

RATM: Recurrent Attentive Tracking Model

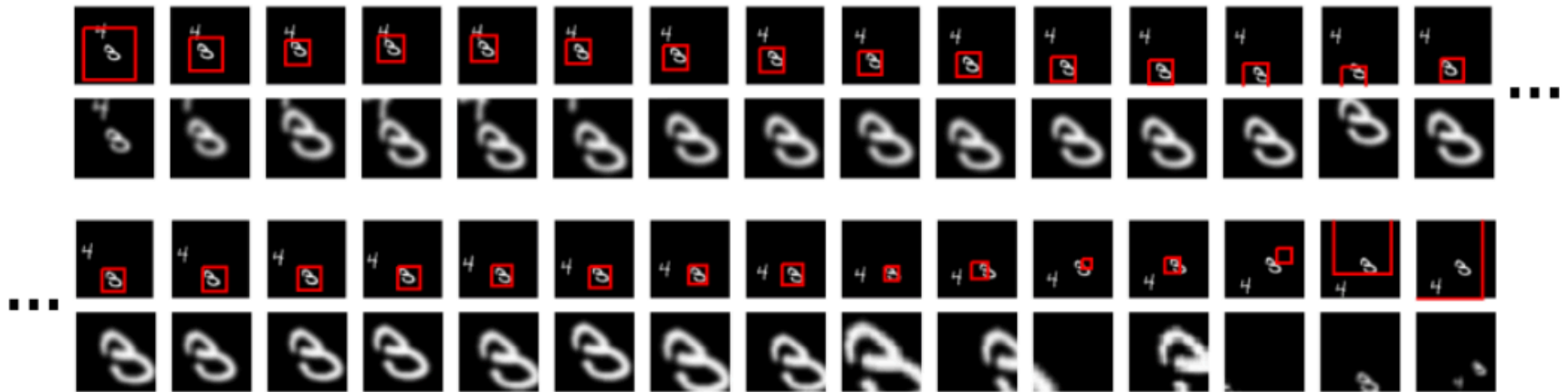
Samira Ebrahimi Kahou, Vincent Michalski, Roland Memisevic

2016/1/15

発表者: 金子貴輝

選考理由・概要

- ICLR 2016のunder review
- RNNを使った注意モデル
 - → 新しい再帰構造になるかも
- 物体追跡のための注意モデル
 - → 人工データで亜種の問題を作れそう
- コードは公開予定(まだ無かった)



内容

- 概要
- 導入
- 関連手法
- 提案手法
- 実験
- 比較
- 結論・今後の課題

導入：従来研究

- 物体追跡は古典的なタスク
 - 従来手法は特徴空間の類似性にもとづいた複数の窓候補を選択評価するサンプリング法を使っていた(Smeulder et al., 2014)
- 提案手法はフレーム内の検索なしに場所を予測する
 - Gregor et al. (2015)(DRAW)の注意モデルをベースとする
 - DRAWでは画像生成だが、物体追跡に使える
- (Denil et al., 2011)はDLで物体追跡しているが強化学習を使った複合モデル
 - 対して、提案手法はDRAWと同じく勾配法で学習できる

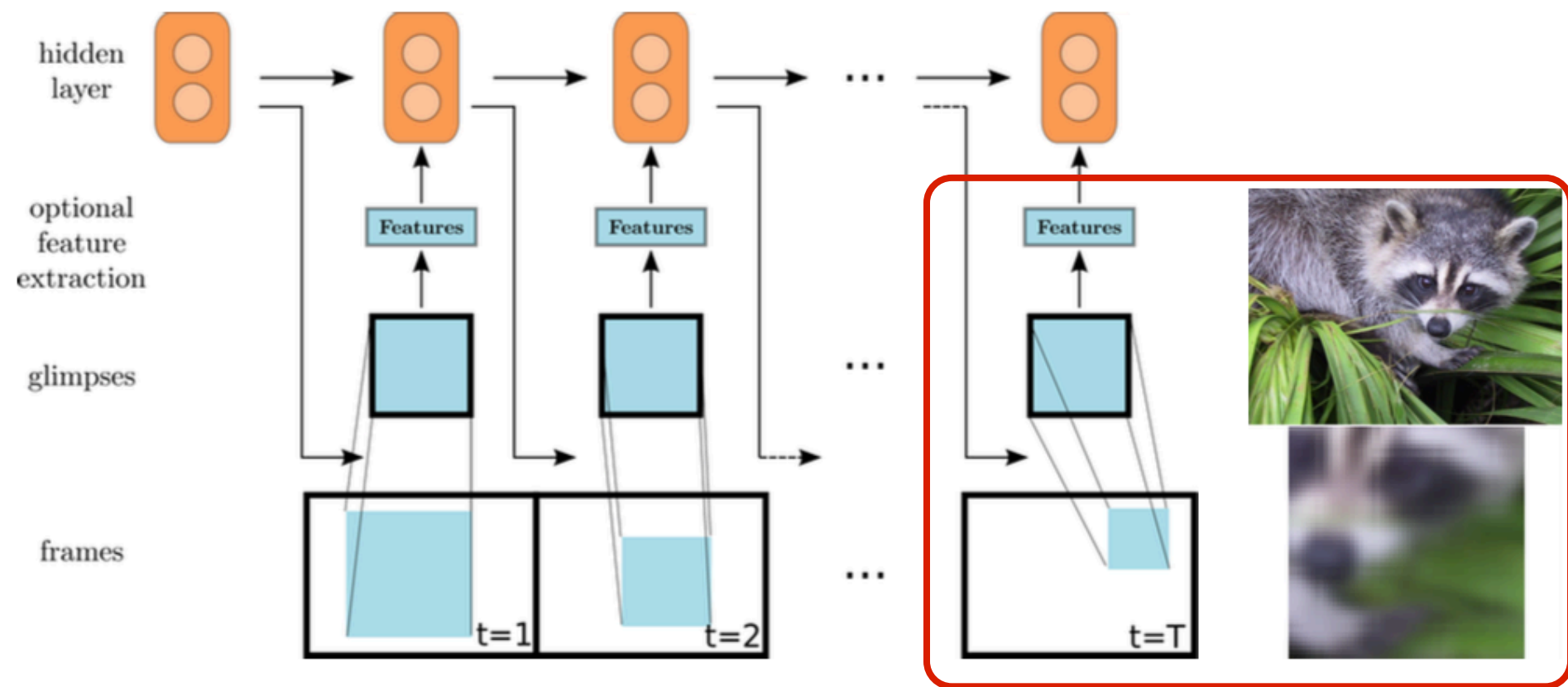


Tracking region



導入: Attention Mechanism(1)

- Attention Mechanism 入力から一部の情報を抜き出す機構
- 画像ではパッチを同じサイズに拡大して取り出す

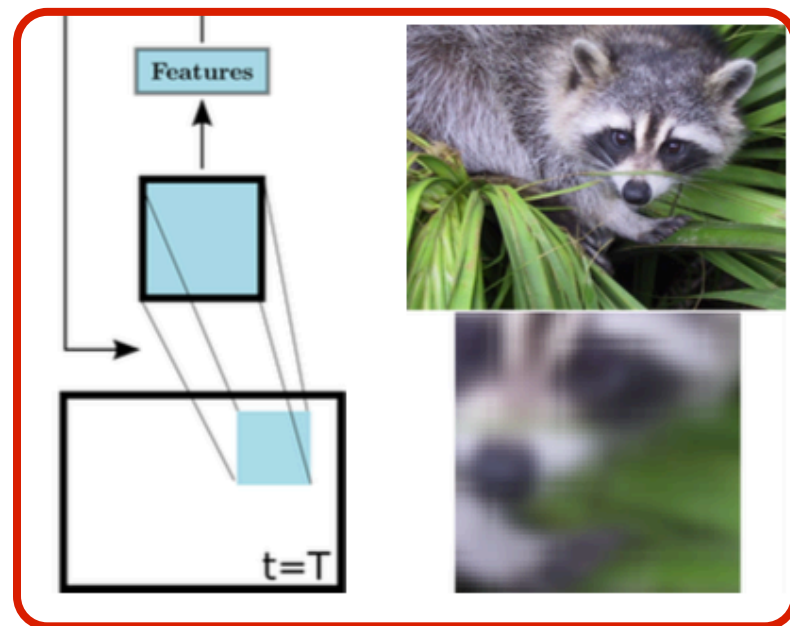


導入: Attention Mechanism(2)

- DRAWでは, フィルタを掛けて固定長のパッチにする
- フィルタのパラメータは中心座標, ぼかし具合, サイズ
- 隠れ層の線形変換で決定
- 提案手法でも同様の方式を用いる

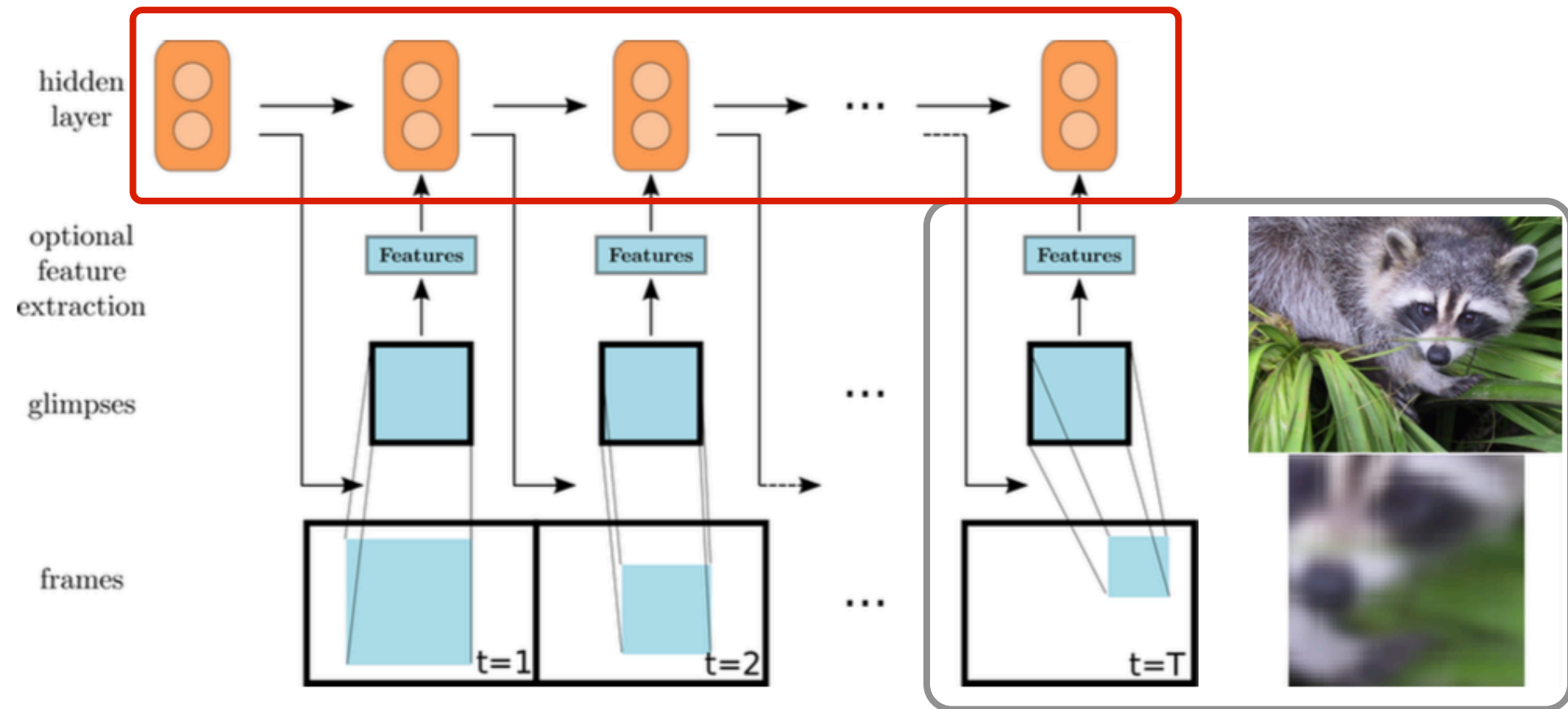
$$\mathbf{p} = \mathbf{F}_Y \mathbf{x} \mathbf{F}_X^T$$

$$(\tilde{g}_X, \tilde{g}_Y, \tilde{\sigma}, \tilde{\delta}) = \mathbf{W} \mathbf{h}$$



提案手法：全体図

- イメージを切り出す注意機構と、どこを注意すべきか予測するRNNからなる



提案手法：RNN

- 再帰重みをスケールされた単位行列で活性化関数をReLUにした, IRNNを使用

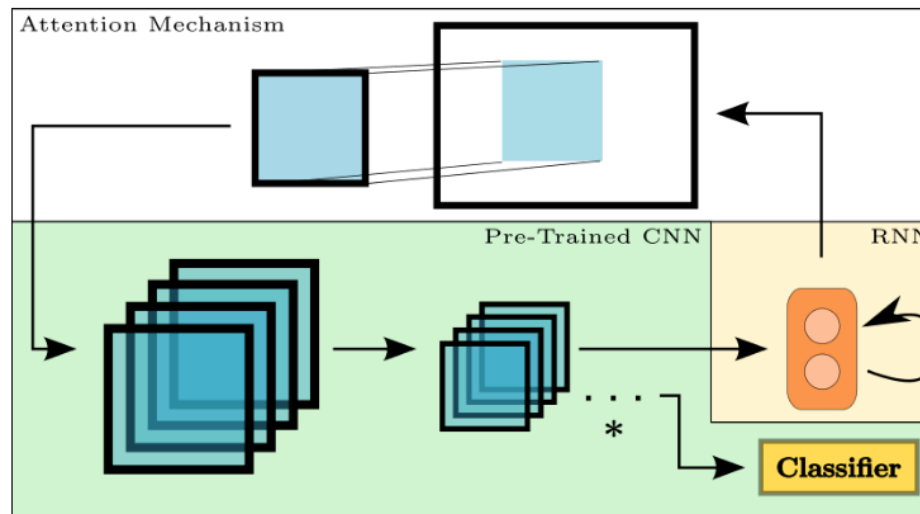
$$\mathbf{h}_t = \sigma(\mathbf{W}_{in}\mathbf{x}_t + \mathbf{W}_{rec}\mathbf{h}_{t-1})$$

$$\sigma(x) = \max(0, x)$$

- Ebrahimi Kahouらの実装したTheano-based IRNNを使用

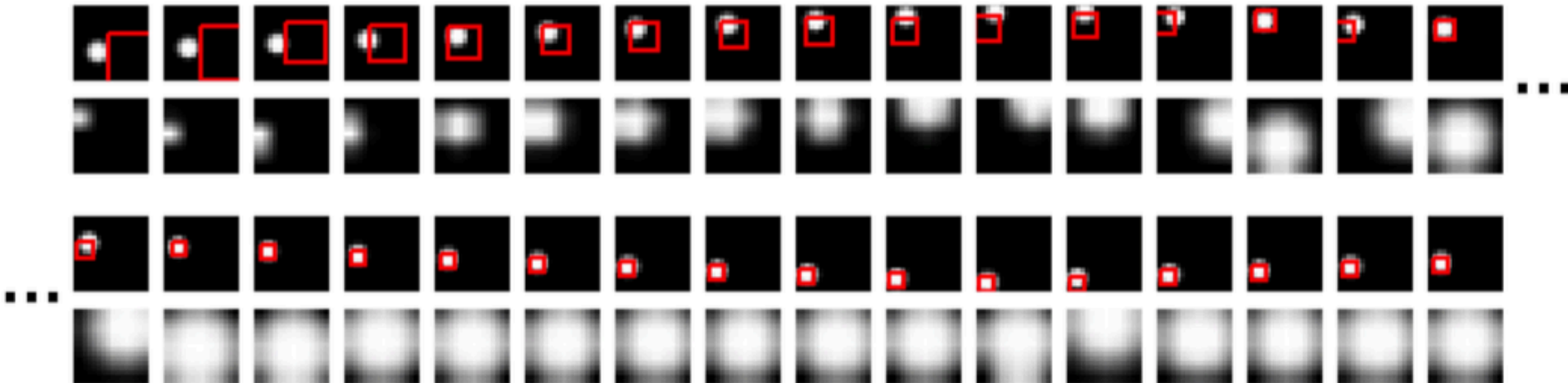
提案手法：RATM

- 学習されたCNNを使い画像から特徴量を作る
- CNNはターゲットのクラス分類もする
- 過去の軌道は追跡に役立つので注意した場所をRNNの隠れ状態で表現する
- コスト関数の作り方は後述



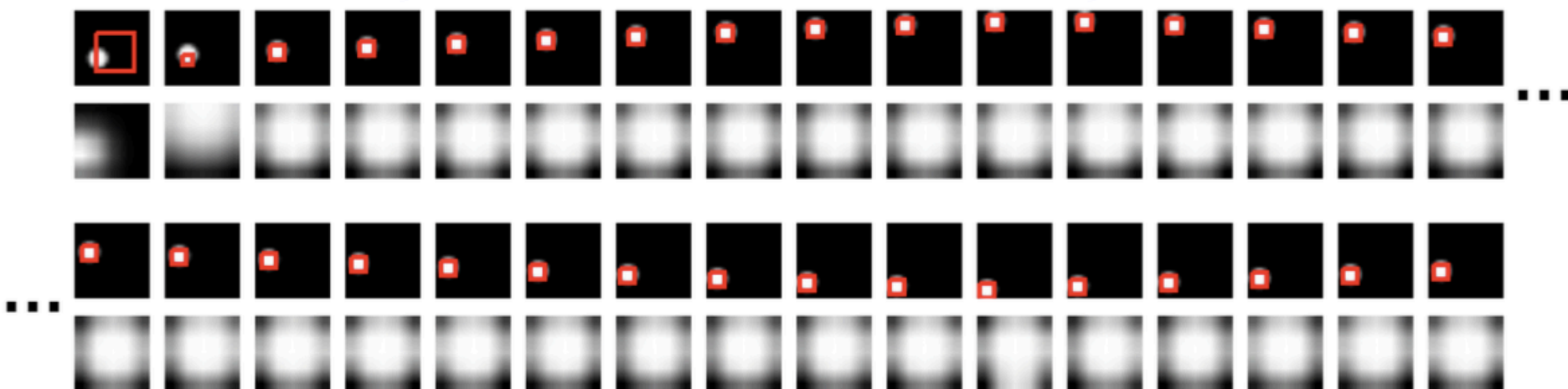
実験1 : Bounding ball

- 跳ねるボールデータセット(Sutskever et al.2009)
 - 32フレーム, サイズ20x20
- 隠れユニット数=64, 注意窓のピクセル数=5x5
- 学習はSGD, 学習率=0.01, epoch数=200
- 初期の注意窓は全体を見る
- 最終フレームで注意されたパッチと真のパッチの平均2乗誤差を最小化
- 訓練100,000枚, テスト10,000枚, 勾配制限1.0, ミニバッチ数は32 or 16
 - この後の実験でも共通



実験1 : Bounding ball

後述する窓の座標についての平均二乗誤差の最小化でも実験



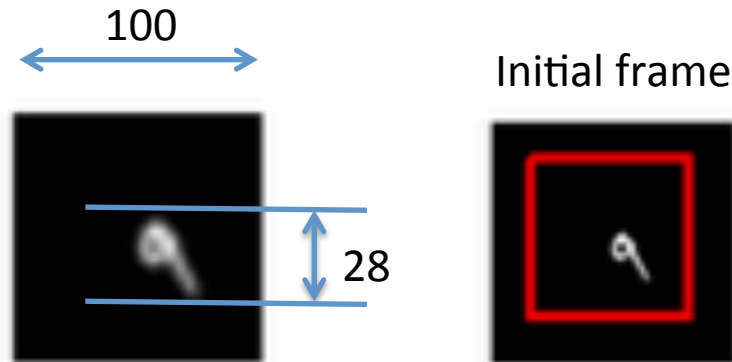
注目しなくても、RNNがボールの単純な動きを覚えただけ？



物体と動きを複雑にして実験

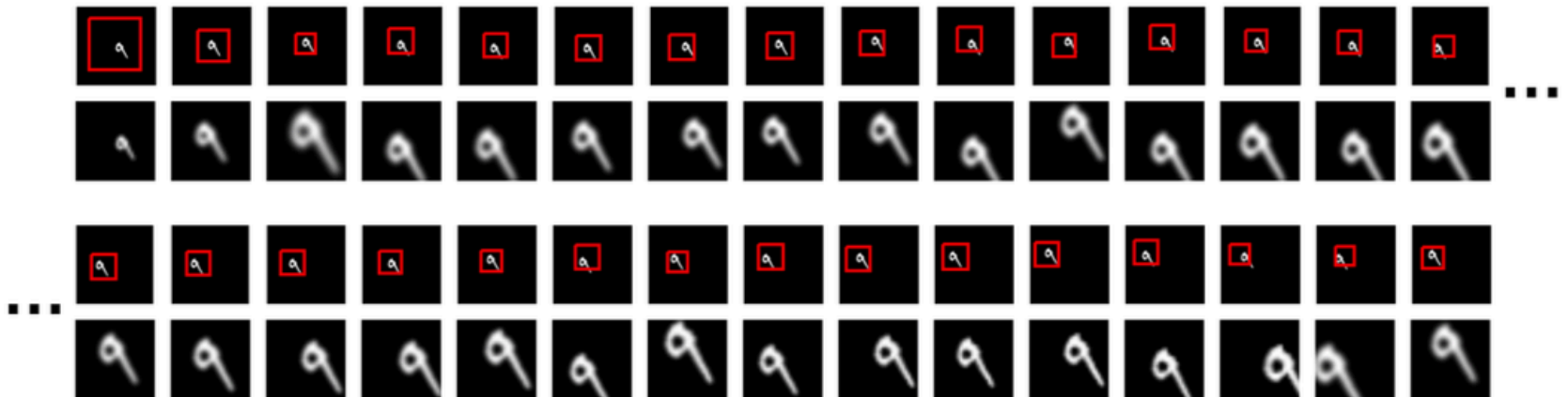
実験2,3 : MNIST digits

- 物体の種類を増やしダイナミクスを予測しにくくする
- ランダムに取り出してきた1つの数字を, 慣性付きのランダムウォークさせる
 - 取り出す数字は訓練, テストで分ける
- 隠れユニット数=100
- 学習はSGD, 学習率=0.001, momentum=0.9, エポック数=500
- 初めはおおまかに全体に注目
- CNNの事前学習
 - CNNはMNIST digitsで訓練され, RATMの学習では固定
 - CNNは32個の5x5filtersの畳み込み層と2x2maxpooling層の2セット, dropout率=0.25, ReLUを使用, 最終層は10classのsoftmax



実験2 : MNIST digits (1文字)

- 事前学習された浅いCNNの平均クラス分類誤差をコスト関数に
 - 失敗
- 正しい窓と注目窓の角の距離をコストに追加
 - 失敗
- 窓の中心座標の2乗誤差に変更
 - 成功
 - 分類誤差のコストがズームしようとするのに対し, このコストが数字をなぞらせる
- 訓練は10フレーム
- テスト時は30フレーム, CNN(のおそらくコスト関数)をoffにする



実験3 : MNIST digits(2文字)

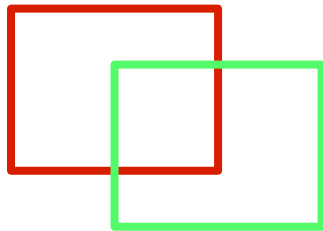
- 他の数字が動いている中で追えるか
- 設定は1文字の時と同じ
 - 初期の窓が追うべき数字にフォーカスされている
- 10フレームで学習, 15フレームでテスト



評価

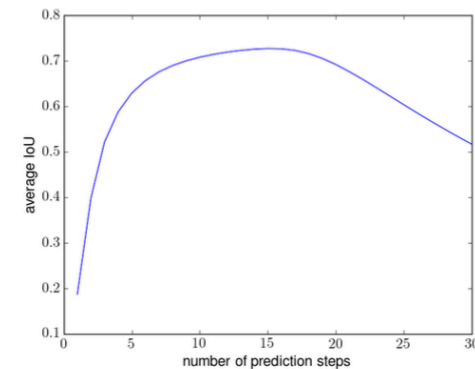
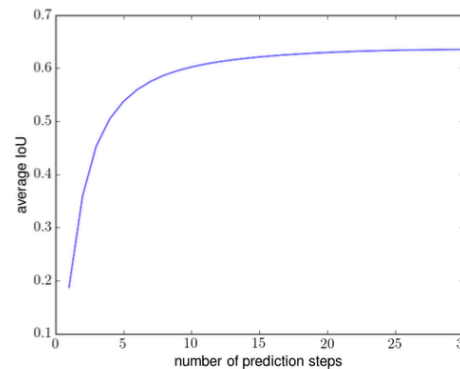
- IoU(領域のAND/領域のOR)で評価
- フレーム平均の表と, フレーム数での変化のグラフ

$$IoU = \frac{|B_{gt} \cap B_{pred}|}{|B_{gt} \cup B_{pred}|}$$



Experiment

Bouncing Balls (training penalty only on last frame)
Bouncing Balls (training penalty only on last frame)
Bouncing Balls (training penalty on all frames)
MNIST (single-digit)
MNIST (multi-digit)



Average IoU (over # frames)

69.15 (1, only last frame)
54.65 (32)
66.86 (32)
63.53 (30)
51.62 (30)

結論・今後の課題

- RNNの注意のメカニズムで人工データで物体追跡できた
- クラス分類誤差だけでは学習できず, 注意窓の座標の2乗誤差を必要とした
- 実世界に対応させたい
 - Caltech Pedestrian Dataset を予定している
- 複数物体の追跡と学習された表現の調査をしたい
 - 特に重なりを扱えるか, 物体の同一性と軌道を圧縮できるかを調べるのが今後の課題

参考文献

- Denil, Misha, Bazzani, Loris, Larochelle, Hugo, and de Freitas, Nando. Learning where to attend with deep architectures for image tracking. *CoRR*, abs/1109.3737, 2011. URL <http://arxiv.org/abs/1109.3737>.
- Ebrahimi Kahou, Samira, Michalski, Vincent, Konda, Kishore, Memisevic, Roland, and Pal, Christopher. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMi '15*, pp. 467–474, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3912-4. doi: 10.1145/2818346.2830596. URL <http://doi.acm.org/10.1145/2818346.2830596>.
- Gregor, Karol, Danihelka, Ivo, Graves, Alex, and Wierstra, Daan. DRAW: A recurrent neural network for image generation. *CoRR*, abs/1502.04623, 2015. URL <http://arxiv.org/abs/1502.04623>.
- Smeulders, Arnold W M, Chu, Dung M, Cucchiara, Rita, Calderara, Simone, Dehghan, Afshin, and Shah, Mubarak. Visual tracking: An experimental survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1442–1468, July 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.230.