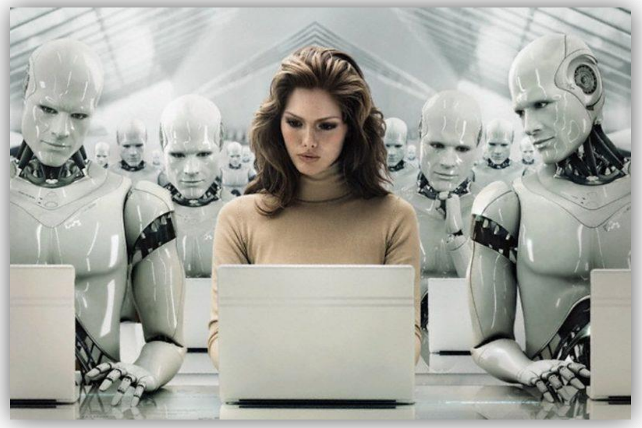
Example

Example

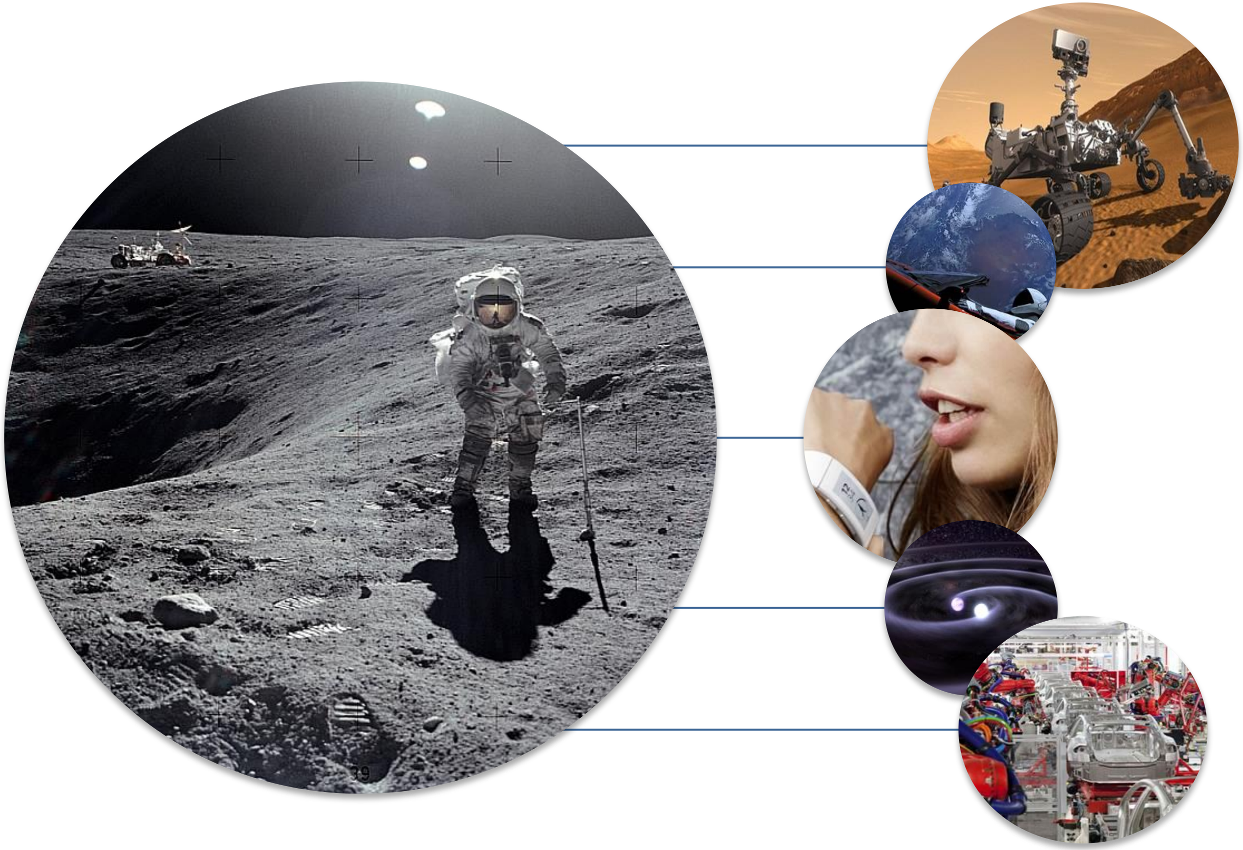
Intelligent Machines



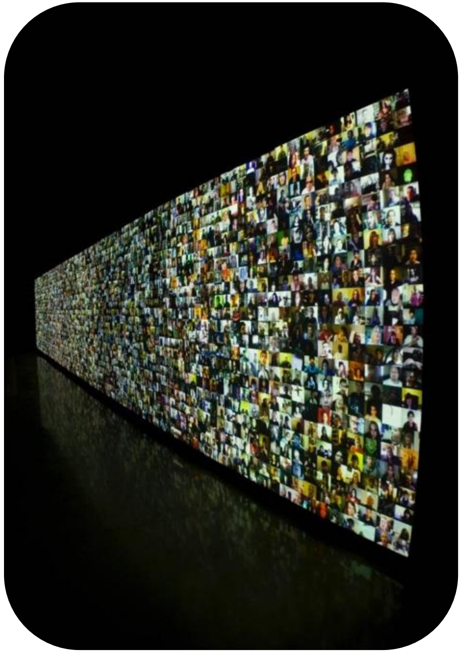
Deep Representation **Learning** for **Multimedia** Data Analysis

Dr. Haojin Yang

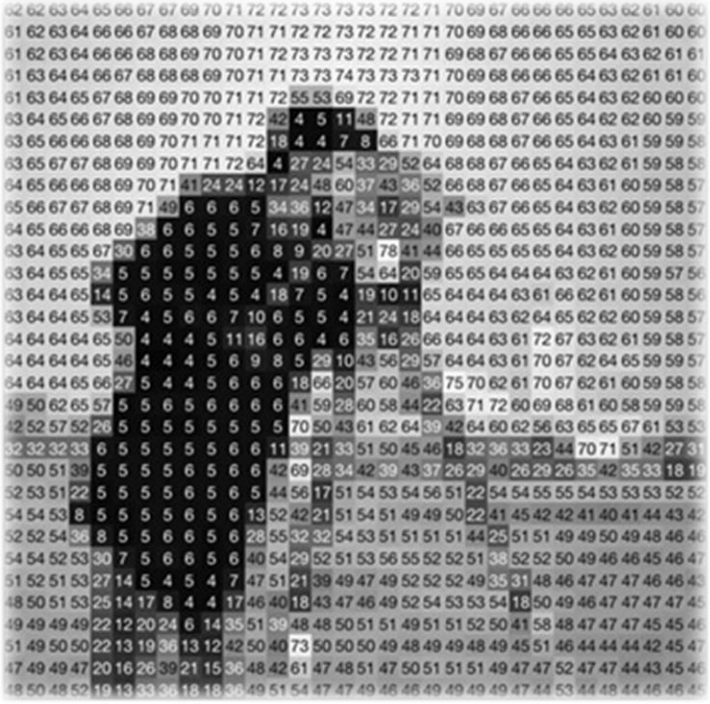
Technologies



Technologies

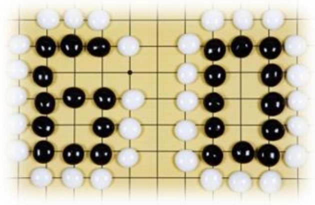


Why Machine Vision so Hard?



Representative Features

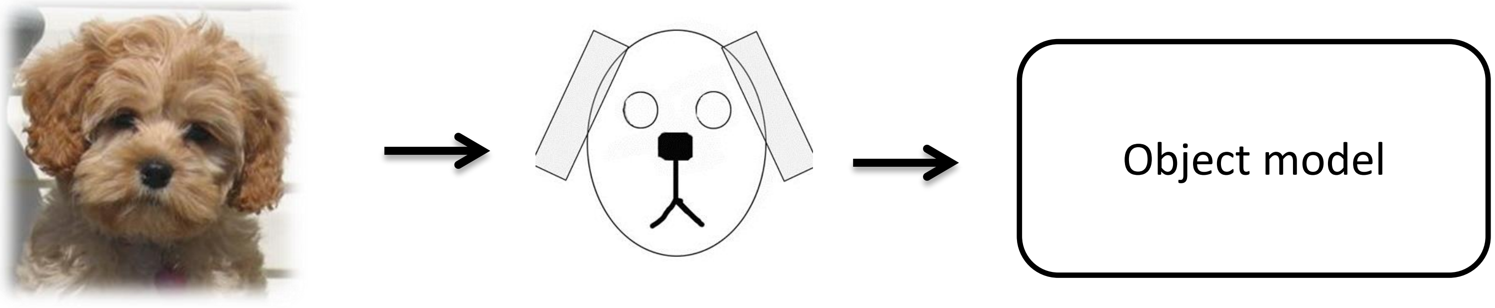
- Raw representations
  - Speech: phoneme
  - Language: letter
  - Image: pixel



3^{361} states $>$ sum of the universe's atoms

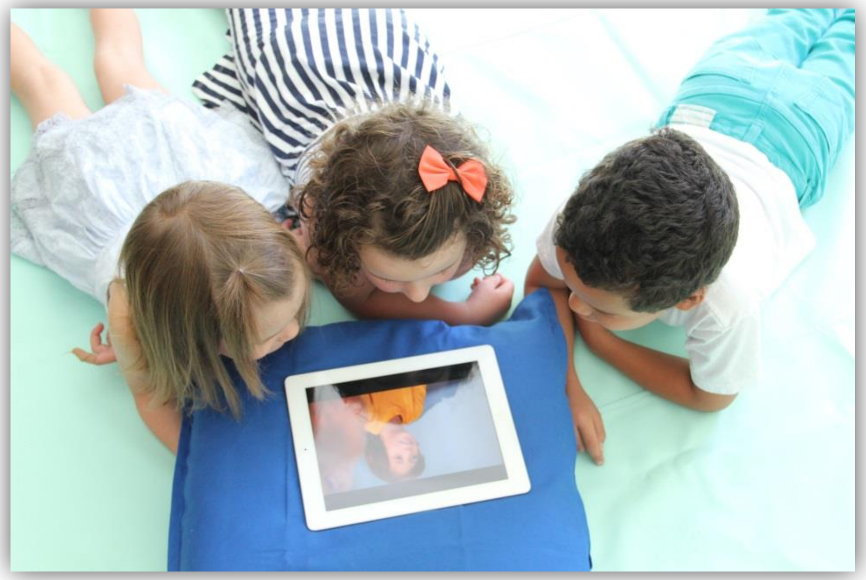
$256^{3\times640\times480}$ states by using pixel representation

Representative Features



Representative Features

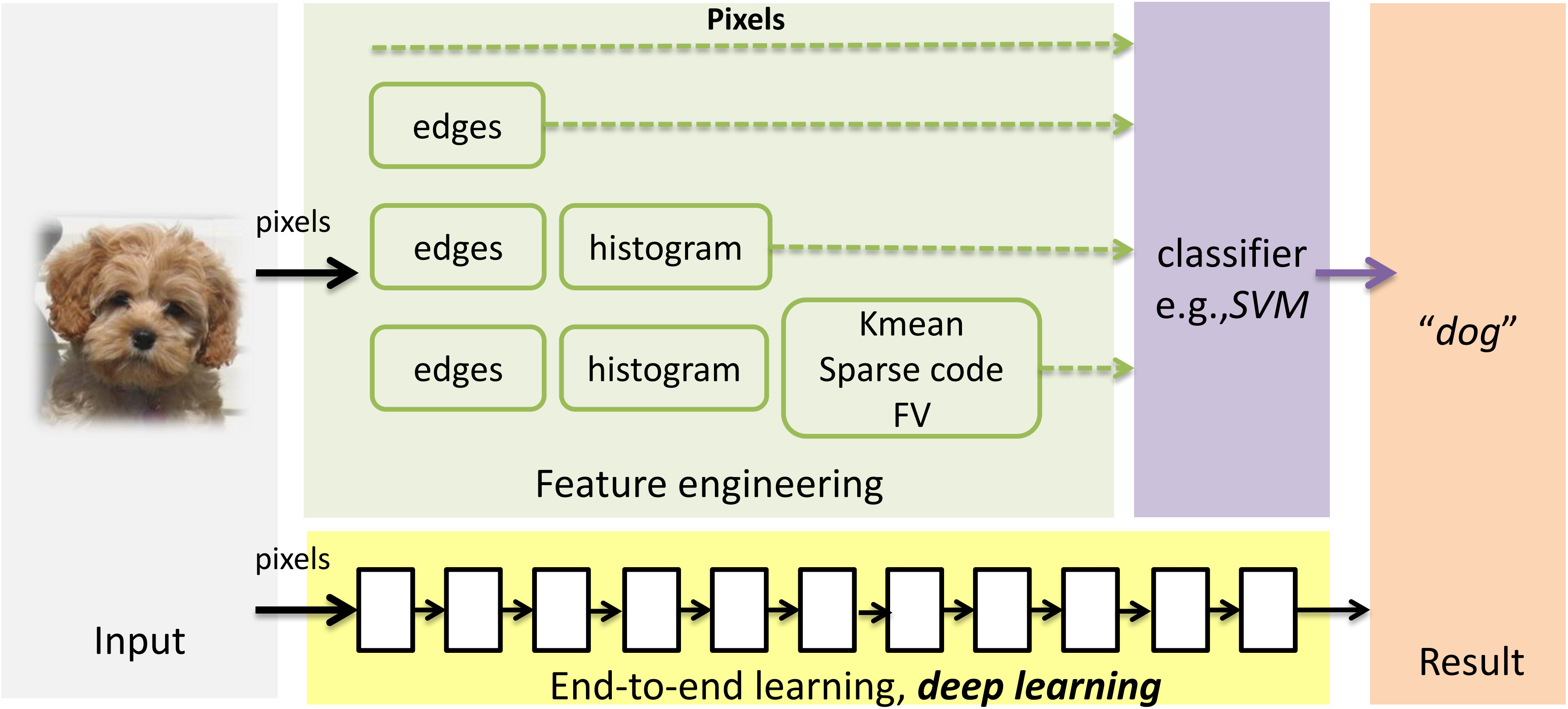
How Kids Know This World



Why has Deep Learning Been so Successful Lately?

- **Largescale annotated data sets** (e.g., ImageNet: 14 million images in 22k categories; YouTube-8M)
- Deep learning algorithms
- Significant improvement in computational power (**GPU**, distributed computing)

Working Ideas on Algorithms

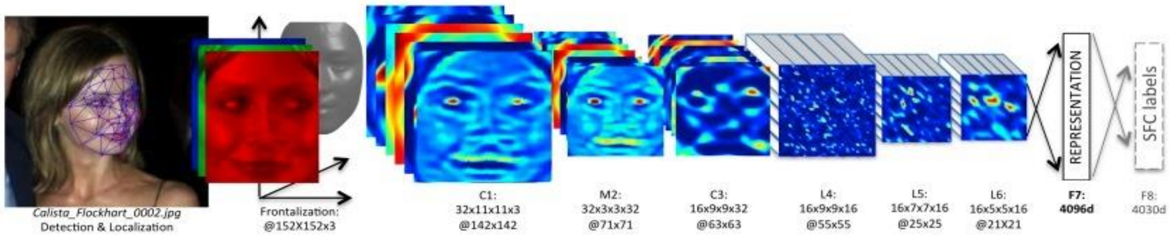


Why has Deep Learning Been so Successful Lately?

- Largescale annotated data sets (e.g., ImageNet, 14 million images in 22k categories)
- **Deep learning algorithms**
- Significant improvement in computational power (**GPU**, distributed computing)

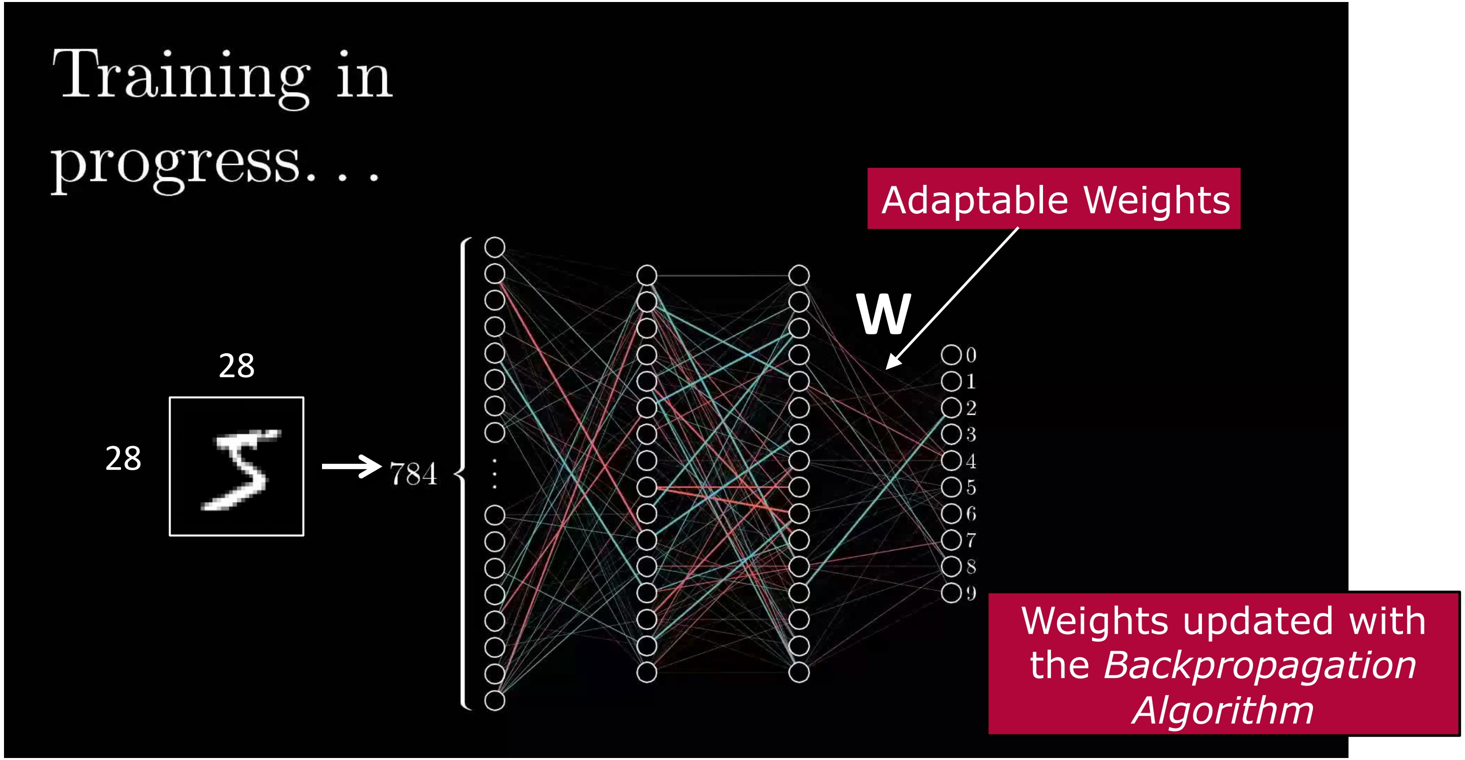
Deep learning

as human beings



(Taigman et al. 2014)

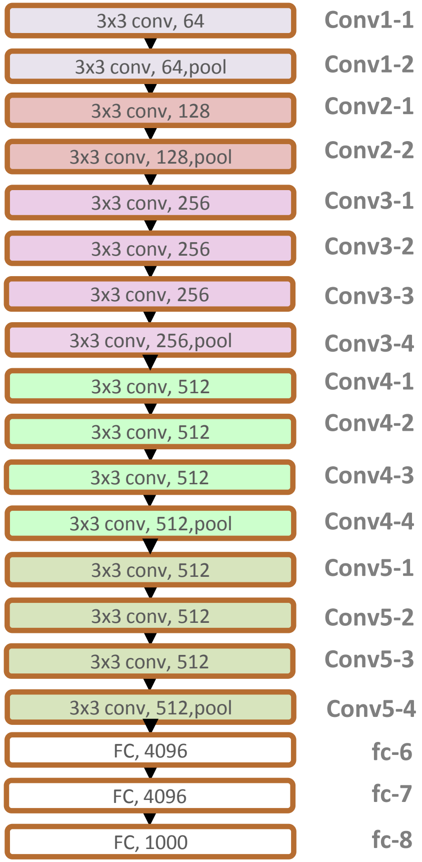
Artificial Neural Networks



Source: gfycat.com

ILSVRC'14 Winner: VGG-Net

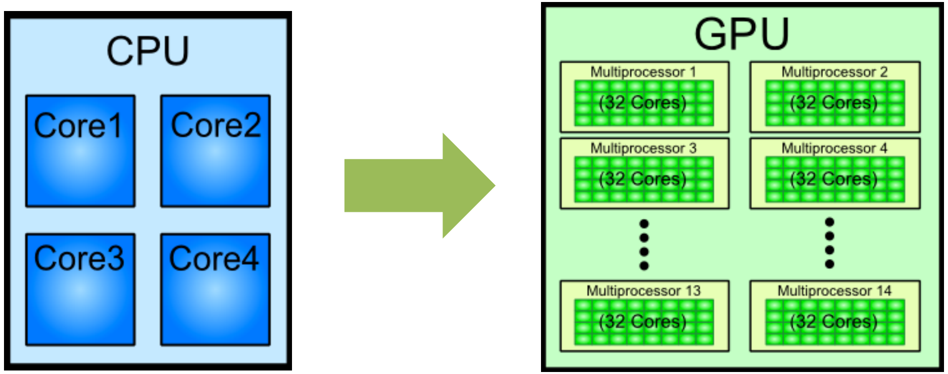
- VGG-Net has 16/19 layers, **24M** nodes, **14M** parameters and, **15B** connections
  - model size **550MB**
  - memory: 24M $*$ 4 bytes $\approx$ **96MB / image** (only forward)



Simonyan et al. VGG-Net, ICLR'15

Why has Deep Learning Been so Successful Lately?

- Largescale annotated data sets (e.g., ImageNet, 14 million images in 22k categories)
- Deep learning algorithms
- **Significant improvement in computational power** (GPU, distributed computing)



(*Roelof Pieters* 2015)

Computational Power

Rapid development of hardware acceleration and massive amounts of computational power

- Applying GPUs/TPUs in neural network computation
- Training time of a very deep model:

  10 years ago: several months → Today: ?
- Cloud computing, distributed system

Powered by 2048 GPUs

[v1] Fri, **29 Mar 2019** 17:55:31 UTC (61 KB)

Yet Another Accelerated SGD: ResNet-50 Training on ImageNet in 74.7 seconds

Masafumi Yamazaki, Akihiko Kasagi, Akihiro Tabuchi, Takumi Honda, Masahiro Miwa, Naoto Fukumoto, Tsuguchika Tabaru, Atsushi Ike, Kohta Nakashima

Fujitsu Laboratories Ltd.

{m.yamazaki, kasagi.akihiko, tabuchi.akihiro, honda.takumi, masahiro.miwa, fukumoto.naoto, tabaru, ike, nakashima.kouta}@fujitsu.com

Limitations of Deep Learning

- The main achievements are in supervised and reinforcement learning
  - **Requiring more annotated data**
  - **Semi-supervised and weakly supervised methods do not perform well**
- **Computationally expensive**
- Difficult to engineer with, architecture engineering
- Deep models have very limited interpretability
- Other issues such as adversarial attack, ethical issue, inability to distinguish causation from correlation, not well being integrated with prior knowledge, and other potential risks

Research Questions

Q1.1:“**SceneTextReg**”

- *Q1*: How can we alleviate the reliance on substantial data annotations of *DL*?
  - Through synthetic data?
  - Through unsupervised or semi-supervised learning method?

Q1.2,Q2:“**SEE**”

- *Q2*: How can we perform multiple computer vision tasks with a uniform end-to-end neural network?
- *Q3*: How can we apply *DL* models on low power devices as e.g., smart phones, embedded devices

Q3:“**BMXNet**”

- *Q4*: Can *DL* models gain multimodal and cross-modal representation learning tasks?
- *Q5*: Can we effectively apply multimedia analysis and *DL* algorithms in real-world applications?

Q4:“**Neural Captioner**”

Q5: “**Automatic Online Lecture Highlighting**”
“**Medical Image Segmentation**”

Publications

- During my Ph.D. study (2010-2013): 13 papers
  - Ph.D. thesis: *automatic video indexing and retrieval using video OCR technology* (***summa cum laude***)
- After Ph.D. (2014-preset): > 45 papers

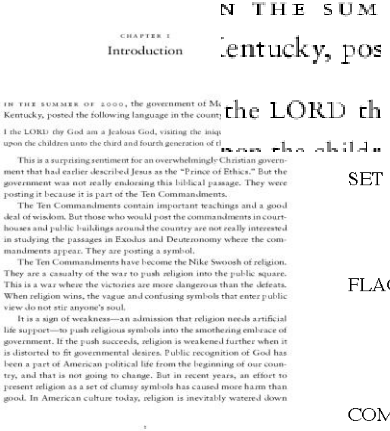
Selected Publications

- SceneTextReg: *A real-time video ocr system*, ACM Multimedia 2016
- SEE: *Towards semi-supervised end-to-end text recognition*, AAAI 2018
- BMXNet
  - Bmxnet: *An open-source binary neural network implementation based on mxnet*, ACM Multimedia 2017
  - *Back to simplicity: How to train accurate BNNs from scratch?* ICCV 2019 (*under review*)
- Neural Captioner: *Image captioning with deep bidirectional LSTMs and multi-task learning*, ACM Trans. Multimedia Computing 2018
- RE-DNN*: A deep semantic framework for multimodal representation learning,* Multimedia Tools and Applications 2016
- *Recurrent generative adversarial network for learning imbalanced medical image semantic segmentation,* Multimedia Tools and Applications 2019
- *Automatic online lecture highlighting based on multimedia analysis,* IEEE Trans. Learning Technology 2018

Selected Publications

- SceneTextReg: *A real-time video ocr system*, ACM Multimedia 2016
- SEE: *Towards semi-supervised end-to-end text recognition*, AAAI 2018
- BMXNet
  - Bmxnet: *An open-source binary neural network implementation based on mxnet*, ACM Multimedia 2017
  - *Back to simplicity: How to train accurate BNNs from scratch?* ICCV 2019 (*under review*)
- Neural Captioner: *Image captioning with deep bidirectional LSTMs*, ACM Trans. Multimedia Computing, 2018
- RE-DNN*: A deep semantic framework for multimodal representation learning,* Multimedia Tools and Applications 2016
- *Recurrent generative adversarial network for learning imbalanced medical image semantic segmentation,* Multimedia Tools and Applications 2019
- *Automatic online lecture highlighting based on multimedia analysis,* IEEE Trans. Learning Technology 2018

Print OCR *vs.* OCR in Multimedia

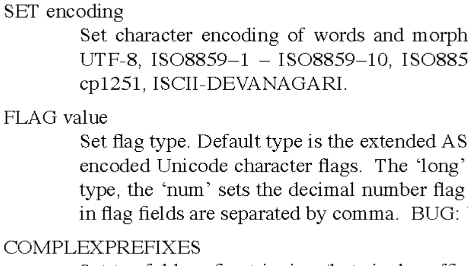

N THE SUM
CHAPTER 1
Introduction
:entucky, pos
IN THE SUMMER OF 2000, the government of Mc
Kentucky, posted the following language in the count
the LORD th
I the LORD thy God am a Jealous God, visiting the iniqu
upon the children unto the third and fourth generation of tl

This is a surprising sentiment for an overwhelmingly Christian government that had earlier described Jesus as the "Prince of Ethics." But the government was not really endorsing this biblical passage. They were posting it because it is part of the Ten Commandments.

The Ten Commandments contain important teachings and a good deal of wisdom. But those who would post the commandments in courthouses and public buildings around the country are not really interested in studying the passages in Exodus and Deuteronomy where the commandments appear. They are posting a symbol.

The Ten Commandments have become the Nike Swoosh of religion. They are a casualty of the war to push religion into the public square. This is a war where the victories are more dangerous than the defeats. When religion wins, the vague and confusing symbols that enter public view do not stir anyone's soul.

It is a sign of weakness—an admission that religion needs artificial life support—to push religious symbols into the smothering embrace of government. If the push succeeds, religion is weakened further when it is distorted to fit governmental desires. Public recognition of God has been a part of American political life from the beginning of our country, and that is not going to change. But in recent years, an effort to present religion as a set of clumsy symbols has caused more harm than good. In American culture today, religion is inevitably watered down

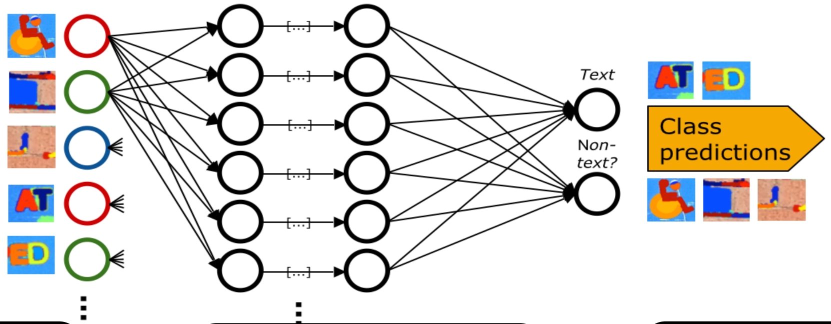

SET encoding
Set character encoding of words and morph
UTF-8, ISO8859–1 – ISO8859–10, ISO885
cp1251, ISCII-DEVANAGARI.

FLAG value
Set flag type. Default type is the extended AS
encoded Unicode character flags. The 'long'
type, the 'num' sets the decimal number flag
in flag fields are separated by comma. BUG:

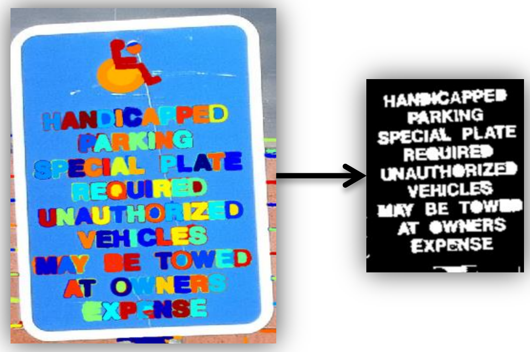
COMPLEXPREFIXES

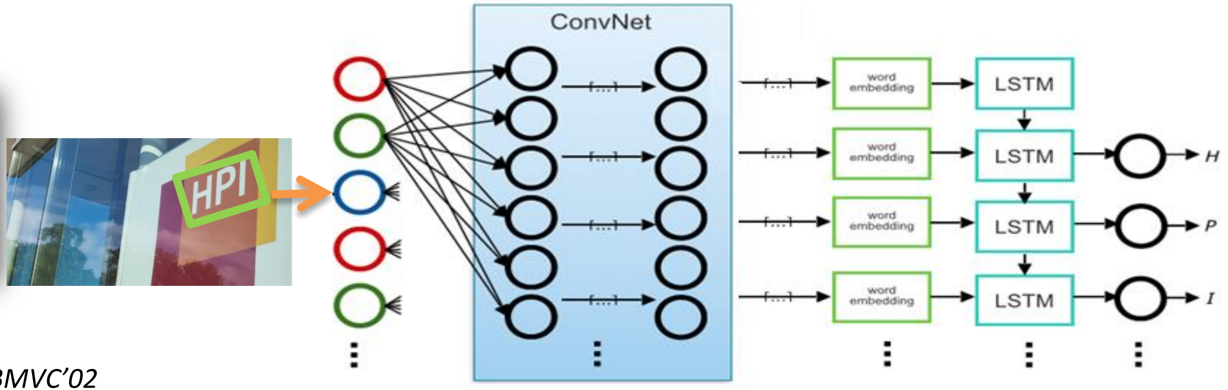
SceneTextReg

SceneTextReg: real-time scene text recognition, Yang, Wang, Bartz and Meinel, ACM MM'16



MSER Component detector → **CNN** Text candidate validator → **RCNN** Word recognizer





Matas et al. Maximally Stable Extremal Regions, BMVC'02

SceneTextReg - Data Generator

Features

- Various fonts (>1500)
- Different colors, sizes, shadows, borders with varying displacements to the rendered texts
- Transformations: distortion, rotation
- Random blur, reflection
- Background blending (nature scene images)

generated samples **real word images**

ICDAR data set (right column)

SceneTextReg - Evaluation

ICDAR'13/15 data set (IAPR International Conference on Document Analysis and Recognition) on focused scene word recognition:

Method	WRA
Google's PhotoOCR	**0.8283**
SceneTextReg	**0.8237**
PicRead	0.5799
NESP	0.642
PLT	0.6237
MAPS	0.6274
PIONEER	0.537
ABBY OCR SDK10	0.453

WRA*: Word Recognition Accuracy*
*(**Case sensitive with punctuation, special chars**)*

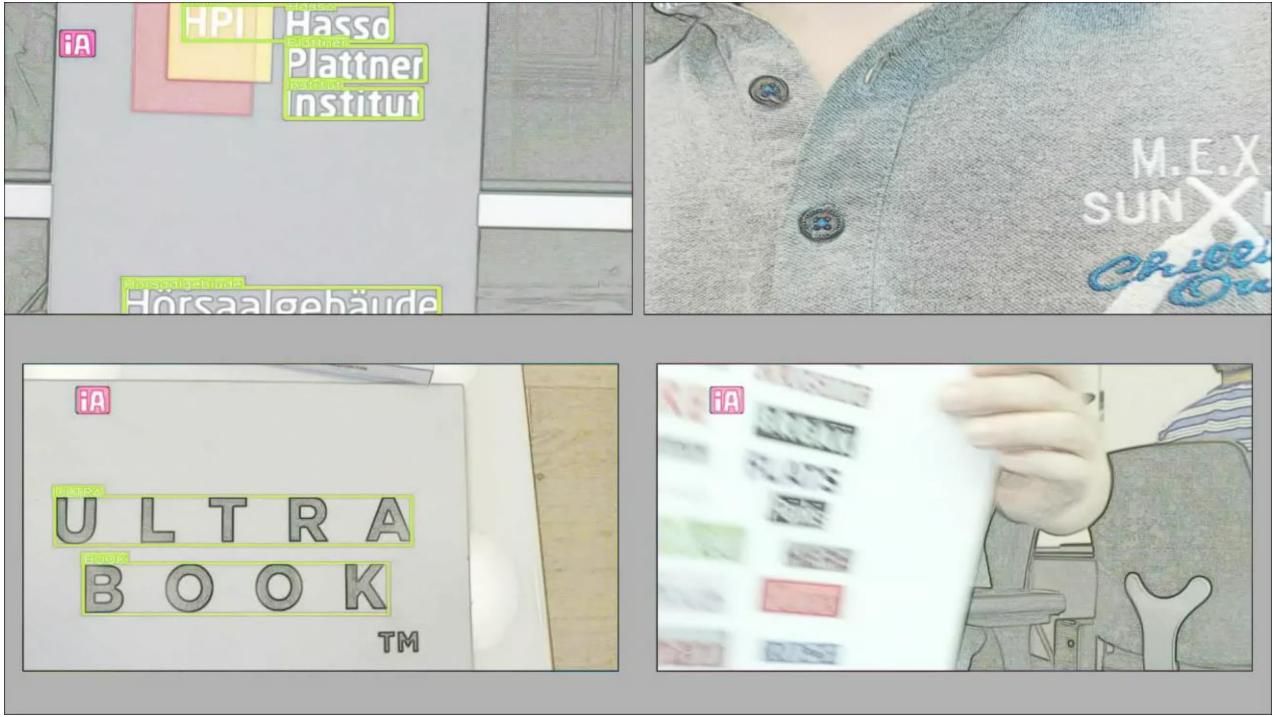
Bissacco et al. PhotoOCR, ICCV'13

62-way char classification (on ICDAR'03 data set):

Method	Classification Accuracy
Our result	**0.872**
Jaderberg et al. (*ECCV'14*)	0.868
Alsharif et al. (*ICLR'14*)	0.86
Wang et al. (*ICPR'12*)	0.839
A. Coates et al. (*ICDAR'11*)	0.817

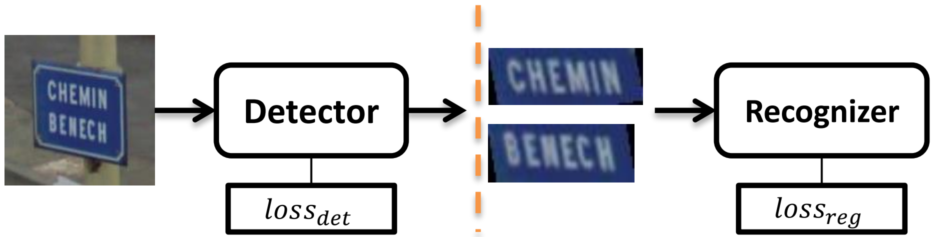
Only synthetic data used for training!

SceneTextReg - Demo

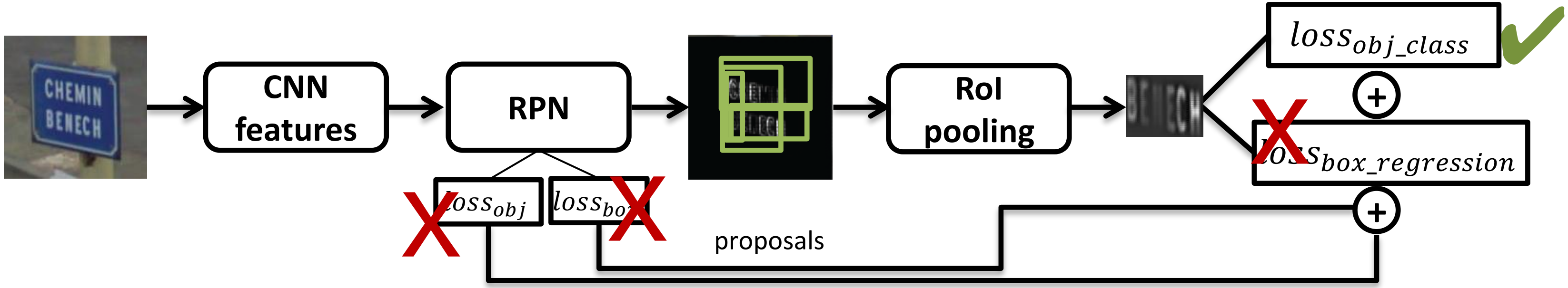


Scene Text Recognition with NN

- Two stage system as e.g., *SceneTextReg*



- End-to-end system as e.g., *Faster RCNN*

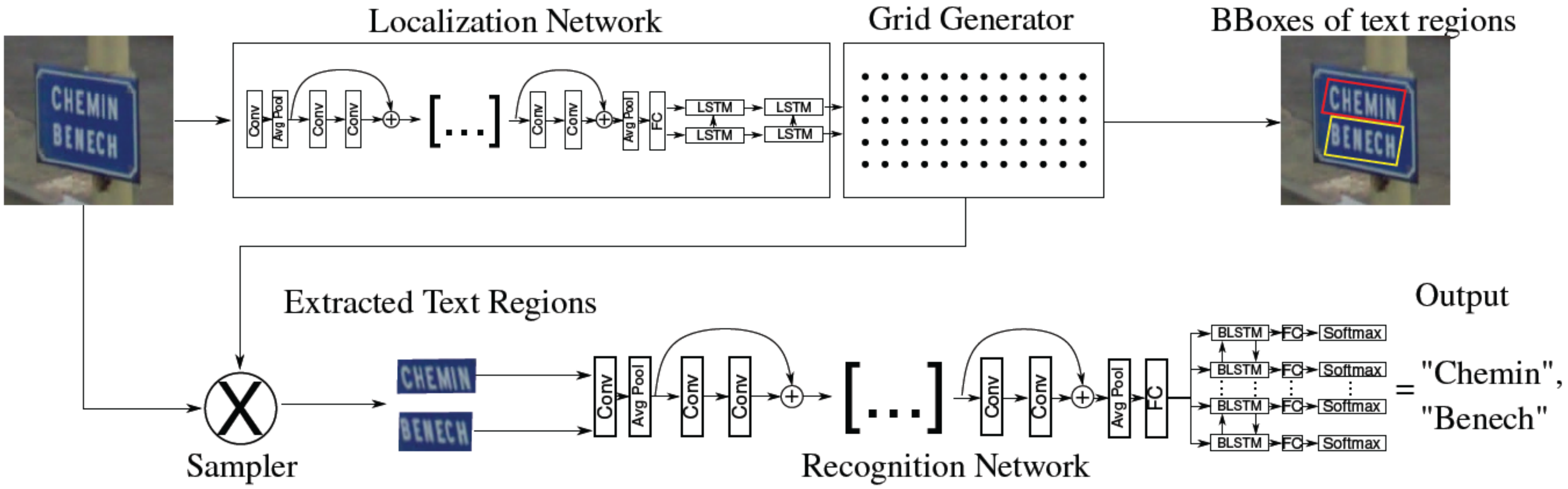


Ren et al. Faster r-cnn, NIPS'15

SEE

SEE: Towards Semi-Supervised End-to-End Scene Text Recognition, Bartz, Yang, Meinel, AAAI 2018

Affine transformation matrices



SEE - Evaluation

Method	Accuracy
Maxout CNN, *(ICLR'14)*	0.96
ST-CNN, *(NIPS'15)*	0.963
SEE	**0.952**

SVHN house number data set

Method	IC13/15	SVT	IIIT5K
Google's PhotoOCR, *(ICCV'13)*	0.876	0.78	-
CharNet, *(ECCV'14)*	0.818	0.717	-
CRNN, *(TPAMI'16)*	0.867	0.808	0.782
RARE, *(CVPR'16)*	0.875	**0.819**	0.819
SEE	**0.903**	0.798	**0.86**

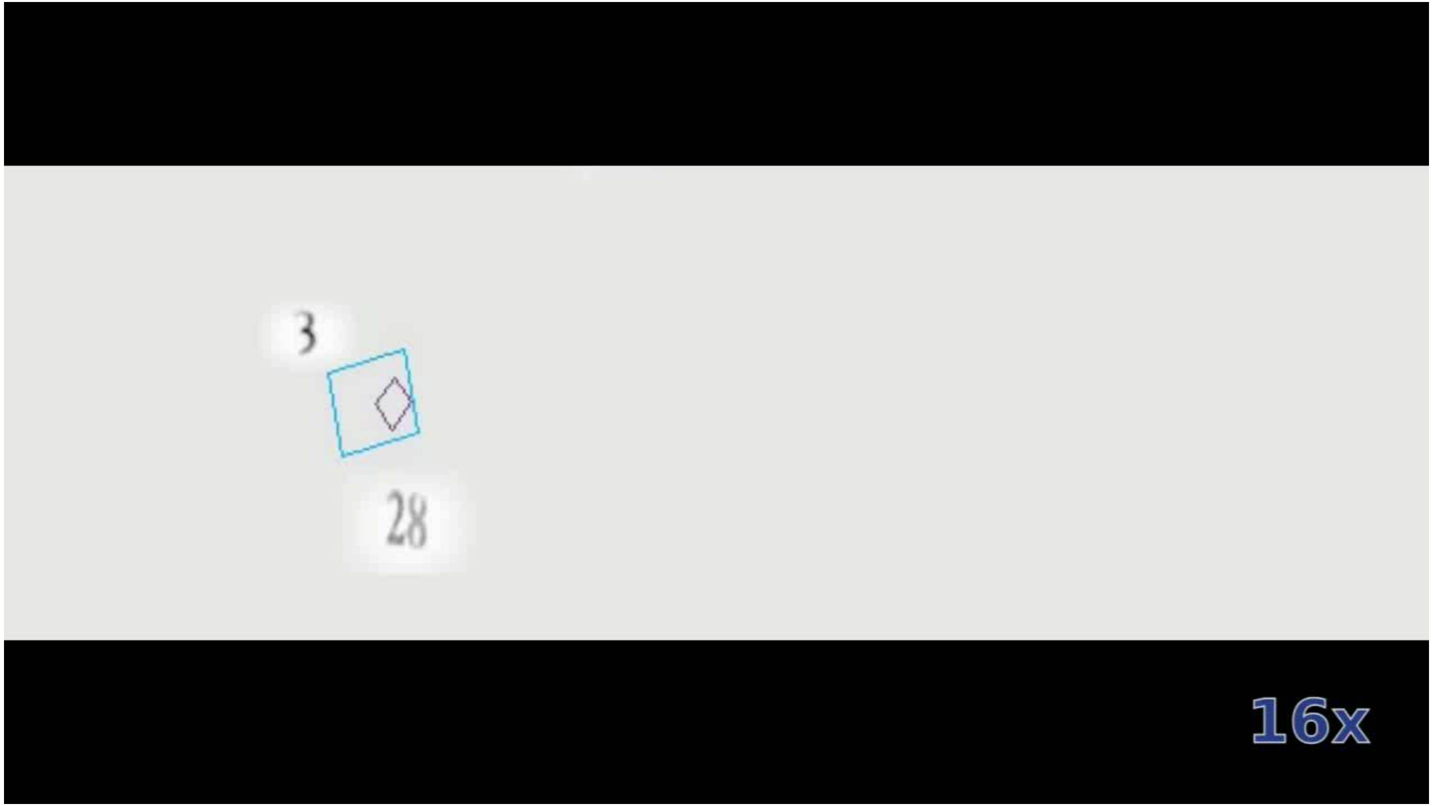
ICDAR’13/15, SVT, IIIT5K data set

SEE - Evaluation

Method	Accuracy
Smith et al.(Google) (*ECCV’16)*	0.725
Wojna et al.(Google) (*ICDAR’17)*	**0.842**
SEE	0.78

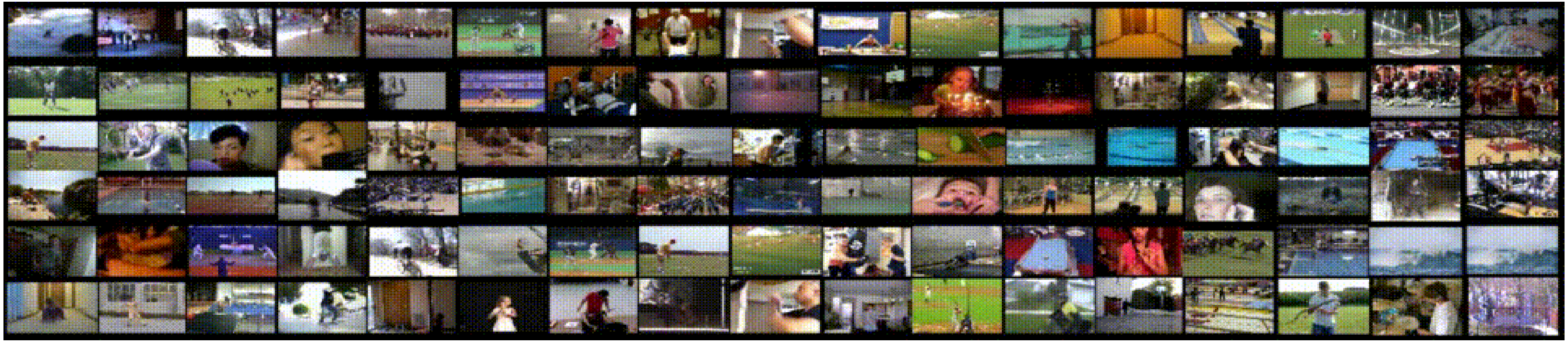
French street name signs data set

SEE - Demo



Multimodal Retrieval

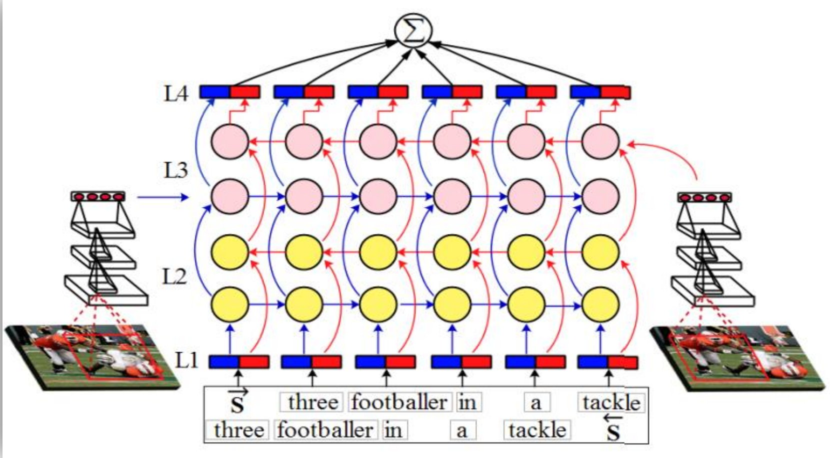
- Image captioning
- Video classification
- Human action recognition in surveillance video

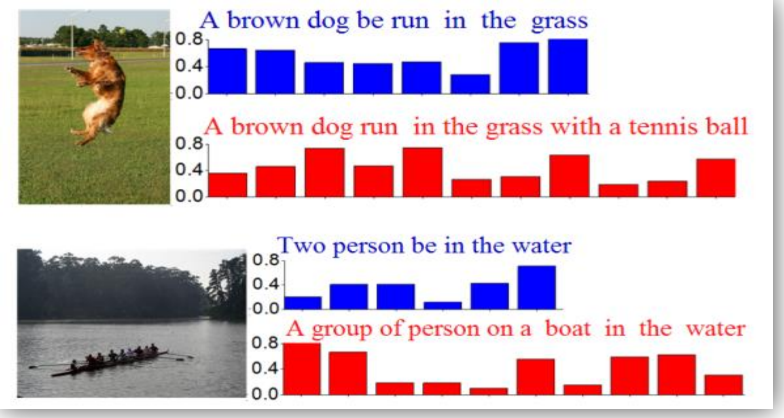


Neural Captioner

Image Captioning with Deep Bidirectional LSTMs, Wang, Yang, Bartz and Meinel, ACM MM'16

- Visual representation → CNN model
  - Transfer learning from ImageNet models
- Visual to sentence (language) embedding
  - **Bi-directional LSTM** (Long Short-Term Memory)
- Data augmentation: random cropping, mirroring, shifting



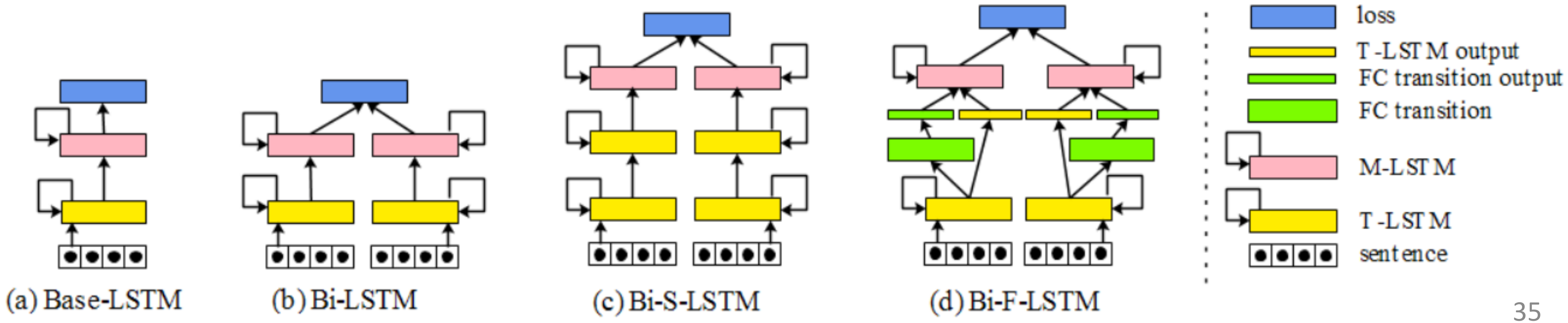


Forward Backward

Neural Captioner

The proposed architectures

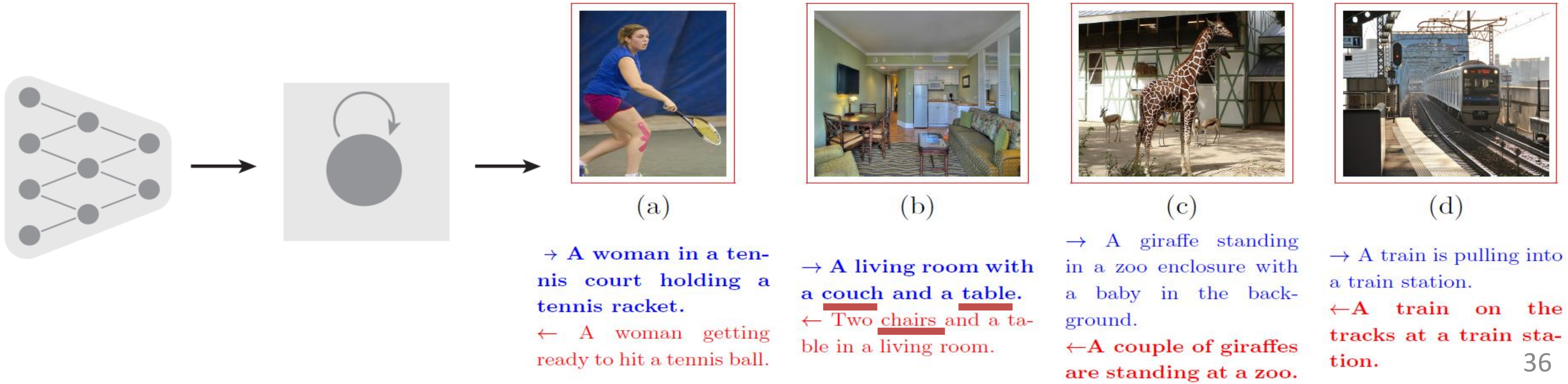
- baseline model (a)
- bidirectional LSTM (b)
- bidirectional Stacked LSTM (c)
- bidirectional LSTM with fully connected (FC) transition layer (d)



Neural Captioner

Contributions

- Cover more semantics by Bi-LSTM
- Great portion of generated sentences not appear in training set
- Achieved state-of-the-art on Flickr8K, Flickr30K, MSCOCO and Pascal1K image captioning data sets



Neural Captioner - Demo

Deep Learning on Low Power Devices

A state-of-the-art ResNet-152 (152 layers) surpasses human performance on the image classification task.

Number of **operations**:

- AlexNet (240MB), 720 MFLOPs,
- VGG19 (550MB), 19.6 BFLOPs
- **ResNet-152 (240MB), 11.3 BFLOPs**

Inference time on CPU:

- AlexNet: 3 fps,
- VGG19: 0.25 fps
- **ResNet-152: 0.63 fps**

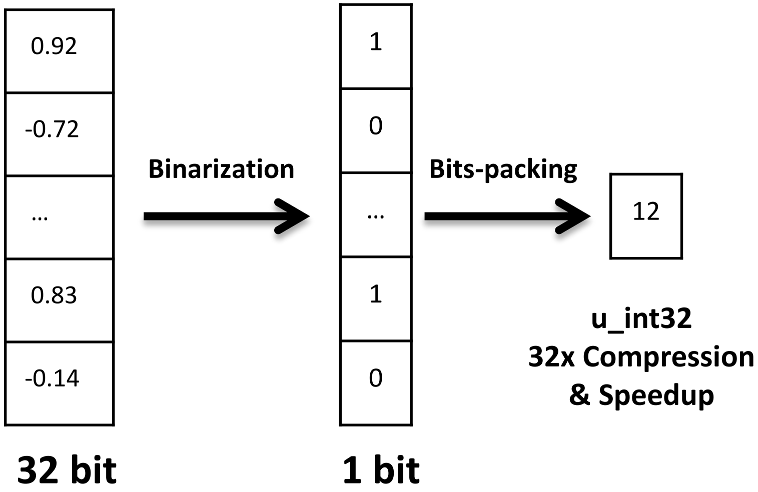
Deep Learning on Mobile Devices

Autonomous driving

Assistance apps

Low power devices

Binary Neural Networks



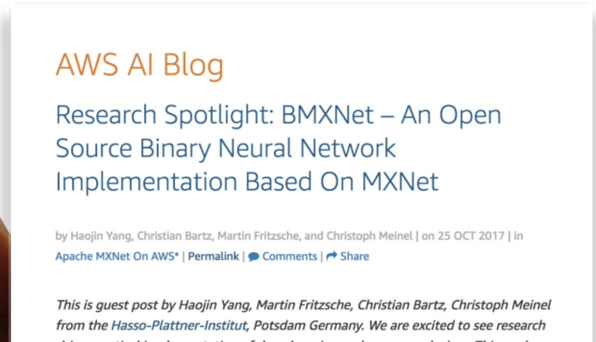
Benefits

- 32x smaller model size
  - e.g., FPGAs with <10MB on-ship **memory**
- 32x less memory access → much less **energy** consumption
- Bitwise operator e.g., *XNOR, bitcount* instead of arithmetic operations in NN
  - It allows for a **speedup** factor of up to 32 by combining multiple operations in one CPU cycle
- On devices, offline prediction → better **privacy** protection

BMXNet

An open-source binary neural network implementation based on mxnet, Yang, Fritzsche, Bartz and Meinel, ACM MM'17

- Flexible design and fully compatible with standard neural network components
- Source code: https://github.com/hpi-xnor
- E.g., ResNet-18 for image classification on Cifar-10 data set
  - 45MB (full precision) → 1.5MB (binary)


AWS AI Blog

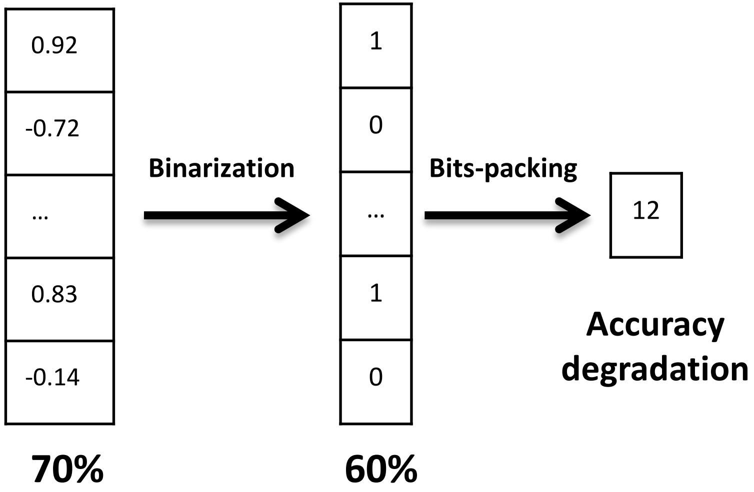
Research Spotlight: BMXNet – An Open Source Binary Neural Network Implementation Based On MXNet

by Haojin Yang, Christian Bartz, Martin Fritzsche, and Christoph Meinel | on 25 OCT 2017 | in Apache MXNet On AWS* | Permalink | Comments | Share

This is guest post by Haojin Yang, Martin Fritzsche, Christian Bartz, Christoph Meinel from the Hasso-Plattner-Institut, Potsdam Germany. We are excited to see research

BMXNet

***Back to Simplicity**: How to Train Accurate Binary Neural Network from Scratch?* Bethge, Yang, Borstein and Meinel, ICCV'19 (*submitted*)



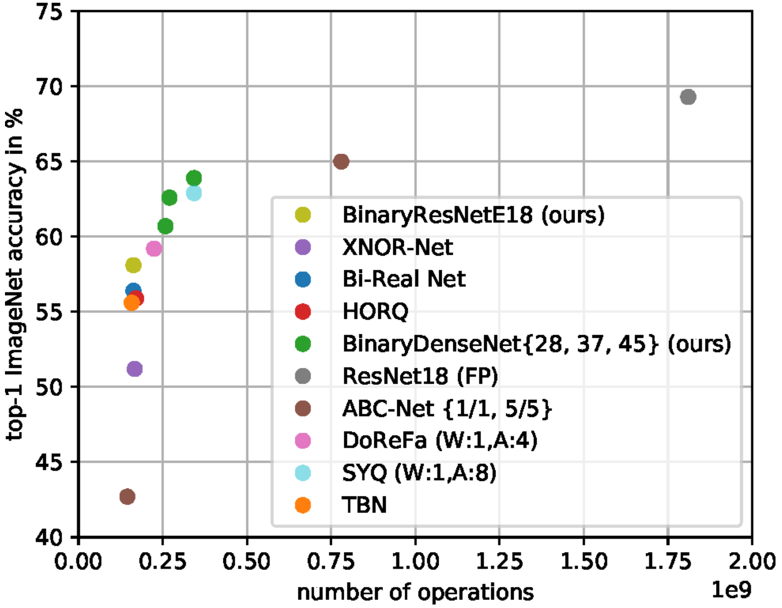
Contributions:

- Challenging conventional wisdom: Highly accurate BNNs can be trained by using standard training strategy.
- We suggest general **design principles** for BNNs
- Our ***BinaryDenseNet*** significantly surpasses all existing BNNs for image classification without tricks.
- We provide codes to facilitate follow-up studies

BMXNet - Evaluation

Model size	Method	Top-1/Top-5 accuracy
~4.0MB	XNOR-ResNet18 (*ECCV'16*)	51.2%/73.2%
	TBN-ResNet18 (*ECCV'18*)	55.6%/74.2%
	Bi-Real-ResNet18 (*ECCV'18*)	56.4%/79.5%
	BinaryResNetE18 (ours)	58.1%/80.6%
	BinaryDenseNet28 (ours)	**60.7%/82.4%**
~5.1MB	TBN-ResNet34 (*ECCV'18*)	58.2%/81.0%
	Bi-Real-ResNet34 (*ECCV'18*)	62.2%/83.9%
	BinaryDenseNet37 (ours)	62.5%/83.9%
	BinaryDenseNet37-dilated (ours)	**63.7%/84.7%**
7.4MB	*BinaryDenseNet45 (ours)*	63.7%/84.8%
46.8MB 249MB	Full-precision ResNet18 Full-precision AlexNet	69.3%/89.2% 56.6%/80.2%

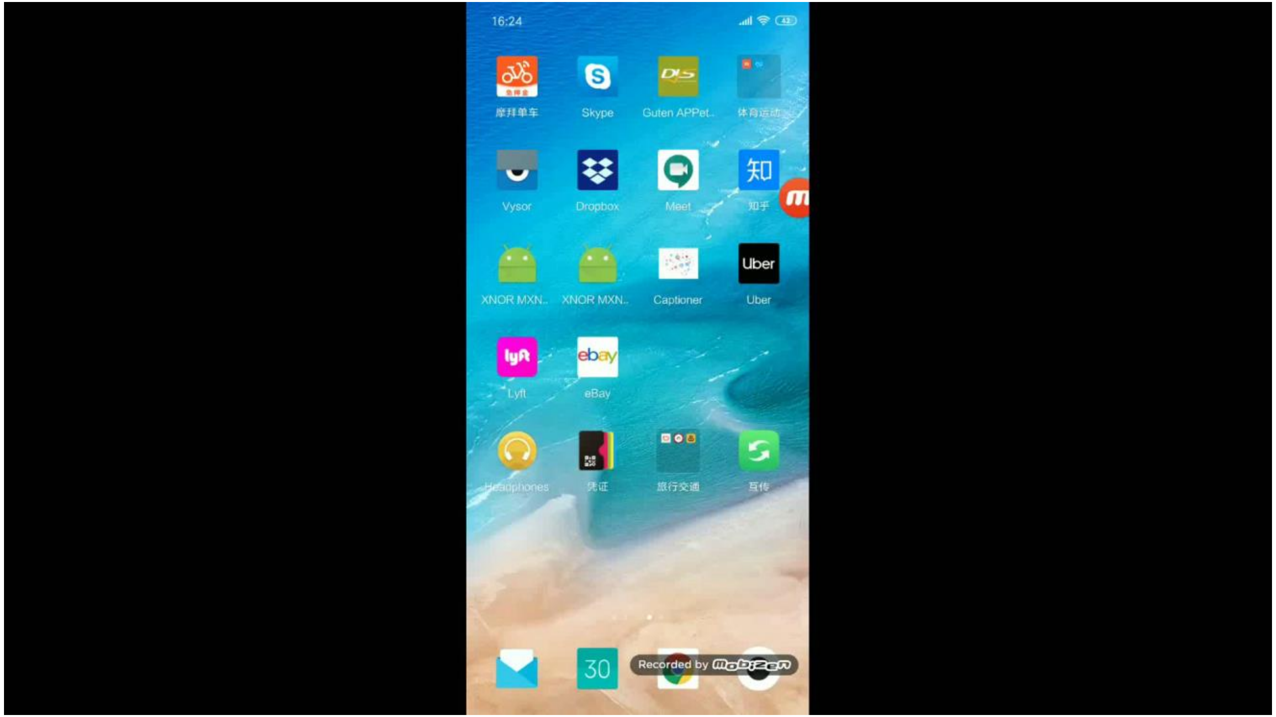
Comparison to state-of-the-art BNNs on ImageNet



The trade-off of top-1 validation accuracy on ImageNet and number of operations. All the binary/quantized models are based on ResNet18 except *BinaryDenseNet*.

XNOR-Net ECCV'16, TBN ECCV'18, Bi-Real Net ECCV'18, AlexNet NIPS'12, ResNet CVPR'15, ABC-Net NIPS'17, HORQ ICCV'17, DoReFa-Net CoRR'16, SYQ CVPR'18

BMXNet - Demo



16:24
摩拜单车
Skype
Guten APPet..
Vysor
Dropbox
Meet
知乎
XNOR MXN..
XNOR MXN..
Captioner
Uber
Uber
Lyft
eBay
Headphones
凭证
旅行交通
互传
30
Recorded by

Thank you for your Attention!

0 to **3** →

"Medical Image Segmentation"

"Automatic Online Lecture Highlighting"

"SEE"

"Neural Captioner"

"BMXNet"

"SceneTextReg"

→ Beyond!