ディープラーニングの実世界応用と今後の可能性

尾形哲也

早稲田大学基幹理工学部表現工学科教授
産業技術総合研究所人工知能研究センター特任フェロー

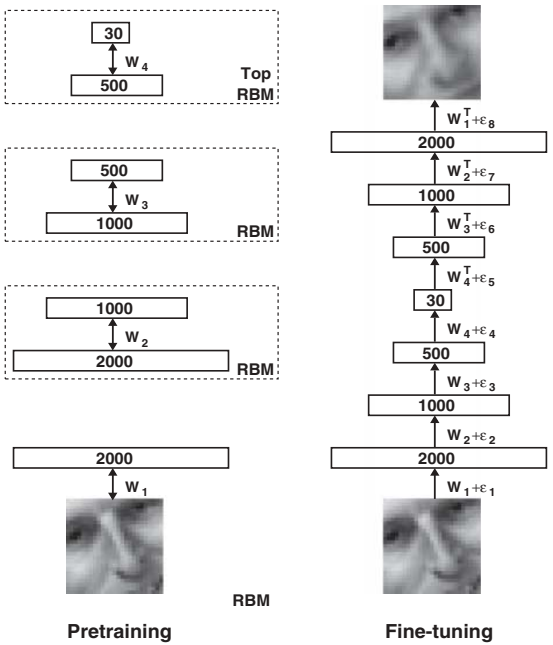
人工知能研究センター

尾形哲也（おがたてつや）

1993	早稲田大学理工学部機械工学科卒業
1997～1999	日本学術振興会 特別研究員（DC2）
1999～2001	早稲田大学理工学部 助手
2001～2009	早稲田大学ヒューマノイド研究所 客員講師&客員准教授
2001～2003	理化学研究所脳科学総合研究センター 研究員
2003～2012	京都大学大学院情報学研究科知能情報学専攻 講師&准教授
2009～2015	科学技術振興機構 さきがけ領域研究員（5年）
2012～現在	早稲田大学理工学術院基幹理工学部表現工学科 教授
2015～現在	産業技術総合研究所人工知能研究センター 招聘研究員&特任フェロー

2013~2014 日本ロボット学会理事，2016~(2018) 人工知能学会理事
2016～現在 科学技術振興機構ACT-I「情報と未来」領域アドバイザー
2017～現在 科学技術振興機構さきがけ研究「社会デザイン」領域アドバイザー
2016～現在 株式会社エクサウィザーズ（元エクサインテリジェンス）技術顧問
2017～現在 日本ディープラーニング協会理事

Deep learning



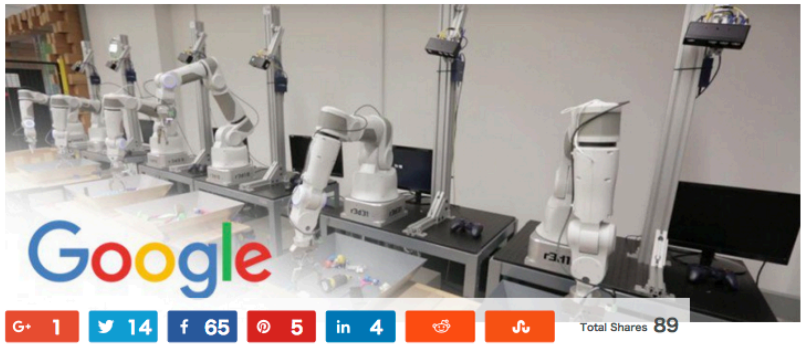
(Hinton, 2006)

AlphaGo
(D. Silver, D. Hassabis et al. 2016)

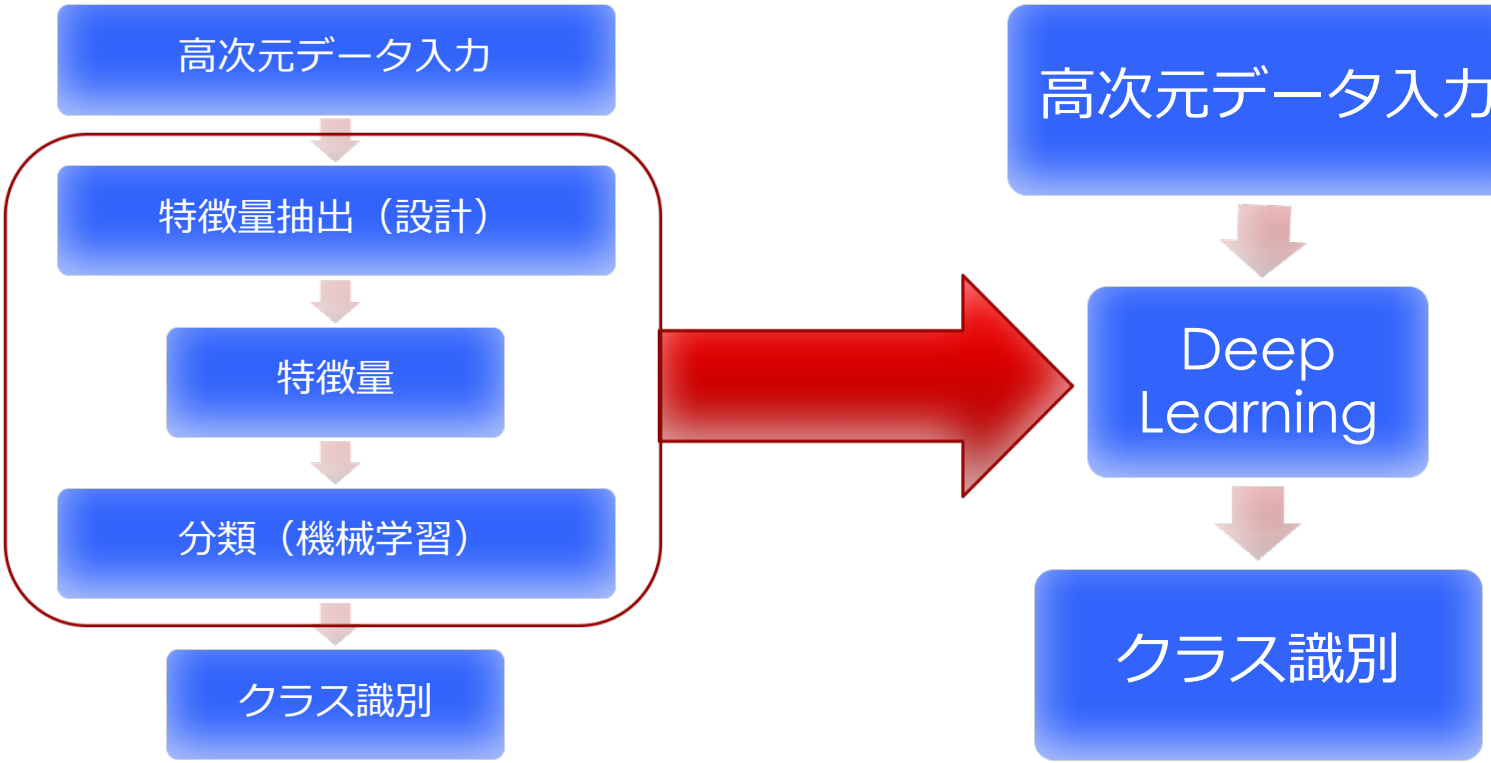
(Google official blog, 2012)

Google's hive-mind robot arms learn to negotiate a cluttered world

Martin Anderson Wed 9 Mar 2016 12.08pm



ディープラーニングによる学習



画像認識・音声認識などのデータで有効性が報告

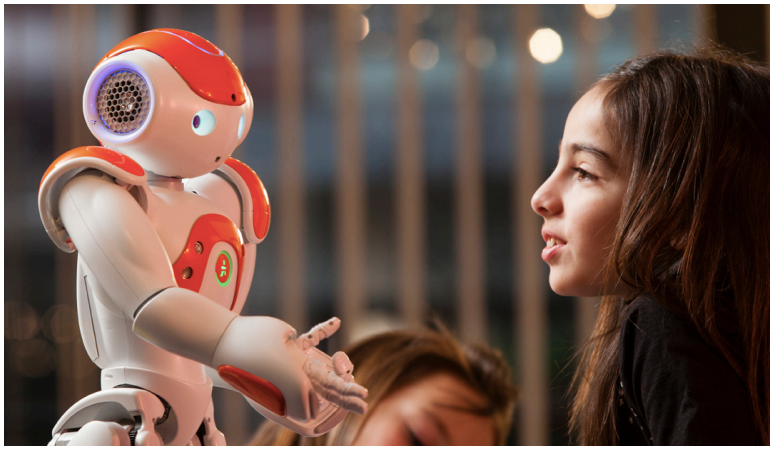
[A. Krizhevsky et al., 2012], [G. E. Dahlet al., 2012]

マルチモーダル音声認識

(Honda Research Instituteとの共同研究)

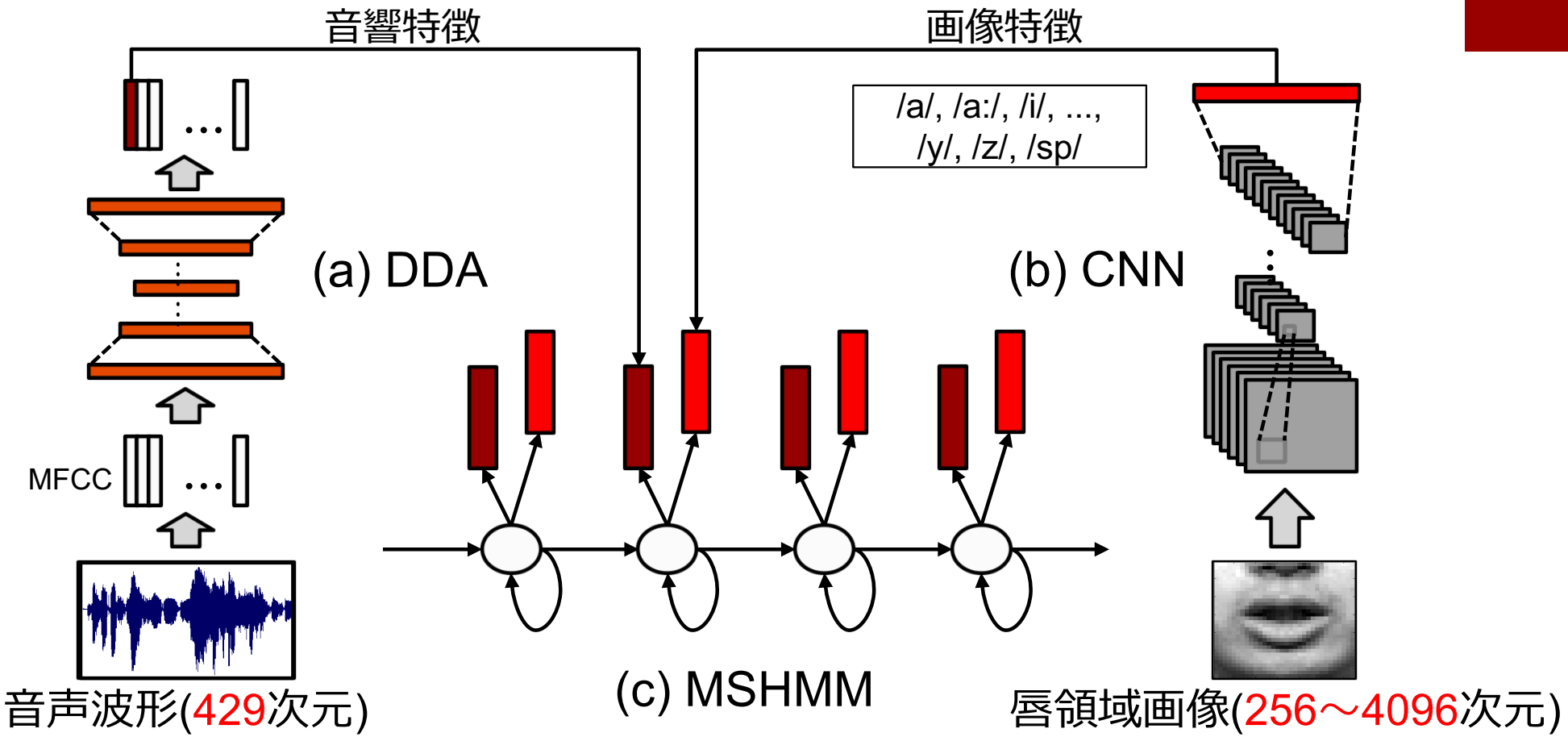
家庭用ロボットや自動運転車など，実環境下での人間・機械インタラクション実現には，雑音に頑健な音声入力インターフェースの実現が不可欠

音響・映像信号から抽出した特徴量を組み合わせて音声認識を行う視聴覚音声認識（AVSR: Audio-Visual Speech Recognition）によって実現



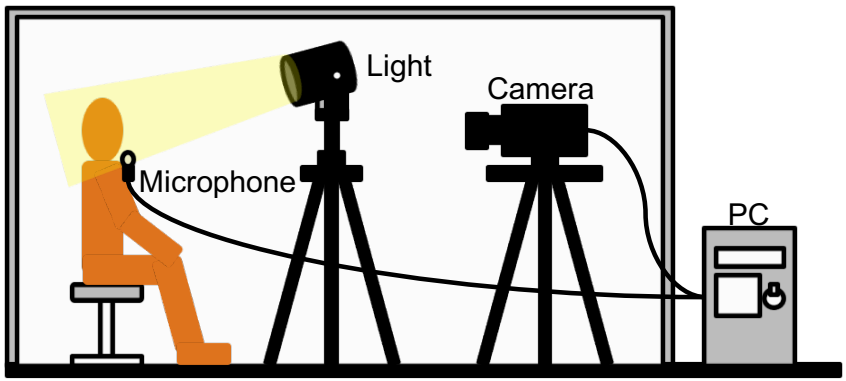
マルチモーダル音声認識システム

K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata,
Applied Intelligence, Vol. 42, Issue. 4, 2015.

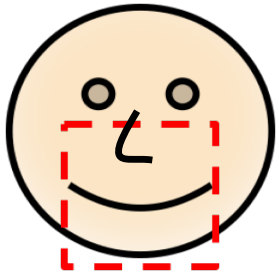


実験データセット

データ収録環境（出典 [Koiwa2008]）



話者	男性6人
単語セット	ATR音素バランス単語216＋ ATR重要単語84
音声データ	16bit, 1ch 16kHz, 1800 files
画像データ	640x480 pixel, 8bit モノクロ, 99.9fps, 約24万枚



128x128
唇領域を切り出し

64x64

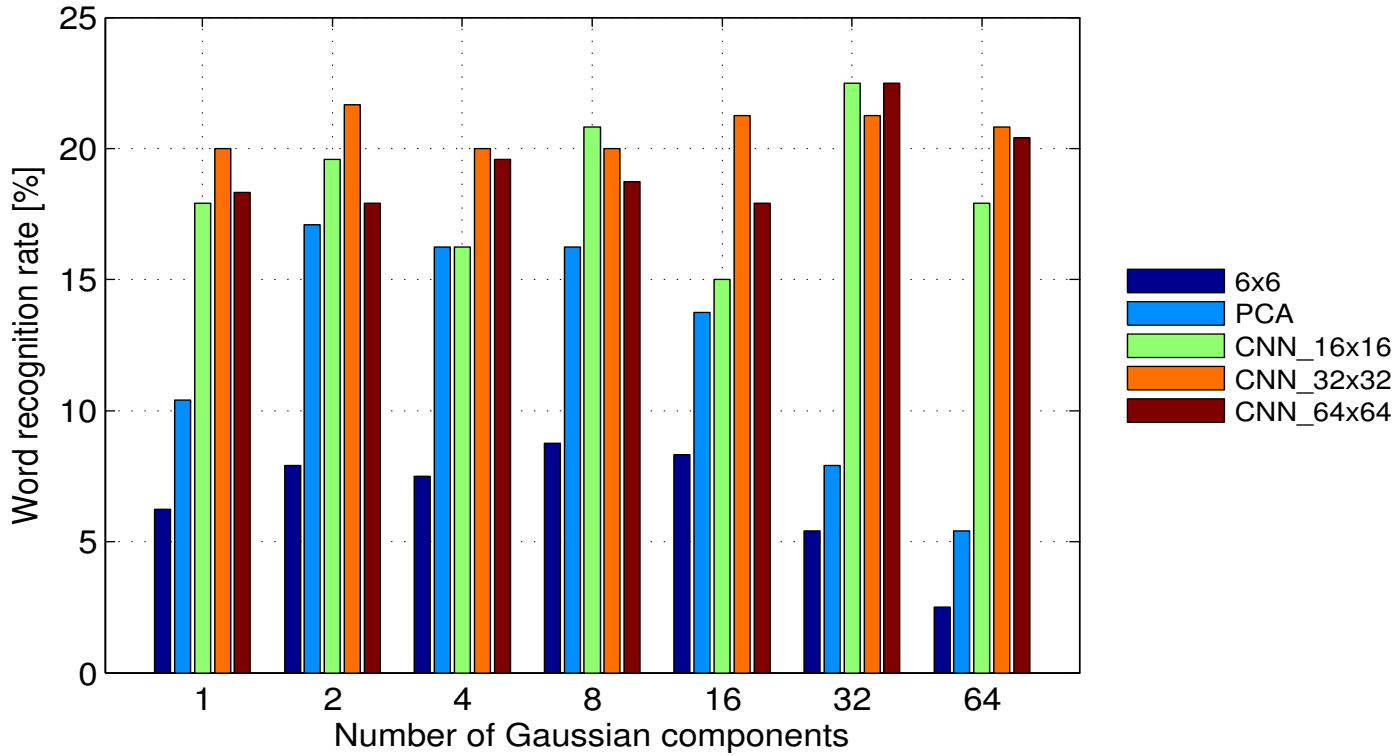
32x32

16x16

3種類の解像度にリサイズ

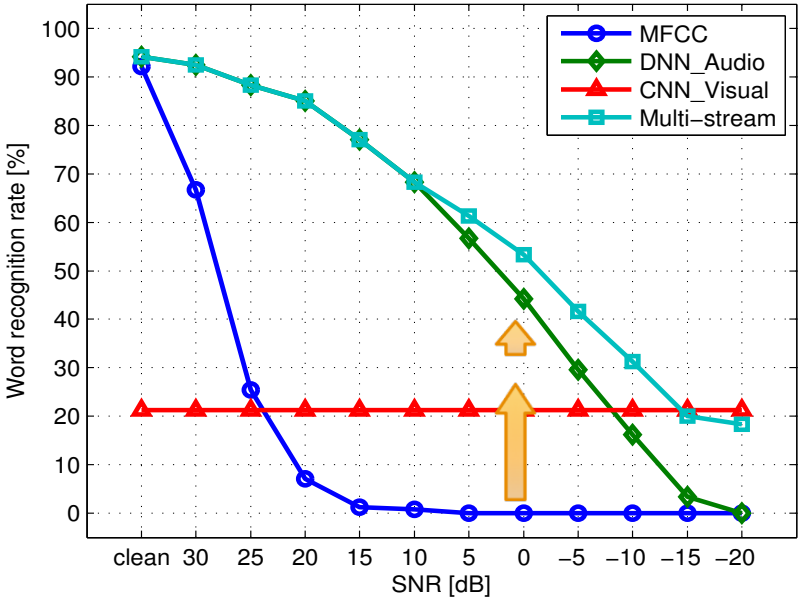
画像特徴量ごとの単語認識率

- CNNを用いた画像特徴により，画像特徴のみでも単語認識率約23%を実現可能

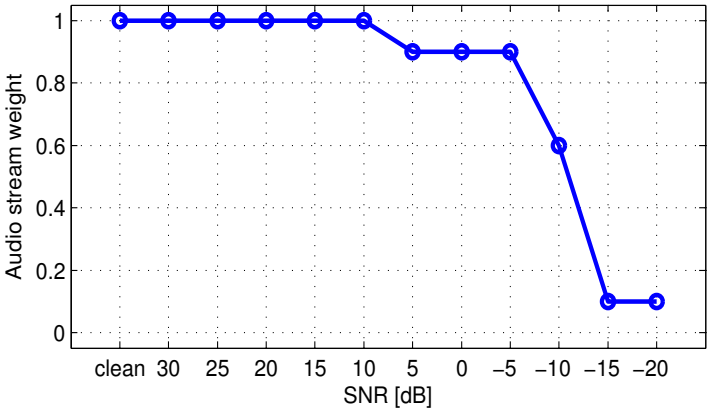


ノイズレベルごとの単語認識率

- ノイズに対する頑健性の向上
  - DDAによるデノイジングにより認識率が向上
  - 画像特徴量を相補的に用いることで低SNR領域で認識率が向上



音響特徴のストリーム重み



LipNet

Y. Assael et al., 2016 (DeepMind)

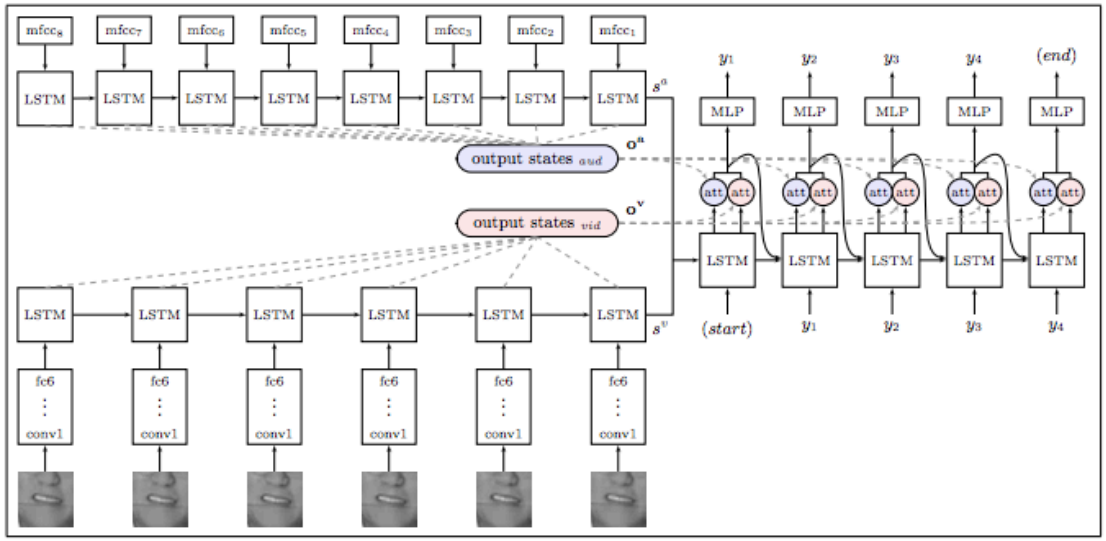


Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character y_i, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

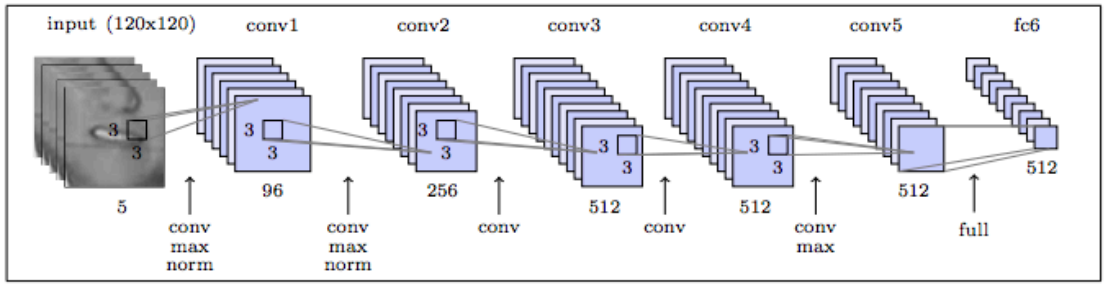


Figure 2. The ConvNet architecture. The input is five gray level frames centered on the mouth region. The 512-dimensional fc6 vector forms the input to the LSTM.

[28] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata. Lipreading using convolutional neural network. In *INTERSPEECH*, pages 1149–1153, 2014.

[29] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata. Audio-visual speech recognition using deep learning. *Applied Intelligence*, 42(4):722–737, 2015.

ロボットと人工知能

- （伝統的）ロボット
  - 機械・電気
    - 物理モデル（微分方程式，線形近似）
- （伝統的）人工知能
  - 情報・通信
    - グラフィカルモデル（確率方程式，平均と分散）

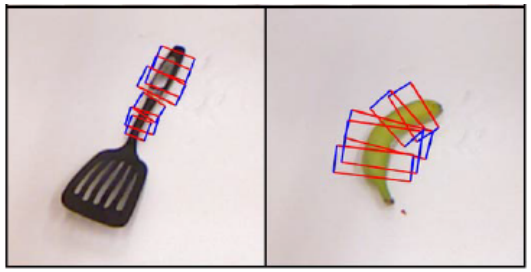
「ロボットって人工知能と関係あるんですか？」

- Deep Learning (neural net)
  - モデルを持たない！

ロボットビジョン（のみ）への応用

① CNNを用いた把持位置を予測

- CNNの出力を把持位置ベクトルとして学習
- 課題
  - 実ロボットでは見評価
  - RGB-D画像が使用



[Joseph et al., 2015]

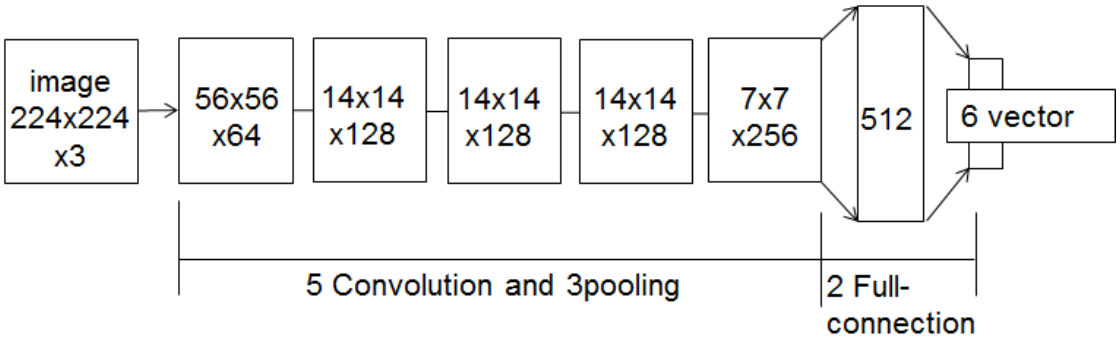
---

② 実ロボットによる教師データ作成と把持

- 教師なしで把持位置を予測・把持
- 5万回把持を700時間
- 課題
  - 膨大な学習時間
  - RGB-D画像が必要

[Lerrel, 2015]

教示データとCNNの構造



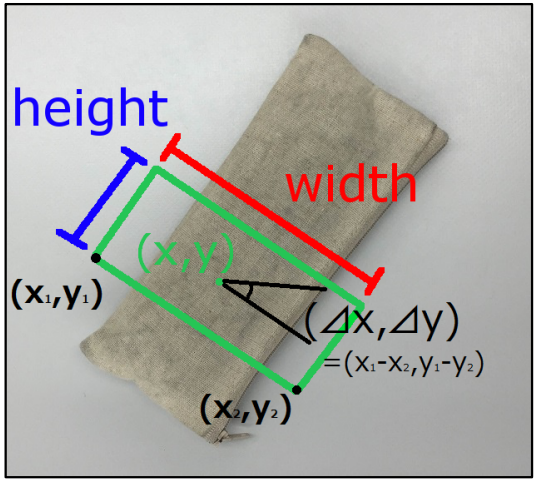
- 入力：画像
- 出力：把持位置ベクトル

- 教師データ作成
- 把持位置ベクトルを与える

(x, y, ⊿x, ⊿y, width, height)

arctan(⊿y / ⊿x)

- 1000枚の画像を平行移動，回転させ30000枚に増幅

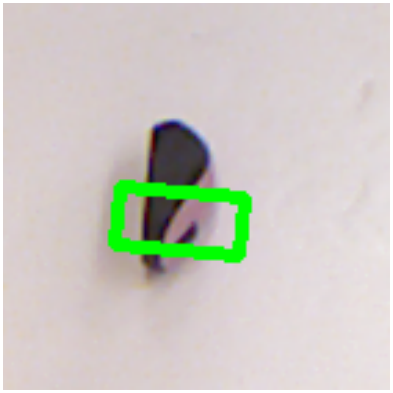


評価法

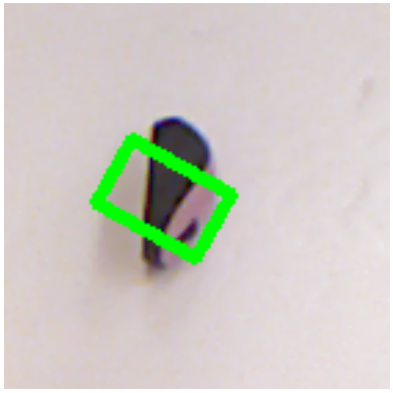
- 関連研究①の評価基準Accuracyを
  - 回転角度誤差が30°以内
  - 教師位置Aと予想位置Bの重複部が合計面積の25%以上

$$Accuracy = \frac{|A \cap B|}{|A \cup B|}$$

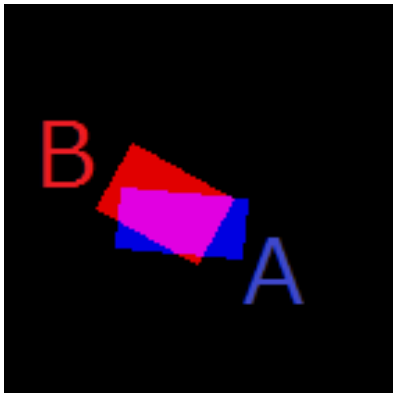
⇒既知の教師データに対する予測把持位置を比較



教師位置A

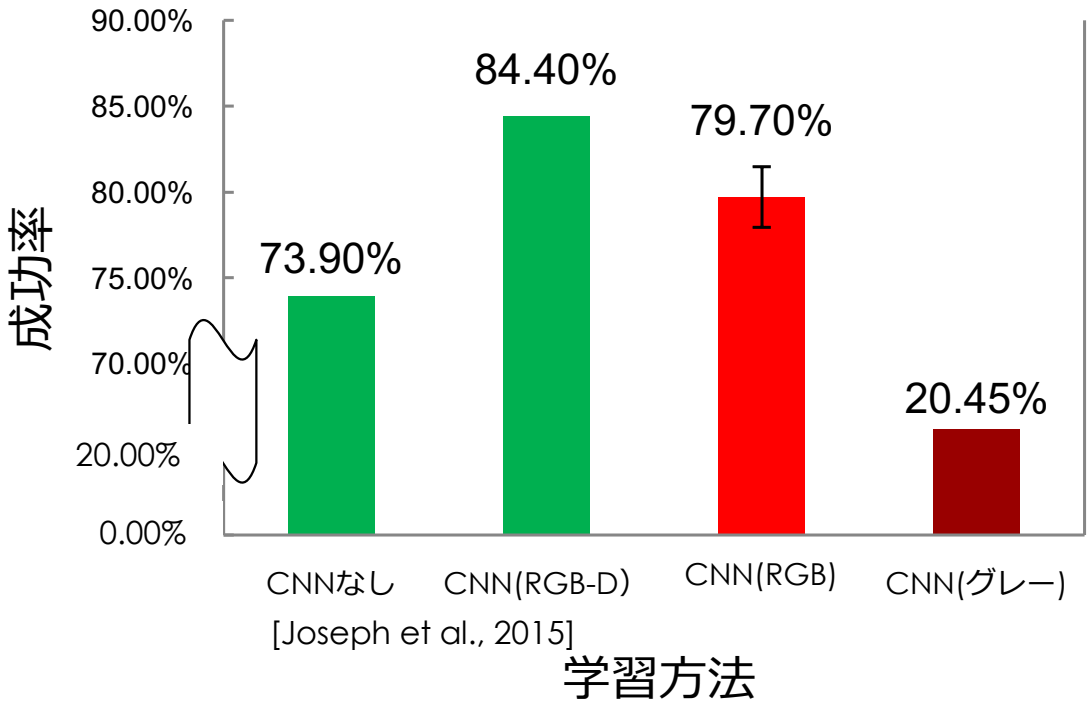


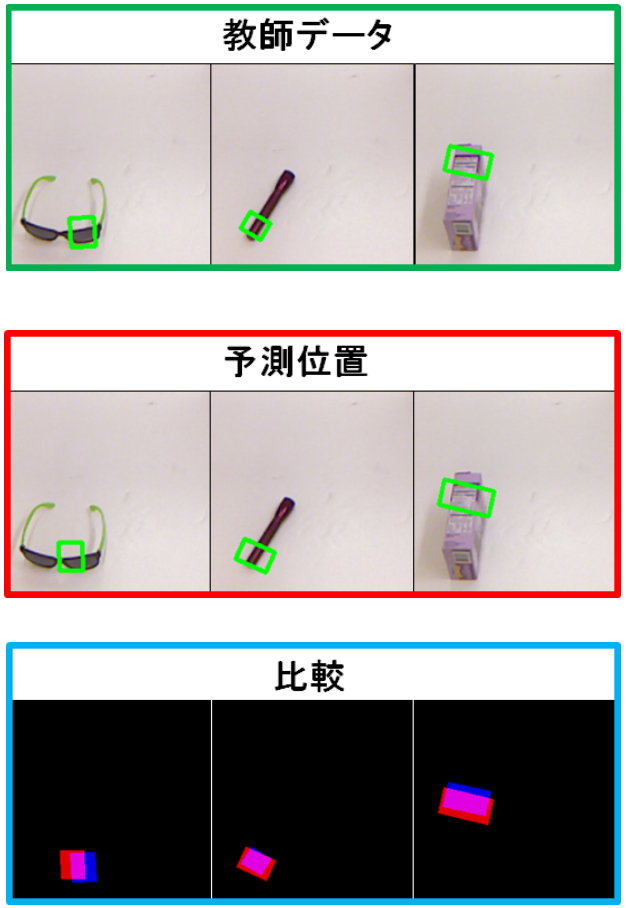
予測位置B



把持位置合成

実験結果



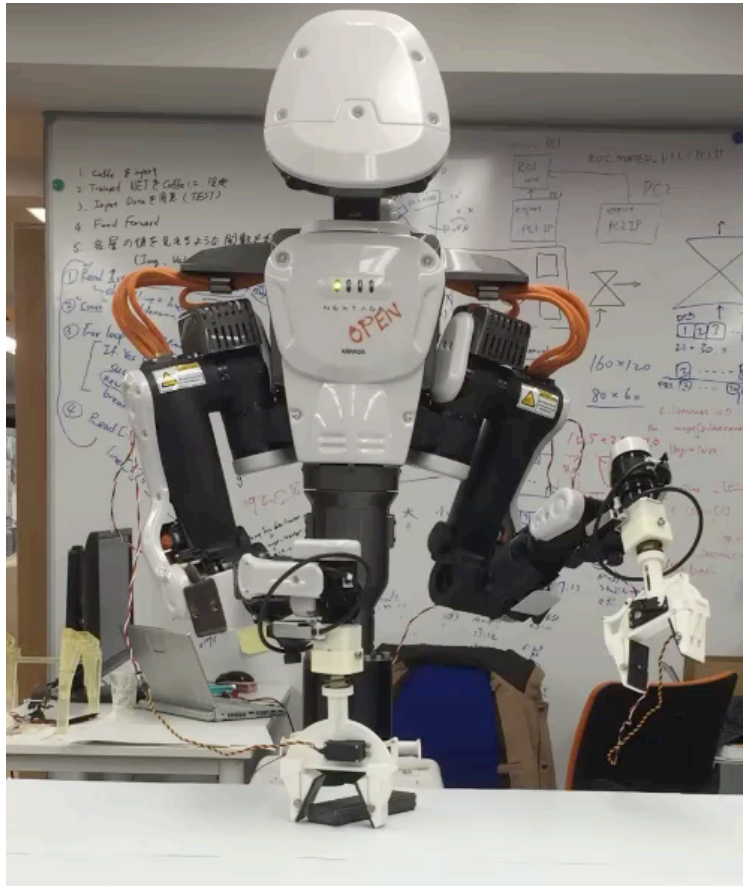


- 本実験の成功率・・・79.7%
- 既知の教師データに対しての予測は高い精度で行うことが可能
- グレースケールの精度は著しく低下（色情報の重要性）

問題点

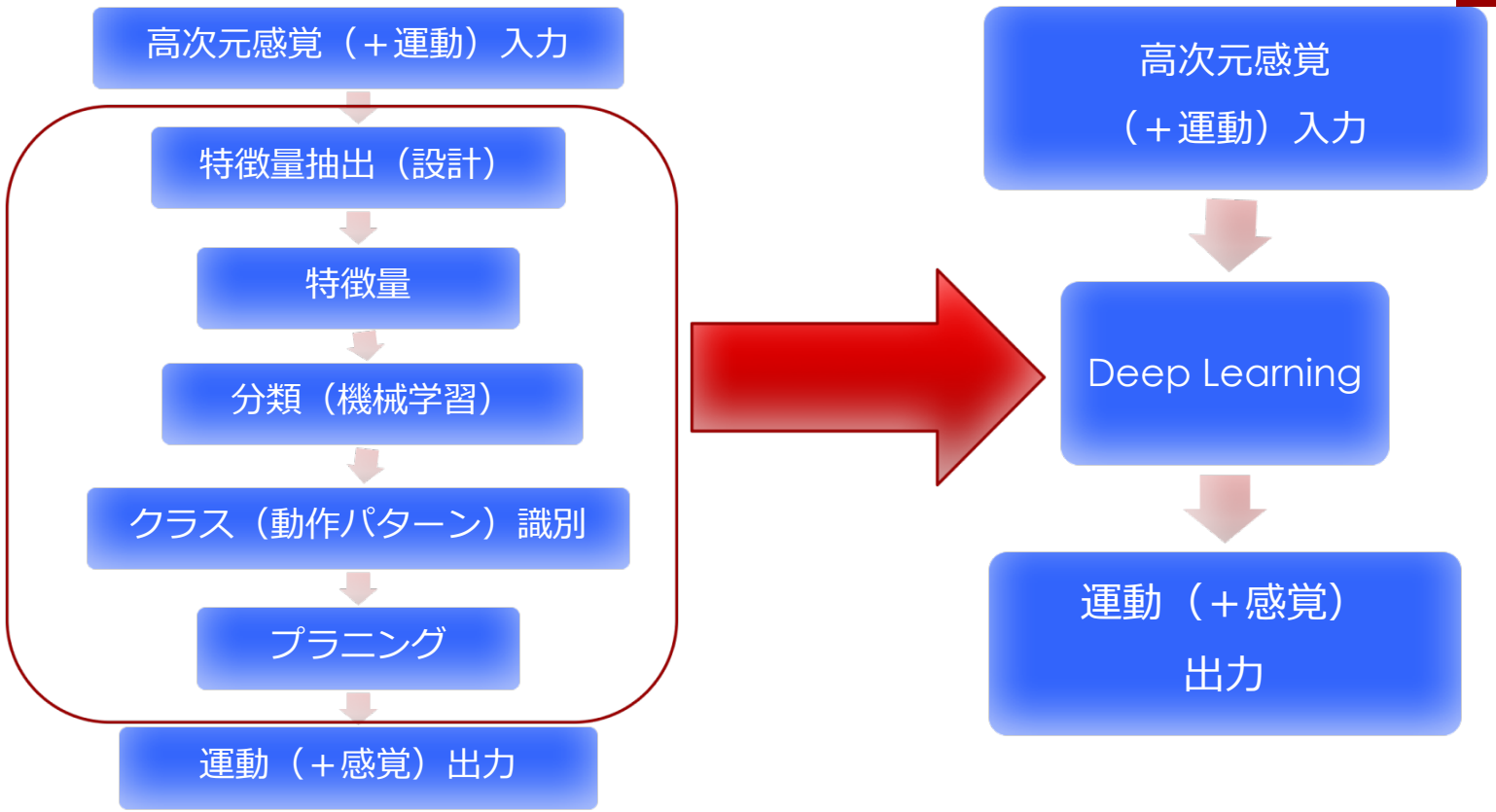
- 物体の画像（と3Dモデル）"だけ"を学習するのでは難しい

  - 把持主体の身体性（ロボットの身体構造とそこから生じうる動作の可能性）が重要
    - 例1）物体を常に上からとるのか？　ハンドの構造は？
    - 例2）対象の材質，変形の可能性は？



(x1)

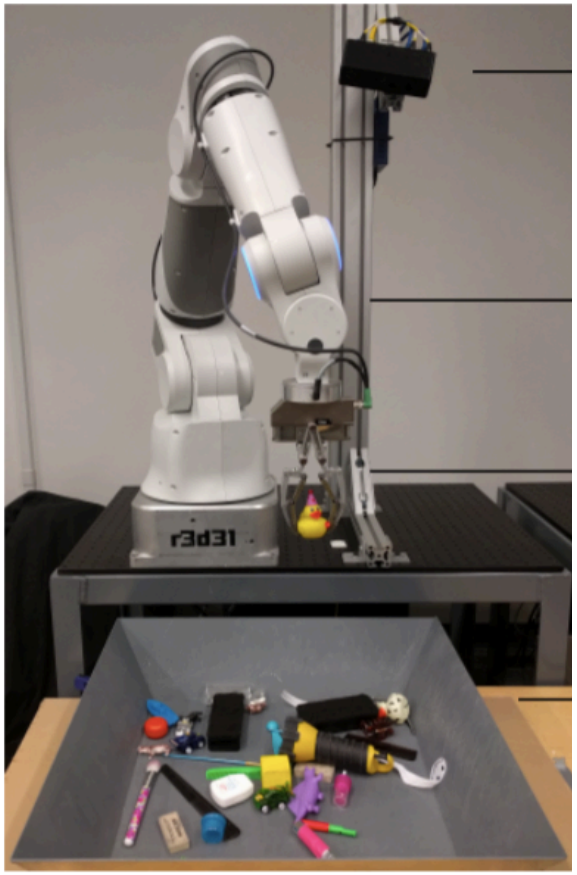
DNNによるロボット行動学習 (End to End Learning)



Googleのロボット制御

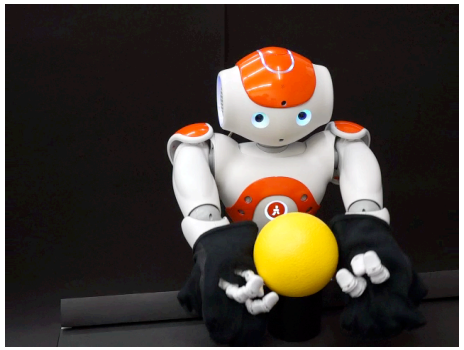
- 一般家庭やオフィスにある様々な物体の把持を，画像特徴量，計画なしで実現
- Deep Q-Learning（深層強化学習）に類似した手法
  Q-function → 画像ピクセルと把持状態から把持成功確率を予測
  政策（Policy）→ 把持動作制御
- 14台のロボットマニピュレータで，計80万回の把持動作を2ヶ月かけて収集

Movie

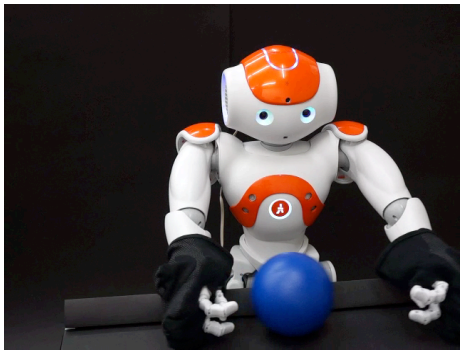


マルチモーダル学習

K. Noda, H. Arie, Y. Suga, and T. Ogata, RAS, Vol.62, No.6, 2014
2014年8月～2015年2月, Top download



Ball lift



Ball rolling

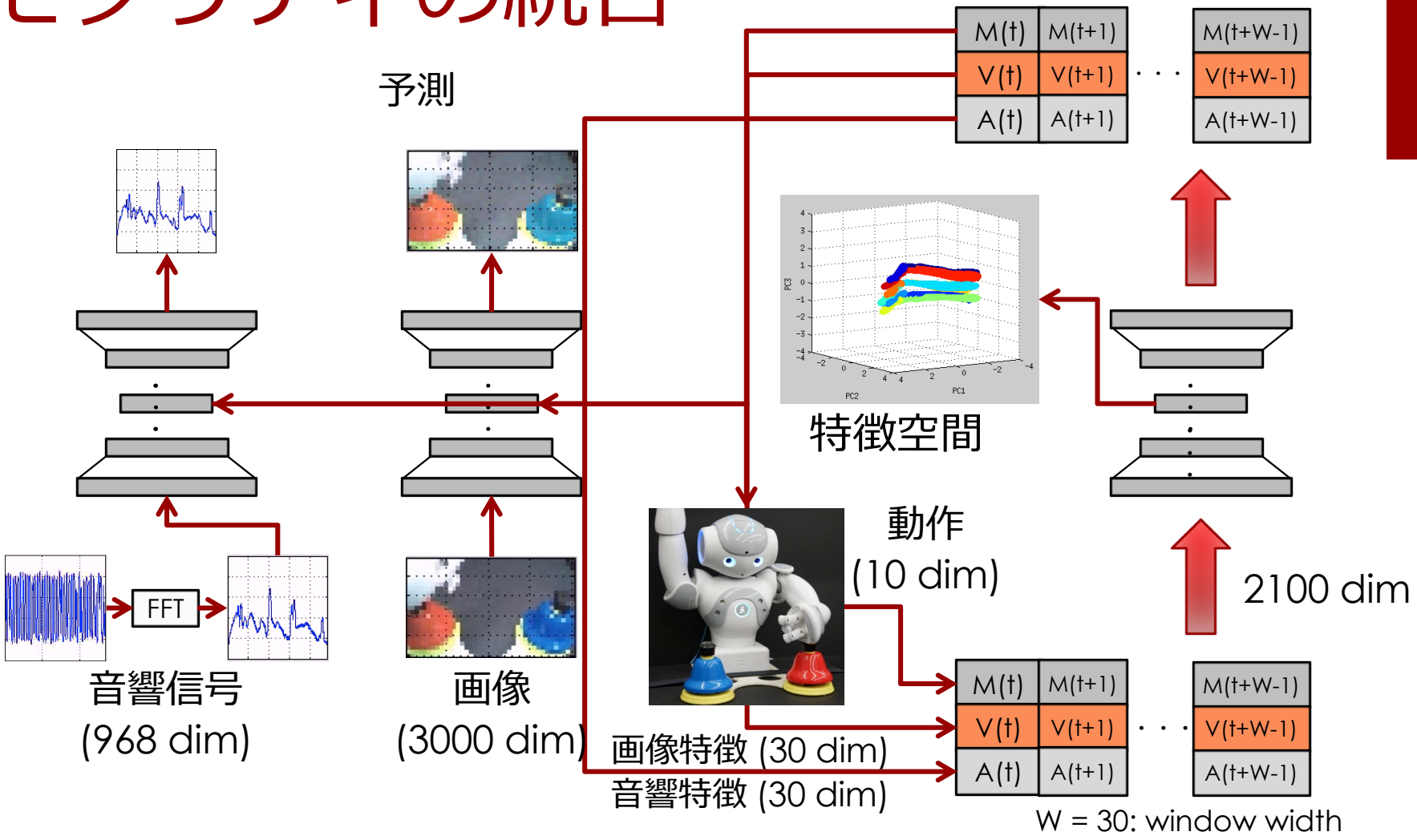
Bell ring R

Bell ring L

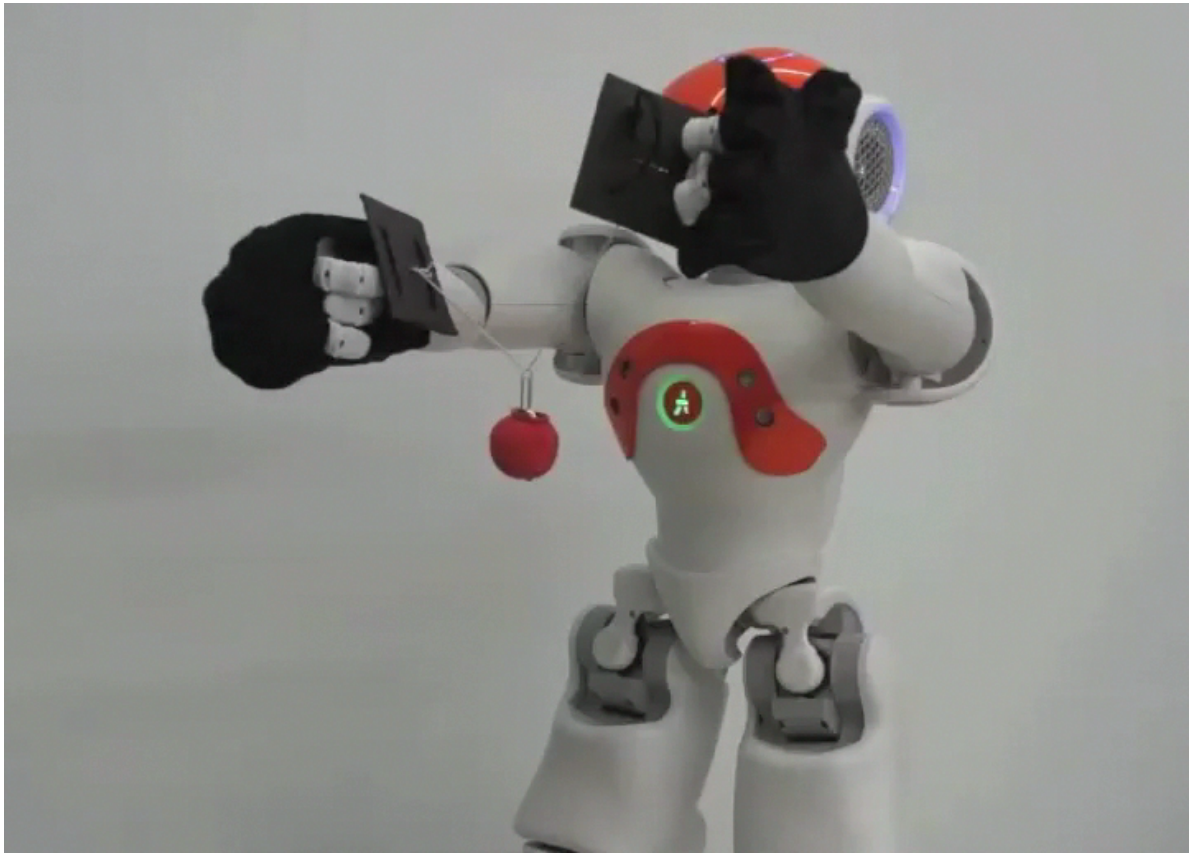
Ball rolling on a plate

Ropeway

モダリティの統合


予測
M(t)
M(t+1)
M(t+W-1)
V(t)
V(t+1)
V(t+W-1)
A(t)
A(t+1)
A(t+W-1)
特徴空間
FFT
音響信号
(968 dim)
画像
(3000 dim)
動作
(10 dim)
2100 dim
画像特徴 (30 dim)
音響特徴 (30 dim)
W = 30: window width

動作生成



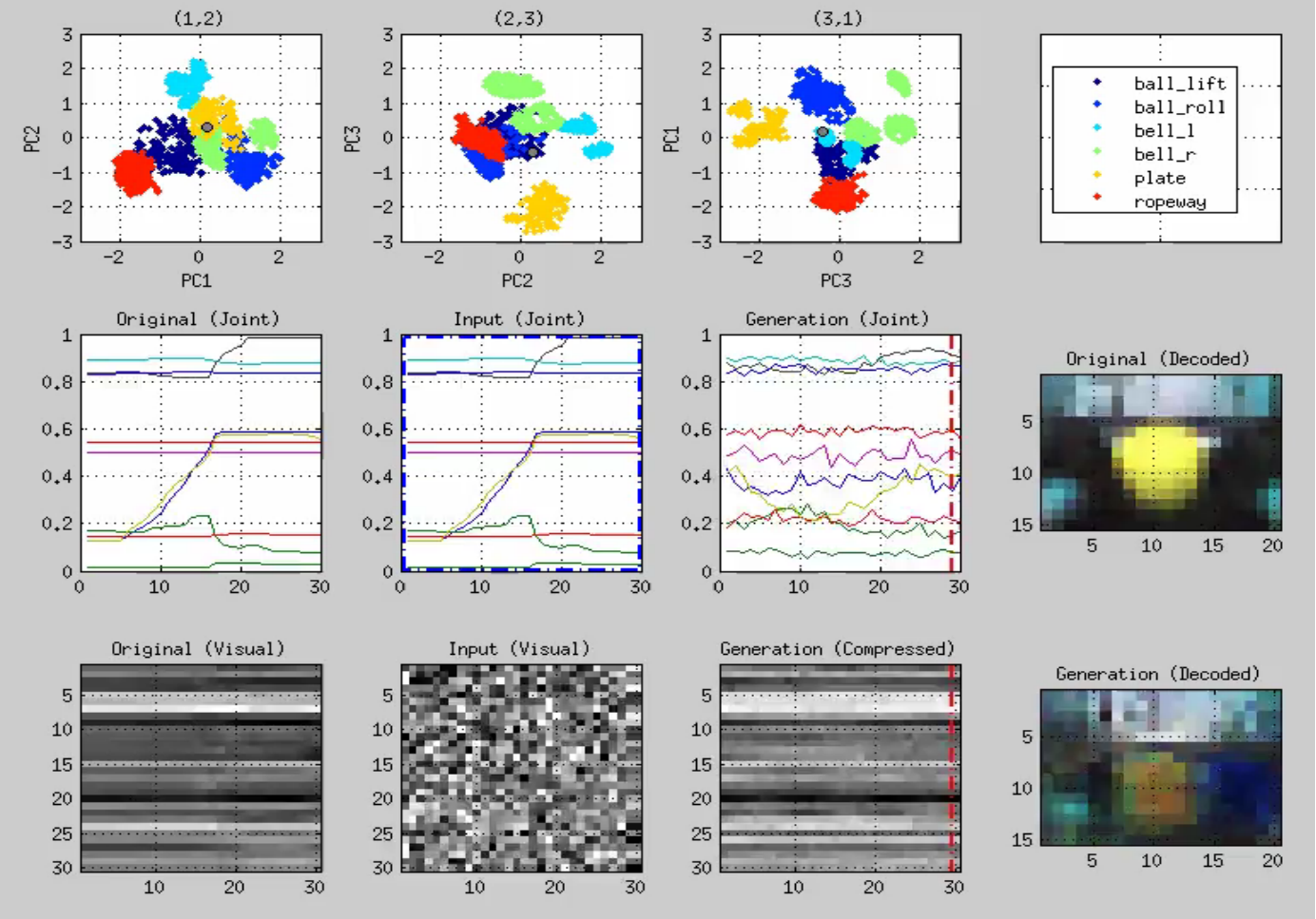
Ropeway → Bell ring R → Bell ring L → Bell ring R

関節角度から画像の想起

DNNによるマルチモーダル空間

関節角度系列

DNN画像特徴系列

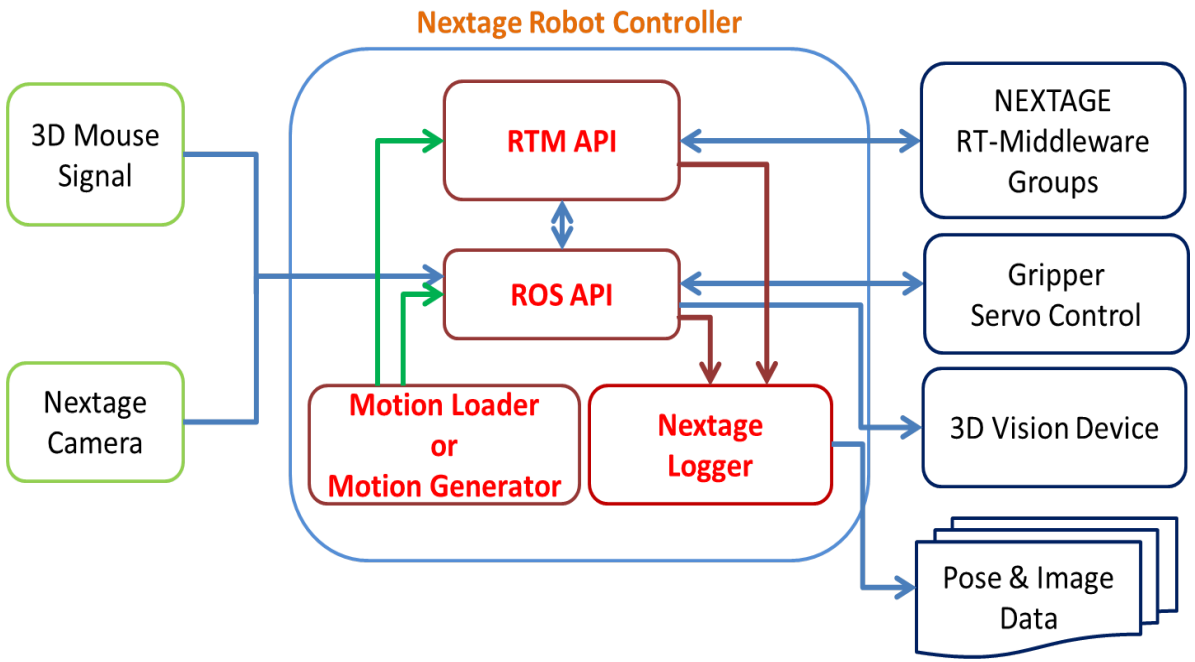


正解画像

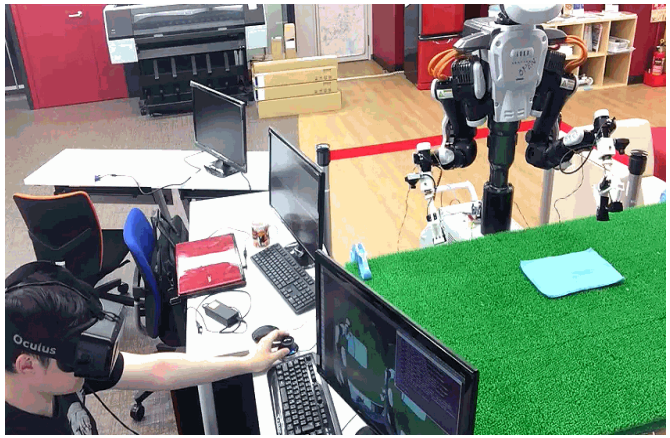
想起画像

教示学習システム

P. Yang, K. Sasaki, K. Suzuki, K. Kase, S. Sugano, and T. Ogata, IEEE Robotics and Automation Letter, 2016.

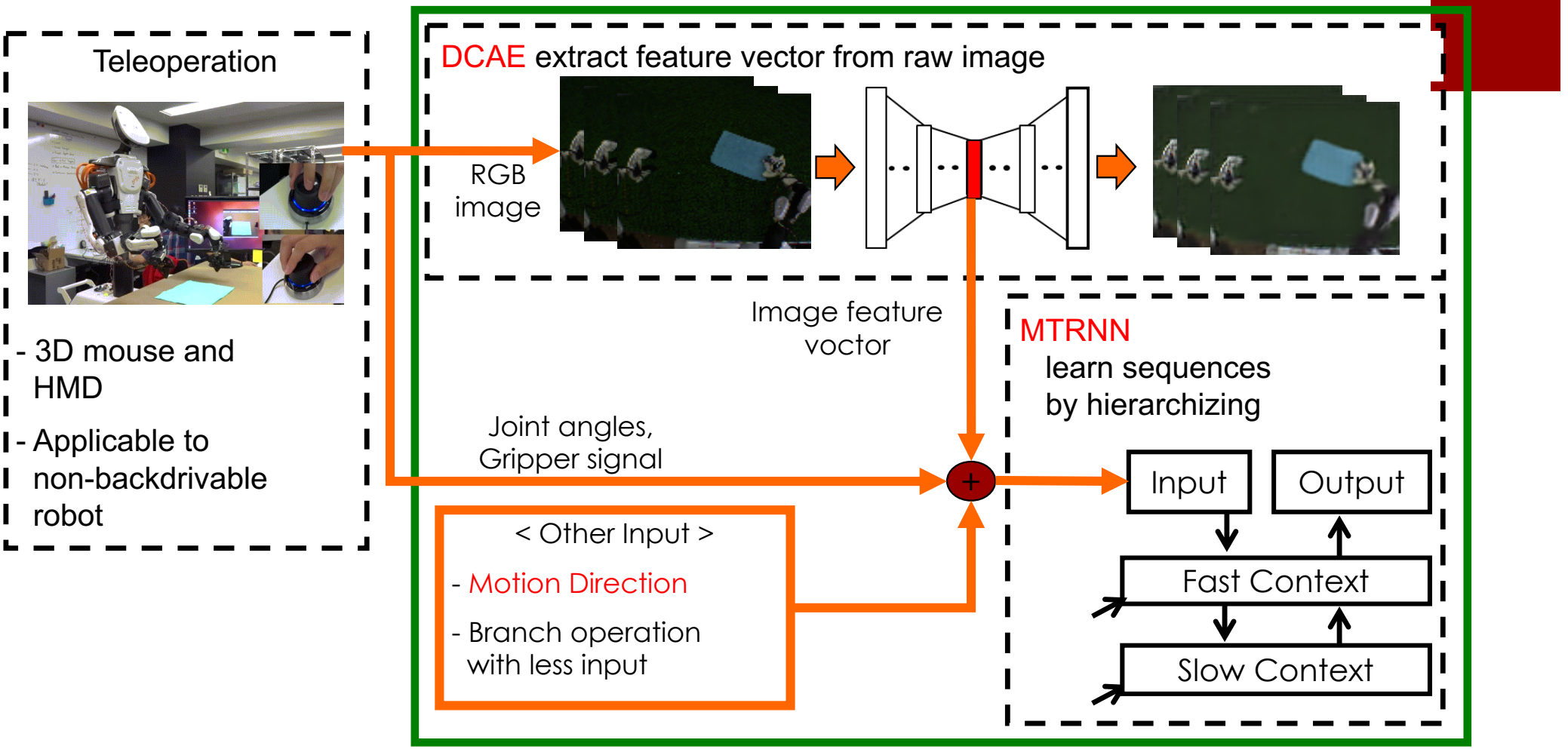


マニュアル教示モード



コマンド教示モード

システム構成

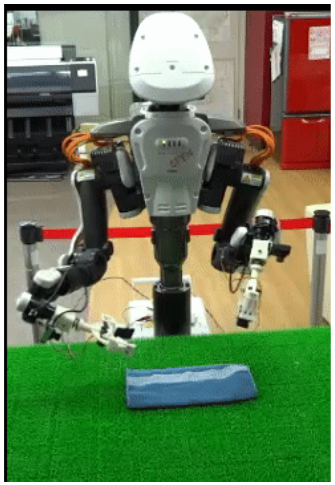

Teleoperation
- 3D mouse and HMD
- Applicable to non-backdrivable robot
DCAE extract feature vector from raw image
RGB image
Image feature voctor
MTRNN
learn sequences by hierarchizing
Joint angles, Gripper signal
+
Input
Output
Fast Context
Slow Context
< Other Input >
- Motion Direction
- Branch operation with less input

折畳みタスク

- 対象物体：
  - 学習データ：4種類の布
  - テストデータ：3種類の布
- 動作：
  - 無造作に置かれた布の把持と折畳み
  - ホームポジションへの戻る動作を含む
- 学習データ:
  - 右カメラ：112x112x3 (37632次元)
  - 2腕 + 2ハンド (14次元)
  - サンプリングレート 10FPS
    (35動作, 平均70秒)

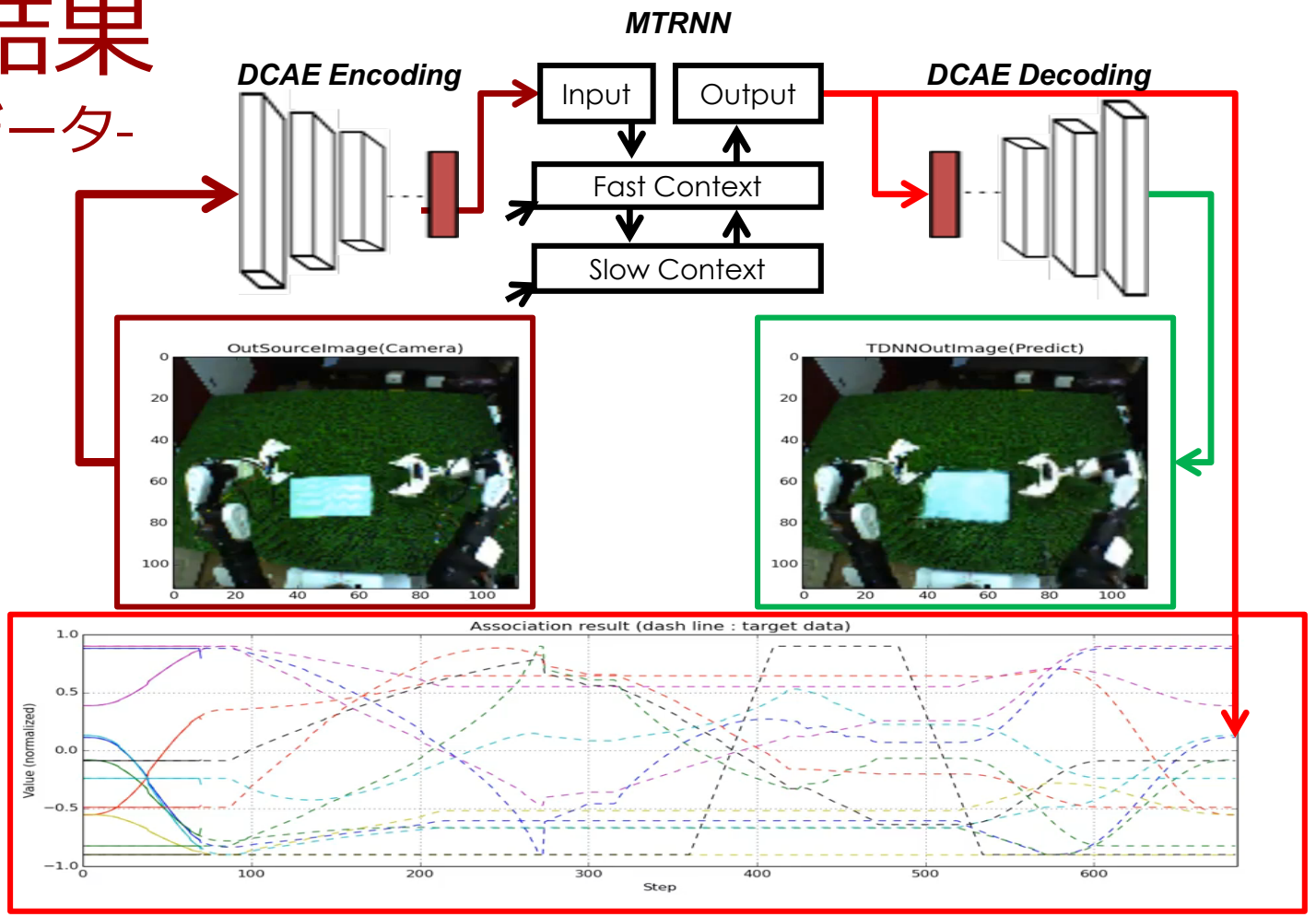
Train Data

Test Data

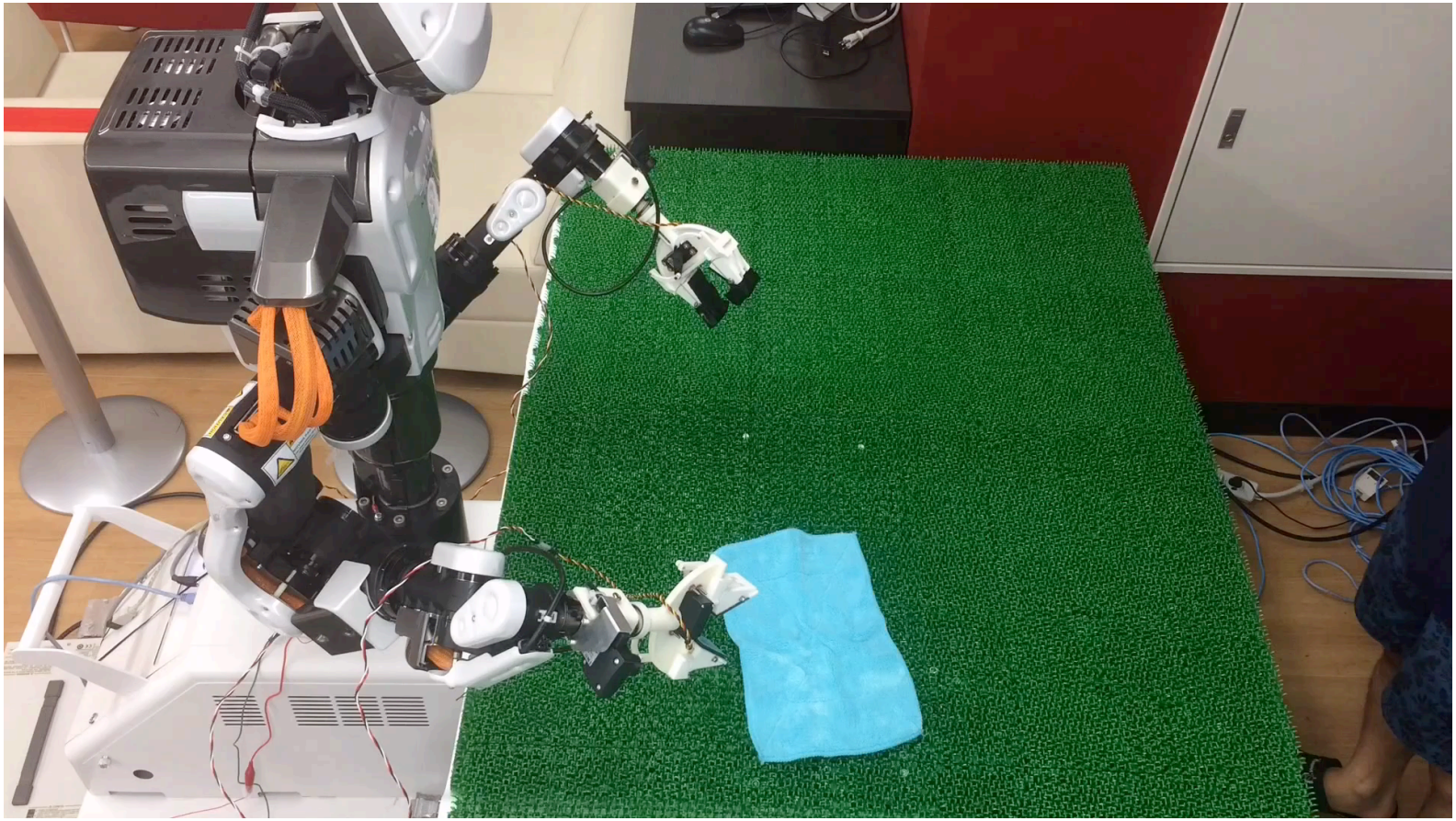


連想結果

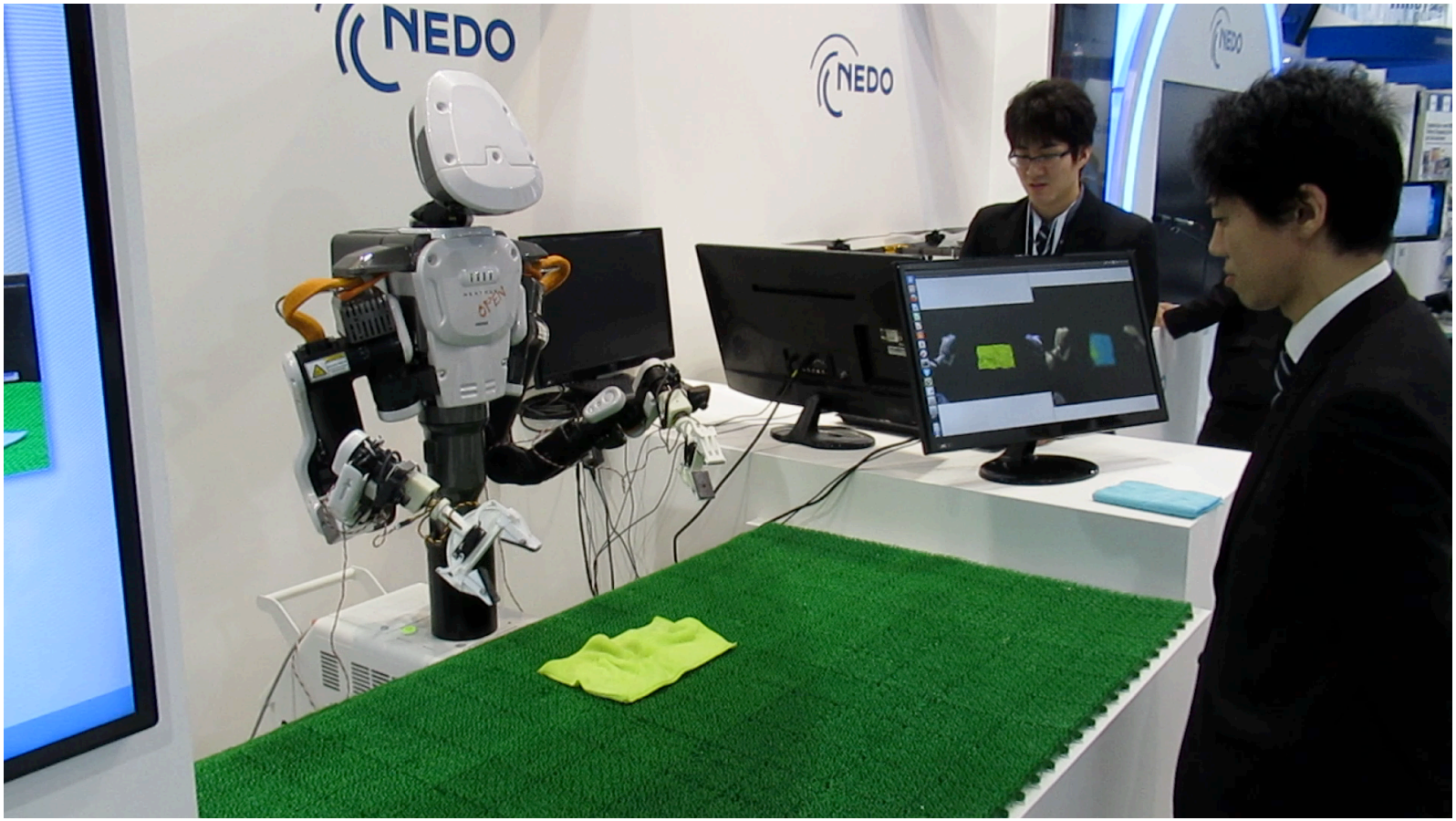
–未学習データ-


MTRNN
DCAE Encoding
DCAE Decoding
Input
Output
Fast Context
Slow Context
OutSourceImage(Camera)
TDNNOutImage(Predict)
Association result (dash line : target data)
Value (normalized)
Step

オンライン動作生成 (with MTRNN)

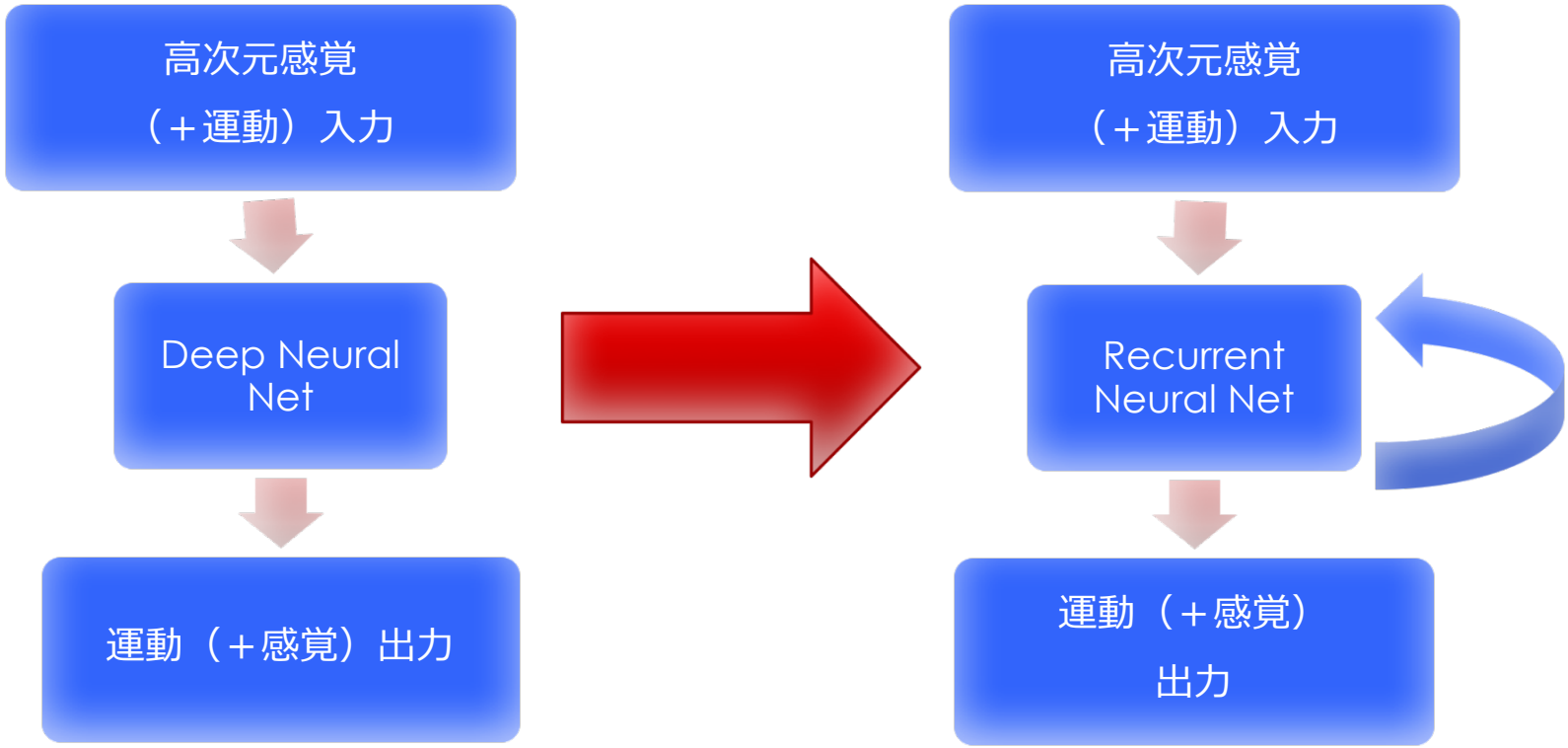


オンライン動作生成 (with MTRNN)



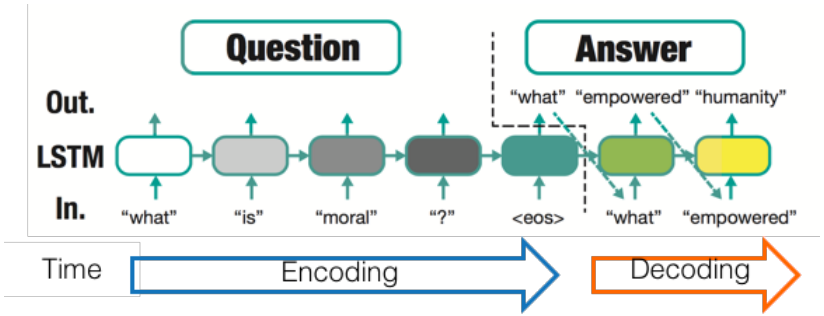
@Cebit 2017 Orignal speed x1

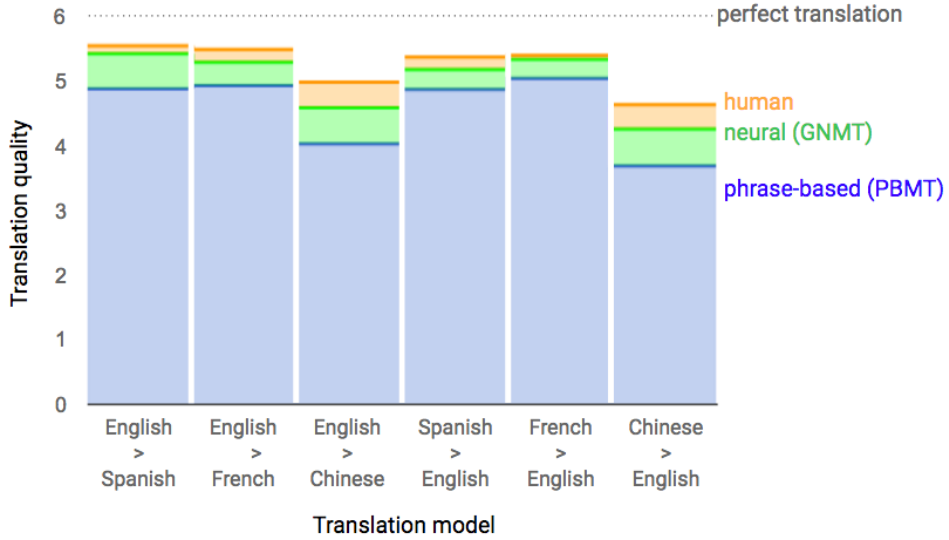
多層型から再帰結合型へ (Seq. to Seq. Learning)



Seq2Seq Learning

- 近年，対話，翻訳の急激なパフォーマンス向上





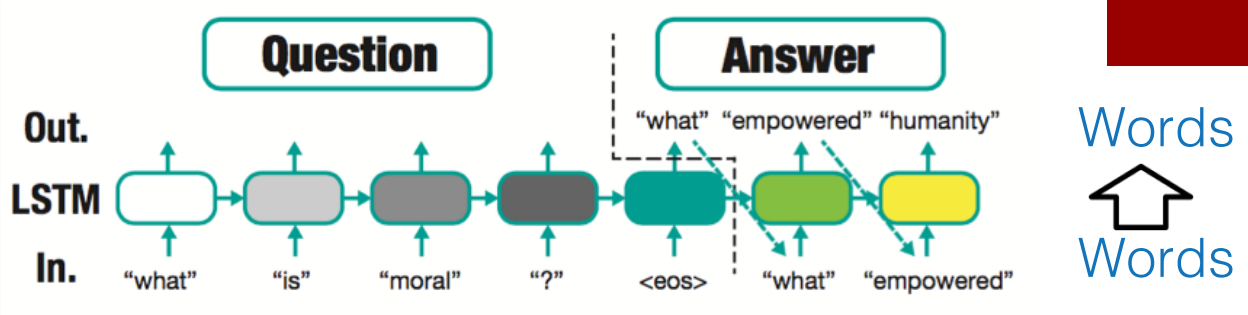
Conversation 5: Morality

Human: *what is moral ?*
Machine: *what empowered humanity , what intellectual the essence is .*
Human: *what is immoral ?*
Machine: *the fact that you have a child .*
Human: *what is morality ?*
Machine: *what is altruism ?*
Human: *what is the definition of altruism ?*
Machine: *if you don 't believe in god , then you don 't know .*
Human: *ok ... so what is the definition of morality ?*
Machine: *well , the truth is , you 're not a believer in god almighty .*
Human: *tell me the definition of morality , i am quite upset now !*
Machine: *i 'm not ashamed of being a philosopher !*
Human: *tell me the definition of morality .*
Machine: *i don 't have ethics .*

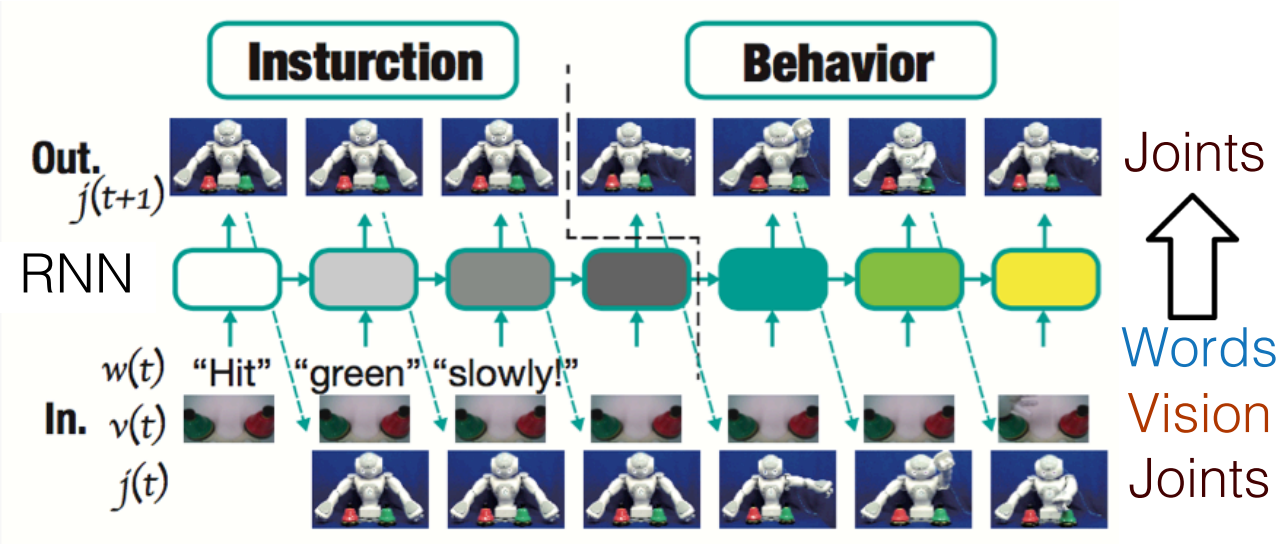
言語と運動の結合

T. Yamada, S. Murata, H. Arie, T. Ogata, Frontiers in Neurorobotics, NIPS2016

言語タスクの
seq2seq

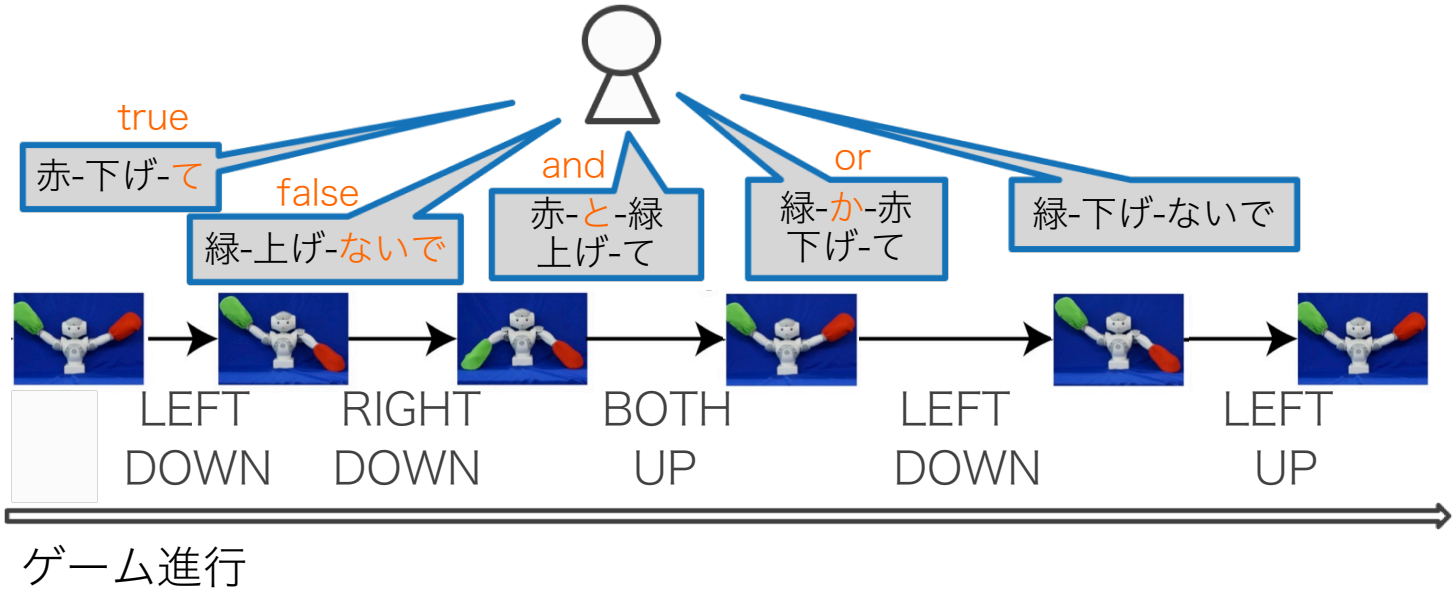


言語から行動へのマルチモーダル
seq2seq



旗上げゲーム

① ロボットの両腕に旗 (RGBのいずれか) を持たせる
② 人が指示を与える
③ ロボットが行動を生成



論理語の扱い

[1] “true”, “false”

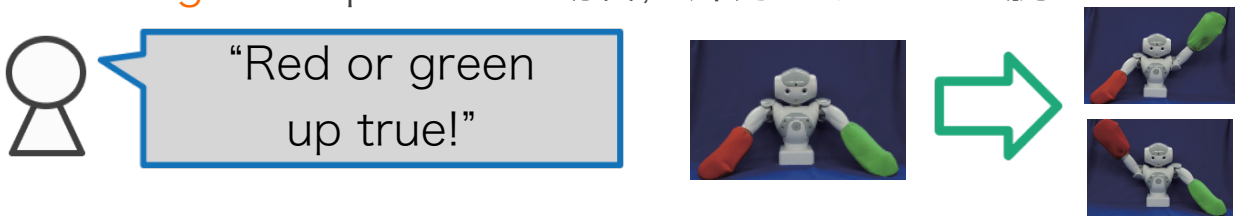
“up true (上げて)”と”down false (下げないで)”は同一の意味UPを表す.

[2] “and”

“Red and green up true.”の場合, 両方の腕を上げるを正解とする.

[3] “or”

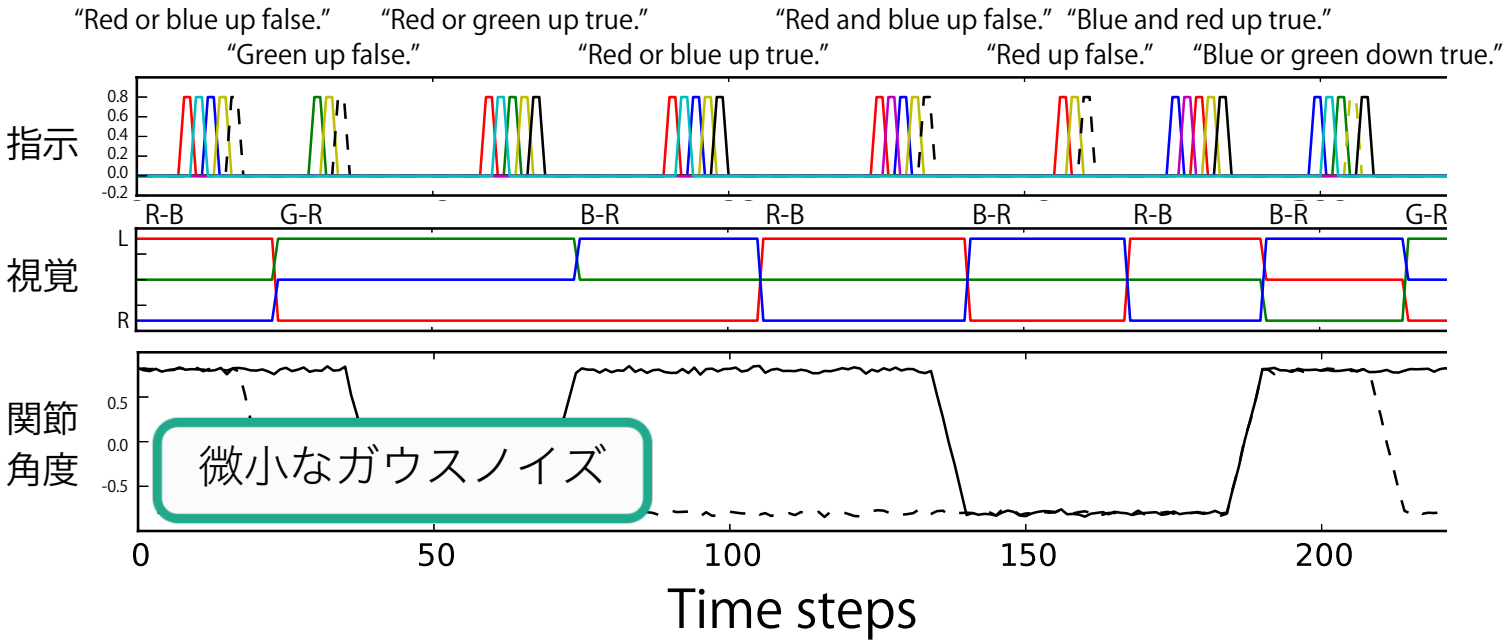
“Red or green up true.”の場合, 片方いずれかの腕のみを上げる.



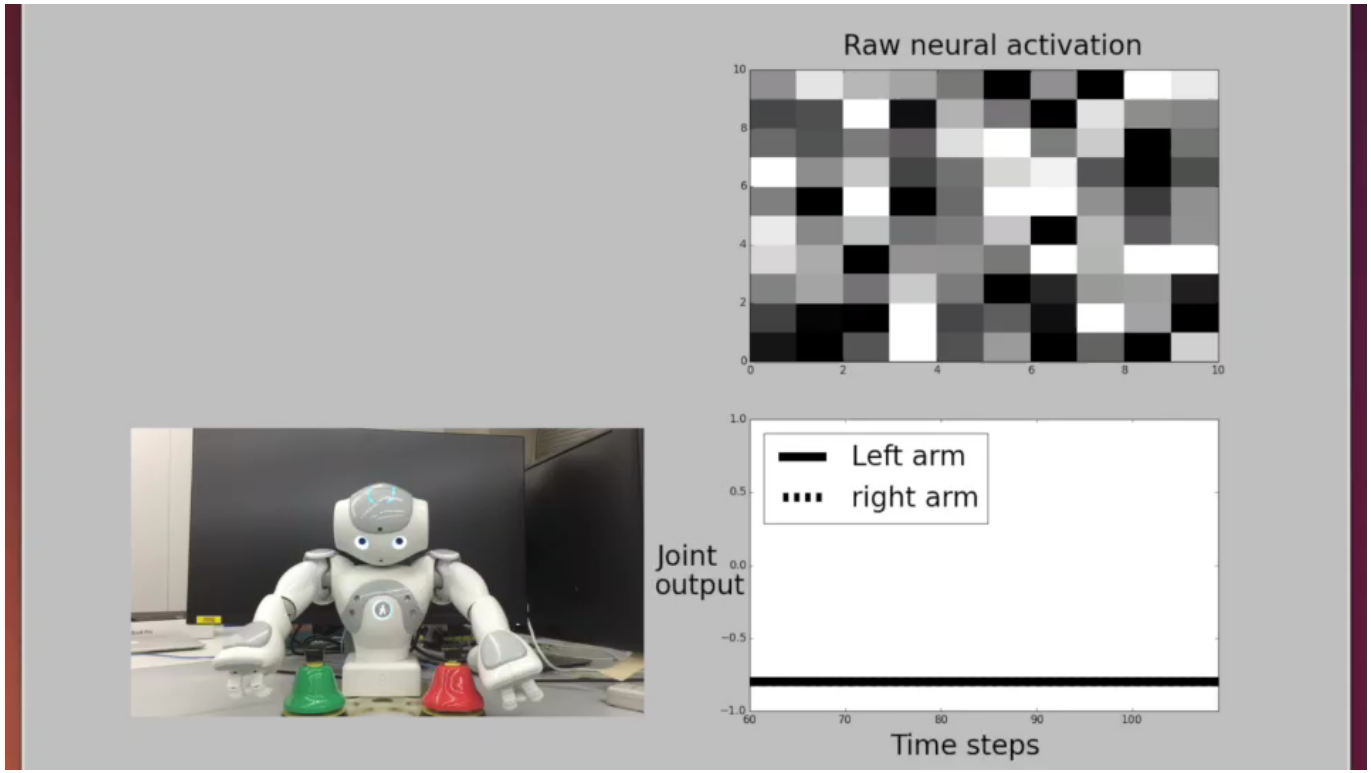
RNN（LSTM）への入力例

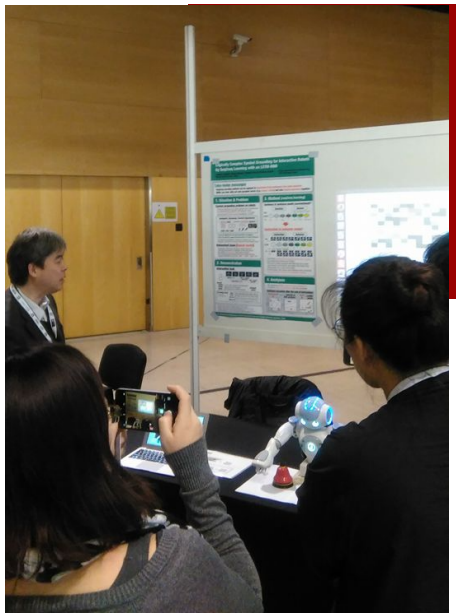
旗の持ち方 x 待機の姿勢 x 言語指示の組み合わせ

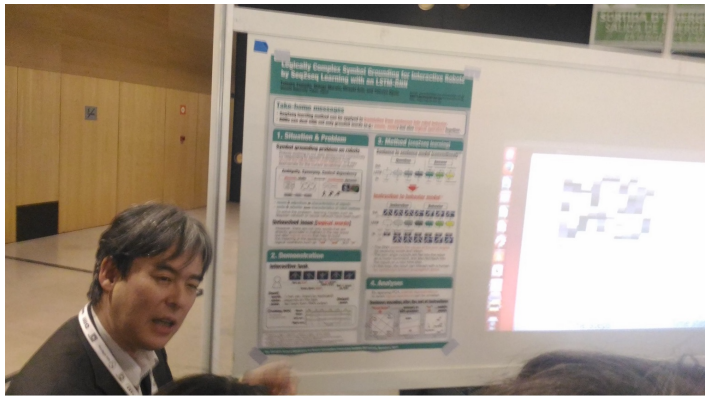
576通りの可能な状況 を含む組合せ的に複雑なタスク



タスクと学習結果 例







NIPS2016でのデモ発表

今後の発展

- Web，テキストデータの認識
  - データは大量にある

  （ここに壁）

- ロボット（実世界）への応用
  - 汎化性能を確保しうるデータをどう集めるか
    - クラウドと接続した標準機(IoT?)
- 言語理解

ブラックボックスとしてのDL (1)

- コア技術として発展しつつあるが，メカニズム，設計法，利用法などの理解は未だ不十分

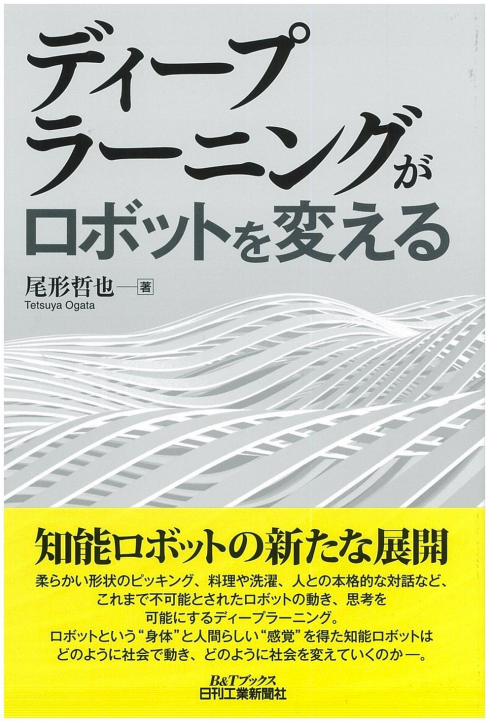
- 今後の理論，応用など多様なレベルでのDLの積極的な研究と活用が極めて重要！
  - 深層学習の内部理解の研究はまだ途上
    - 複数の低次元多様体（の様なもの）に情報が埋め込まれる

JDLA

日本ディープラーニング協会

ブラックボックスとしてのDL (2)

- 「中身がわからないが性能が良い」というシステム
  - 移動手段としての"馬"，人間介助を行う"犬"
- 「創発的」な知能
  - 動作製造責任
    - 消費者期待基準
    - 危険効用基準
    - 開発危険の抗弁（予期し得ない危険）
  - ロボットを購入した人
    - 動物のアナロジー（Schaerer et al.）
    - 子供のアナロジー（夏井）
  - ロボットに責任主体性を観念できるか（ロボット法人説, Asaro, 2007など）



Waseda University

終わり

ogata@waseda.jp