Single Object Tracking using the Siamese Family of Trackers

Object Tracking is an inherently challenging computer vision task. The tracker is given an initial "seed" or bounding box at the beginning of the video sequence and then asked to associate and localize the target objects in consecutive video frames with a certain level of accuracy.

Advances in deep learning, especially in the realm of Convolutional Neural Networks (CNN), have created exponential progress in the domain of image classification and subsequently object detection. This rapid progress on the Object Detection front, particularly data driven approaches and algorithmic thinking for understanding localization, is leading to slower but still crucial progress for Object Tracking.

There are two major industry initiatives (called challenges) that are pushing forward research and benchmarks for accurate Object Tracking:

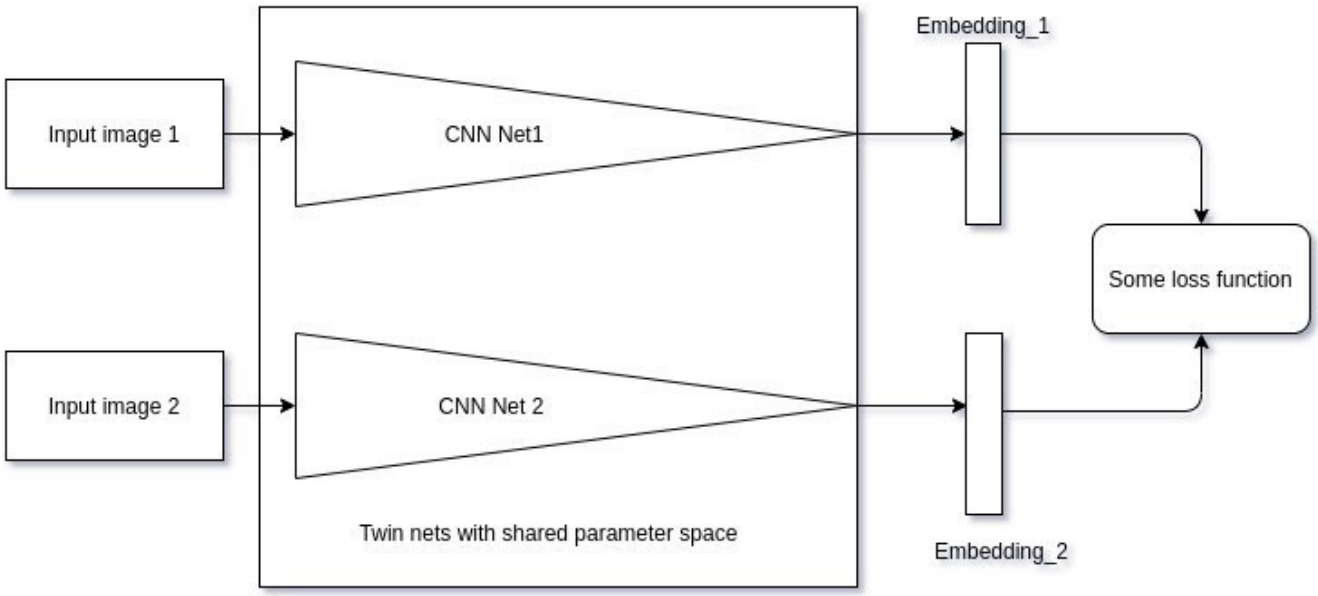
1. Single Object Tracking — comprised of short and long term tracking challenges and hosted yearly by VOT challenge.
2. Multi Object Tracking — hosted yearly by MOT challenge.

In this article we focus predominantly on Single Object tracking, specifically the SiamFC tracker, the first in line under the Siamese family of trackers. Future articles will do a deep dive on its successors, SiamRPN and SiamRPN++.

Siamese Networks

First, we need to explore what a Siamese network is.

A Siamese Neural Network is a class of neural network architectures that contain two neural identical subnetworks running in tandem. These parallel subnetworks share the same weights and parameter space. Siamese Networks get their name from the phenomenon of co-joined or "Siamese" twins and generally have a Y-shaped neural architecture indicating a comparative approach.



In all Siamese architectures, we have two input vectors we want to compare, so we pass both of them through the same subnetwork configuration to obtain a multi-dimensional embedding representation. These embeddings are then trained on a certain loss function (like L2 loss or triple loss) to measure the semantic similarity between them.

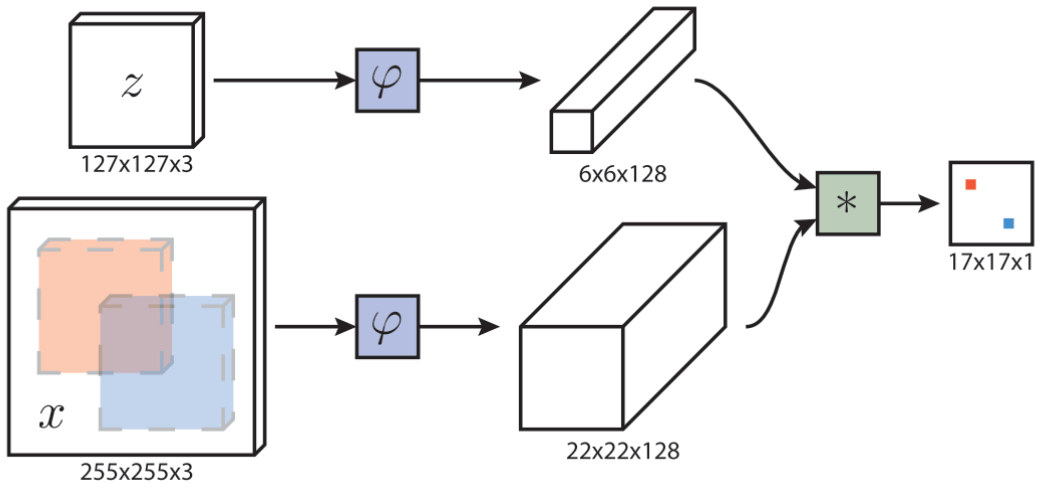
The SiamFC Architecture

Traditionally, the object tracking scene has been dominated by kernel based tracking (like KCF, mean-shift, etc.) and contour based tracking like **C**onditional **D**ensity **P**ropogation (the Condensation algorithm).

These algorithms deterministically learned the features online but were too data deprived to do any offline learning (most of them were proposed in the pre-deep learning era). Without enough data to learn from, these algorithms were not able to localize objects accurately, especially in instances of occlusion, change of camera angle, and illumination. It was hard to predict the state space of the objects of interest even when combined with filtering approaches like Kalman and Particle filters.

The beauty of the Siamese-styled tracking approach is that it leverages the juxtaposition of two results in order to localize more accurately a Region-of-Interest (ROI).

Given an object and its location in the current frame, find the location of the same object in the next frame.



SiamFC architecture

SiamFC uses two identical CNN's to address an offline similarity learning problem and then this function is then evaluated online during an inference phase. As shown in the above figure, we are essentially training the network to locate an exemplar image, denoted by z, within the larger search image x. In this example, the red and blue pixels in the score map contain the similarities for the corresponding sub-windows. The FC in SiamFC stands for Fully Convolutional architecture with respect to the search image, so there is no restriction on the size of the test image.

Mathematically, the central idea here is to learn a function $f(z, x)$ that compares the exemplar image z with the candidate image x and outputs a high scalar valued score map if they are similar and a low score if they are not similar. We are essentially building a class-agnostic similarity scoring function between 2 image patches. $f(z, x)$ can be viewed as a composite function $g(\varphi(z), \varphi(x))$ where φ is the identical mapping function that takes the variable sized input vector and creates an embedding of both the exemplar and candidate vectors. g in this case can be considered as a similarity metric or just a distance metric.

Training and Testing

The ImageNet 2015 Video dataset was used to train the SiamFC model. This dataset contained more than 30 different classes for animals and vehicles, with more than 4500 videos and more than 1 million annotated images. During training the exemplar image size is fixed to 127 x 127 while the candidate size is 255 x 255 pixels. Positive and negative training pair images are carefully curated (i.e they are extracted from the same video and are usually at most $\mathbb{T}$ frames apart) before training, and a logistic loss is used as the cost function.

$$(y, v) = log(1 + exp(-yv))$$

Here v is the score of the exemplar-candidate and y is the ground truth either +1 or -1. Stochastic Gradient Descent optimizer is applied on the loss function to find the best parameters for the model.

Conclusion

The **SiamFC** tracker, released in 2016, and its improved version, the **SiamFC conv5 baseline**, won the VOT-17 real-time challenge and was the state-of-the-art solution in 2016–2017. However, there were still some fundamental flaws in its design which kept it from being a really stellar object tracking method.

After the initial release, there were rapid changes and improvements to this tracker. We will explore the details of these changes and improvements in future posts on SiamRPN.

Sign up for 150 hours of free annotation time in Alegion Control to see this Siamese family of trackers at work.