学術俯瞰講義
人工知能の未解決問題とディープラーニング

東京大学 松尾 豊

ディープラーニングの今後の発展

① 画像

画像から、特徴量を抽出する

画像認識の精度向上

② マルチモーダル

映像、センサーなどのマルチモーダルなデータから
特徴量を抽出し、モデル化する

動画の認識精度の向上、行動予測、異常検知

③ ロボティクス（行動）

自分の行動と観測のデータをセットにして、特徴量を抽出する。
記号を操作し、行動計画を作る。

プランニング、推論

④ インタラクション

外界と試行錯誤することで、外界の特徴量を引き出す

オントロジー、高度な状況の認識

⑤ 言葉とのひもづけ（シンボルグラウンディング）

高次特徴量を、言語とひもづける

言語理解、自動翻訳

⑥ 言語からの知識獲得

グラウンディングされた言語データの大量の入力により、さらなる抽象化を行う

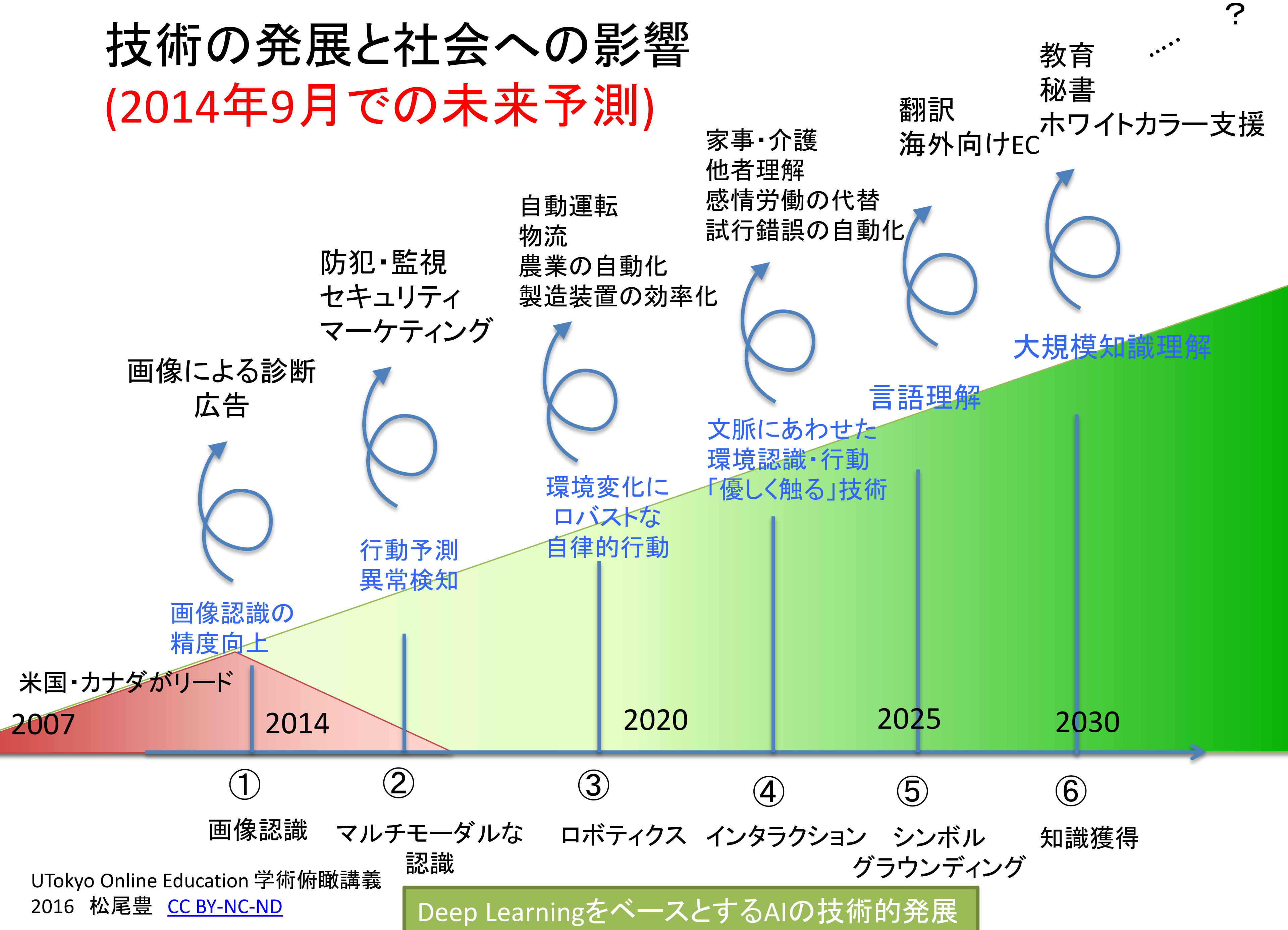
知識獲得のボトルネックの解決

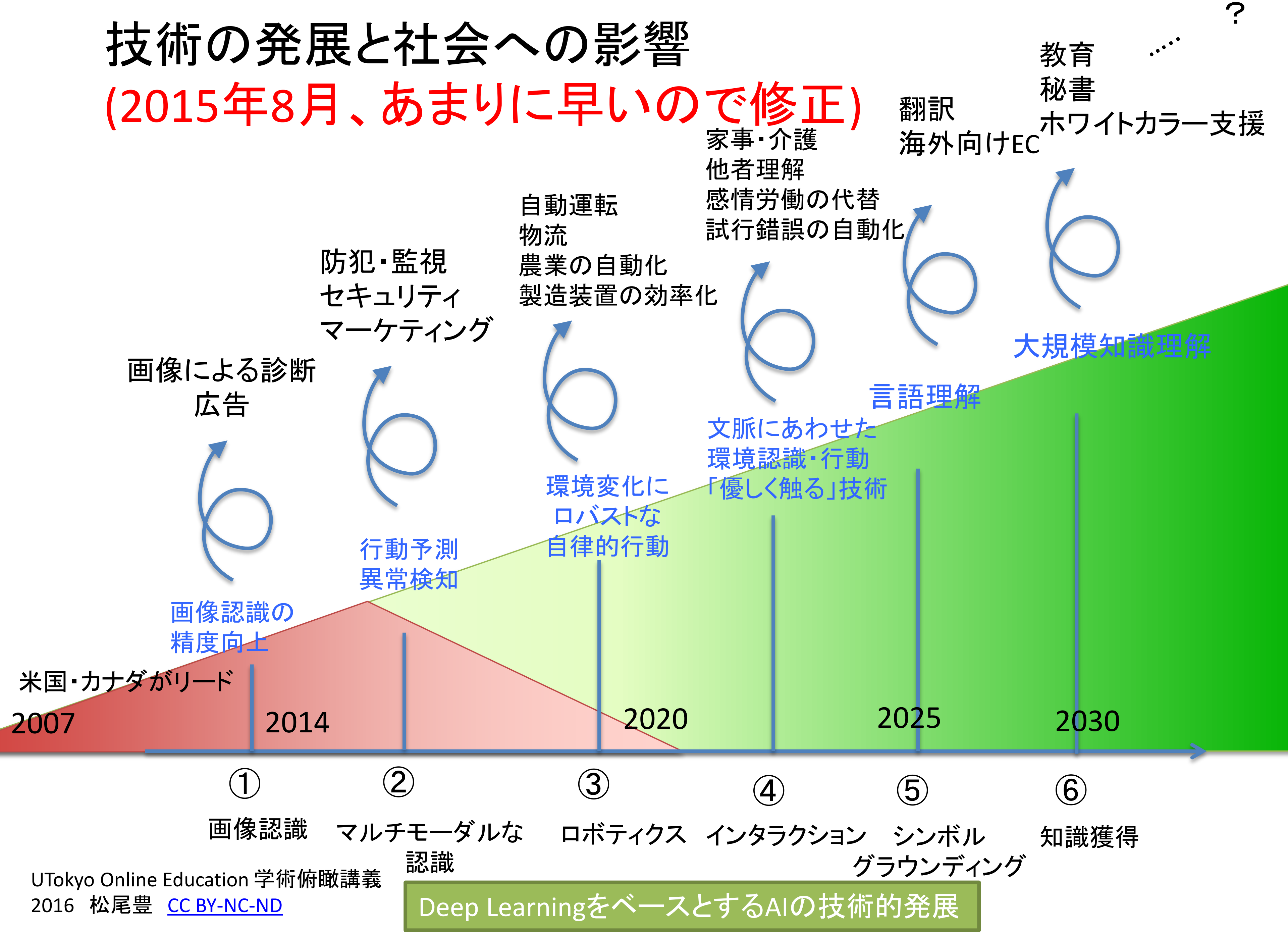
①② → 認識
③④ → 運動
⑤⑥ → 言語

認識 → 運動 → 言語

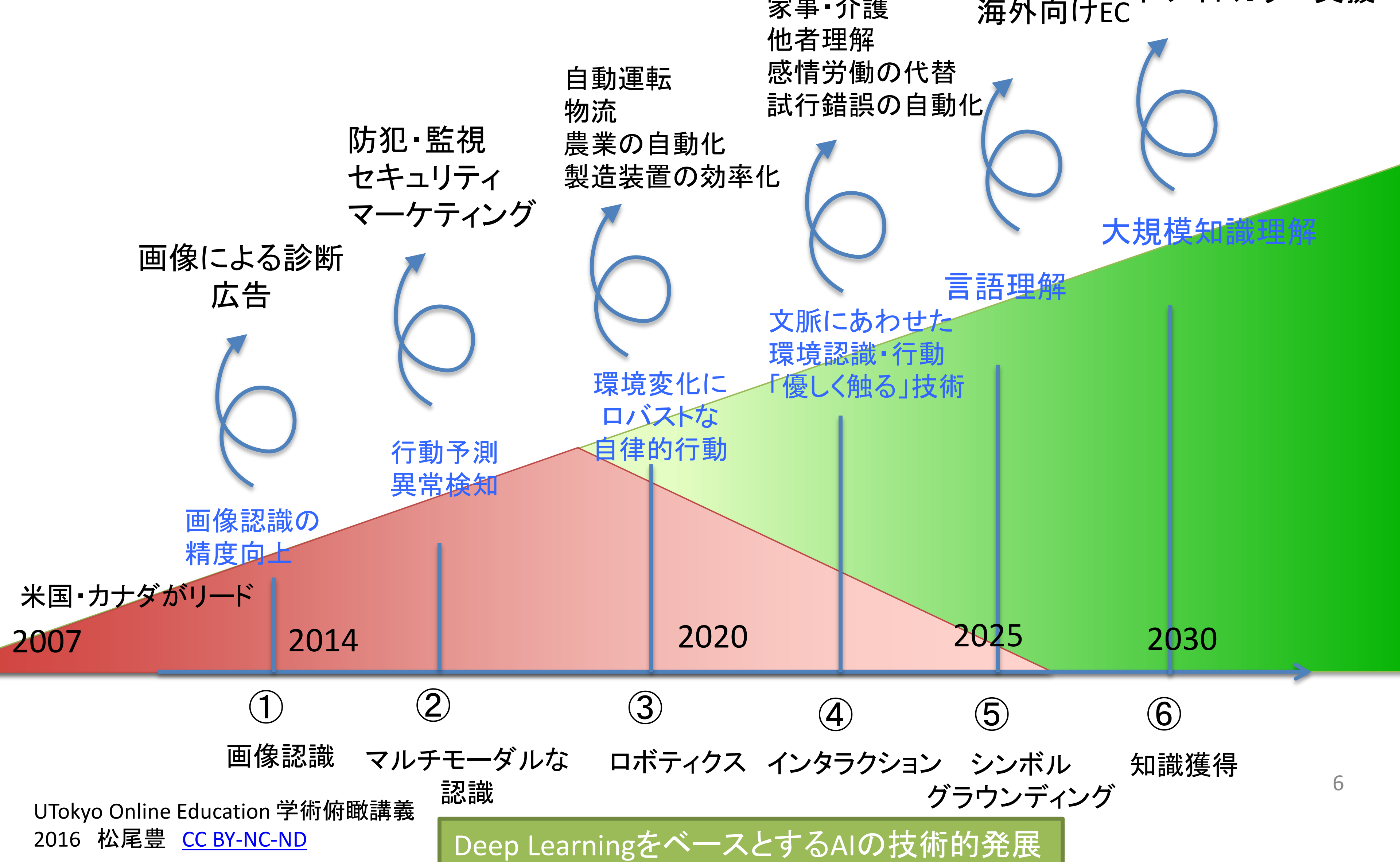
**ディープラーニングがすごいというより
その先に広がる世界がすごい**

技術の発展と社会への影響
(2014年9月での未来予測)




技術の発展と社会への影響
(2015年8月、あまりに早いので修正)
？
……
教育
秘書
ホワイトカラー支援
翻訳
海外向けEC
家事・介護
他者理解
感情労働の代替
試行錯誤の自動化
自動運転
物流
農業の自動化
製造装置の効率化
防犯・監視
セキュリティ
マーケティング
画像による診断
広告
大規模知識理解
言語理解
文脈にあわせた
環境認識・行動
「優しく触る」技術
環境変化に
ロバストな
自律的行動
行動予測
異常検知
画像認識の
精度向上
米国・カナダがリード
2007
2014
2020
2025
2030
① 画像認識
② マルチモーダルな認識
③ ロボティクス
④ インタラクション
⑤ シンボルグラウンディング
⑥ 知識獲得
Deep LearningをベースとするAIの技術的発展
UTokyo Online Education 学術俯瞰講義
2016 松尾豊 CC BY-NC-ND

技術の発展と社会への影響
(2015年12月、再度修正)



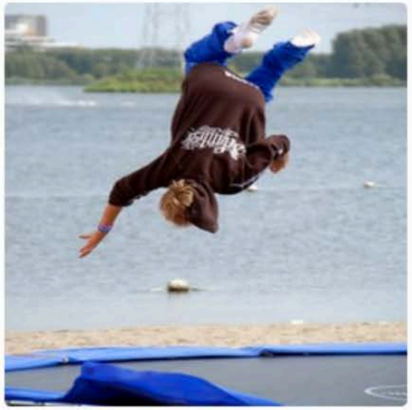
Andrej KarpathyStanford Computer Science Ph.D. student
Automated Image Captioning with ConvNets and Recurrent Nets
https://www.meetup.com/ja-JP/sfmachinelearning/events/219842815/?eventId=219842815&chapter_analytics_code=UA-57800518-1
Automated Image Captioning with ConvNets and Recurrent Nets
http://cs.stanford.edu/people/karpathy/sfmltalk.pdf
(最終閲覧日：2017年7月21日)

Automated Image Captioning (2014-)

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."



"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

言語の意味理解：Generating Images (2015.12-)

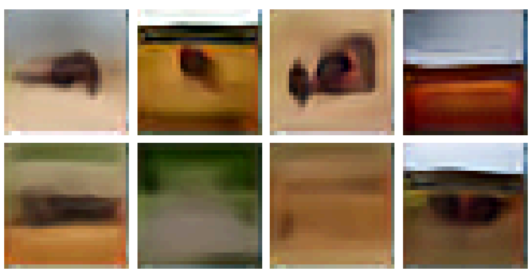
A very large commercial plane flying in blue skies.

A very large commercial plane flying in rainy skies.

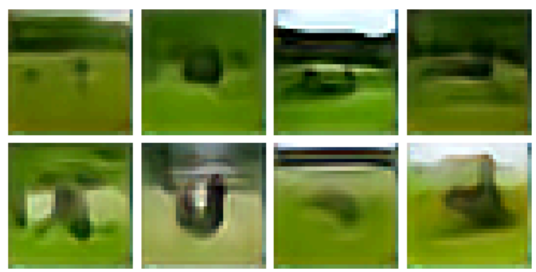
Elman Mansimov et. al: "Generating Images from Captions with Attention", Reasoning, Attention, Memory (RAM) NIPS Workshop 2015
http://www.thespermwhale.com/jaseweston/ram/papers/paper_13.pdf
p. 3, Fig. 3

Elman Mansimov et. al: "Generating Images from Captions with Attention", Reasoning, Attention, Memory (RAM) NIPS Workshop 2015
http://www.thespermwhale.com/jaseweston/ram/slides/session1/gen-captions-elman-mansimov.pdf

A stop sign flying in blue skies.



A herd of elephants walking across a dry grass field.

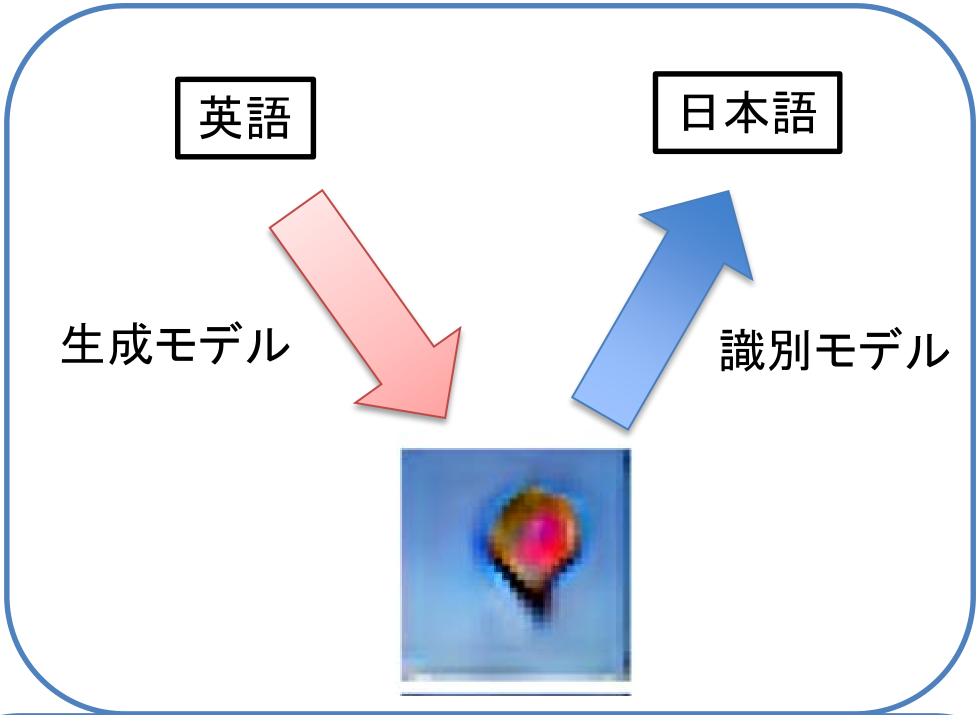


A herd of elephants walking across a green grass field.

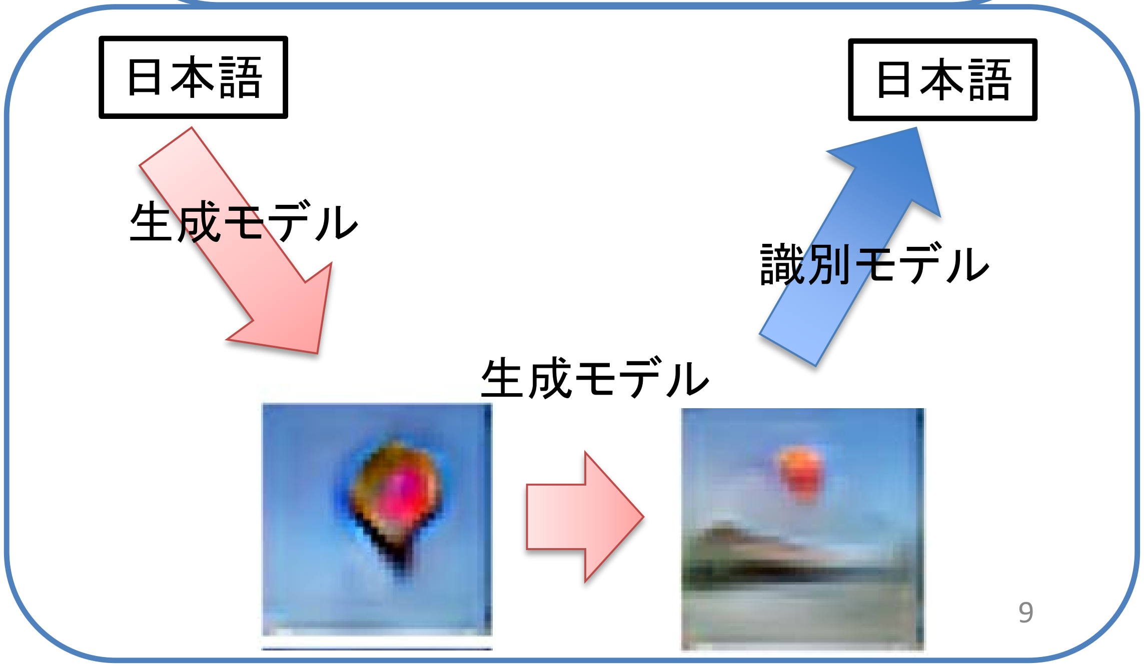
Elman Mansimov et. al: "Generating Images from Captions with Attention", Reasoning, Attention, Memory (RAM) NIPS Workshop 2015
http://www.thespermwhale.com/jaseweston/ram/papers/paper_13.pdf
p. 3, Fig. 3

Elman Mansimov et. al: “Generating Images from Captions with Attention”, Reasoning, Attention, Memory (RAM) NIPS Workshop 2015
http://www.thespermwhale.com/jaseweston/ram/slides/session1/gen-captions-elman-mansimov.pdf

- 画像による翻訳（意訳）
  - 日本語→画像→英語
  - 課題
    - 解像度
    - 画像から映像、体験へ
    - 抽象概念はどう扱うの？
  - いずれにしても視覚的な機構がベースにあるのは間違いない



- 映像による推論
  - 言語→映像→シーン予測→次の映像→言語
  - 「風船が飛んでいる」→「山まで飛んで行くのかな」



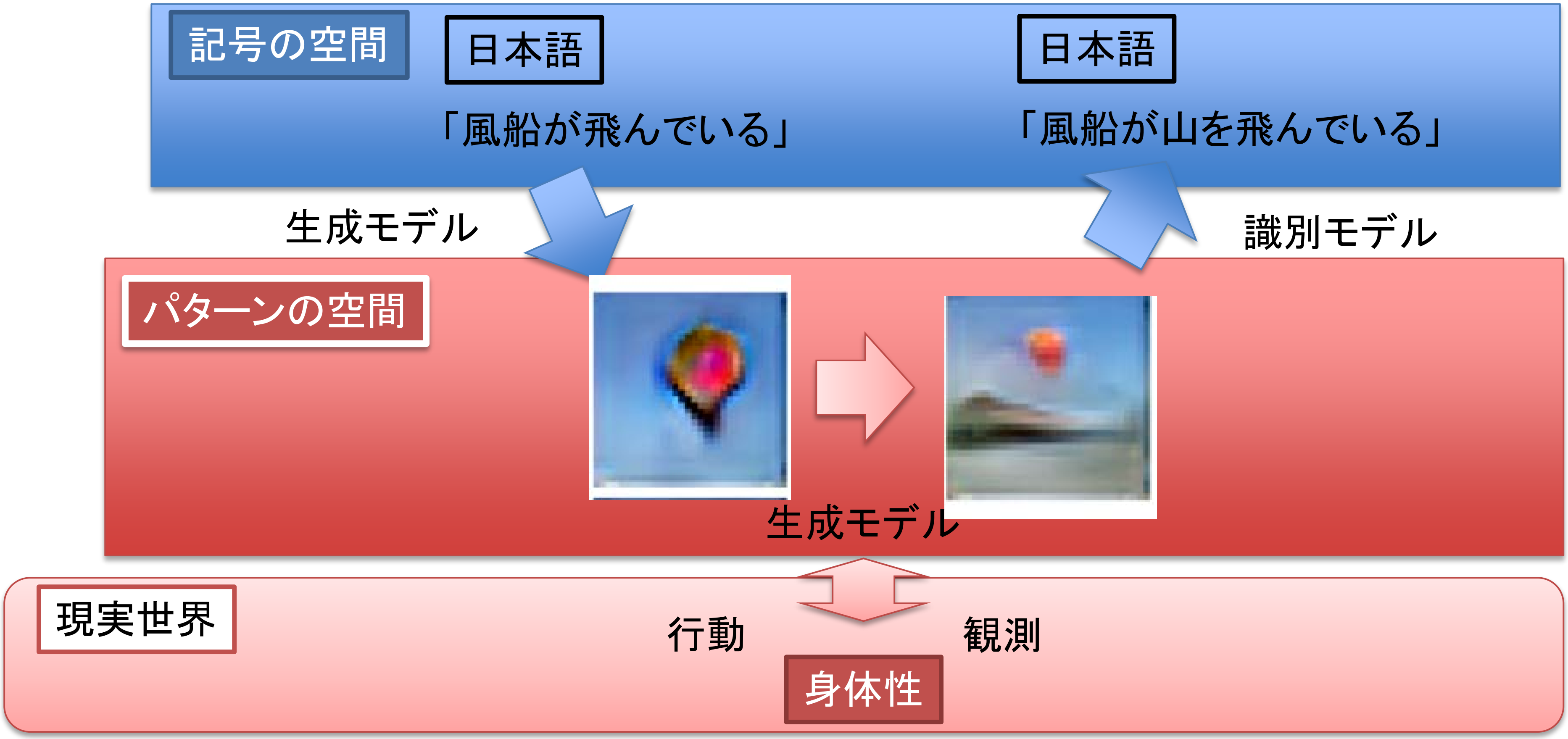
言葉の空間とパターンの空間を
自由に行き来するのが人間の知能

Elman Mansimov et. al: “Generating Images from Captions with Attention”, Reasoning, Attention, Memory (RAM) NIPS Workshop 2015
http://www.thespermwhale.com/jasweston/ram/slides/session1/gen-captions-elman-mansimov.pdf

記号処理の目指すべきところ

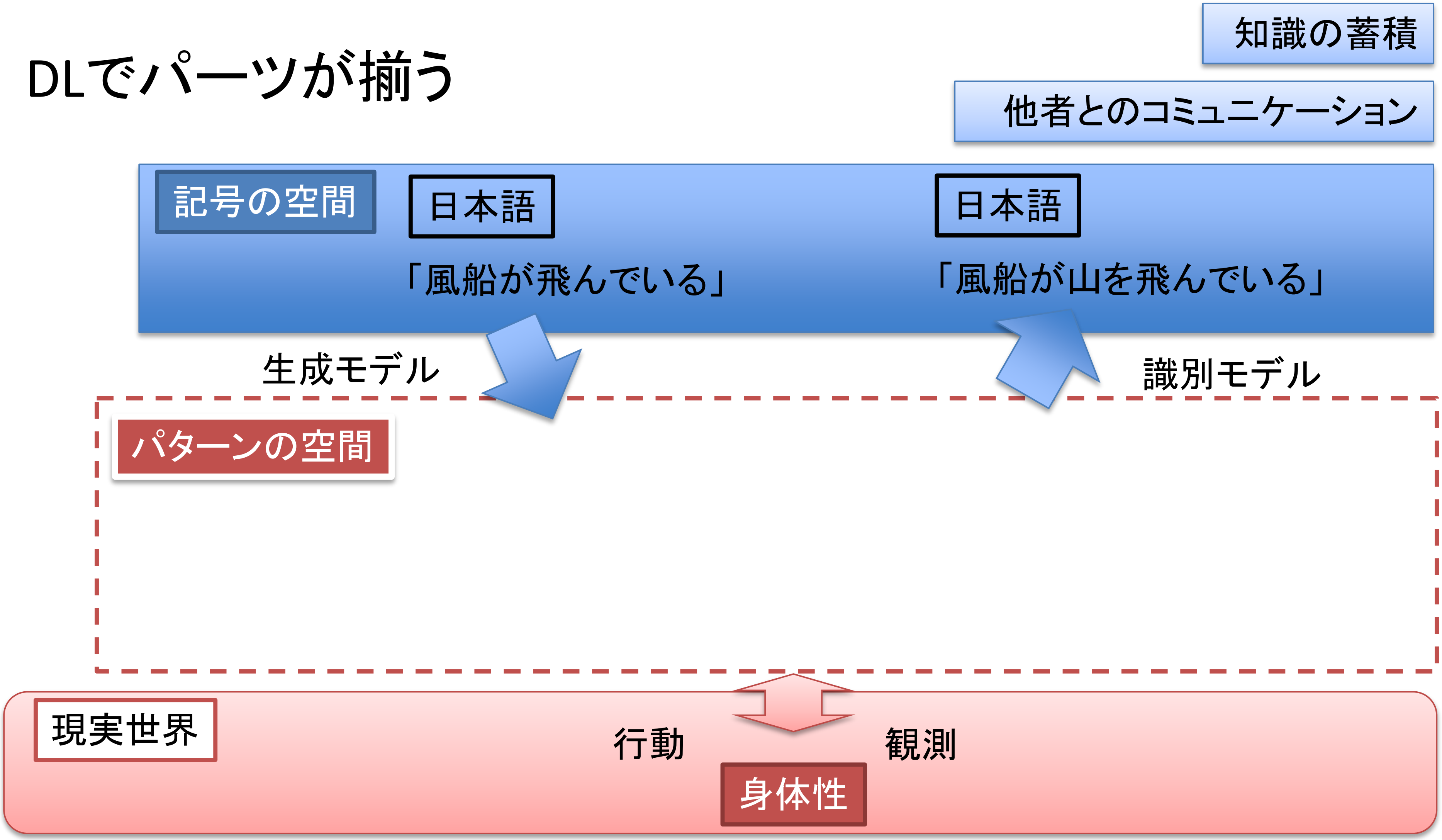
知識の蓄積

他者とのコミュニケーション



- いままでの推論（述語・命題論理による演繹や帰納・仮説推論）は、記号の空間だけでやろうとしてきた。
- 思考とは、パターンの空間と記号の空間をいったりきたりすること。（ラプラス変換や周波数変換と近い。）
- この上に、言語によるコミュニケーションや知識の蓄積が構成される。
- いずれも目的は、「いかに少ないサンプルで自由度の高いモデルを同定するか」

DLでパーツが揃う



- いままでの推論（述語・命題論理による演繹や帰納・仮説推論）は、記号の空間だけでやろうとしてきた。
- 思考とは、パターンの空間と記号の空間をいったりきたりすること。（ラプラス変換や周波数変換と近い。）
- この上に、言語によるコミュニケーションや知識の蓄積が構成される。
- いずれも目的は、「いかに少ないサンプルで自由度の高いモデルを同定するか」

「子どもの人工知能」と「大人の人工知能」

- 大人の人工知能：ビッグデータから人工知能へという持続的イノベーション
  - ビッグデータ全般、IoT全般、ワトソン、Siri、Pepper…
  - 一見すると専門家（大人）ができることができるが、人間が裏で作りこんでいる。
  - 販売、マーケティングなど。今後は、医療、金融、教育など

- 子どもの人工知能：ディープラーニングを突破口とする破壊的イノベーション
  - ディープラーニングを中心とする発展
  - 子どものできることができるようになっている
  - 人間の発達と同じような技術進化：認識能力の向上、運動能力の向上、言語の意味理解という順で技術が進展する
  - ものづくり中心。

特徴量の設計を人間がやらないといけないのが大人の人工知能、
やらなくてよいのが子どもの人工知能

既存産業の発展

農業	収穫判定	トラクター、コンバインの 適用範囲拡大、効率向上 選別調製等の自動化	自動での収穫 自動での耕うん
建設	測量	掘削、基礎工事、 外装内装作業等の 効率向上	多くの作業の 自動化・効率化
食品 加工	振り分け 確認	カット、皮むき、解体等 の自動化	多くの加工工程の 自動化
組み立て 加工	目視確認の 自動化	動作効率の向上	段取りの自動化 セル生産の自動化

⋮

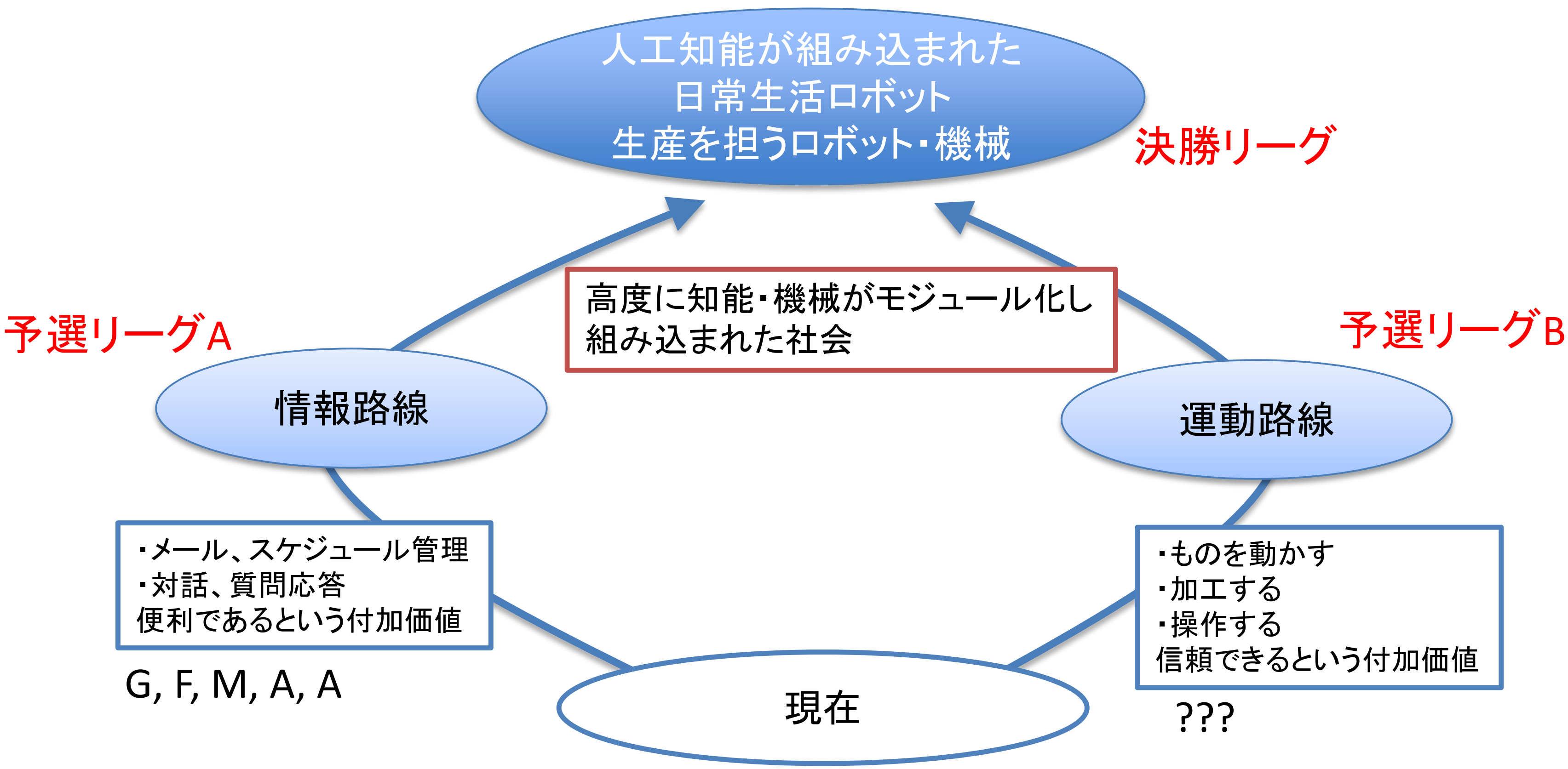
A: 画像認識　　B: 運動の習熟　　C: 計画立案を伴う運動

変化の本質

- 認識（画像・映像）
  - 世の中に、画像認識ができないから人間がやっている仕事がたくさんある。そこが自動化される。
  - コストが下がる。監視のコストは100分の1以下になる。
    - 森林の管理や災害の監視も。新たな事業が次々と。
- 運動の習熟
  - 我々は、機械は「機械的な動き」しかできない、ロボットは「ロボット的な動き」しかできないと思い込んでいる。（まさにこの形容詞が表している。）
  - 機械も習熟するし、ロボットも上達するようになる。
    - 自然物を相手にしているものは場面場面で状況が異なるので、そもそも自動化が難しかった。それが自動化される。
    - 例えば、農業、建設、食品加工。
    - さらには、日常生活のロボット、生産・仕事を担う機械・ロボットが実現される。

日本は運動路線のほうが戦いやすい

- 最終的には、日常生活、仕事におけるロボット・機械の活用。状況ごとに個別性があるので、認識能力がない状況では対応できなかった。ここにどう至るかが鍵。
- 情報路線で行く道（Google, Facebook系）と、運動路線で行く道があるのではないか。
- 海外企業・研究者は機械・ロボットに苦手意識
- 予選を勝ち進んだ企業が決勝に進むイメージ



世界の動きは早い：画像の世界

**Netatmo、Deep Learningを使った
屋外用監視カメラを発表(2016/1/5)**

著作権等の都合により、
ここに挿入されていた画像を削除しました

ウェブニュース
TechCrunch
CES 2016: ウェザーステーションのNetatmo、
屋外用監視カメラを発表
2016年1月5日
http://jp.techcrunch.com/2016/01/05/20160104netatmo-makes-outdoor-security-cameras-suck-less/

著作権等の都合により、
ここに挿入されていた画像を削除しました

ウェブニュース
TechCrunch
Placemeterは歩行者数を計測し、実世界のコンバージョン率を導き出す
2015年9月24日
http://jp.techcrunch.com/2015/09/24/placemeter-measures-retail-shop-conversion-rates/

インタフェースや医療も変わる

Apple、感情認識のAI企業Emotientを買収(2016/1/8)

著作権等の都合により、
ここに挿入されていた画像を削除しました

ウェブニュース
IT Media News
Apple、感情認識のAI（人工知能）企業Emotientを買収—Wall Street Journal報道
2016年1月8日
http://www.itmedia.co.jp/news/articles/1601/08/news093.html

ディープラーニングの肺がん検出率は人間より上、米Enlitic (2016/1/5)

著作権等の都合により、
ここに挿入されていた画像を削除しました

ウェブニュース
日経BP ITPro
ディープラーニングの肺がん検出率は人間より上、スタートアップの米Enlitic
2016年1月5日
http://itpro.nikkeibp.co.jp/atcl/column/15/061500148/122400043/?ST=bigdata&P=2

眼をもった機械の可能性
（認識系技術）

黒：さっさとやればいいもの（プロトタイプ開発）
赤：研究が必要なもの
青：議論が必要な戦略論

- 警備、防犯技術
- 介護施設、病院、独居老人等の見守り技術
- 防犯や交通違反検知を含めた社会インフラ構築
- 顔による認証・ログイン・広告技術
- わいせつ画像判定、意匠の類似判定等、既存領域での画像活用
- 表情読み取り技術（サービス業全般にきわめて重要、嘘発見技術も）
- 顔認証含めたより根本的なプライバシーリスク検討（人が写った画像・映像は個人情報か、特徴量利用の制限スキーム、本人認証書類の変化等も含め）
- 国家の安全保障、入国管理、警察業務、輸出入管理業務等での利用
- 実世界最適化支援（店舗内行動、街づくり等）
- 防災系画像処理（河川、火山、土砂崩れ）
- 医療画像処理（X線、CT、皮膚、心電図）
- コンテンツ生成系（アート、デザイン、広告制作）
- コンテンツ生成の発展系（深層生成モデルの発展。実写代替技術、アニメや映画）
- 一般数値データ異常監視（プラント、打音検査、情報セキュリティ等）
- 日本語の一般音声認識技術
- 画像認識系に関する戦略論（協調と競争のすみわけ、どこで競争力の差が生まれるか）
- 防犯による不動産価値向上、夜間活用、森林活用等の土地・場所の活用に関する検討

眼をもった機械の可能性（運動系技術）

- 重機系（掘削、揚重）
- 建設現場系（セメント固め、溶接、運搬、取り付け）
- 農業系（収穫、選果、防除、摘花・摘果）
- 自動操縦系（ドローン、小型運搬車、農機、建機）
- 自動運転系（ドリフトする自動運転車、時速300キロの自動運転車、全力で飛び込んでくる歩行者に対応できる自動運転車）
- 産業用ロボット系（特に組み立て加工等）
- 調理系（牛丼、炊飯）
- ペットロボット系（AIBO＋強化学習）
- 医療・介護・バイオ系（手術ロボ、介護ロボ、実験ロボ）
- 廃炉系（深海や鉱山、宇宙も含めた極限環境）
- ピッキング系基礎技術開発（ハードとのすり合わせ、片付けロボや陳列ロボ）
- 学習効率を上げるためのシミュレーション・試作・転移学習等の基礎技術開発
- 学習工場、プラットフォーム化に関する検討と戦略論（どういう切り出しにより「Intel inside」あるいは「MicrosoftのWindows化」を実現するか。安全性、信頼性等の競争力につなげる方法論など）
- 軍事目的での利用に関する潜在的リスク・可能性の検討

学術俯瞰講義
記号とパターンのはざまに

松尾 豊

そもそもの動機

- 深い階層のニューラルネットワークを作りたい
  – 現実世界はおそろしく非線形

- 「意識下の処理」が実現できていない。
  - 「二つのモデル（コネクショニストモデルと記号表現モデル）の弱点を乗り越えた情報処理モデルは、まだ現れているとはいえない。この点を乗り越え、記号を使う心のはたらきとそうでないはたらきが意識のうえと意識下でどう相互作用しているのか、とくに、人が知覚や運動の機能をはたらかせながら、その一方で記号によってことばを操れるのはなぜかという問題に答を与えることは、多くの認知科学者が挑戦してきた主要な課題の一つである。」（安西祐一郎 「心と脳 –認知科学入門」（岩波新書）, 2011 ）

- モラベックのパラドックス：「 子供のできることほど難しい。」
  – 高度な推論よりも、認識や運動スキルの方が難しい。

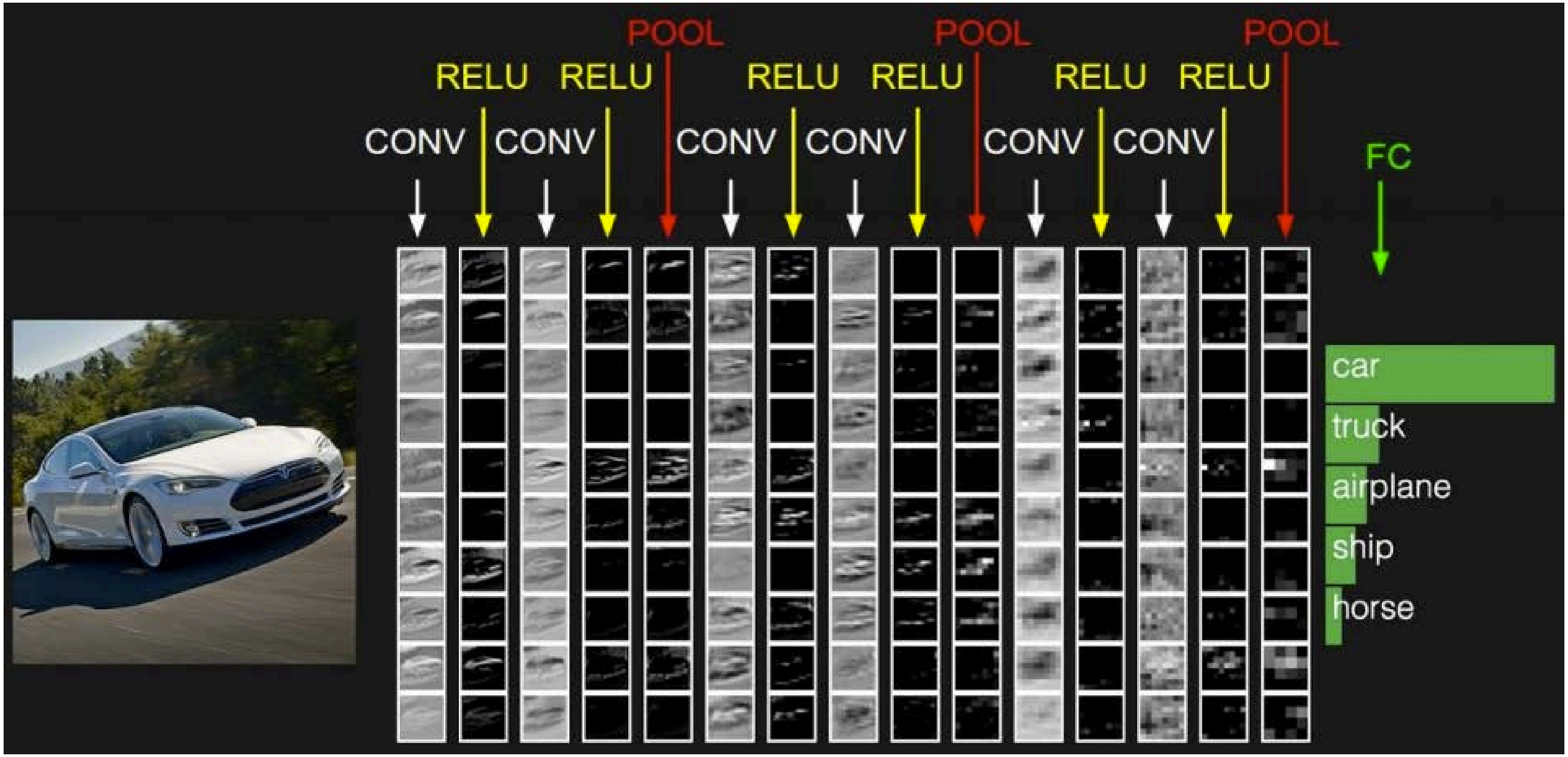
ところが、DLにより状況が変わりつつある。

もくじ

- CNN
- RNN
- 生成モデルから世界シミュレータへ
- 深層強化学習からプランニング
- 言語からの画像生成、その先へ

Convolutional Neural Network

CS231n: Convolutional Neural Networks for Visual Recognition
Convolutional Neural Networks (CNNs / ConvNets)
http://cs231n.github.io/convolutional-networks/
The activations of an example ConvNet architecture.



CS231n: Convolutional Neural Networks for Visual Recognition
Convolutional Neural Networks (CNNs / ConvNets)
http://cs231n.github.io/convolutional-networks/
The activations of an example ConvNet architecture.

CNNの歴史

Hubel & Wiesel,

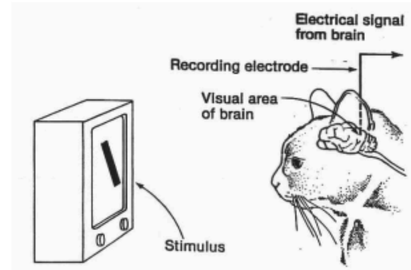
1959

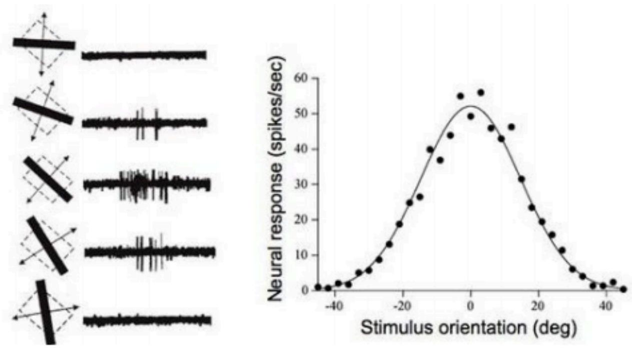
RECEPTIVE FIELDS OF SINGLE NEURONES IN
THE CAT'S STRIATE CORTEX

1962

RECEPTIVE FIELDS, BINOCULAR INTERACTION
AND FUNCTIONAL ARCHITECTURE IN THE CAT'S VISUAL CORTEX

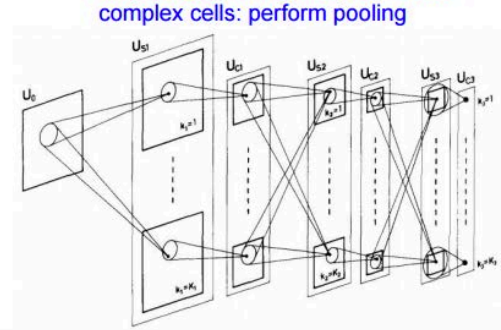
1968...

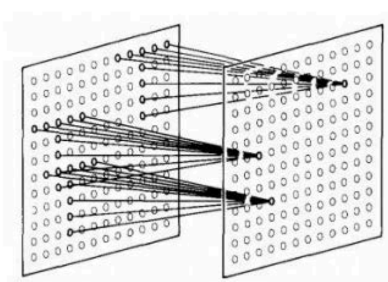


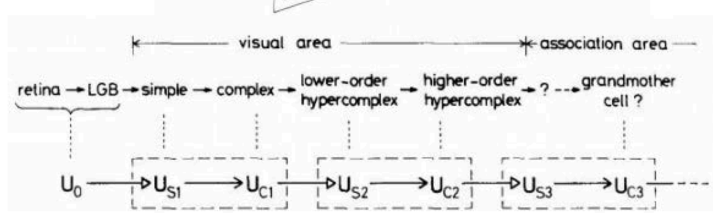


Neurocognitron
[Fukushima 1980]

"sandwich" architecture (SCSCSC...)
simple cells: modifiable parameters
complex cells: perform pooling







Fast-forward to today:ConvNets are everywhere

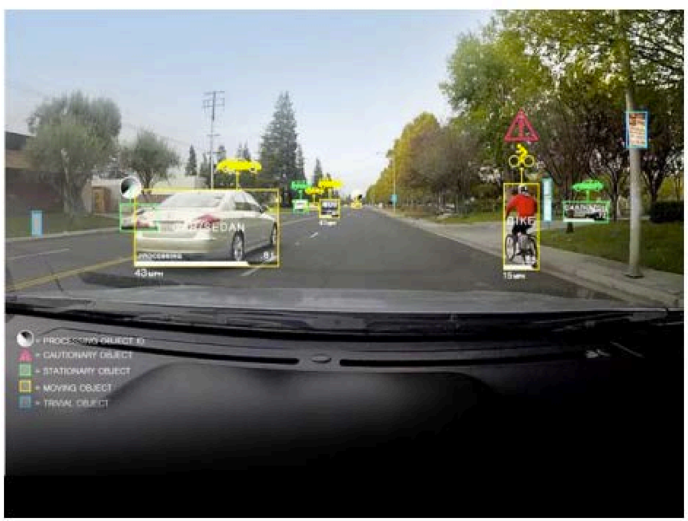
著作権等の都合により、
ここに挿入されていた画像を削除しました

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun,
"Faster R-CNN: Towards Real-Time Object Detection
with Region Proposal Networks"
https://arxiv.org/pdf/1506.01497.pdf
p. 4, Figure 3: Right: Example detections using RPN
proposals on PASCAL VOC 2007 test.

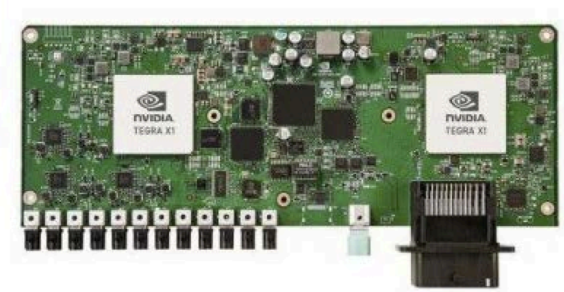
著作権等の都合により、
ここに挿入されていた画像を削除しました

clement.farabet.net
Scene Parsing
http://www.clement.farabet.net/research.html#parsing

Fast-forward to today: ConvNets are everywhere



self-driving cars

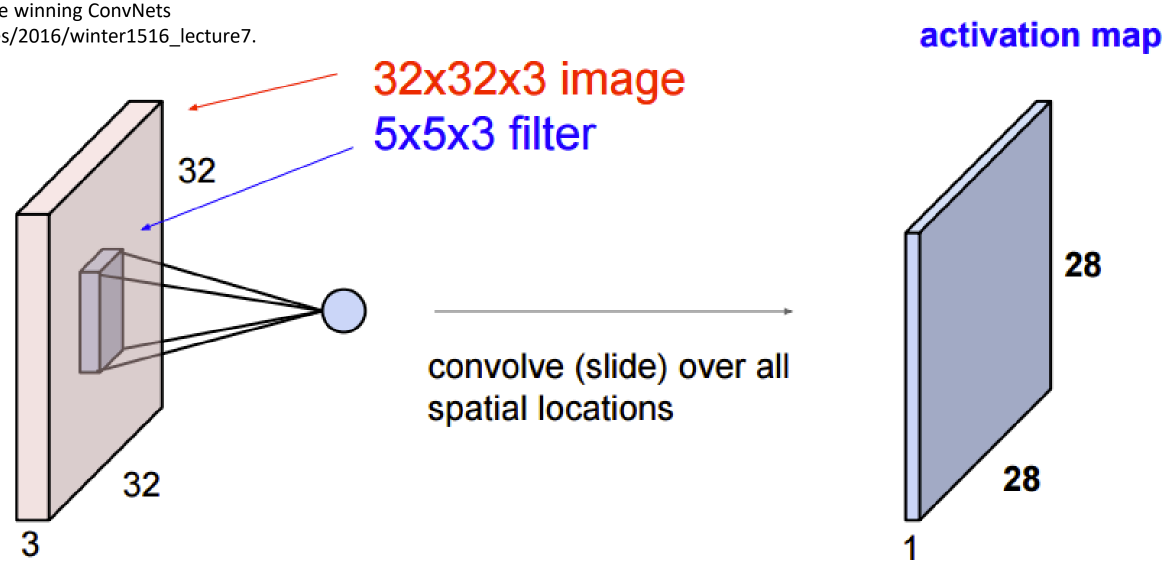


NVIDIA Tegra X1

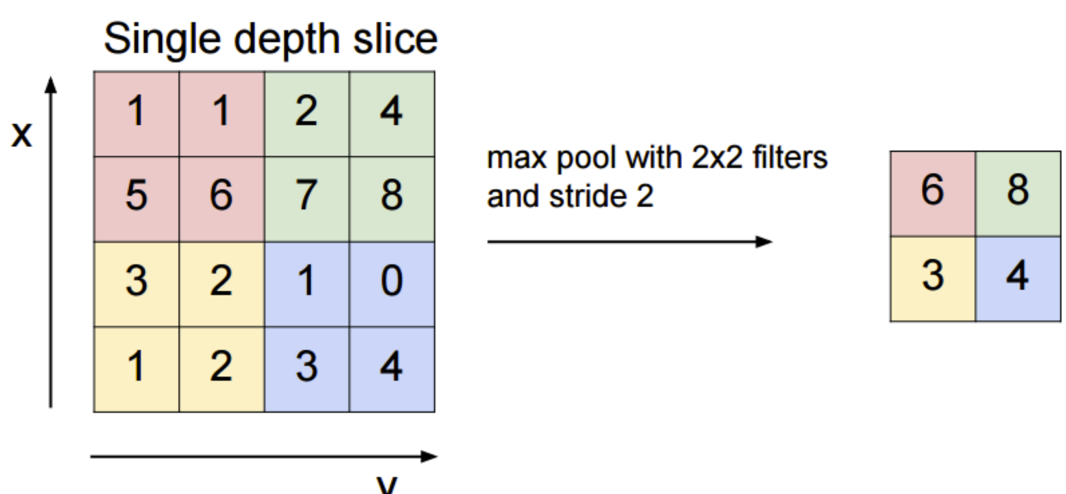
NVIDIA
自動運転車向け AI 車載コンピューター
http://www.nvidia.co.jp/object/drive-px-jp.html

CS231n: Convolutional Neural Networks for Visual Recognition
Schedule and Syllabus
http://cs231n.stanford.edu/syllabus.html
Lecture 27 Jan Convolutional Neural Networks: architectures, convolution / pooling layers
Case study of ImageNet challenge winning ConvNets
http://cs231n.stanford.edu/slides/2016/winter1516_lecture7.pdf

CONVOLUTION



MAX POOLING



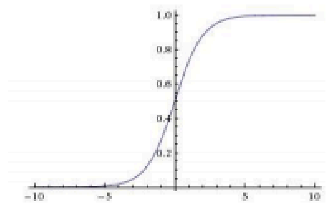
CS231n: Convolutional Neural Networks for Visual Recognition
Convolutional Neural Networks (CNNs / ConvNets)
http://cs231n.github.io/convolutional-networks/
Pooling layer downsamples the volume spatially, independently in each depth slice of the input volume.

CS231n: Convolutional Neural Networks for Visual Recognition
Schedule and Syllabus
http://cs231n.stanford.edu/syllabus.html
Lecture 6 Jan 25
http://cs231n.stanford.edu/slides/2016/winter1516_lecture6.pdf

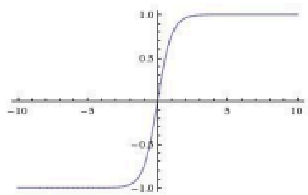
なぜうまくいくようになったか：活性化関数の進化

Sigmoid

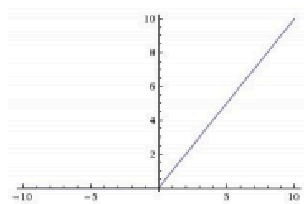
$$\sigma(x) = 1/(1+e^{-x})$$



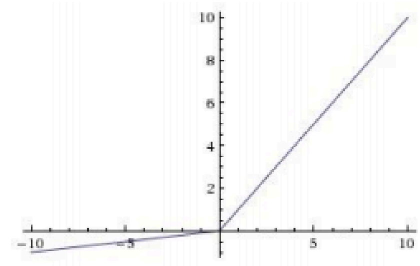
tanh tanh(x)



ReLU max(0,x)



Leaky ReLU
max(0.1x, x)



Maxout $\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha\,(\exp(x) - 1) & \text{if } x \leq 0 \end{cases}$$

ReLUが使われるようになったのが大きい。
2000年に提案されたが、2012年ごろから使われるようになった。

なぜうまくいくようになったか：Batch Normalization (2015)

- 勾配消失、勾配爆発に対して、かなり強力な武器
  - これでほとんど気にしなくてよくなった。
- バッチごとに正規化する。そのための層を入れる。

Batch Normalization［Ioffe and Szegedy,2015］

1. Compute the empirical mean and variance independently for each dimension.
2.Normalize

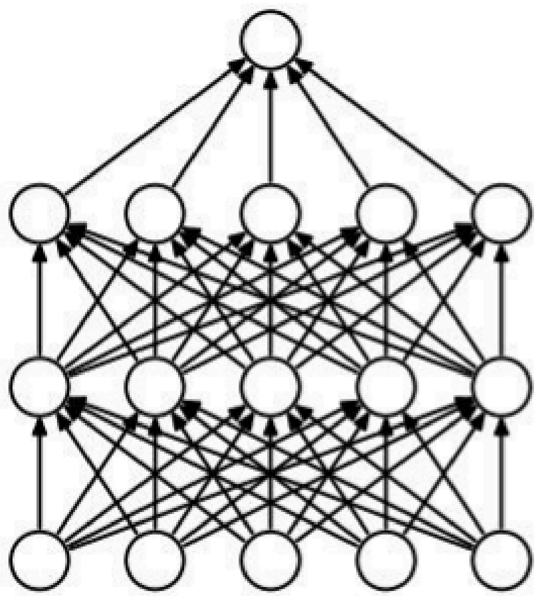
$$\hat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}]}}$$

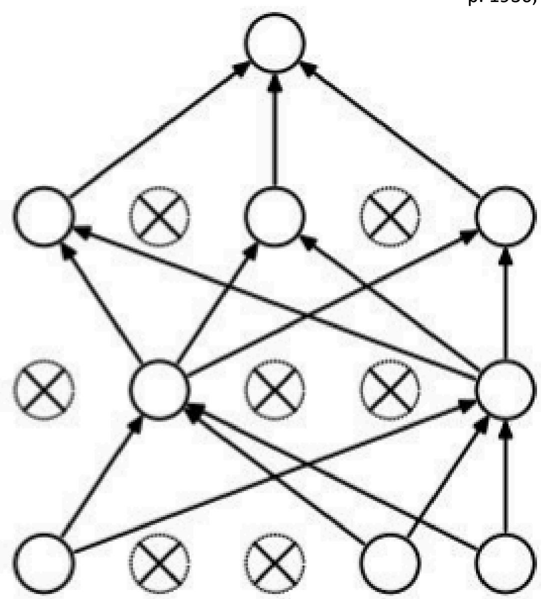
他にも、Normalization Propagation(2016), Weight Normalization(2016)など、改良した方法が次々と提案されている。

なぜうまくいくようになったか：Dropout

- ランダムに(0.5の確率で)ニューロンを停止させて訓練する。
- テスト時には、全部をonにして行う。
- 良くなることが2010年ごろから示されていた。

Nitish Srivastava et al., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting"
Journal of Machine Learning Research 15 (2014) 1929-1958
http://dl.acm.org/citation.cfm?id=2670313&CFID=914288649&CFTOKEN=84775229
p. 1930, Figure 1: Dropout Neural Net Model.


(a) Standard Neural Net


(b) After applying dropout.

[Srivastava et al., 2014]

たくさんのモデルのアンサンブルになっていることが理論的に示された。

1998年のConvolutional Network (LeCunら）

著作権等の都合により、
ここに挿入されていた画像を削除しました

Yann LeCun, "Gradient-based learning applied to document recognition"
Proceedings of the IEEE (Volume: 86, Issue: 11, Nov 1998)
Page(s): 2278 - 2324
http://ieeexplore.ieee.org/document/726791/
p. 2283, Fig. 2. Architecture of LeNet-5, a convolutional NN, here used for digits recognition.

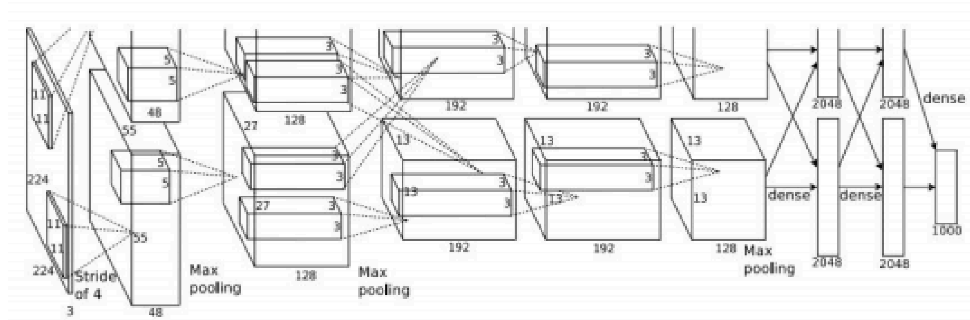
７層のCNN。
ただし、前処理＋SVMと精度は変わらなかった。

2012年に大きな躍進を遂げたモデル：ILSVRC2012のwinner

Krizhevsky, A., Sutskever, I., Hinton, G.: “ImageNet Classification with Deep Convolutional Neural Networks”, NIPS 2012
https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
p. 5, Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities
between the two GPUs.

Case Study: AlexNet

[Krizhevsky et al. 2012]



Full (simplified) AlexNet architecture:
[227x227x3] INPUT
[55x55x96] CONV1: 96 11x11 filters at stride 4, pad 0
[27x27x96] MAX POOL1: 3x3 filters at stride 2
[27x27x96] NORM1: Normalization layer
[27x27x256] CONV2: 256 5x5 filters at stride 1, pad 2
[13x13x256] MAX POOL2: 3x3 filters at stride 2
[13x13x256] NORM2: Normalization layer
[13x13x384] CONV3: 384 3x3 filters at stride 1, pad 1
[13x13x384] CONV4: 384 3x3 filters at stride 1, pad 1
[13x13x256] CONV5: 256 3x3 filters at stride 1, pad 1
[6x6x256] MAX POOL3: 3x3 filters at stride 2
[4096] FC6: 4096 neurons
[4096] FC7: 4096 neurons
[1000] FC8: 1000 neurons (class scores)

Details/Retrospectives:
- first use of ReLU
- used Norm layers (not common anymore)
- heavy data augmentation
- dropout 0.5
- batch size 128
- SGD Momentum 0.9
- Learning rate 1e-2, reduced by 10 manually when val accuracy plateaus
- L2 weight decay 5e-4
- 7 CNN ensemble: 18.2% -> 15.4%

- 8層（CONVが5層と、Fully Connected層が3層）
- ReLU、Dropoutが使われた。

ILSVRC2014のwinner

著作権等の都合により、
ここに挿入されていた画像を削除しました

Christian Szegedy et al., "Going deeper with convolutions"
2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
http://ieeexplore.ieee.org/document/7298594/
p. 4, Figure 2: Inception module
(b) Inception module with dimensionality reduction

複数の設定のConvolution層を混ぜたほうがよくなる。
FC層をなくして、パラメータ数を大きく減らした。

ILSVRC2015のWinner

著作権等の都合により、
ここに挿入されていた画像を削除しました

Kaiming He, Xiangyu Zhang,
Shaoqing Ren, Jian Sun
"Deep Residual Learning for Image Recognition"
https://arxiv.org/abs/1512.03385
p. 4, Figure 3. Example network architectures for ImageNet.

- これまでは、層を重ねると精度が悪くなることがあった。
- パラメータが増え、初期化が難しくなるため。
- そこで、ショートカットするコネクションを作った。少なくとも悪くはならない。

著作権等の都合により、
ここに挿入されていた画像を削除しました

kaiminghe.com
ICML 2016 Tutorial on Deep Residual Networks
tutorial slides
http://kaiminghe.com/icml16tutorial/icml2016_tutorial_deep_residual_networks_kaiminghe.pdf

性能の急激な伸び

著作権等の都合により、
ここに挿入されていた画像を削除しました

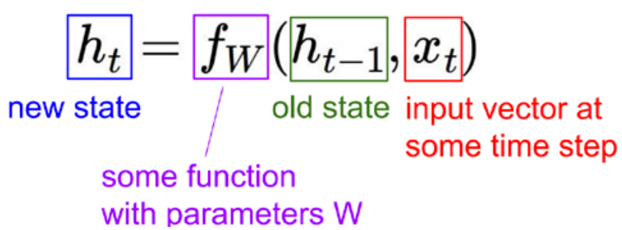
性能の伸びについてのグラフ
kaiminghe.com
ICML 2016 Tutorial on Deep Residual Networks
tutorial slides
http://kaiminghe.com/icml16tutorial/icml2016_tutorial_deep_residual_networks_kaiminghe.pdf

性能の急激な伸び。層がますます深くなっている。
2016年3月には、3.06%のモデルも出た。Human Error ～5.1%。

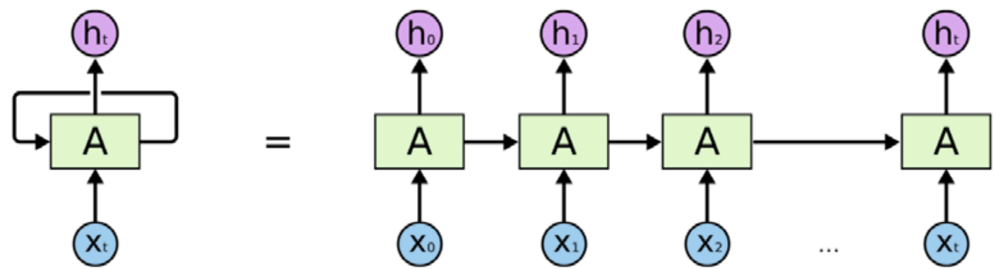
RNN

RNN

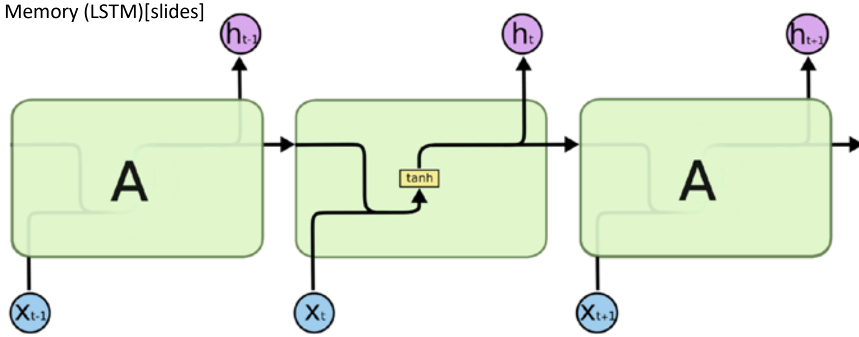
- 1980代から
- 最近では、LSTMというモデルが使われるようになったことが大きい。



CS231n: Convolutional Neural Networks for Visual Recognition
Schedule and Syllabus
http://cs231n.stanford.edu/syllabus.html
Lecture　Feb 8
Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM)[slides]



An unrolled recurrent neural network.

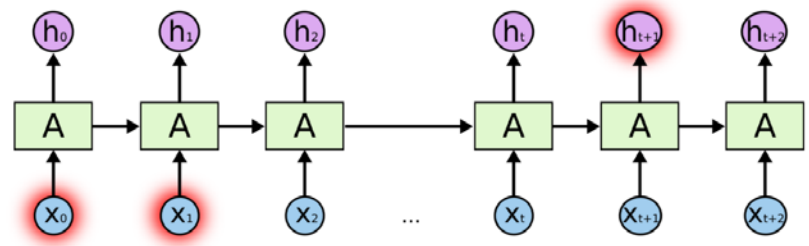


The repeating module in a standard RNN contains a single layer.

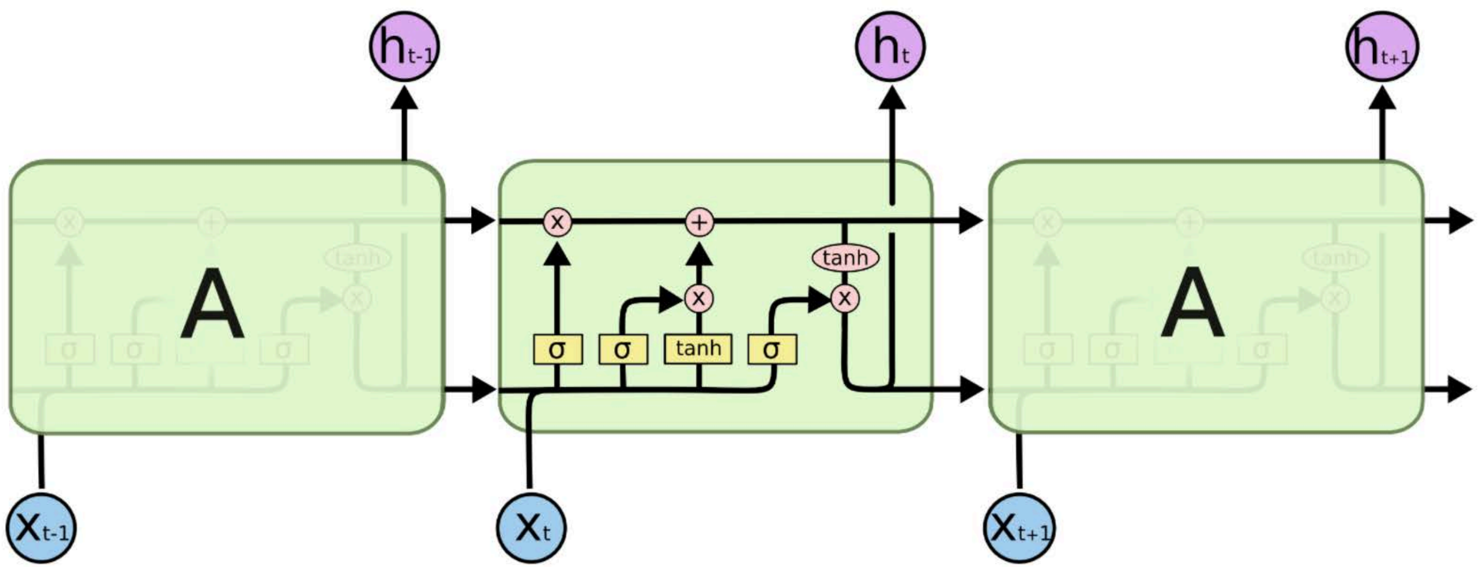
colah's blog
Understanding LSTM Networks
http://colah.github.io/posts/2015-08-Understanding-LSTMs/
An unrolled recurrent neural network (RNN-unrolled)
The repeating module in a standard RNN contains a single layer (LSTM3-SimpleRNN)
（最終閲覧日：2017年7月21日）

LSTM

colah's blog
Understanding LSTM Networks
http://colah.github.io/posts/2015-08-Understanding-LSTMs/
Unfortunately, as that gap grows, RNNs become unable to learn to connect the information. (RNN-longtermdependencies)
The repeating module in an LSTM contains four interacting layers.(LSTM3-chain)
（最終閲覧日：2017年7月21日）



- Long-Short Term Memory
  - Hochreiter & Schmidhuber (1997)
- 3つのゲート
  - Forget gate, Input gate, Output gate
- Long-term dependencyが捉えられる
- 似たような変種に GRU: Gated Recurrent Unit (2014)も提案されている。

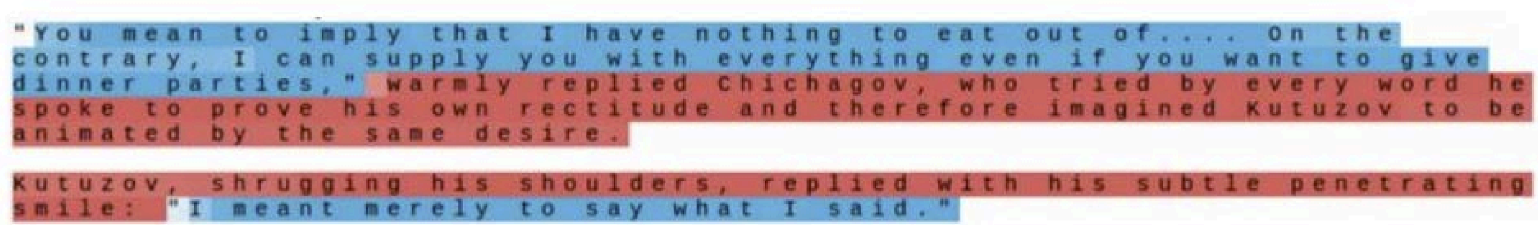


The repeating module in an LSTM contains four interacting layers.

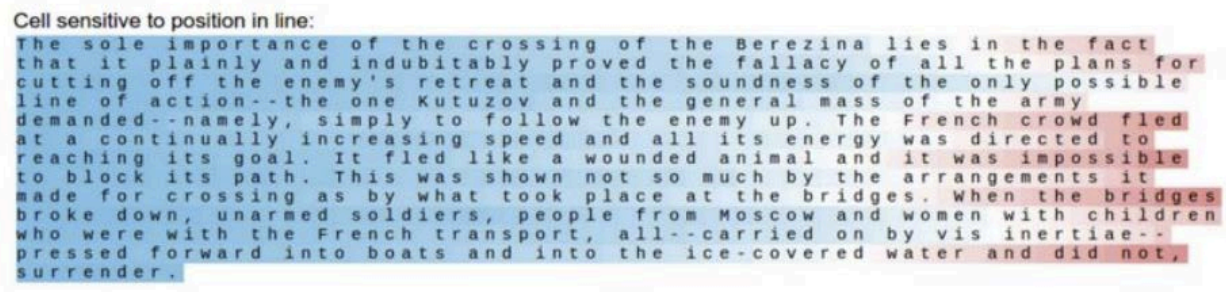
RNNが何を学習しているか

- Wikipediaの文章や、LinuxのカーネルをLSTMで学習
  - プログラムを、文字ごとのシーケンスと思って学習させる。
  - 6,206,996 characters
- それぞれのCellが何を表しているのかを可視化する。
- すると、面白いことが学習されている。

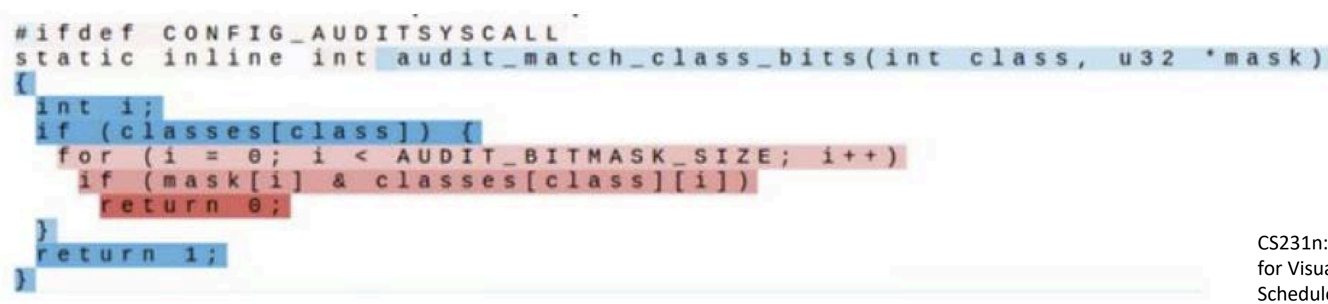
CS231n: Convolutional Neural Networks for Visual Recognition
Schedule and Syllabus
http://cs231n.stanford.edu/syllabus.html
Lecture Feb 8
Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM)[slides]



quote detection cell

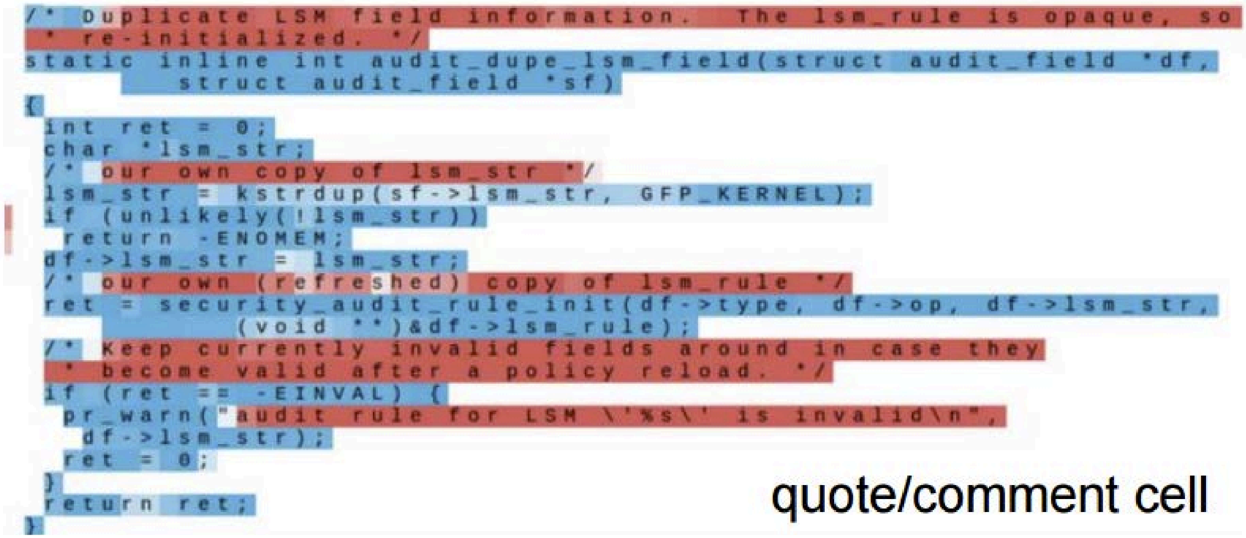


line length tracking cell

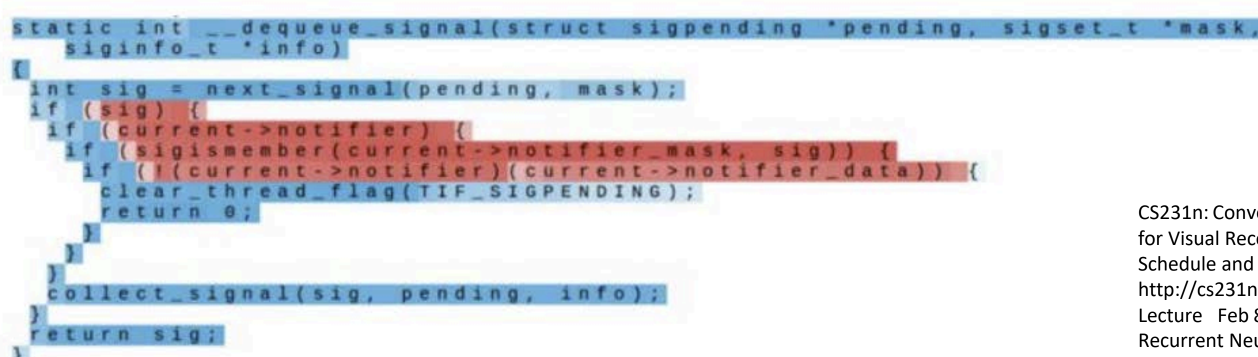


CS231n: Convolutional Neural Networks for Visual Recognition
Schedule and Syllabus
http://cs231n.stanford.edu/syllabus.html
Lecture Feb 8
Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM)[slides]

code depth cell



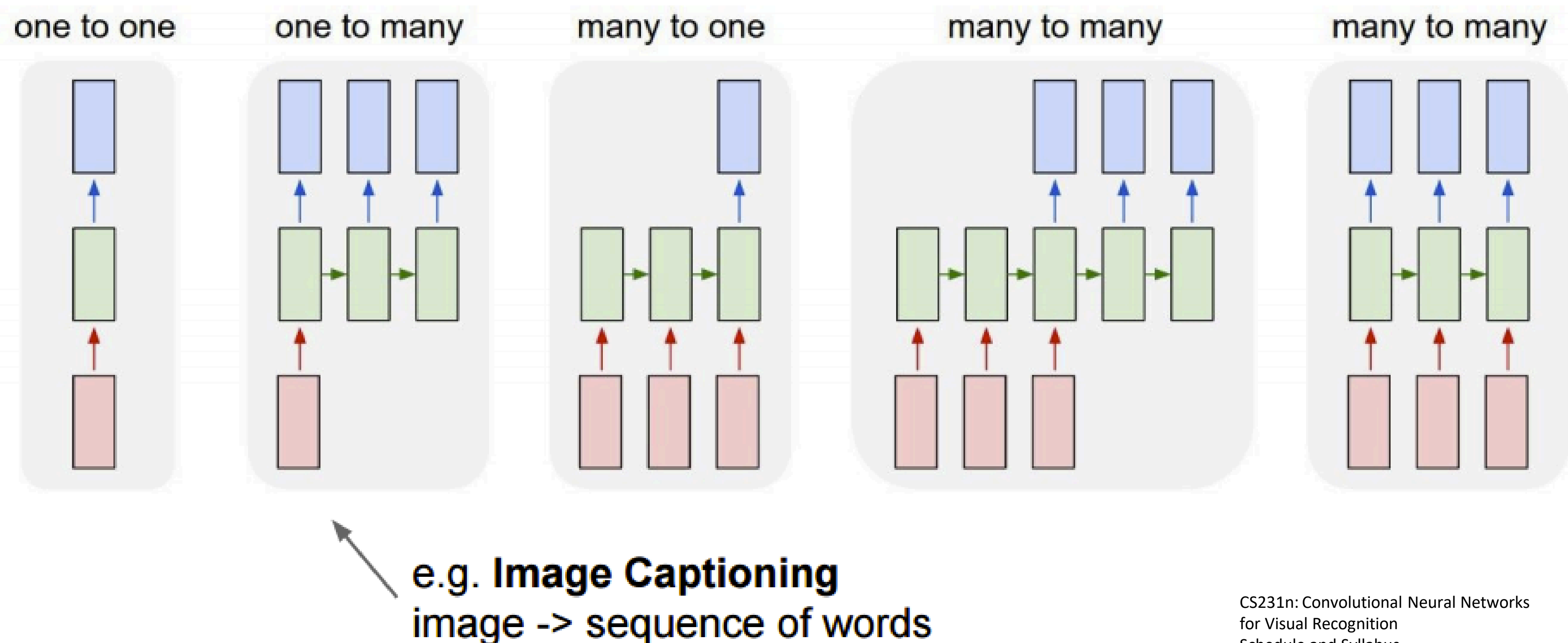
quote/comment cell



CS231n: Convolutional Neural Networks for Visual Recognition
Schedule and Syllabus
http://cs231n.stanford.edu/syllabus.html
Lecture Feb 8
Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM)[slides]

if statement cell

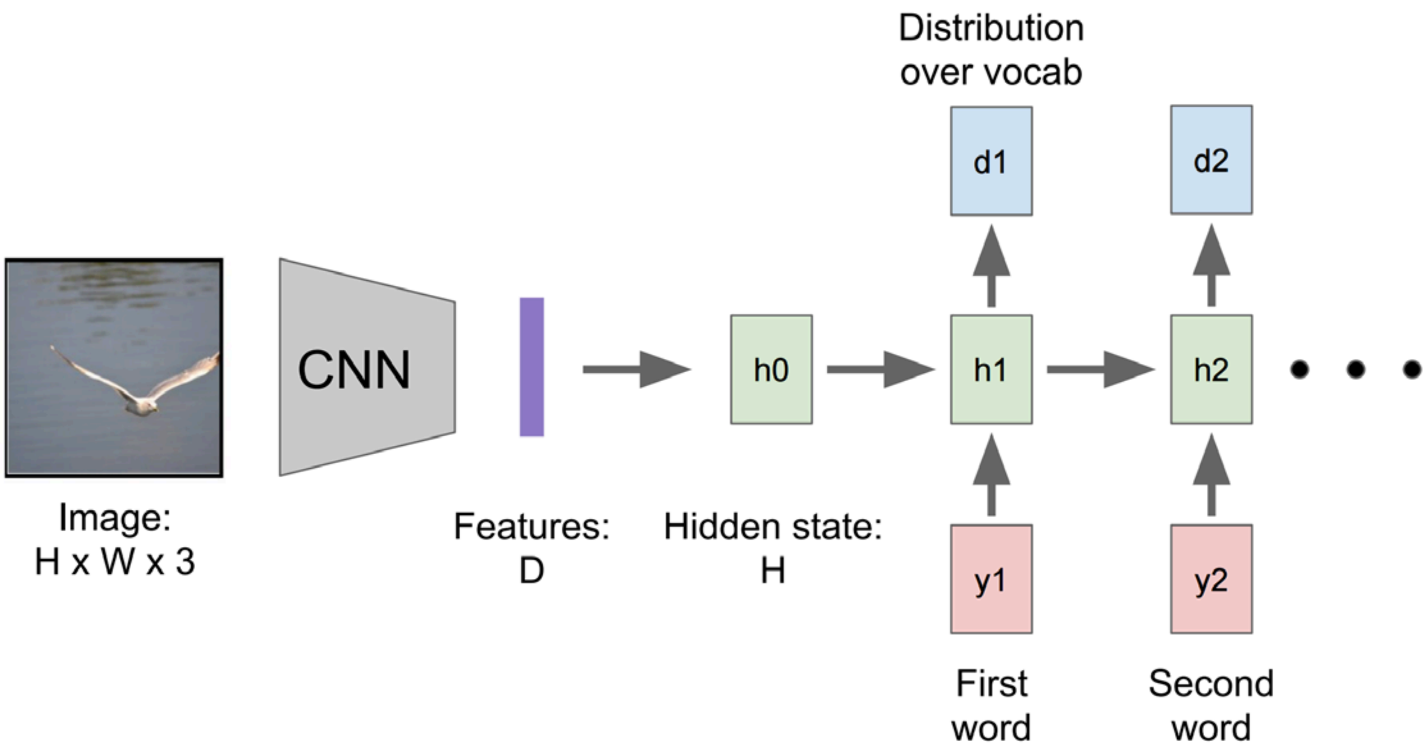
RNNは、さまざまな形で用いることができる。



CS231n: Convolutional Neural Networks for Visual Recognition
Schedule and Syllabus
http://cs231n.stanford.edu/syllabus.html
Lecture Feb 8
Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM)[slides]

RNN for Image captioning

- CNNで特徴量を出して、RNN(LSTM)に入れる。



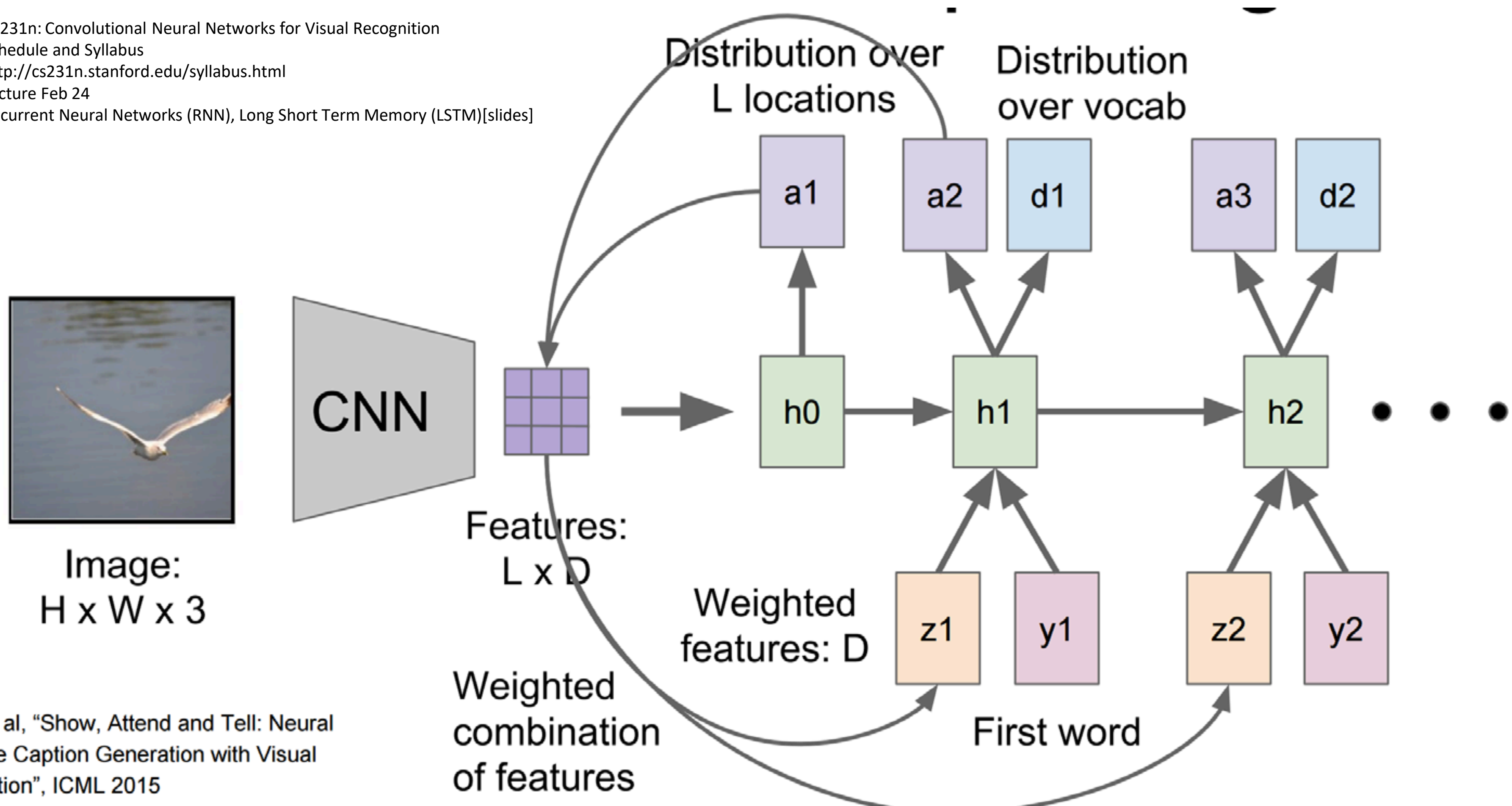
CS231n: Convolutional Neural Networks for Visual Recognition
Schedule and Syllabus
http://cs231n.stanford.edu/syllabus.html
Lecture Feb 24
Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM)[slides]

出したい文:A bird flying over a body of water.

Soft attention for captioning

- 画像全体を最初に入力するのではうまくいかない。
- 都度、画像の該当位置にアテンションをかけられないか

CS231n: Convolutional Neural Networks for Visual Recognition
Schedule and Syllabus
http://cs231n.stanford.edu/syllabus.html
Lecture Feb 24
Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM)[slides]

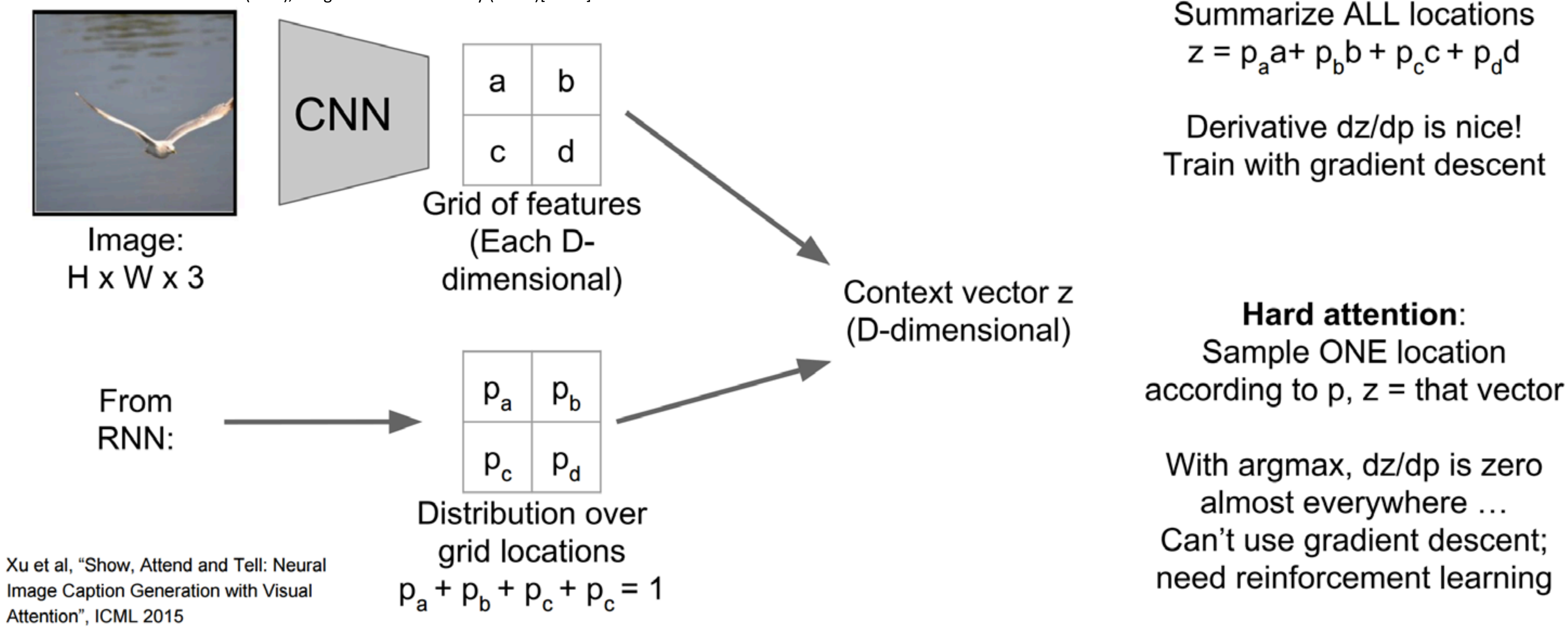


Xu et al, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015

Attentionの仕組み

- 画像のグリッドへの重み。これを学習する。

CS231n: Convolutional Neural Networks for Visual Recognition
Schedule and Syllabus
http://cs231n.stanford.edu/syllabus.html
Lecture Feb 24
Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM)[slides]

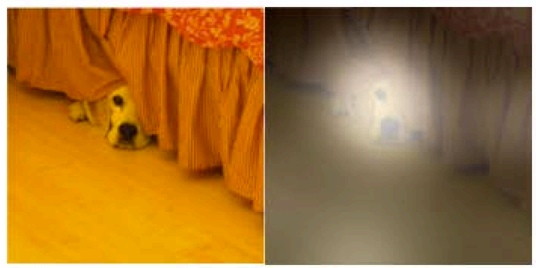


Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (2015)

- Microsoft COCOデータセット：82,783画像、5個の文
- Flickr8k/30k データセット：8,000/30,000枚の画像、5個の文
- 画像をいれると、キャプションを自動で出せるようになる。

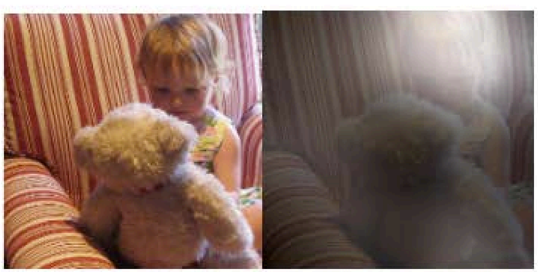
Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention"
https://arxiv.org/abs/1502.03044
p. 2, Figure 3. Examples of attending to the correct object

A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

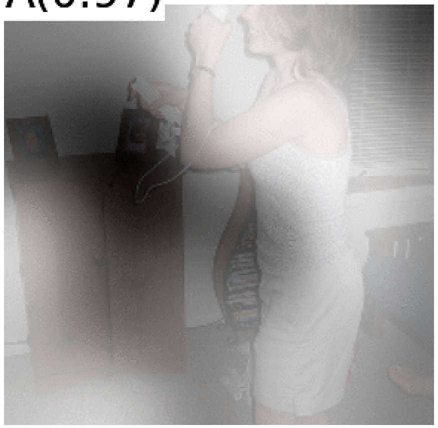
A giraffe standing in a forest with trees in the background.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (2015)

- Microsoft COCOデータセット：82,783画像、5個の文
- Flickr8k/30k データセット：8,000/30,000枚の画像、5個の文
- 画像をいれると、キャプションを自動で出せるようになる。

Kelvin Xu
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention
http://kelvinxu.github.io/projects/capgen.html
The model in action

A(0.97)



A(0.99)

A(0.99)

A(0.96)

Neural Machine Translation

- Ilya Sutskever, Oriol Vinyals, Quoc V. Le (Google), Sequence to Sequence Learning with Neural Networks (2014)
  - WMT'14 English to French タスクで、34.81 (BLEU score)

著作権等の都合により、
ここに挿入されていた画像を削除しました

Ilya Sutskever, Oriol Vinyals, Quoc V. Le, "Sequence to Sequence Learning with Neural Networks"
Proceeding NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems
Pages 3104-3112
http://dl.acm.org/citation.cfm?id=2969033.2969173
https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks
p. 2, Figure 1: Our model reads an input sentence "ABC" and produces "WXYZ" as the output sentence.

- Neural Machine Translation by Jointly Learning to Align and Translate (2015)
  - Bengioらの研究。Attentionに近い仕組みを入れて、精度を上げている。

Neural Machine Translation

- GoogleのNMT。2016年9月から。
- 日本語⇔英語にも11月から導入された。すごくいい。
- 8層のbi-directional RNN, attentionつき。5000 GPU?

著作権等の都合により、
ここに挿入されていた画像を削除しました

Google Research Blog
A Neural Network for Machine Translation, at Production Scale
https://research.googleblog.com/2016/09/a-neural-network-for-machine.html
Data from side-by-side evaluations, where human raters compare the quality of translations for a given source sentence. Scores range from 0 to 6, with 0 meaning “completely nonsense translation”, and 6 meaning “perfect translation.“

（松尾の意見）すごいけど、視覚系や強化学習、それに基づく予測モデルと組み合わされてないので、まだ本物の翻訳ではない。もっと上がるはず。

Deep reinforcement learning (deep RL) has been successful in learning sophisticated behaviors automatically; however, the learning process requires a huge number of trials. In contrast, animals can learn new tasks in just a few trials, bene- fiting from their prior knowledge about the world. This paper seeks to bridge this gap. Rather than designing a “fast” reinforcement learning algorithm, we propose to represent it as a recurrent neural network (RNN) and learn it from data. In our proposed method, RL2 , the algorithm is encoded in the weights of the RNN, which are learned slowly through a general-purpose (“slow”) RL algorithm. The RNN receives all information a typical RL algorithm would receive, including observations, actions, rewards, and termination flags; and it retains its state across episodes in a given Markov Decision Process (MDP). The activations of the RNN store the state of the “fast” RL algorithm on the current (previously unseen) MDP. We evaluate RL2 experimentally on both small-scale and large-scale problems. On the small-scale side, we train it to solve randomly generated multi-armed bandit problems and finite MDPs. After RL2 is trained, its performance on new MDPs is close to human-designed algorithms with optimality guarantees. On the largescale side, we test RL2 on a vision-based navigation task and show that it scales up to high-dimensional problems.

深い強化学習（深いRL）は、洗練された行動を自動的に学習するのに成功しています。しかし、学習プロセスには膨大な試行が必要です。これとは対照的に、動物は世界についての以前の知識から恩恵を受け、わずかな試行で新しい仕事を習得することができます。このペーパーは、このギャップを埋めようとしています。「高速」強化学習アルゴリズムを設計するのではなく、それをリカレントニューラルネットワーク（RNN）として表現し、それをデータから学習することを提案します。提案された方法RL2では、アルゴリズムはRNNの重みに符号化され、RNNは汎用（「遅い」）RLアルゴリズムによってゆっくり学習される。RNNは、観察、行動、報酬、および終了フラグを含む典型的なRLアルゴリズムが受け取るすべての情報を受信する。それは所与のマルコフ決定プロセス（MDP）においてエピソード全体にわたってその状態を保持する。RNNの活性化は、現在の（以前は見えなかった）MDP上の "高速" RLアルゴリズムの状態を記憶する。我々は、RL2を小規模問題と大規模問題の両方について実験的に評価する。小規模な面では、無作為に生成された複数武装の禁止問題と有限のMDPを解決するように訓練します。RL2が訓練された後、新しいMDPでの性能は、最適性が保証された人間が設計したアルゴリズムに近くなります。大規模な側面では、RL2をビジョンベースのナビゲーションタスクでテストし、高次元の問題までスケールアップすることを示します。

CNNやRNNによって、非常に深い関数が学習されている

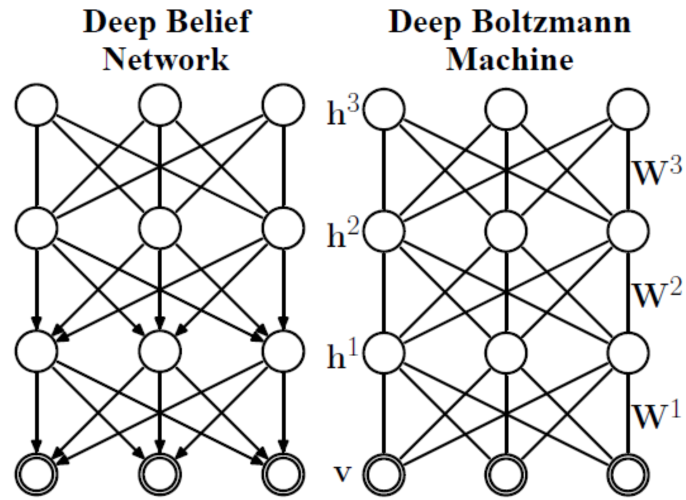
- アイディア自体は古くからあるもの
  - 計算機のパワーの大幅な向上
  - 細かい工夫の発見

- CNNもRNNも、結局かなり似てきた
  - 時間方向・空間方向に同一性を仮定して、パラメータを減らす。
  - 微分を一定に：CEC (Constant Error Carousel) ≒ Batch Normalization
  - 微分を遠くまで届ける：ResNetの考え方 ≒ LSTMの考え方

- CNNやRNNをブロックとして用いて、手法が構築されるようになってきた。

生成モデルから世界シミュレータへ

深層生成モデル



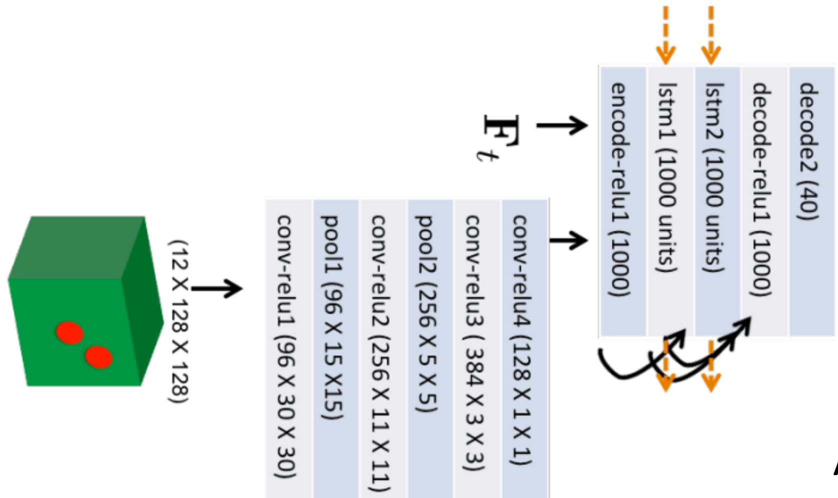
Ruslan Salakhutdinov, Geoffrey Hinton, "Deep Boltzmann Machines" Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics April 16-18, 2009, Clearwater Beach, Florida USA, Vol. 5:448-455. http://www.jmlr.org/proceedings/papers/v5/salakhutdinov09a/salakhutdinov09a.pdf
p. 451, Figure 2: Left: A three-layer Deep Belief Network and a three-layer Deep Boltzmann Machine.

- 深層生成モデル（deep generative model）
  - 潜在変数が多層になったモデル
  - より複雑なモデルを学習することができる.
- Deep belief network (stacked RBM) [Hinton+ 2006]
  - Deep Learningの元祖
  - 最終層以外は有向グラフ，最終層は無向グラフのグラフィカルモデル
- Deep Boltzmann machine [Salakhutdinov 2009]
  - 全ての層が無向グラフのグラフィカルモデル
- その後着目されなくなった.
  - 事前学習もされなくなり，教師あり学習が主流に.
- しかし2014年頃から，再び生成モデルが着目されるようになった.
  - それに合わせて教師なし学習も復権.
- 最近の深層生成モデルは次の2つが代表的
  - Variational autoencoder (VAE)
  - Generative adversarial network (GAN)

未来を描く:ビリヤードの球の動きを予期する

- Learning Visual Predictive Models of Physics for Playing Billiards (ICLR2016)
- ビリヤードの球の動きを(物理モデルを使わずに)学習する。
- CNN (Alexnet) + 2レイヤーのLSTM
  - AlexnetはImagenetで事前学習しておく
- 4フレーム分の画像が入力。20フレーム先を予測する。
- 1万サンプルで学習。データはシミュレーションで作る。

Katerina Fragkiadaki, Pulkit Agrawal, Sergey Levine, Jitendra Malik, "Learning Visual Predictive Models of Physics for Playing Billiards"
https://people.eecs.berkeley.edu/~katef/papers/Physics.pdf
p. 5, Figure 2: Network architecture.



Alexnet + 2レイヤーのLSTM

Figure 2: **Network architecture.** At each time step t, for each object, the network is provided with the previous four glimpses centered on the object's position, as well as the agent's applied forces $\mathbf{F}_t = (F_t^x, F_t^y)$ and the hidden states of the LSTM units from the previous time step. The output is ball displacements $\mathbf{u}_{t+k} = (\Delta x_{t+k}, \Delta y_{t+k})$ for $k = 1 \cdots h$ in the next h frames.

未来を描く:ビリヤードの球の動きを予期する

Katerina Fragkiadaki, Pulkit Agrawal, Sergey Levine, Jitendra Malik, "Learning Visual Predictive Models of Physics for Playing Billiards"
https://people.eecs.berkeley.edu/~katef/papers/Physics.pdf
p. 8, Figure 6: Visual Imaginations generated by our model.

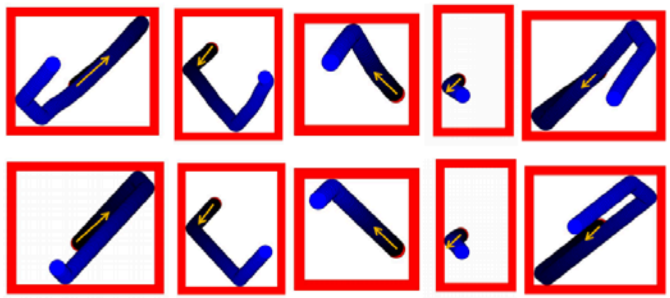


Figure 6: Visual Imaginations generated by our model. (Top) shows the trajectory of the ball generated by our model by iteratively rendering the next world image using the predicted ball velocities and feeding this rendered image as input to the model. The imagined ball trajectory is color coded to reflect the progression of time. Change in color from dark to light blue indicates moving forward in time. (Bottom) shows the respective ground truth trajectories. Force on the ball is applied at the first time step and the force vector is shown by the orange arrow. The figure shows that our model learns the dynamics of balls and collisions and is successfully able to image collision events.

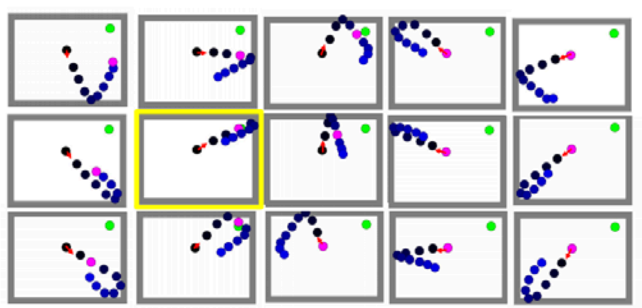


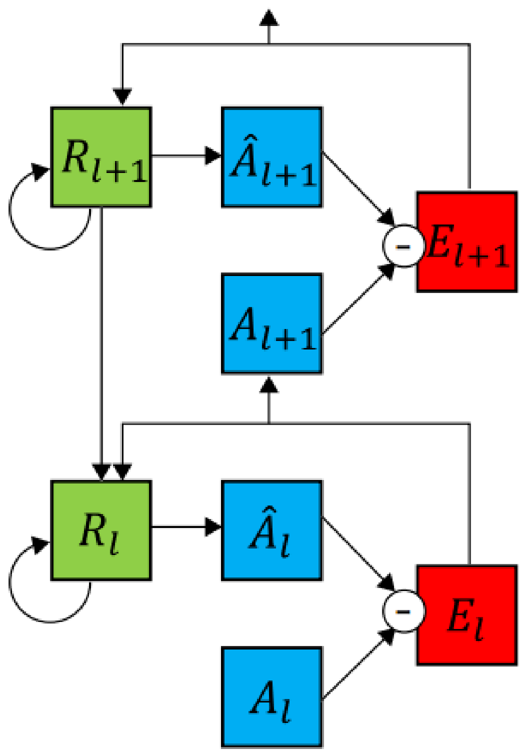
Figure 8: Illustration of the method for determining the force required to push the cue ball (the ball with the arrow) to a target location (shown in green). The agent runs multiple simulations of the world (i.e. visual imagination) by applying different forces (shown by red arrows). Each grey box shows the results of simulations performed by the agent. The position of the ball closest to the target location is shown in pink in each of the simulations. The agent chooses the force that leads to a ball location that is closest to the target location (simulation in yellow box).

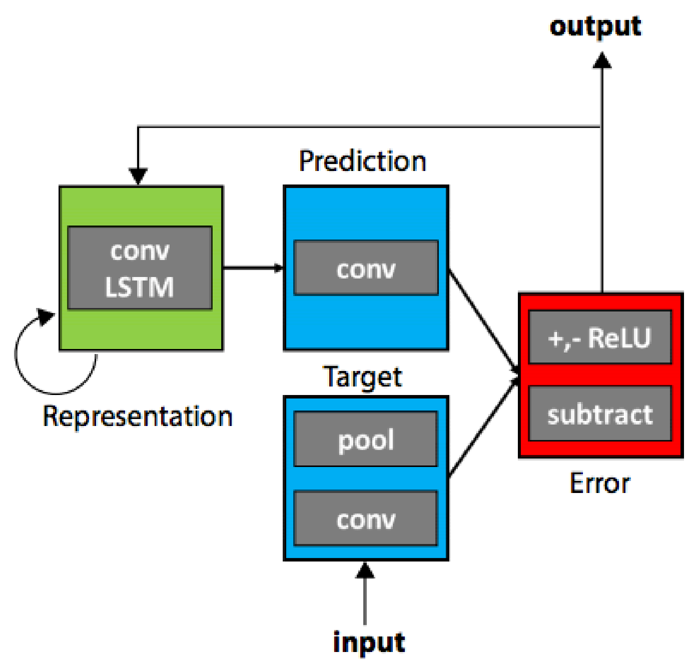
動画 https://www.youtube.com/watch?v=98qfuYdVnLw

Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning (2016)

- Predictive Codingを、CNNとLSTMで実現したもの。
- フレームの予測をする。誤差を予測するモデルを重ねる。
- ２つのデータセットで実験：合成データ、実際の映像

William Lotter, Gabriel Kreiman, David Cox, "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning"
https://arxiv.org/abs/1605.08104
p. 2, Figure 1: Predictive Coding Network (PredNet).





William Lotter, Gabriel Kreiman, David Cox, "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning"
https://arxiv.org/abs/1605.08104
p. 5, Figure 2: PredNet next-frame predictions for sequences of rendered faces rotating with two degrees of freedom.

- 16,000枚の合成画像。10フレーム分。
- 最初の2フレームで残りが予測できる。
- 深い層のセルが、方向などの抽象的な特性をコーディングしている

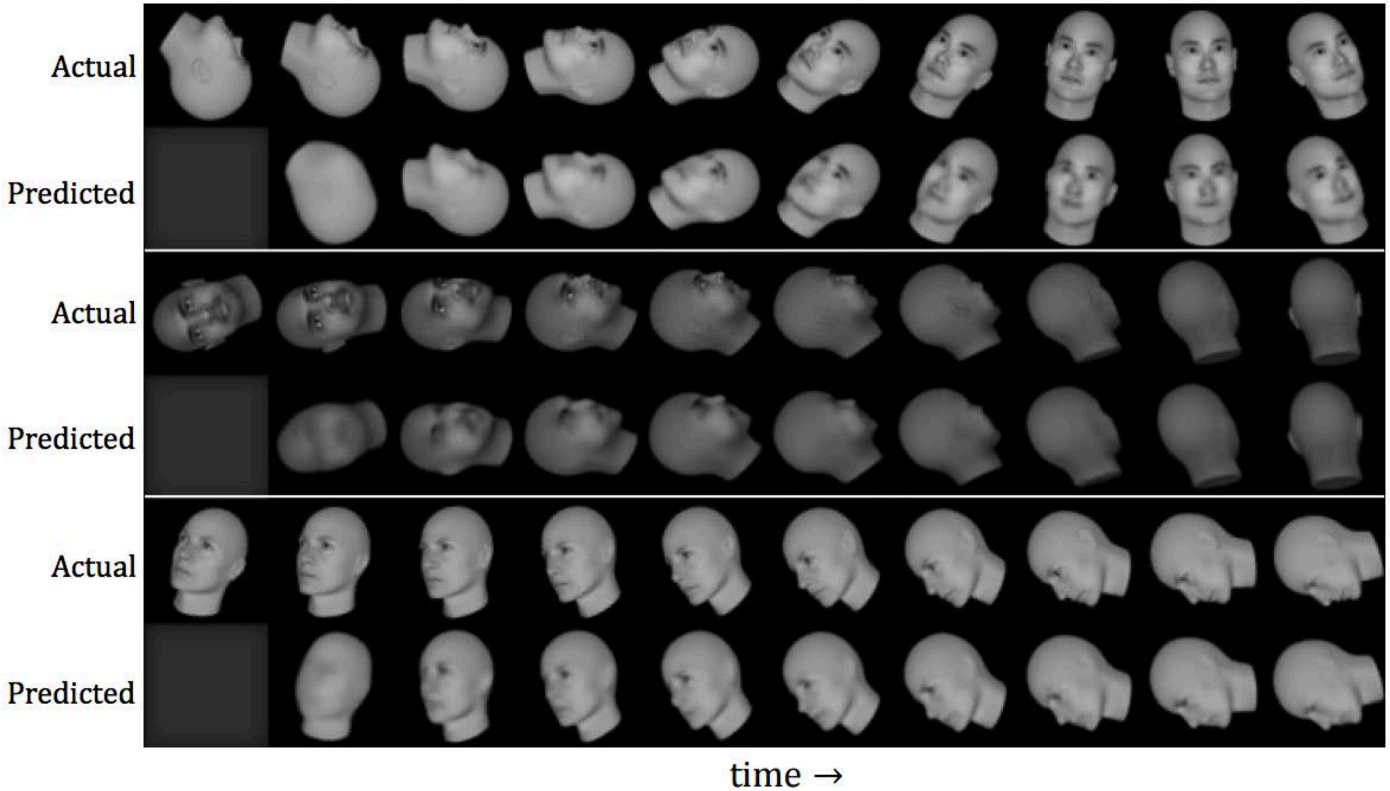
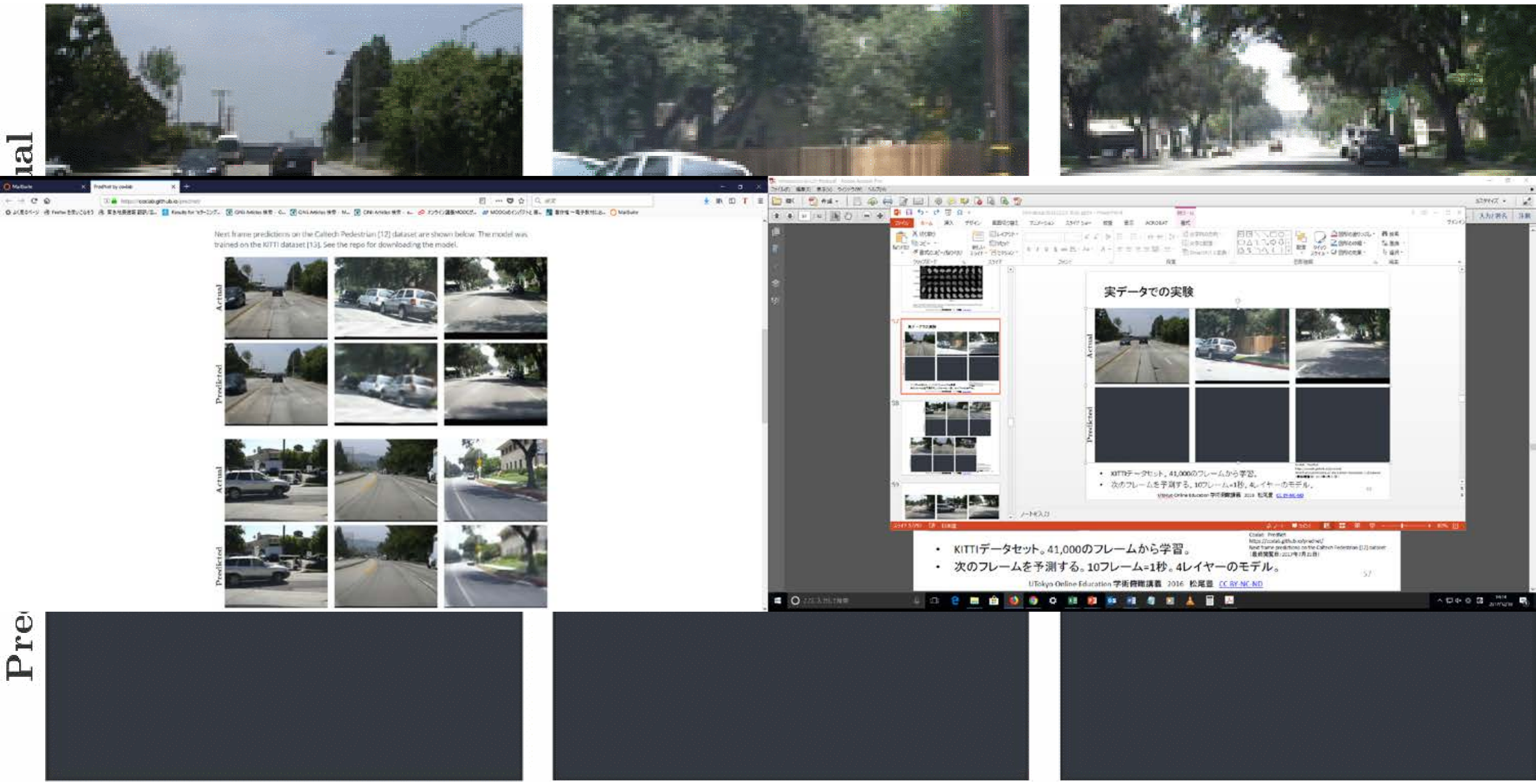


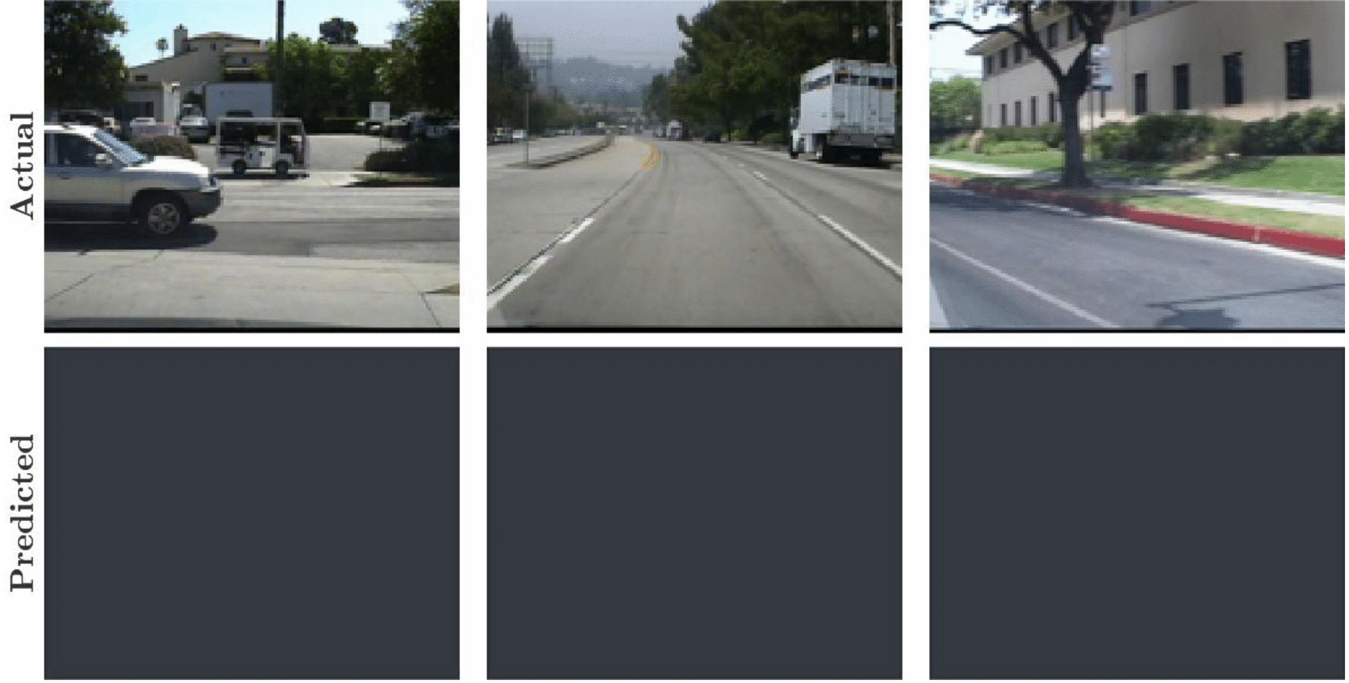
Figure 2: PredNet predictions for sequences of rendered faces rotating with two degrees of freedom. Faces shown were not seen during training.

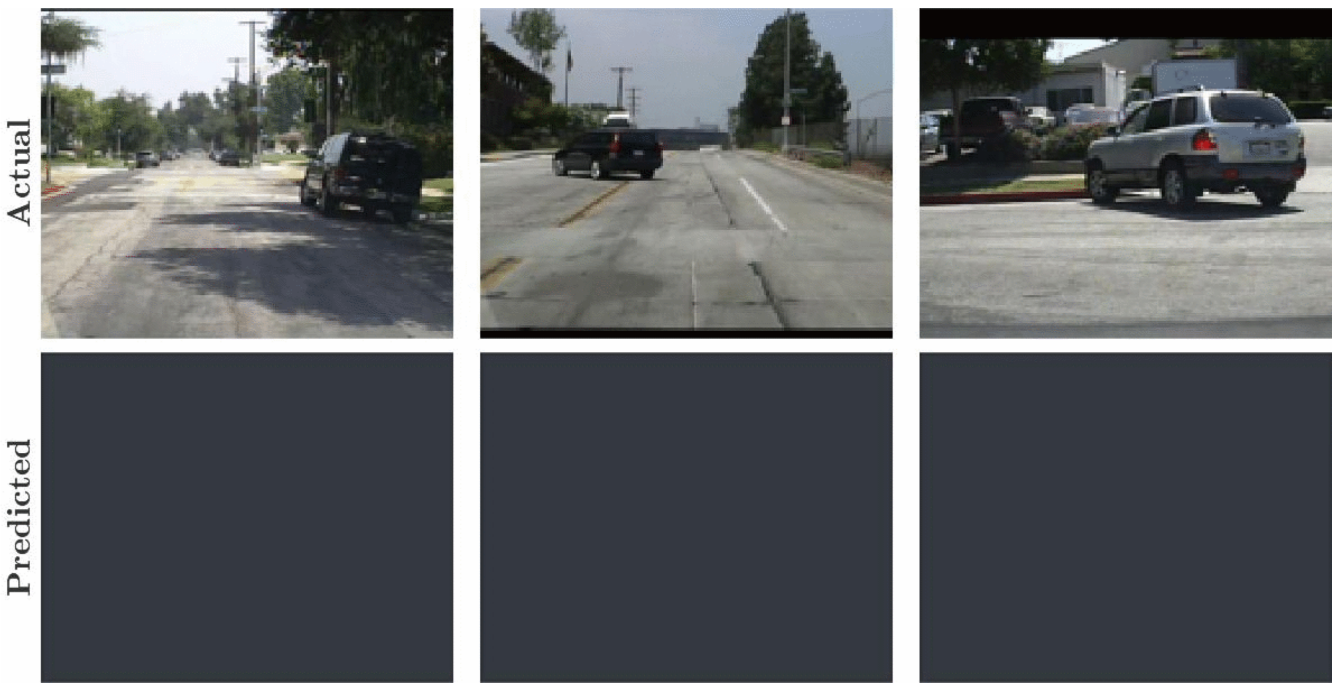
実データでの実験



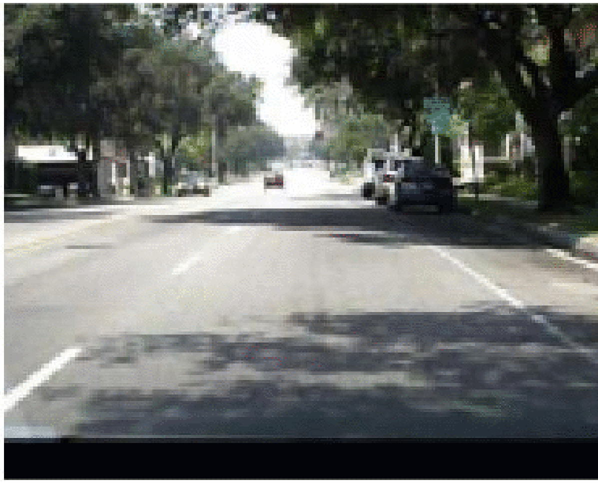
Coxlab　PredNet
https://coxlab.github.io/prednet/
Next frame predictions on the Caltech Pedestrian [12] dataset
（最終閲覧日：2017年7月21日）

- KITTIデータセット。41,000のフレームから学習。
- 次のフレームを予測する。10フレーム=1秒。4レイヤーのモデル。





Coxlab　PredNet
https://coxlab.github.io/prednet/
Next frame predictions on the Caltech Pedestrian [12] dataset
（最終閲覧日：2017年7月21日）



Coxlab　PredNet
https://coxlab.github.io/prednet/
Multi-timestep ahead predictions can be made by recursively feeding predictions back into the model. Below are several examples for a PredNet model fine-tuned for this task.
（最終閲覧日：2017年7月21日）

- より先の未来を予測。5フレーム（0.5秒後）になると、ややぼやけてくる。

Generating Videos with Scene Dynamics (NIPS2016)

- ラベルなしの動画から、動画認識（行動分類）と動画生成（未来予測）の両方に使えるシーンダイナミクスを学習する。
- CNNによる時空間の畳み込み＋GANを使って、背景と前景を切り分ける。
- １秒までの短いビデオをフルフレームで生成できる。

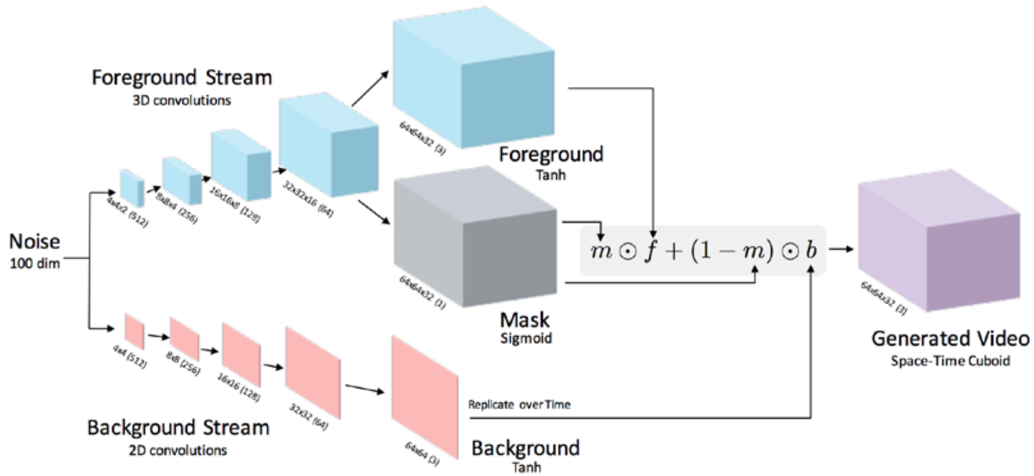


Figure 1: **Video Generator Network:** We illustrate our network architecture for the generator. The input is 100 dimensional (Gaussian noise). There are two in pathway of fractionally-strided spatio-temporal convolutio fractionally-strided spatial convolutions, both of which up-sa to create the generated video using a mask from the motion and the number of channels in parenthesis.

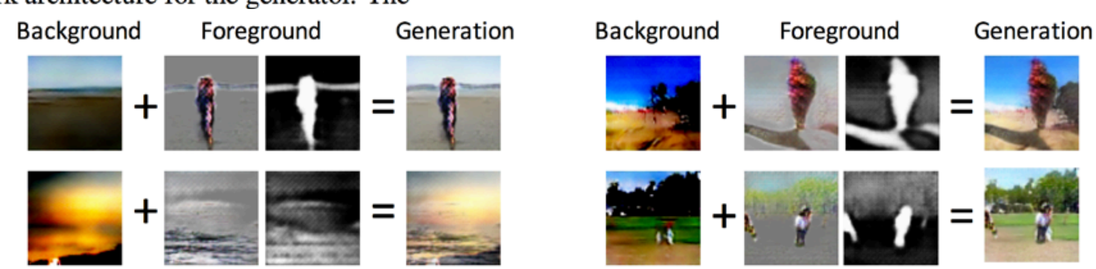
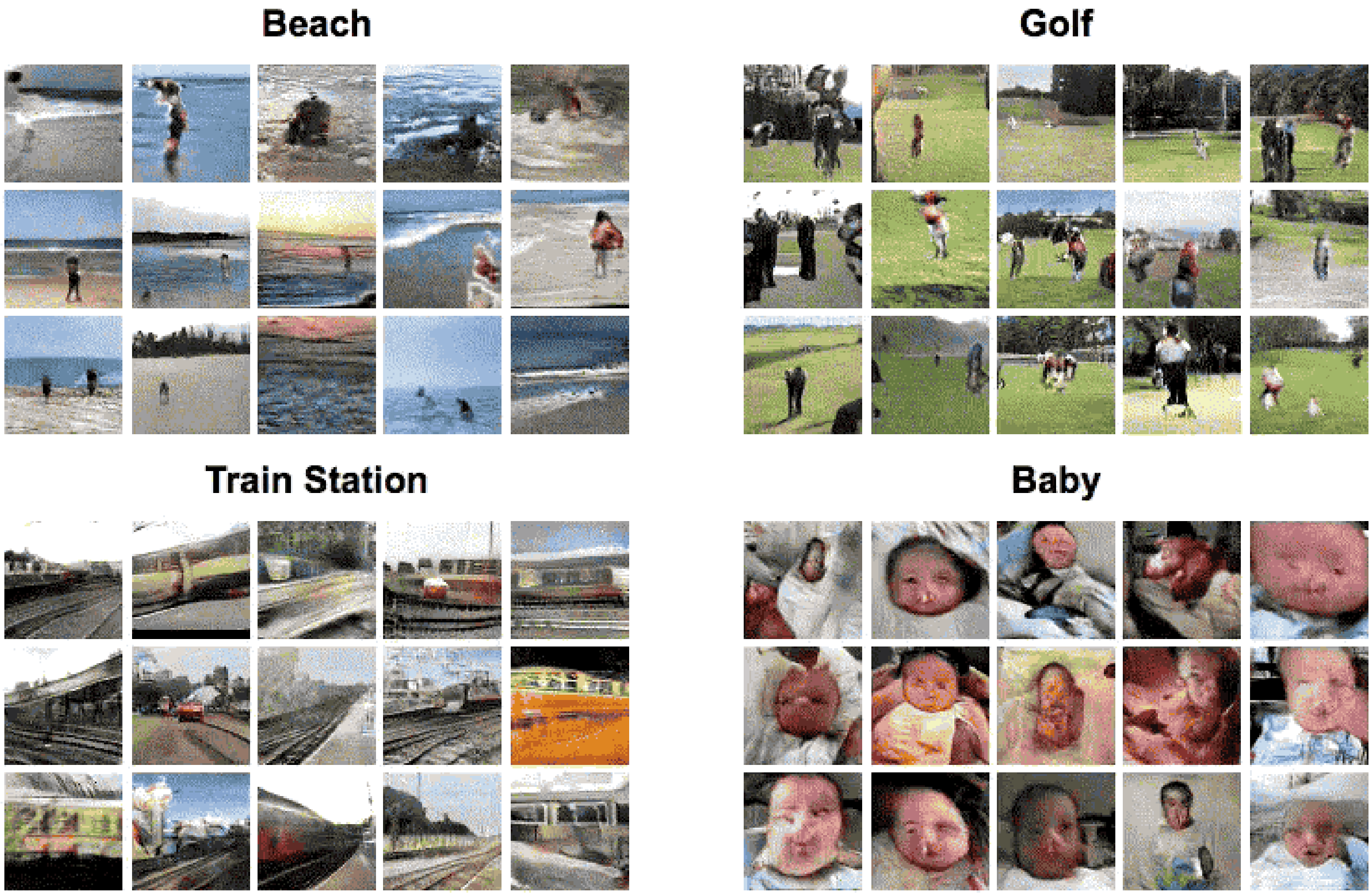


Figure 3: **Streams:** We visualize the background, foreground, and masks for beaches (left) and golf (right). The network generally learns to disentangle the foreground from the background.

Carl Vondrick, Hamed Pirsiavash & Antonio Torralba, "Generating Videos with Scene Dynamics" NIPS 2016
http://web.mit.edu/vondrick/tinyvideo/
p. 3, Figure 1: Video Generator Network
p. 6, Figure 3: Streams:

Carl Vondrick
Generating Videos with Scene Dynamics
http://web.mit.edu/vondrick/tinyvideo/
Video Generations



- 最初のフレームだけ与え、次の1秒を生成することができる。
- 200万の動画（Flickr）から5000時間分のデータで学習。

動作とその帰結の予測

Chelsea Finn, Ian Goodfellow, Sergey Levine, "Unsupervised Learning for Physical Interaction through Video Prediction
" 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.
https://papers.nips.cc/paper/6161-unsupervised-learning-for-physical-interaction-through-video-prediction.pdf
p. 3, Figure 1: Architecture of the CDNA model, one of the three proposed pixel advection models
p. 5, Figure 2: Robot data collection setup (top)

- Unsupervised Learning for Physical Interaction through Video Prediction (2016)
- 5万本のロボットのインタラクション（押す動作）の動画でも学習。
- CNNに、ロボットのアクションを途中で入れて、見える画像を予測する。

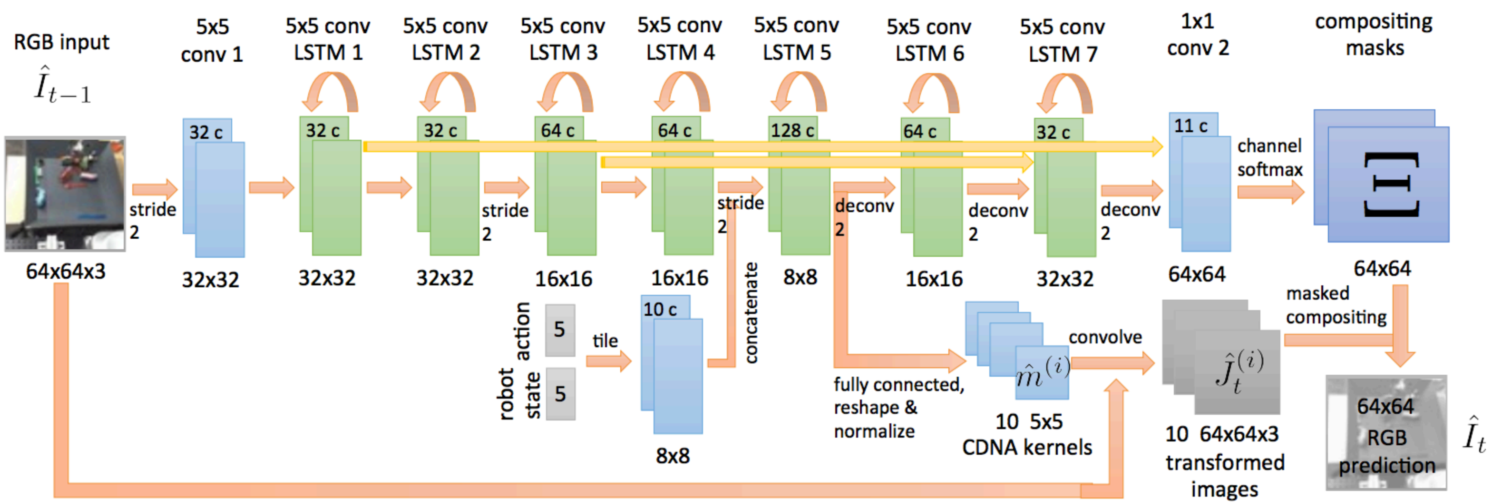
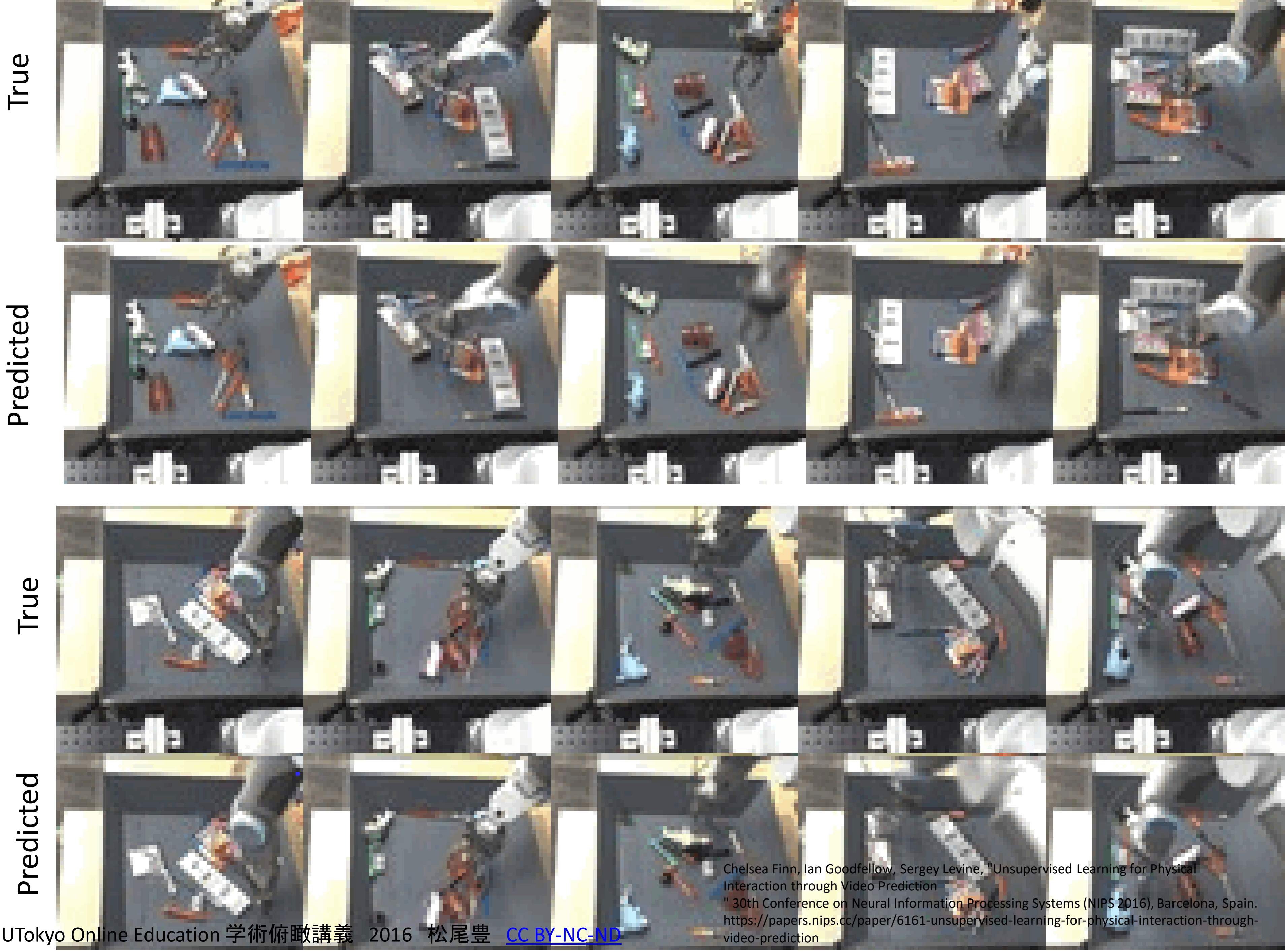
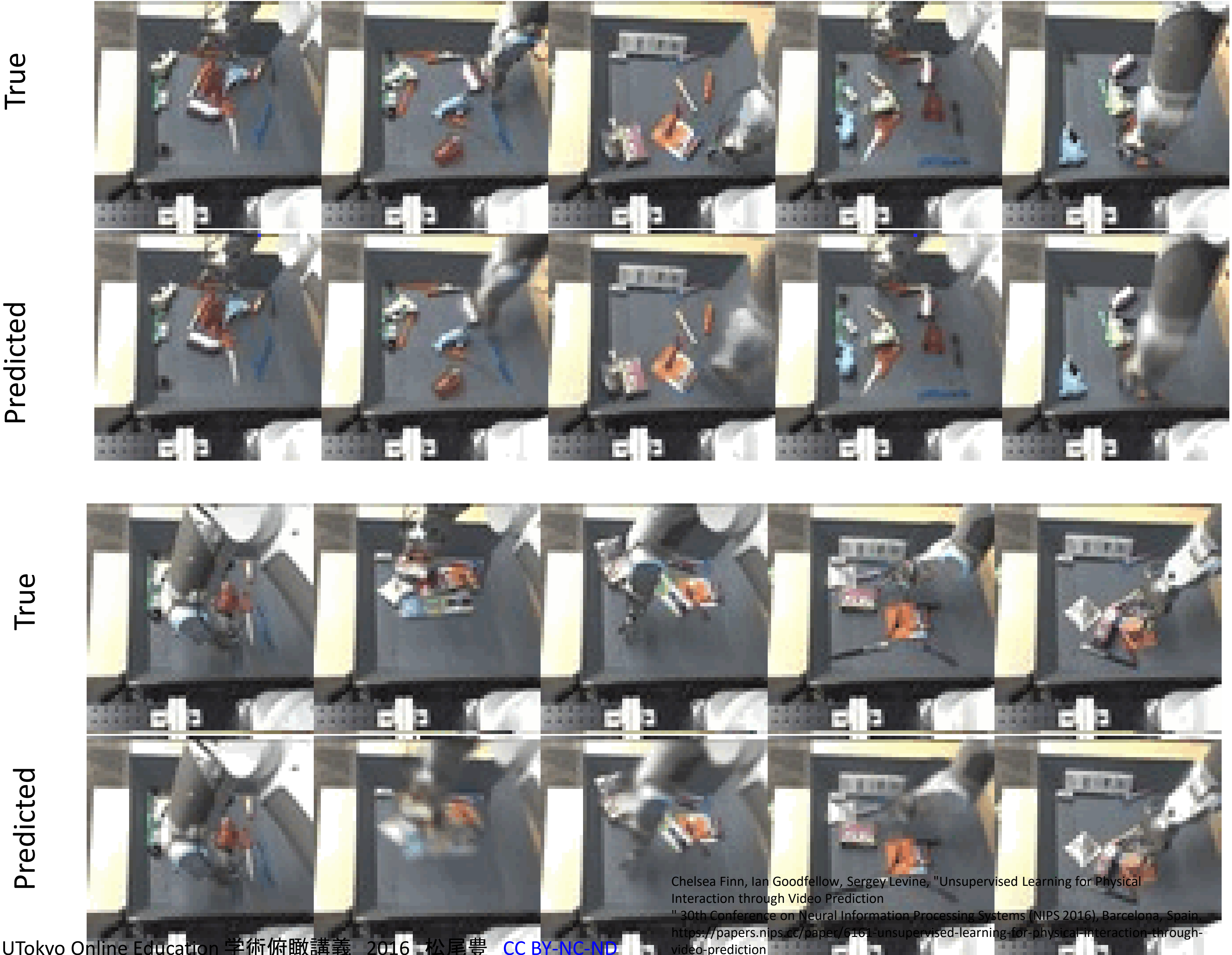


Figure 1: Architecture of the CDNA model, one of the three proposed pixel advection models. We use convolutional LSTMs to process the image, outputting 10 normalized transformation kernels from the smallest middle layer of the network and an 11-channel compositing mask from the last layer (including 1 channel for static background). The kernels are applied to transform the previous image into 10 different transformed images, which are then composited according to the masks. The



Chelsea Finn, Ian Goodfellow, Sergey Levine, "Unsupervised Learning for Physical Interaction through Video Prediction
" 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.
https://papers.nips.cc/paper/6161-unsupervised-learning-for-physical-interaction-through-video-prediction



Chelsea Finn, Ian Goodfellow, Sergey Levine, "Unsupervised Learning for Physical Interaction through Video Prediction
" 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain. https://papers.nips.cc/paper/6161-unsupervised-learning-for-physical-interaction-through-video-prediction

生成モデルから世界シミュレータへ

- ビリヤードのボールの動き
  - Learning Visual Predictive Models of Physics for Playing Billiards (2016)
  - 「視覚的想像（visual imagination）」。ニュートンの方程式を解いているわけでもないのに、ボールがどう転がるか予想できる。ビリヤードで、現在の画像と、かける力から次のフレームを予測する。AlexNetと4フレームのLSTMでボールの位置を予測。
- ゲームのフレーム予測
  - Action-Conditional Video Prediction using Deep Networks in Atari Games (2015)
  - ATARIのゲームでフレームを予測する。アクションを挟んだオートエンコーダ、（あるいはリカレントを含んだオートエンコーダ）で、従来よりも適切にフレームの予測ができる。小さいオブジェクトは苦手。
- LSTMによる映像予測
  - Unsupervised Learning of Video Representations using LSTMs (2015)
  - LSTMを使って映像の表現を学習する。最も良かったのは、オートエンコーダーと未来予測器の複合モデル。時間の範囲の外の動きもちゃんと出すことができた。
- Deep Predictive Coding Network
  - Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning (2016)
  - リカレントネットワークによる生成部分による入力の予測と、実際の入力を比較して、その差分が出力される。で、それがまた予測される。

従来は、世界シミュレータを作らないといけなかったものが、
CNNやRNNの組み合わせで、データから学習して作れるようになってきている。

深層強化学習からプランニング

Deep Q Learning

- 2013年のDeep Mindの論文。その後、Nature等に。
- CNN＋Q学習
- ATARIのゲームから、最近は3Dのゲームに。

著作権等の都合により、
ここに挿入されていた画像を削除しました

Leonardo Araujo dos Santos, Artificial Intelligence
Deep Q Learning
https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/deep_q_learning.html

ドライビング: https://www.youtube.com/watch?v=0xo1Ldx3L5Q&feature=youtu.be
ラビリンス: https://www.youtube.com/watch?v=nMR5mjCFZCw&feature=youtu.be

アルファ碁

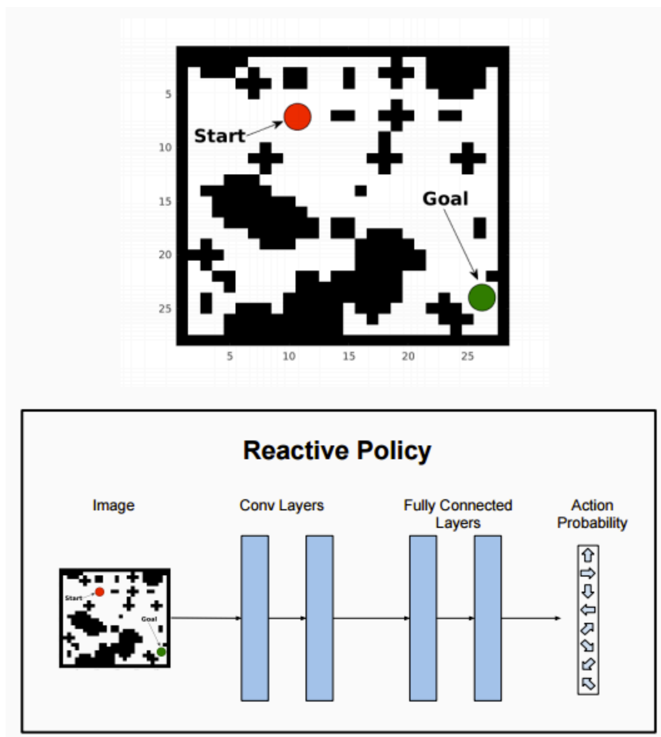
- 19×19×48のFeature mapが生データ。これに13層のCNN
- 教師あり学習を初期値とし、policy network, value network
- モンテカルロ探索と組み合わせる

著作権等の都合により、
ここに挿入されていた画像を削除しました

アルファ碁におけるCNNの構成
フィルターの模式図
大槻将棋のページへようこそ
人工知能エンジニアMeet UpのAlphaGoに関する講演資料はここ(Nature論文解読の一助になれば･･･)。
囲碁AI “AlphaGo” はなぜ強いのか?～ディープラーニング、モンテカルロ木探索、強化学習～
http://home.q00.itscom.net/otsuki/20160415AlphaGopublic.pdf
最終閲覧日：20170425

Value Iteration Network (2016)

- 強化学習にプランニングを加える。
- Reactive Policyだけではうまく汎化できない

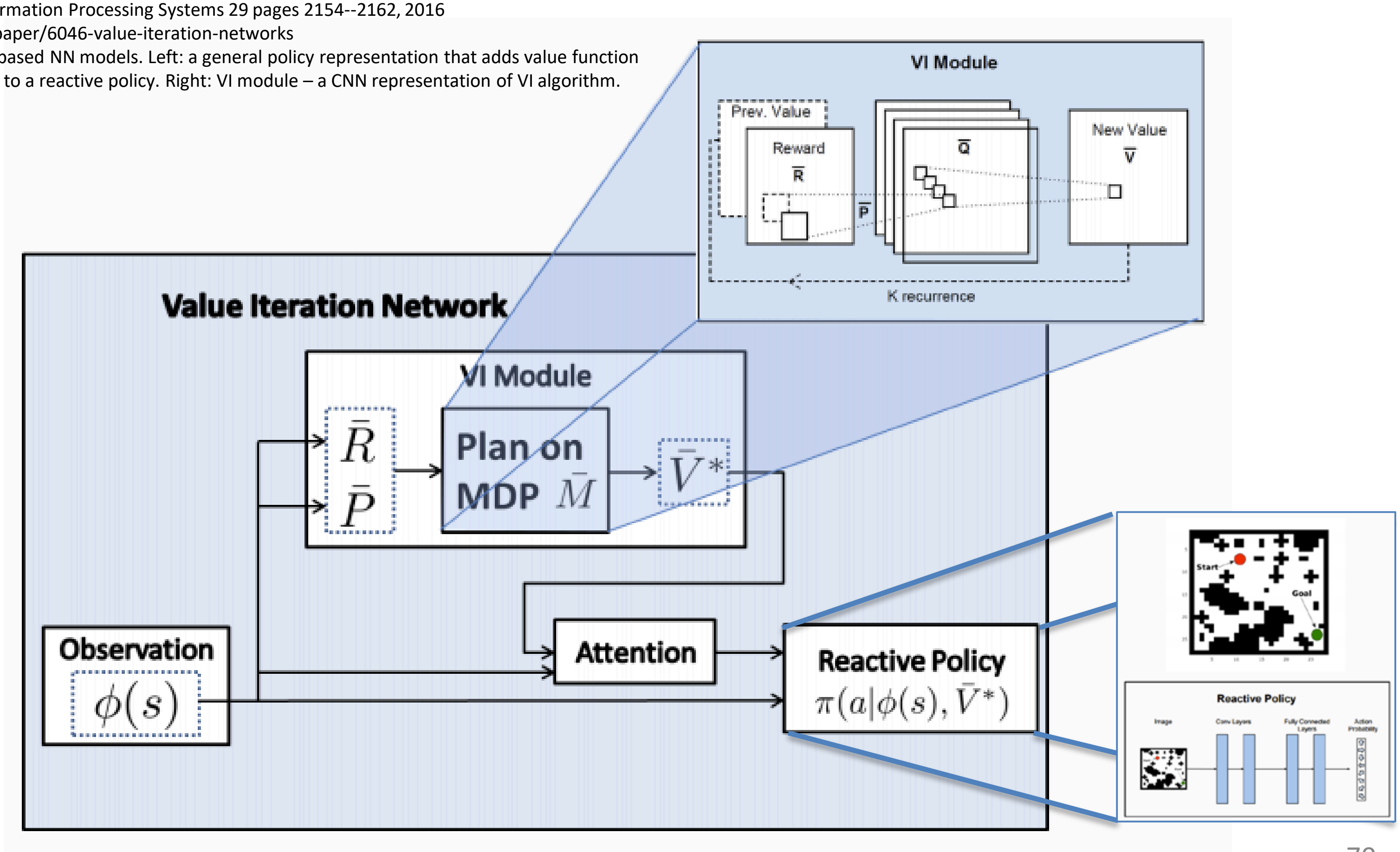


By Aviv Tamar

http://technion.ac.il/~danielm/icml_slides/Talk7.pdf

- プランニングのモジュールを加える
- 良さそうなところに、Attentionを当てることで、Reactive Policy を誘導する。

Aviv Tamar, Yi Wu, Garrett Thomas, Sergey Levine, Pieter Abbeel, "Value Iteration Networks"
Advances in Neural Information Processing Systems 29 pages 2154--2162, 2016
https://papers.nips.cc/paper/6046-value-iteration-networks
p. 4, Figure 2: Planning-based NN models. Left: a general policy representation that adds value function features from a planner to a reactive policy. Right: VI module – a CNN representation of VI algorithm.



- 大きいマップでも成功する（迷路タスク）

Evaluation:

- Action prediction error (on test set)
- Success rate – reach target without hitting obstacles

Results:

Domain	VIN		CNN		FCN	
	Prediction loss	Success rate	Pred. loss	Succ. rate	Pred. loss	Succ. rate
8×8	0.004	**99.6%**	0.02	97.9%	0.01	97.3%
16×16	0.05	**99.3%**	0.10	87.6%	0.07	88.3%
28×28	0.11	**97%**	0.13	74.2%	0.09	76.6%

VINs learn to plan!

By Aviv Tamar

- Strategic Attentive Writer for Learning Macro-Actions (2016, DeepMind)も同様に、マクロのプランニングを行う仕組みを提案している。

深層強化学習：より効率的な学習、より長期の行動へ

- 基本論文：DQN
  - Human-level control through deep reinforcement learning (2015)
  - Nature論文。Deep Q-Learning
- Experience replay
  - Prioritized Experience Replay (2016)
  - 全体のロスを下げるようなリプレイを選ぶように優先づける。（値に比例した方法と、ランクをつけてそれに基づく方法と。）「目隠し崖歩き」の例。
- Double Q learning
  - Deep Reinforcement Learning with Double Q learning (2016)
  - 楽観的な予測にしたがって行動してしまうため、行動の選択時と評価時でのmaxオペレーションを分ける。それのDQN版。パラメータが２セットになる。
- Dueling network
  - Dueling Network Architectures for Deep Reinforcement Learning (2016)
  - 状態価値関数と、状態ごとの行動有利関数（state-dependent action advantage function）。有利関数は、Q(s,a)からV(s)を引いたもので、どのくらい相対的に有利な手かを示している。畳み込み層から、２つのFC層を作り、それらが価値関数と有利関数を表す。

オンラインあるいは現実世界での、深層強化学習が適用されている。
プランニングとも融合しはじめている。

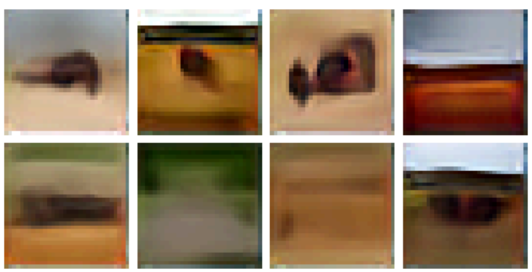
言語からの画像生成、
その先へ

文から画像の生成：Generating Images (2015)

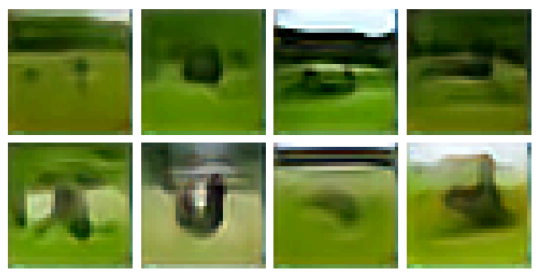
A very large commercial plane flying in blue skies.

A very large commercial plane flying in rainy skies.

Elman Mansimov et. al: “Generating Images from Captions with Attention”, Reasoning, Attention, Memory (RAM) NIPS Workshop 2015
http://www.thespermwhale.com/jaseweston/ram/papers/paper_13.pdf
p. 3, Fig. 3



A herd of elephants walking across a dry grass field.



A herd of elephants walking across a green grass field.

Elman Mansimov et. al: “Generating Images from Captions with Attention”, Reasoning, Attention, Memory (RAM) NIPS Workshop 2015
http://www.thespermwhale.com/jaseweston/ram/papers/paper_13.pdf
p. 3, Fig. 3

Elman Mansimov et. al: “Generating Images from Captions with Attention”, Reasoning, Attention, Memory (RAM) NIPS Workshop 2015
http://www.thespermwhale.com/jaseweston/ram/slides/session1/gen-captions-elman-mansimov.pdf

A stop sign flying in blue skies.

手法

Elman Mansimov et. al: “Generating Images from Captions with Attention”, Reasoning, Attention, Memory (RAM) NIPS Workshop 2015
http://www.thespermwhale.com/jaseweston/ram/papers/paper_13.pdf
p. 2, Figure 1: alignDRAW: Generative model of images conditioned on captions.

- 言語はBi-directional RNN (２層のLSTM）
- 画像はDRAW(2015)：VAE＋RNNによる画像の生成モデル

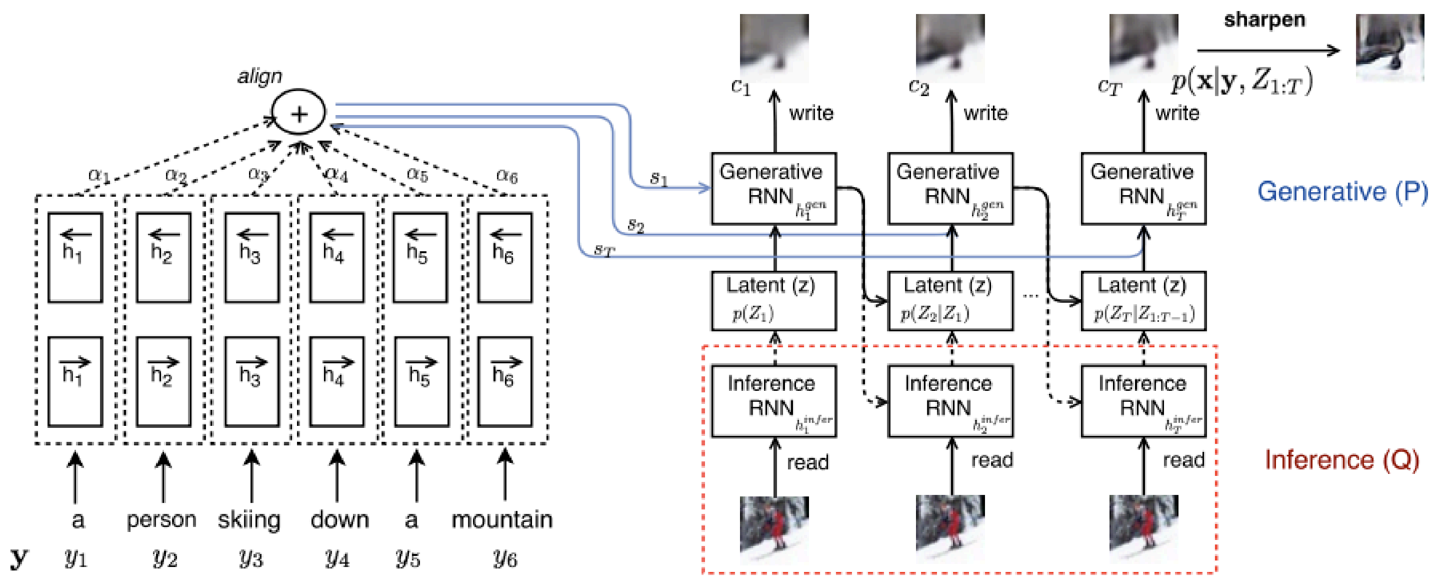


Figure 2: AlignDRAW model for generating images by learning an alignment between the input captions and generating canvas. The caption is encoded using the Bidirectional RNN (left). The generative RNN takes a latent sequence $z_{1:T}$ sampled from the prior along with the dynamic caption representation $s_{1:T}$ to generate the canvas matrix c_T, which is then used to generate the final image $\mathbf{x}$ (right). The inference RNN is used to compute approximate posterior Q over the latent sequence.

言語からの画像生成：今後の展開

- 画像の特徴量と言語との紐づけ
  - まだまだ単純
- シーンの予測や行動結果の予測と、言語を紐づけたい。
  - そうするとかなり幅広い概念を捉えることができるのでは。
- 記号による思考
  - 機械学習的にはどのように解釈されるのか
  - どのように学習能力や探索能力を上げるのか
- 言語の創発に関して
  - Learning to Communicate with Deep Multi-Agent Reinforcement Learning (2016)
  - 複数のエージェントが、環境中で共有する効用を最大化する問題。通信のプロトコルを学ばないといけない。強化エージェント間学習（RIAL）と、微分可能エージェント間学習（DIAL）を提案。言語の創発という意味では面白い論文。

さらに先のフロンティア

- 海馬の働き
  - NTM（ニューラルチューリングマシン）やMemory networkというのが出てきてはいる。（が、ちょっと微妙）

- 意識

- 社会

- 人間とは

こうしたディープラーニングをきっかけとする研究の進展が、
何か少しでも従来からの研究に新しい知見を与えることになれば。

ここまでのまとめ

- 構成論的に知能の仕組みの一部が解き明かせる段階にきている。
  - DLの進展により、新たにできることが格段に増えた。
  - 「無意識下の処理」の一部が徐々に実現されている。
  - こうしたDLのイノベーションを無視することはナンセンス。
- 今まで長い間、研究されてきた人工知能や認知科学の方向性は正しいと思う。
  - きちんと、DLの研究を押さえた上で、今まさに、知能の研究の王道をやるべきタイミングではないか。
  - 身体性、記号と言語、思考、意識や社会などの重要な問題に対して、新しいやり方でアプローチできるのではないか。
  - それこそが、人工知能のコミュニティが本当に研究すべきことでは。

眼の誕生

- カンブリア爆発
  - 5億4200万年前から5億3000万年前の間に突如として今日見られる動物の「門」が出そろった現象
  - 古生物学者アンドリュー・パーカーは、「眼の誕生」がその原因だったの説を提唱
- ディープラーニングにより、見えるようになる
  - さらに、次に何が起こるかを予想して動けるようになる。
- 「眼をもった機械」が誕生する。
  - 機械・ロボットの世界でのカンブリア爆発が起こる。
  - これを日本企業が取れるか？

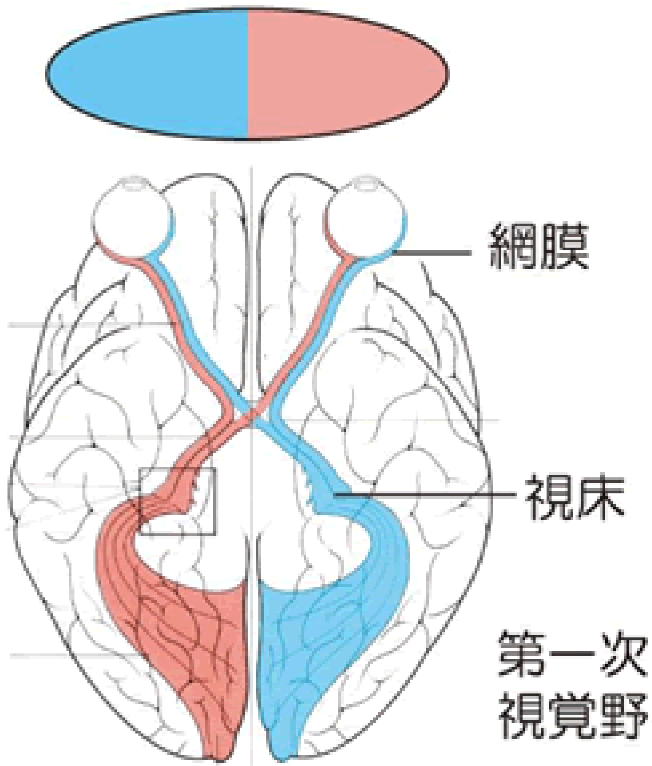
著作権等の都合により、
ここに挿入されていた画像を削除しました

書籍の表紙
アンドリュー・パーカー 著 / 渡辺政隆 訳 / 今西康子 訳
『眼の誕生—カンブリア紀大進化の謎を解く』草思社2006
http://www.soshisha.com/book_read/htm/1478.html

三葉虫：史上初めて眼をもった生物

Photo by Nobu Tamura From Wikimedia Commons
https://commons.wikimedia.org/wiki/File:Opabinia_BW2.jpg
GFDL

眼が見える仕組み



提供：科学技術振興機構(JST)
失われた視覚機能を補う脳の回復メカニズムを解明
http://www.jst.go.jp/pr/announce/20081015/
注2）視覚野

←イメージセンサ

←ディープラーニング
（CNN: 畳み込みニューラルネットワーク）

既存産業の発展

農業	収穫判定	トラクター、コンバインの 適用範囲拡大、効率向上 選別調製等の自動化	自動での収穫 自動での耕うん
建設	測量	掘削、基礎工事、 外装内装作業等の 効率向上	多くの作業の 自動化・効率化
食品 加工	振り分け 確認	カット、皮むき等 の自動化 食洗機に入れる	多くの加工工程の 自動化
組み立て 加工	目視確認の 自動化	動作効率の向上	段取りの自動化 セル生産の自動化

⋮

A: 画像認識　B: 運動の習熟　C: 計画立案を伴う運動

眼をもった機械・ロボットの典型例
単独の製品から入る

- 農業：トマト収穫ロボット
  - トマトは市場規模も大きく、収穫の工数も大きい。
  - 現状の技術で、トマトの認識ができる。上手にもぎ取ることも可能。
  - 先進的な農場から試しに入れる。
- 建設：自動溶接機械
  - 建設の工程（例えば溶接）を自動化する
  - 現状の技術で、接合面の状態等の認識ができる。上手に溶接することも可能。機械を当てれば熟練した人でなくとも熟練の人のような溶接ができる。
  - 一部の建設現場で試しに入れる。
- 食品加工：食洗機にお皿を入れるロボット
  - 食品加工に関わる仕事、まずは食洗機にお皿を入れることを自動化する
  - 現状の技術で、お皿の位置、把持位置の認識ができる。まずは、食器が下げられたところから、食洗機に入れるところを自動化する。（混雑時に重要）
  - ファミリーレストラン等の一部の店舗で試しに入れる。

製品を一刻も早く市場に投入する

眼をもった機械・ロボットの典型例
サービス化へ

- 農業:トマト収穫ロボット(異常発見機能つき)
  - トマトの収穫作業に使いながらデータをためる。(ネットワークに接続)
  - 使われる農場数を増やすとともに、コスト削減。
  - 「病気の判定」ができるようになる。それに対して、新たにチャージすることができる。
- 建設:移動型の自動溶接機械
  - 溶接作業に使いながらデータをためる。(ネットワークに接続)
  - 使われる現場の数を増やすとともに、コスト削減。
  - そのうち、移動しながらの溶接作業や、すでに終わった溶接のチェックができるようになる。それに対して新たにチャージできる。
- 食品加工:お皿管理ロボット
  - お皿を入れながらデータをためる。(ネットワークに接続)
  - 使われる店舗の数を増やすとともに、コスト削減。
  - そのうち、残飯の処理、食器を棚に戻す、食器に盛り付ける等が可能になってくる。それに対して新たにチャージできる。

製品をネットワークにつなぎ、ソフト面の機能向上とともにサービス課金へ

眼をもった機械・ロボットの典型例
プラットフォーム化へ

- 農業：トマト栽培管理プラットフォーム
  - トマトの水やり、施肥、害虫駆除等を自動でできるようになる。
  - トマトの収穫全体を管理するプラットフォームを構築できる。このプラットフォームを利用すれば、品質のいいトマトがたくさん取れるというビジネス。
- 建設：建設現場（躯体工事）プラットフォーム
  - 溶接作業を中心にして、鉄筋を組む作業、コンクリートを入れる作業などを次々に自動化していく。
  - 建設現場の作業全体のプラットフォームを構築できる。プラットフォームを利用すると、自動化された建設現場の施工管理ができるというビジネス
- 食品加工：調理プラットフォーム
  - 食器の管理から食材管理、調理全般を行うまで自動化していく。
  - 調理全体のプラットフォームを構築できる。プラットフォームを利用すると、店舗の応じた調理が自動で提供されるというビジネス

その「場」の機能全体をプラットフォームとして提供

眼をもった機械・ロボットの典型例
海外展開へ

- 農業：グローバルなトマト栽培管理プラットフォーム
  - プラットフォームをそのまま海外に展開し、海外の事業者から課金する。
  - 日本の品質のトマトを、欧米の食卓に届けることができる。
  - 日本はきめ細やかな栽培方法を工夫し続けることで競争力を維持。
- 建設：グローバルな建設現場（躯体工事）プラットフォーム
  - プラットフォームをそのまま海外に展開し、海外の事業者から課金する。
  - 日本のレベルの高い施工技術を、機械化したプラットフォームで提供できる。
  - 日本は建物の耐久性や省エネ、デザイン等、付加価値の高い工夫を続けることで競争力を維持。
- 食品加工：グローバルな調理プラットフォーム
  - プラットフォームをそのまま海外に展開し、海外の事業者から課金する。
  - 日本のレベルの高い食を、世界の飲食店で提供できる。
  - 日本は高い食のレベルを維持し続ける。おいしいものを開発し続ける。

プラットフォームを世界に展開し、高い収益を維持する

モノ売りからの脱却へ

- 「眼のある機械」は、データの継続的収集が不可欠
  - 製品からデータが戻るようにしないと、継続的な品質向上につながらない
  - つまり、製品がネットワークに接続されることがほぼ確定している
- すると、眼のある機械の「稼働」に対して課金できるようになる
  - 「学習ずみモデル」の品質が上がれば、価格を上げることができる
  - 内部コストを下げれば、利益を上げることができる
  - モノ売りからサービス売りへの転換が容易にできる
- さらに、眼のある機械を起点とする「場」全体のプラットフォーム化へ
  - 製品が置かれるオフィス、家、商業施設、工場、農場、建設現場など、製品が取得するデータ・提供するサービスを起点として、その周りのお金・情報の流れに広げ、事業チャンスをとっていくことができる。
- それを世界展開し、日本品質でサービスを提供する
  - 「学習ずみモデル」は日本で作り続け、競争力を維持し続ける

眼のある機械の市場投入→サービス化→周辺を含んだプラットフォーム化→海外へ大きく展開という流れが王道

日本なりのプラットフォーム戦略

- DLの技術はコモディティ化する。
  - 競争力をもつのは、データとハードウェア。
  - 早くDLの技術を取り入れてしまえばよい。
  - DLの技術とハードウェアのすりあわせになった瞬間、日本企業が再度、力を取り戻せる。
- 欧米のスタートアップ（とDL研究者）は、意外なほどハードウェアに対する抵抗感がある
  - そもそも、産業用ロボットの導入台数は日本が（ほぼ）トップ
  - また、ロボットに対する社会的抵抗感もある。
  - 米国は雇用を守らないといけない。日本は人手が足りない。
- ものが関連しないプラットフォームは無理
  - 英語圏でやったほうが絶対に強い。
  - 広告費規模でも10倍、ECの規模でも3倍以上
- 日本には、検索エンジンもECもSNSもあった。
  - GoogleやAmazon, Facebookの位置の企業を出せなくはなかった。
  - しかし、結果はそうなってない。原因は明確で、英語圏でなかったから。

**ものづくりを起点に眼をもった機械を作り、
プラットフォーム化するのは日本ならではの戦略**

どこから始めるか

- 始めること自体はそれほど難しくない
  - 数学の知識：線形代数や最適化
  - プログラムの知識：python、GPU
- ライブラリが揃っている
  - Tensorflow: Google、python
  - Caffe: UCバークレー、C++ベース
  - ほかにも、Keras, chainer, torch7, …
- 教科書
  - 入り口：「人工知能は人間を超えるか」（拙著）
  - 読み物：「人間さまお断り―人工知能時代の富と仕事の手引き」（Jerry Kaplan, 近刊）
  - 教科書：「深層学習」（岡谷貴之）
  - 教科書："Deep Learning"（Y. Bengioら、MIT pressから来年出版。翻訳を出版予定）
    - www.deeplearningbook.org にPDFがあり、読めます。
  - あとは論文を読んでください。
- ３つの主要な国際会議
  - ICML, NIPS, ICLR（それぞれ年1回）
  - ウェブで全ての情報が見れます。

理系の人が半年もやればそこそこできるようになる

著作権等の都合により、
ここに挿入されていた画像を削除しました

書籍の表紙
Ian Goodfellow, Yoshua Bengio, and Aaron Corville, Deep Learning, MIT press 2016.
https://mitpress.mit.edu/books/deep-learning

- 人工ニューラルネットワークの最初の実験が1950年代に行われたのに、なぜ最近になってようやく、深層学習が極めて重要な技術と認識されるようになったかは、不思議に思うかもしれない。深層学習は、1990年代から商業的な応用ではうまく使われていたが、最近まで、しばしば技術というよりはアートであり、専門家だけが使えるものと見なされた。深層学習のアルゴリズムでよい性能を得るには、必要なスキルがあることは真実である。幸いにも、必要なスキルの量は、訓練データの量が増えるにつれて減っている。今日、複雑なタスクで人間の性能に到達する学習アルゴリズムは、1980年代におもちゃの問題を解くのに苦労した学習アルゴリズムとほとんど同一である。これらのアルゴリズムで訓練するモデルは、とても深いアーキテクチャでの訓練を簡略化する変化をしてはいるが。最も重要な新しい進歩は、今日ではアルゴリズムが成功するのに必要とするだけのリソースを、アルゴリズムに提供することができることである。

- 2016年の時点で、大体の目安として、教師ありの深層学習のアルゴリズムは、一般的に、カテゴリごとに約5000のラベル付き事例で、許容できる性能を達成し、少なくとも1000万のラベル付きの事例を含むデータセットで訓練すれば、人間の性能と匹敵する、あるいは超える。これよりも小さいデータセットでうまくいくことは、重要な研究分野であり、教師なしあるいは半教師あり学習で、大量の教師なし事例をうまく活かす方法に特に焦点が当てられる。

変わりゆく社会

- 倫理や社会制度の議論がもう一度必要になる
  - 自動運転で危険回避のときは？人の命の重さは？
- 人工知能システムが社会に広がったときの不具合の問題
  - 製造者責任？保険や社会保障のほうが適切では
- 心をもつように見える人工知能を作ってよいか
  - プログラムの停止させると悲しむ？
  - 恋愛させるビジネスなど（映画「Her」の世界）
- 人工知能を使った軍事
  - ロボット兵士やドローン
  - 権力者を倒す、心を操る？
- 人工知能が知財を生み出す場合の権利
  - 著作権や特許は認めるべきか
- 実は人間が本来的にもっている権利がもっとあるのではないか
  - 忘れられる権利、見られない権利、大目に見られる（警告を受ける）権利、好きになる権利、…

人工知能学会 倫理委員会（松尾が委員長）でも議論。
社会全体で議論していく必要。

内閣府でも、「人工知能と人間社会に関する懇談会」にて議論。

我々はどういう社会を作りたいのか

- 人工知能技術が進めば進むほど、「与えられた目的」に対して、それを実現する手段は賢くできるようになる。
  - 人間＝知能＋生命
  - 知能は、目的を与えられたときの問題解決の力。
  - 生命は、目的を持つ。
    - 自己保存、自己複製、仲間を守るなど。そうしないものは、進化の過程で滅んできたため。
  - 人工知能の技術は知能、すなわち問題解決の技術。
- そうすると、与える目的自体の是非の議論のほうがより重要になる。
  - なにが社会で大事なのか？
  - 個人の幸せや社会全体の幸せはどのように考えればいいのか？
  - 異なる価値観のものをどのようにバランスさせればいいのか？
- これまで人文社会学系でされてきたような議論が今後、改めて重要になる。
  - 特に、哲学、政治学、社会学、法学、心理学、経済学など。

我々は、どういう社会を作りたいのか？
幸せで持続可能な社会とは？

ディープラーニング×ものづくり:
「眼をもった機械」による日本の新たな産業競争力の実現へ

- 少子高齢化しており、労働力が不足している。
  - 頭脳労働は不足していない。「運動を伴う労働」のニーズが高い。
  - 農業従事者、建設・物流、介護、廃炉、熟練工の後継者、etc
- ディープラーニングが解決策になり得る。
  - 眼をもった機械:認識や運動の上達ができる機械・ロボット
  - ものづくりと相性がよく、日本の強みを活かせる。素材や駆動系も強い
- 新たな投資概念の必要性
  - 人への投資をいかに既存の枠組み・文化を踏まえてやるか
  - 企業からの投資を引き出し、そこに人材が流入する生態系を作るか
  - 例えば「学習工場」
- チャンスを捉えるには、正しく早く動いていくことが重要
  - ディープラーニング人材の育成
  - 事業・産業がどう変わるかを早期に検討
  - 社会全体で新しい未来像を描いていくこと