Learning Search Path for Region-Level Image Matching

Onkar Krishna, Go Irie, Xiaomeng Wu, Takahito Kawanishi, Kunio Kashino

NTT Communication Science Laboratories, NTT Corporation, Japan

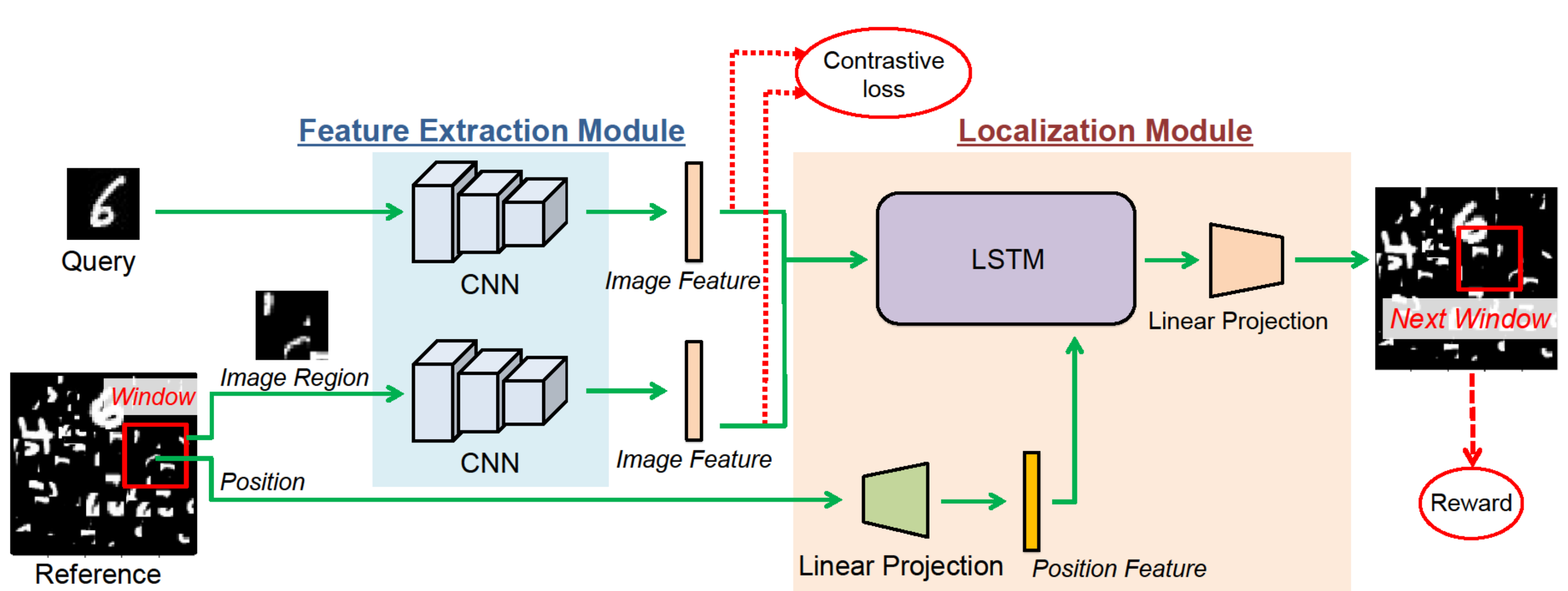
Introduction

- Our task is to find a part of a reference image that matches to a query image.

- A desirable algorithm should be robust and fast.
  - ➢ Existing methods, e.g., pruning based methods, need to evaluate a large number of undesirable candidate regions.
- We proposed a deep-reinforcement learning based image matching method.
  - ➢ Learning efficient search path from data, i.e., use machine learning to pick and evaluate only the highly prospective regions of the reference image.
  - ➢ **Almost 40x faster than the best competitive baseline!**
  - ➢ **Robust to various type of background clutters!**

Model Architecture

Our model has **feature extraction module** and **localization module**



Feature extraction module

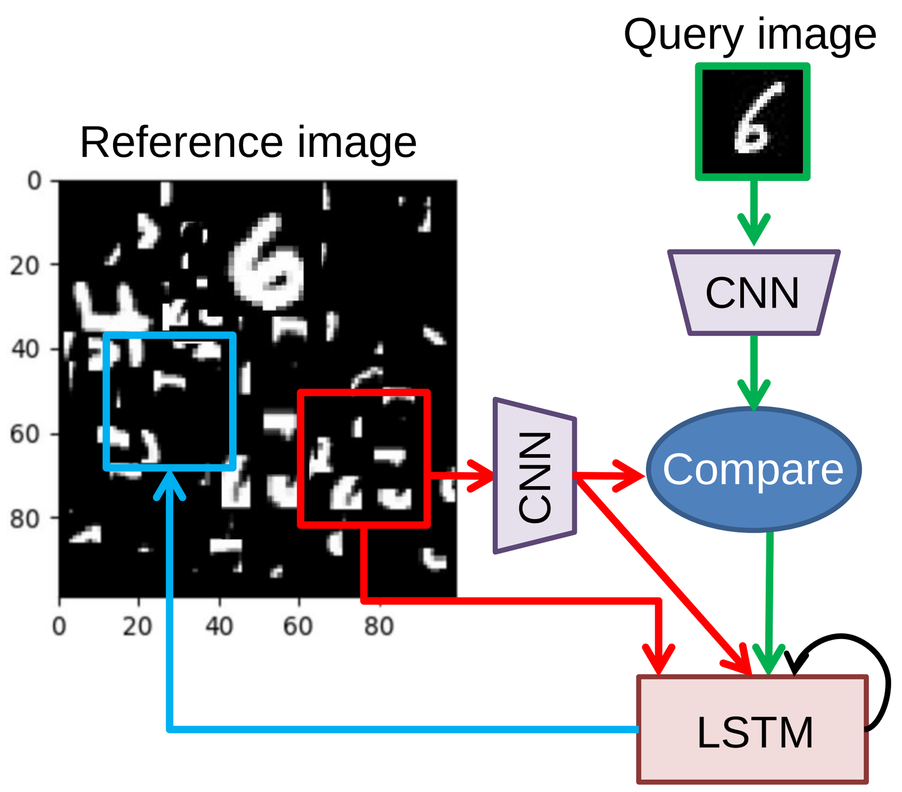
- ➢ Extracts the image features from query and reference image region.
- ➢ Consists of two identical CNNs with the same parameters which have a sequence of five Conv-ReLU layers followed by a global average pooling.

Localization module

- ➢ Has an LSTM that sequentially predicts the next location based on three external inputs including two image features and current window location. .
- ➢ Determines the initial position in a similar way as done in [1]

This design allows us to jointly learn the search path and effective deep features for matching!

Overview of Algorithm Behavior



Step 0. Choose the initial window based on the global reference image.

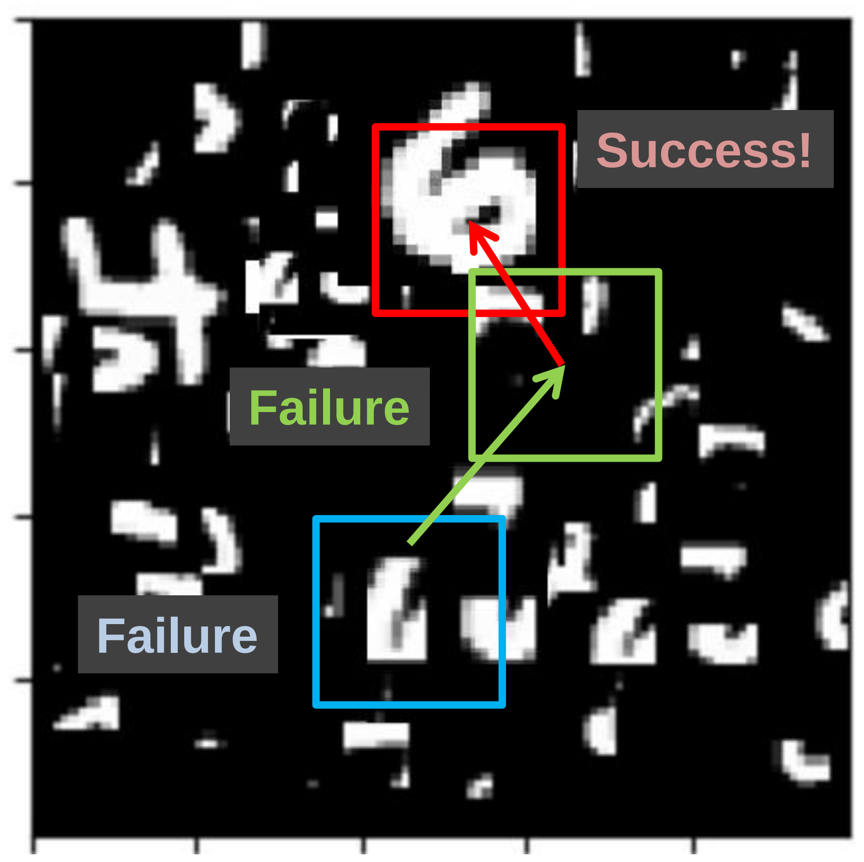
Step 1. Compare the features between the chosen window and the query.

Step 2. Given the features and the previous position of the window, the LSTM determines the next position.

Step 3. If the number of trials reaches the limit, it terminates and outputs the final window. Otherwise back to Step 1.

Learning Strategies

Combination of **reward maximization** and **feature loss minimization**



Reward maximization

- ➢ Get reward 1 if the window finds the query, otherwise 0
- ➢ Maximize the expected reward by policy gradient

Feature loss minimization

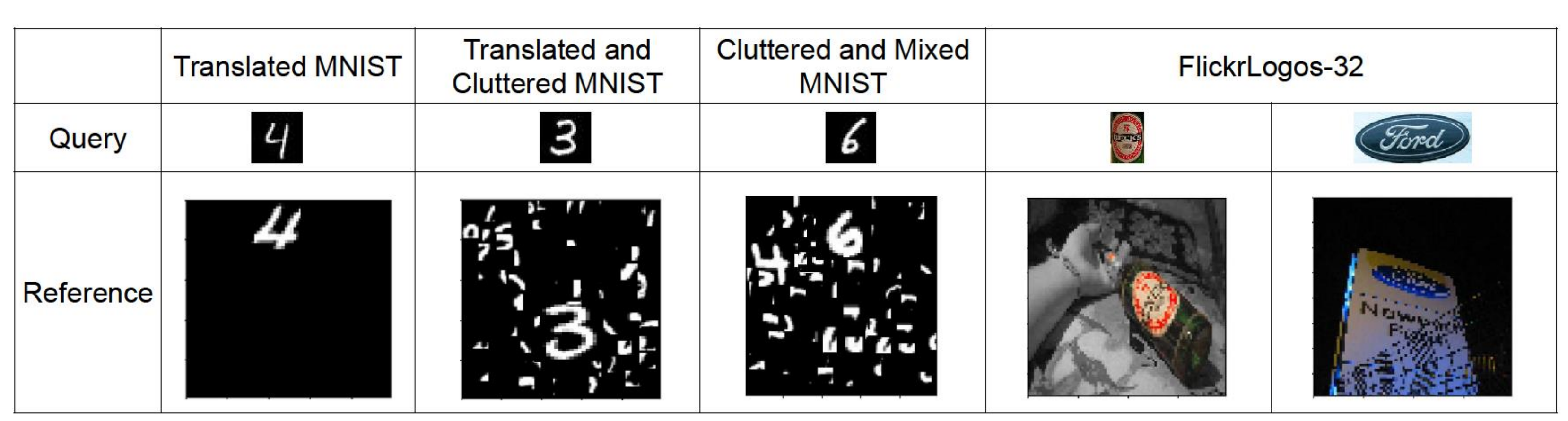
- ➢ If "Success", close the features between the window and the query, otherwise farther
- ➢ Contrastive Loss:

$$L = \begin{cases} d(q,g) & \text{If "Success"} \\ \max\{0, m - d(q,g)\} & \text{otherwise} \end{cases}$$

$d(q,g)$: Euclidian distance between query q reference g

Dataset

Noisy **MNIST** (Translated, Cluttered, and Mixed) and **FlickrLogos-32**



- ➢ Query-reference pair is generated by selecting the query to the same digit/logo as the reference from a set of centered clean digits/logos.

Results

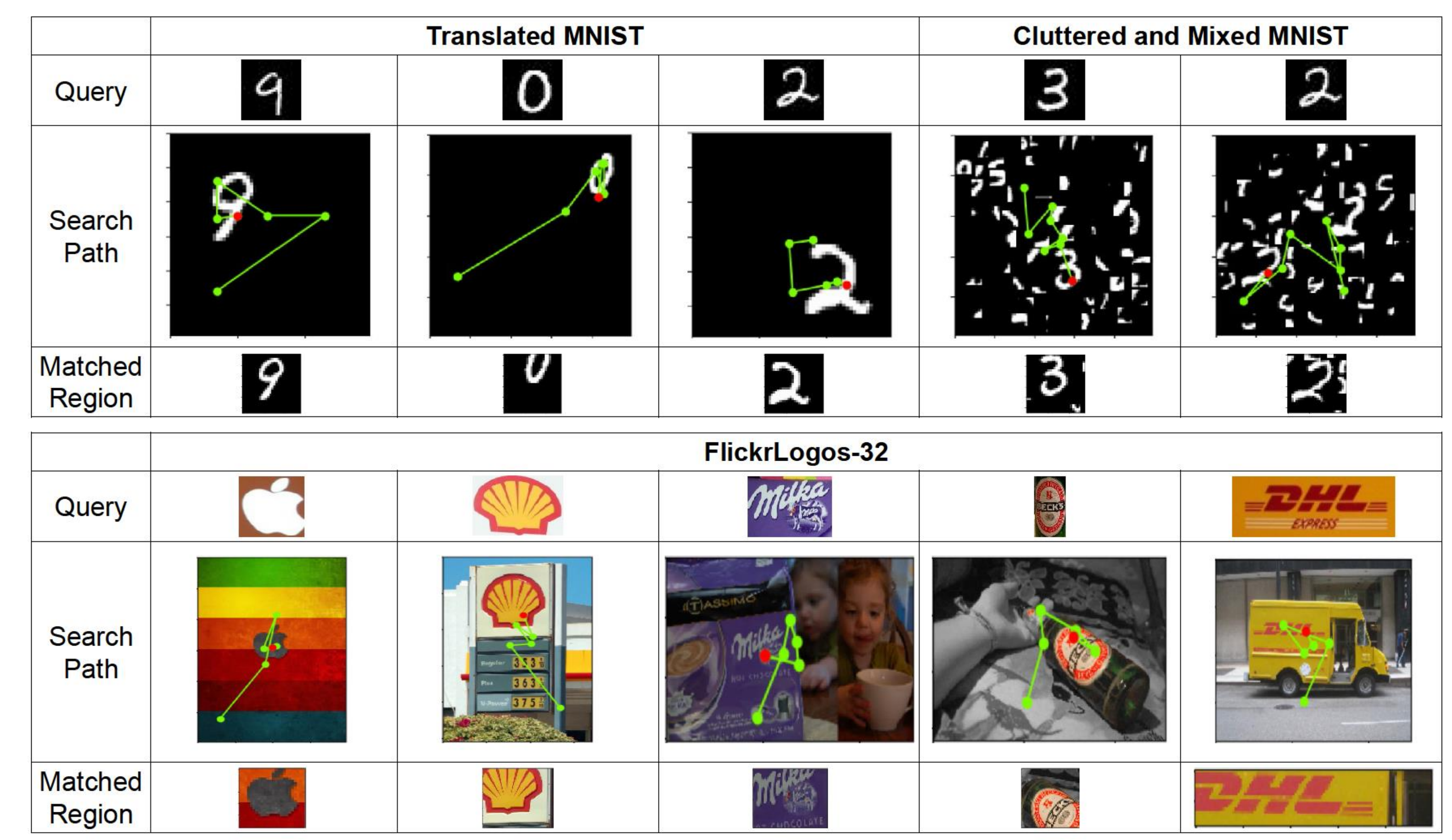
Quantitative results. Matching is successful if the intersection over union (IoU) between the predicted and the ground-truth windows is greater than 0.5,

Success rate (run time in milliseconds)

Dataset	Translated	Cluttered	Mixed	FlickrLogos-32
Ours	**0.95 (1)**	**0.91 (3)**	**0.88 (4)**	**0.39 (6)**
[Yacov+, ICCV11]	0.68 (2)	0.20 (3)	0.15 (5)	0.28 (230)
[Tali+, CVPR15]	0.70 (132)	0.11 (141)	0.08 (148)	0.36 (2390)

Ours can localize the query by processing only a few windows; [Yacov+, ICCV11] takes 230 ms while ours needs only 6 ms!

Qualitative results. For a given query-reference pair, the example shows the search path traced by our model in order to localize the query.



Our method can successfully find the target within only six trials even if they are heavily different in their colors and poses!

Conclusion

We proposed a reinforcement learning approach for image matching that sequentially outputs the next location towards the target region in each iteration.

Key feature:

- Fast: Number of candidate windows processed to localize the query is far smaller than existing methods.
- Robust: Our model is able to localize the query even in severely cluttered reference images.

References:

[1] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu, "Recurrent models of visual attention," in NIPS, 2014.

[2] Artsiom Ablavatski, Shijian Lu, and Jianfei Cai, "Enriched deep recurrent visual attention model for multiple object recognition," in Proc. WACV, 2017.