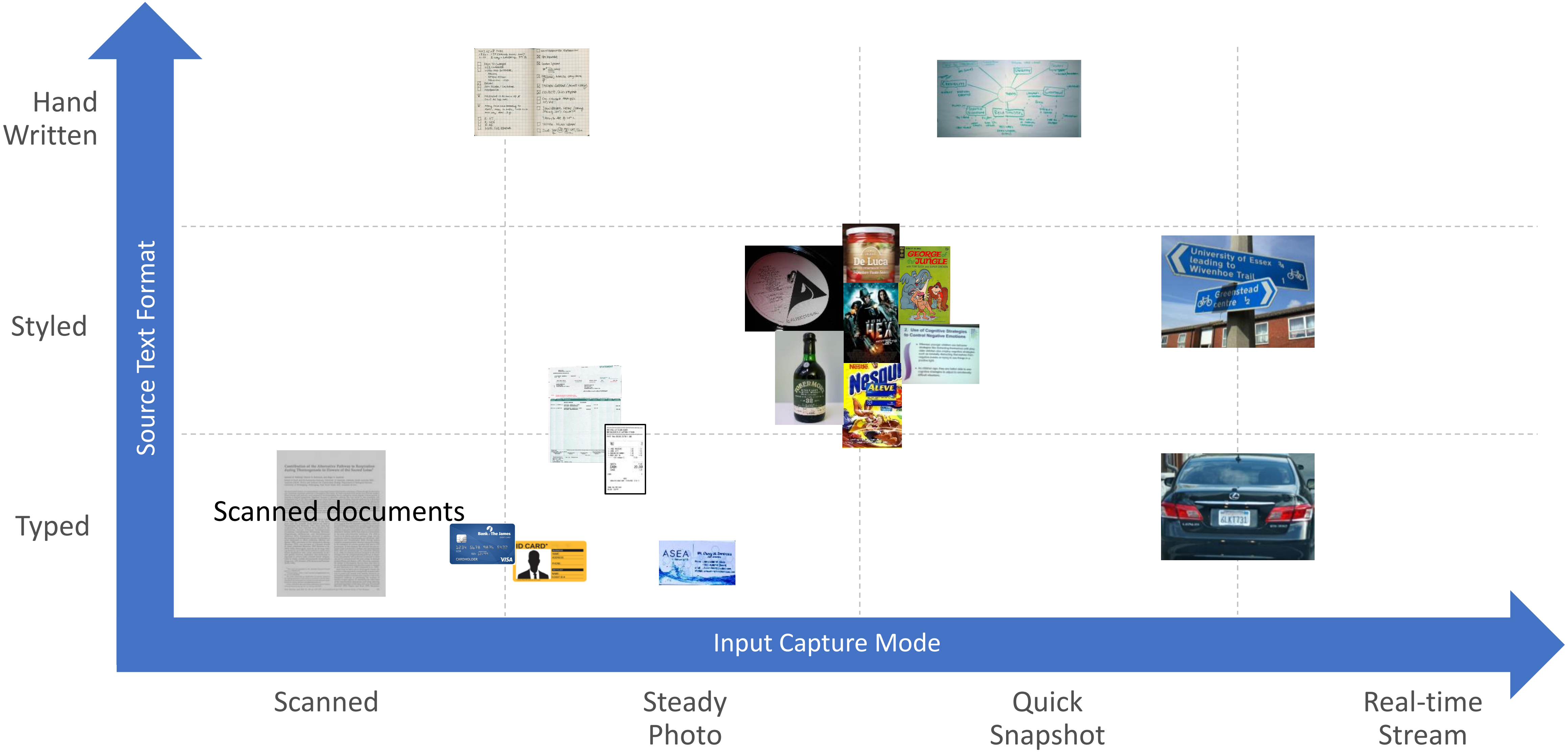
OCR in the Wild



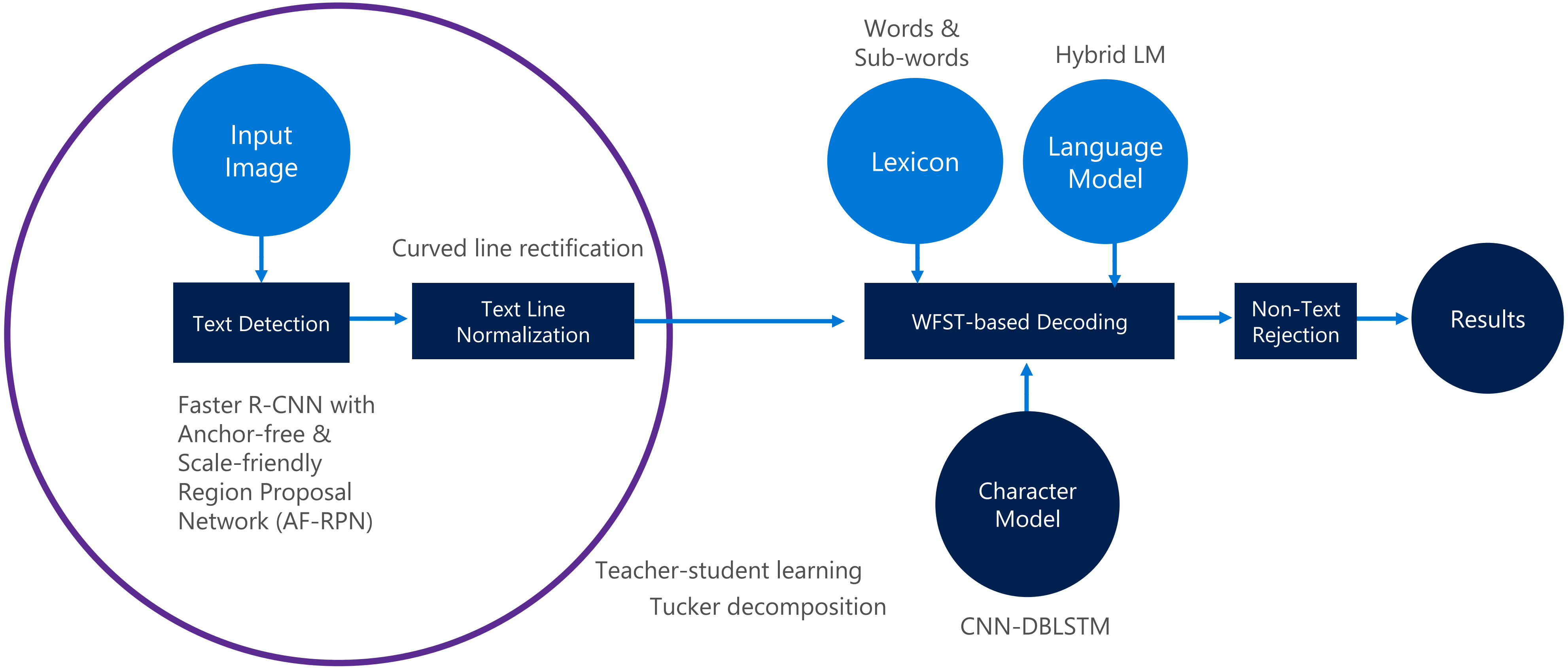




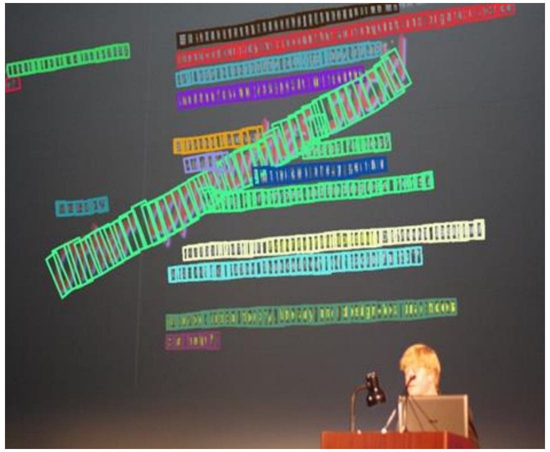
Expanding Scenarios in an Intelligent Cloud/Edge World

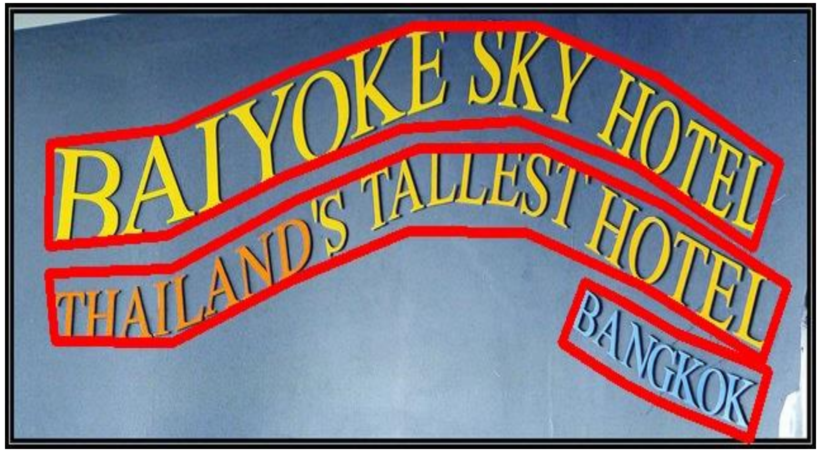


Architecture of New Printed OCR Engine



Variabilities of Text Objects





Complex Backgrounds for Text Objects

Greeting Card

Street View

License Plate

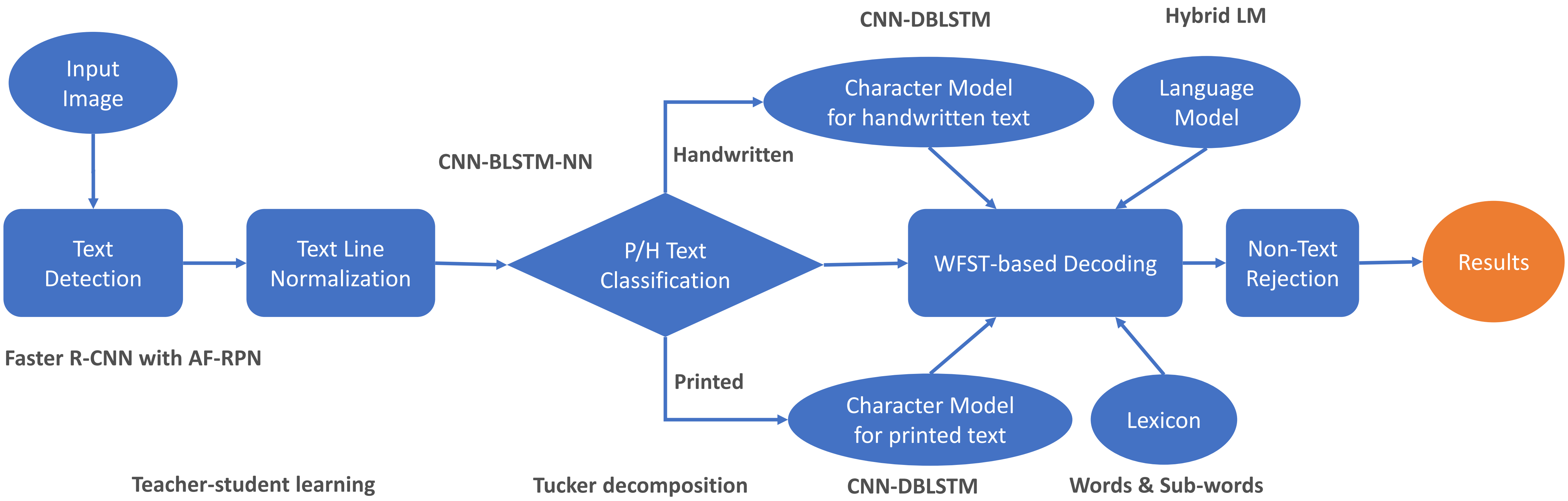
Results of Printed OCR Engine (WER in %)

Scenarios	Prior Industry Leader	Microsoft MS-Old	MS-New	MS-New *vs.* Prior Leader
Document	7.7	14.8	2.8	63.6%
Invoice	11.7	26.8	6.6	43.6%
Receipt	13.7	40.1	11.8	13.9%
Business Card	14.6	41.7	9.2	37.0%
Slide	30.7	56.2	13.6	55.7%
Menu	23.7	38.7	14.7	38.0%
Book Cover	31.7	55.9	14.0	55.8%
Poster	26.6	47.6	15.8	40.6%
GIF/MEME	29.5	53.0	11.8	60.0%
Street View	28.3	61.2	16.8	40.6%
Product Label	42.3	66.7	24.3	42.6%

Progress of Handwritten OCR Engines

			Microsoft		
E2E Evaluation	**Another Industry Player**	**Another Industry Player**	**V1.0**	**V2.0**	**V3.0**
Recall (%)	30.6	52.5	53.0	70.2	74.9
Precision (%)	35.5	53.6	49.8	65.8	70.5
Memory	N/A	N/A	6GB	300MB	350MB
Deployment	Cloud Vision API (2017/04)	Cloud Vision API (2018/03)	OneNote (2016/03)	Cognitive Services (2017/04)	Cognitive Services (2018/05)

Architecture of OneOCR Engine (Ongoing)

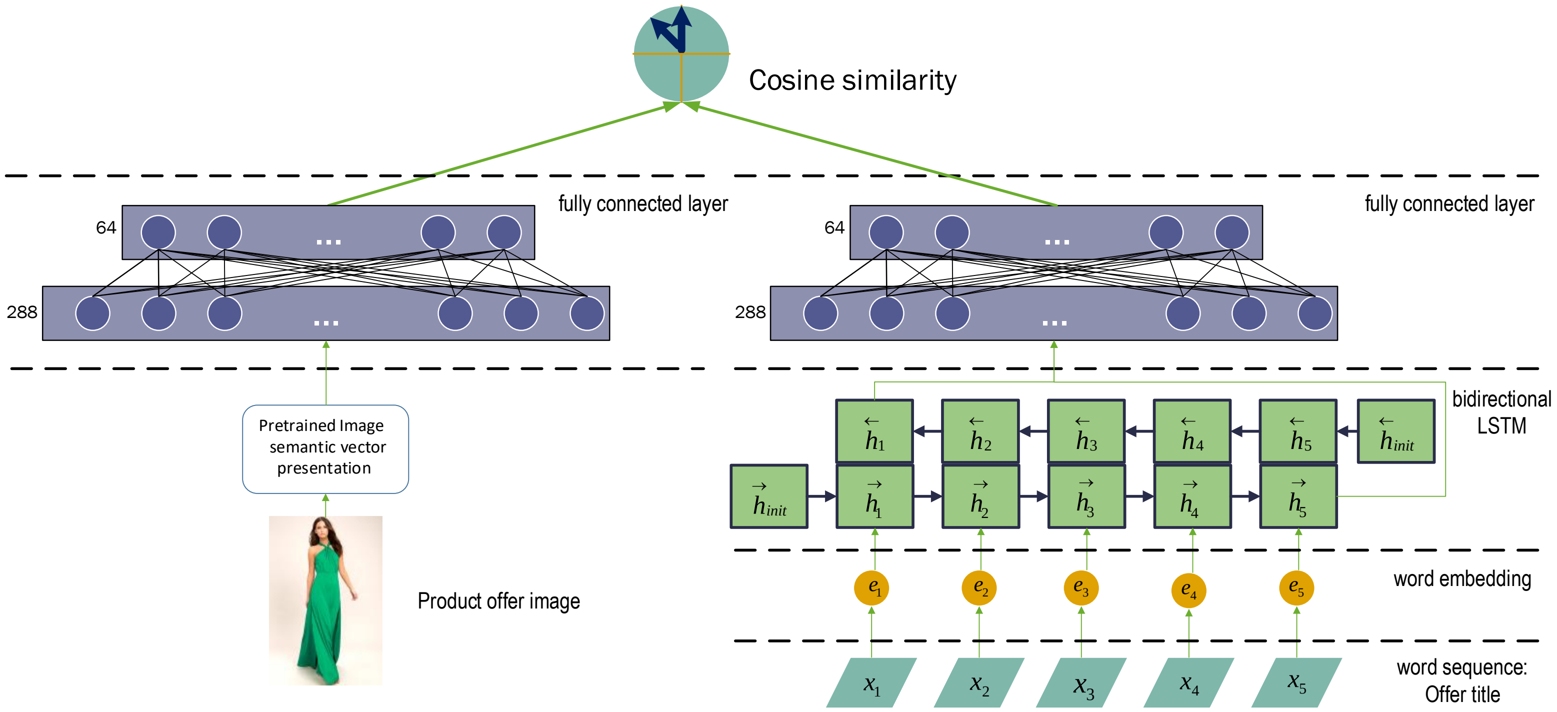


One engine to deal with printed, handwritten, and mixed printed/handwritten OCR

Contact: Dr. Qiang Huo

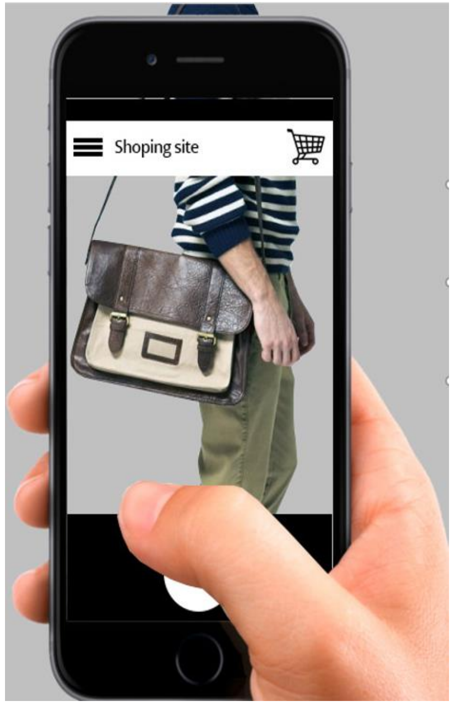
Visual Shopping Assistant

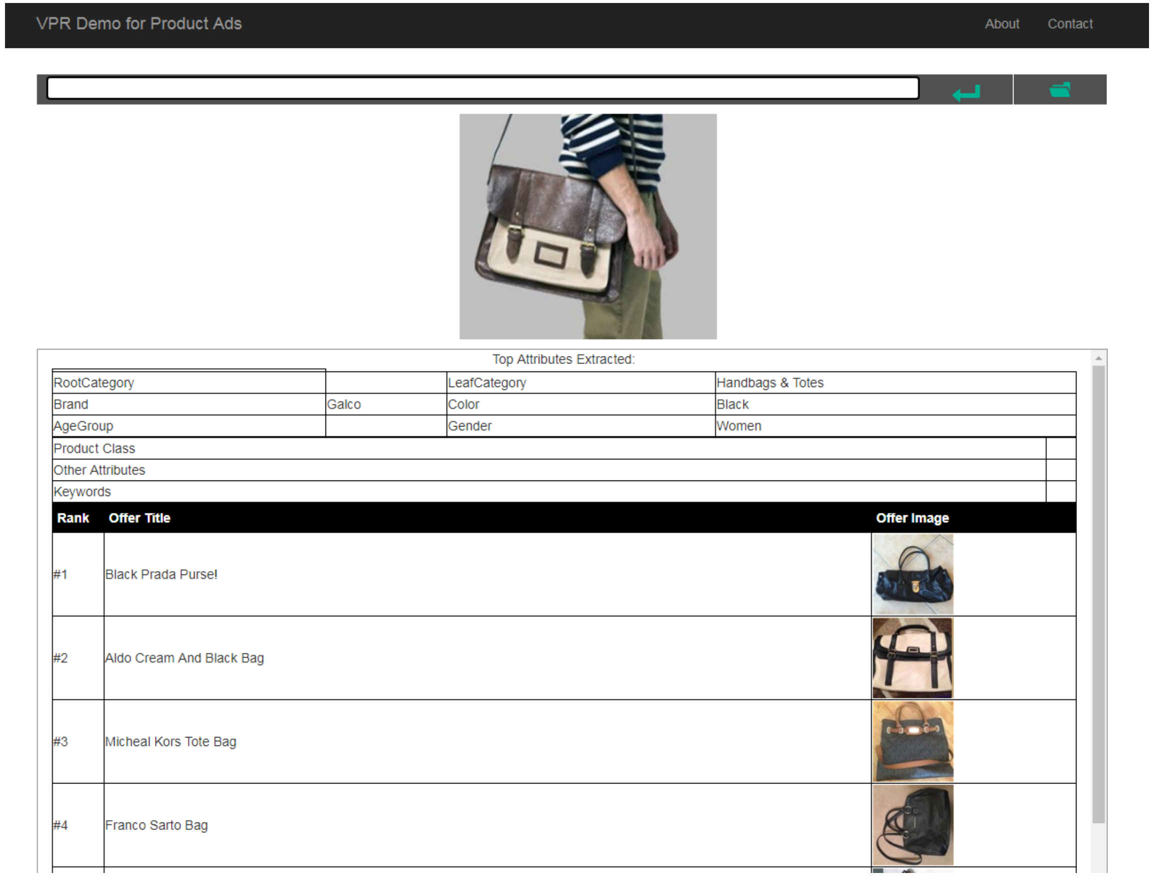
Image-Text Co-modeling



Offer title, Category, Brand, Product class, Model, Gender, Color

Select Object to Shop for in Time





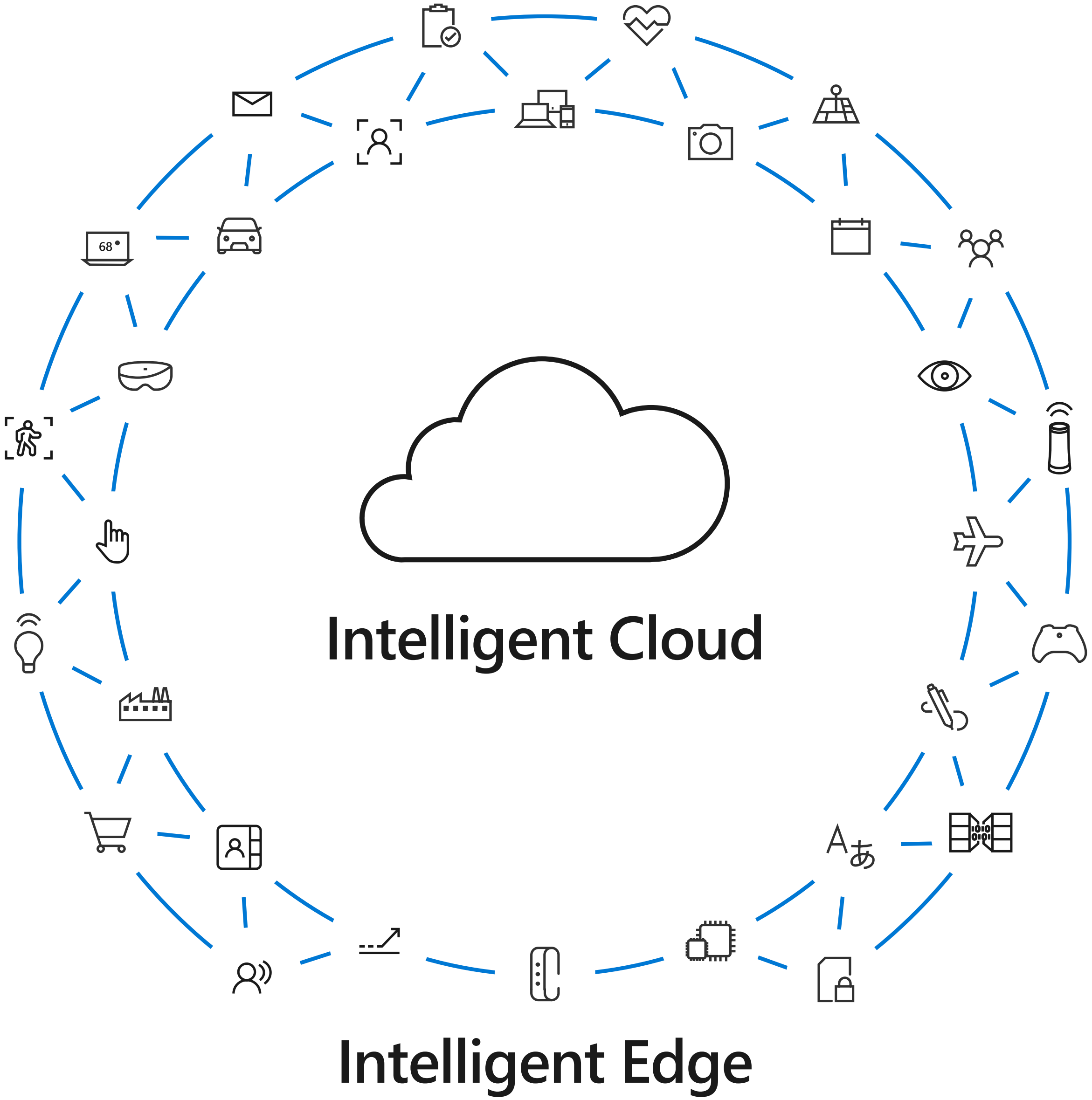
Visual Shopping Assistant Demo

[Camera shooting]

[Photo cropping]

Concluding Remarks

- Accelerated pace of breakthroughs in special AI
- Powerful in solving more real-world problems
- Intelligence everywhere



Acknowledgment

- Harry Shum
- Xue-Dong Huang
- Hsiao-Wuen Hon
- Ming Zhou
- Furu Wei
- Qiang Huo
- Bruce Zhang
- Ying Shan
- Keng-hao Chang
- Other Colleagues at Microsoft AI & Research

Microsoft

Q&A