






Object Tracking Through Residual and Dense LSTMs

Fabio Garcea^(✉) , Alessandro Cucco, Lia Morra^{}, and Fabrizio Lamberti^{}

Dipartimento di Automatica e Informatica, Politecnico di Torino, Turin, Italy
{fabio.garcea, lia.morra, fabrizio.lamberti}@polito.it,
alessandro.cucco@studenti.polito.it

Abstract. Visual object tracking task is constantly gaining importance in several fields of application as traffic monitoring, robotics, and surveillance, to name a few. Dealing with changes in the appearance of the tracked object is paramount to achieve high tracking accuracy, and is usually achieved by continually learning features. Recently, deep learning-based trackers based on LSTMs (Long Short-Term Memory) recurrent neural networks have emerged as a powerful alternative, bypassing the need to retrain the feature extraction in an online fashion. Inspired by the success of residual and dense networks in image recognition, we propose here to enhance the capabilities of hybrid trackers using residual and/or dense LSTMs. By introducing skip connections, it is possible to increase the depth of the architecture while ensuring a fast convergence. Experimental results on the Re³ tracker show that DenseLSTMs outperform Residual and regular LSTM, and offer a higher resilience to nuisances such as occlusions and out-of-view objects. Our case study supports the adoption of residual-based RNNs for enhancing the robustness of other trackers.

Keywords: Object tracking · Recurrent neural networks · Residual networks

1 Introduction

Visual object tracking plays a fundamental role in many applications including, e.g., robotics and video-surveillance. In this paper, we specifically focus on the problem of *generic* object tracking, which can be concisely phrased as follows: “given a bounding box enclosing an arbitrary object at time t , produce bounding boxes for that object in all future frames” [1].

Current generic 2D image tracking systems predominantly rely on training a tracker online according to the *tracking-by-detection* paradigm: an object-specific detector is continuously updated with the new object’s aspect at every frame, to cope with changes in shape and appearance, as well as with occlusions, while the object moves. Compared with trackers trained completely offline, this approach is more robust and flexible, but these advantages are paid with a decrease in

the frame rate that can be achieved. In recent years, hybrid trackers that combine convolutional neural networks (CNNs) for visual feature extraction with Long Short-Term Memory (LSTM) recurrent neural networks have been widely adopted. An example is represented by the Re³ tracker [1]: the CNN is trained completely offline, thus reducing the computational load at inference time, and the LSTM is trained to update and store an object-specific model. This method has shown increased accuracy and robustness against comparable trackers, especially during occlusions, but is still sensitive to changes in the object’s appearance due to occlusions or partially out-of-view targets, to proximity with similar objects, as well as to the presence of background clutter.

A possible way to improve the performance of LSTM-based trackers is to increase the complexity of the recurrent module, e.g., by stacking several LSTMs. This approach, however, can make it harder for the training procedure to converge, due to the increased network depth. Inspired by the success of Residual Networks [2] and Dense Networks [3] in image recognition, few works in literature have explored the use of residual connections in LSTMs, mostly for speech and text analysis [4–8].

In principle, using deeper and more complex LSTM modules should improve the capability of the tracker to model long-term change sequences. Our contribution is thus the design and experimental validation of Dense and Residual LSTM modules for visual object tracking. To assess, *ceteris paribus*, the added benefit of residual connections in object tracking, we modified the established architecture of the Re³ tracker [1]. Our experimental evaluation on the OTB50 and OTB100 benchmarks shows that Dense LSTM modules achieve higher robustness to occlusion and out-of-view targets while maintaining a similar parameter count compared to solutions adopting plain, non-residual layouts.

The rest of the paper is organized as follows. In Sect. 2, related work related to object tracking and residual networks is presented. Afterwards, in Sect. 3, we examine the tracker selected as the baseline and propose two different variations of the original layout involving residual connections in the recurrent module. In Sect. 4, we present the performance obtained by the proposed architectures on different benchmarks and compare them with state-of-the-art trackers. Finally, in Sect. 5, we discuss the main findings of our experiments and give some directions for future works.

2 Related Work

2.1 Object Tracking

Modern trackers can be roughly divided in *offline-trained*, *online-trained*, and *hybrid* [1, 9]. Online trackers operate online, continually learning features to update the object’s appearance during tracking; trackers adopting the well-known tracking-by-detection paradigm belong to this category. This type of tracker must carefully balance adaptation with real-time response abilities.

Recent works have exploited the capabilities of deep neural networks (DNNs) to learn from massive amounts of data by training CNN-based trackers completely offline [10]. These solutions rely on pre-trained CNNs for feature extraction and can operate at faster than real-time speed, but are intrinsically limited in coping with changes in objects' appearance due to movements, occlusions, blurring, etc.

Hybrid solutions like MDNet [11] and Re³ [1] represent an attempt to merge best qualities from both offline- and online-trained solutions. In the Re³ architecture, a CNN is trained offline to perform feature extraction, coupled with an LSTM module that keeps track of the object history over time. A multi-resolution approach is used by combining high-level features derived from the full CNN to low-level features learned by the previous layers, thus increasing the robustness of the feature extraction. This architecture represents a good trade-off between fast, real-time tracking (it achieves speeds of 150 frames per second) and robustness against nuisances such as occlusions.

2.2 Residual Networks

Residual networks and, in more recent times, densely connected networks have shown superior accuracy and training properties than traditional sequential CNNs, and have consistently achieved state-of-the-art results in image classification and other visual tasks [12, 13]. Densely connected CNNs, or DenseNets, represent an extension of the concept of skip connections: the output from each layer is passed as input to all subsequent layers and, as a consequence of the greater flexibility, have proven more effective than ResNets on a variety of visual tasks. A question that naturally arises is whether residual connections can prove as beneficial also for LSTM networks and, by extension, if the performance of hybrid trackers can be improved as well.

The idea of stacking LSTMs in a residual fashion has already been adopted in other fields of study such as distant speech recognition [5], sentiment intensity prediction [8], and object tracking [4]. Dense LSTMs stacking has been recently explored for sentence classification [6] and speech enhancement [7]. In [4], a rule-based residual RLSTM has been applied to tracking, achieving good results compared to other state-of-the-art trackers. However, given the complexity of object tracking networks and considering the role played by the feature extraction part (based on convolutional layers), by the training algorithm and by the training set, we believe that only by conducting controlled experiments the impact of residual connections can be fully appreciated. To the best of our knowledge, the role of Dense LSTMs in the context of tracking has not been investigated yet.

3 ResidualRe³ and DenseRe³

For our experiments, we selected the Re³ tracker as baseline architecture, and propose two alternative LSTM modules: a ResidualLSTM block consisting of two

cascaded LSTMs, and a DenseLSTM block in which four sequential LSTMs are densely connected by applying the same intuition used in DenseNets. The blocks, as well as their position in the overall architecture, are illustrated in Fig. 1. The number and size of layers were carefully chosen to keep the parameter count and the combined depth as similar as possible to the original Re³ architecture. In the following sub-sections, the main characteristics of the two solutions will be illustrated.

3.1 Re³

The Re³ tracker was firstly proposed in [1]. It represents a hybrid solution to the problem of generic object tracking. The layout of this network can be mainly split into three modules solving different tasks. The first module is a stack of convolutional layers used to extract the embeddings from the object being tracked; a concatenation layer is fed with both low-level and high-level information to obtain a more complete representation of the object. In the second module, a recurrent block consisting of a stack of two LSTM layers, each one receiving the features extracted at the previous stage, can keep track of subsequent object's positions and transformations. Finally, a regression layer is used to predict the bounding box of the object in the current frame. The full model is fed with two frame crops from the sequence at each time step; one of them is centered at the object's position in the previous frame, whereas the other is centered at the same position but in the current frame. Both the crops are still large enough to carry some information about the background.

3.2 ResidualLSTM-Based RNN

In the ResidualRe³ version of the tracker, a different architecture has been adopted for the recurrent module. In particular, a sequence of two LSTMs connected in series has been added to the input of the first LSTM module in a residual block fashion. We will refer to this structure as the ResidualLSTM block. The full layout of the recurrent module consists of a stack of three ResidualLSTM blocks. Since the outputs from both the convolutional module and the LSTMs are summed through a merge layer, they need to share the same number of units. In the original version of the tracker, the CNN output is set to use 1024 units, but we decided to downscale it to 768 units to keep the parameter count of the complete network comparable to that in the original version of the tracker. Moreover, a batch normalization layer has been added after the fully connected layer of the CNN to make its output comparable to that of the first ResidualLSTM block.

3.3 DenseLSTM-Based RNN

In the DenseLSTM version of the tracker, we decided to replace the recurrent module with a different structure exploiting dense residual connections; a stack of

four LSTMs has been densely connected through skip connections, thus allowing each subsequent module to be fed from the output of the previous ones. We will refer to this structure as the DenseLSTM block. In this case, the fully connected layer on top of the CNN has been set to use 900 units instead of the original 1024 units to keep a low parameter count for the following recurrent module; moreover, a batch normalization layer has been added on top of this layer to speed up model convergence. In the full network layout, we used a single DenseLSTM block composed of four LSTMs with 512 units each. With these constraints, we were able to maintain a complexity similar to that of the original Re^3 tracker.

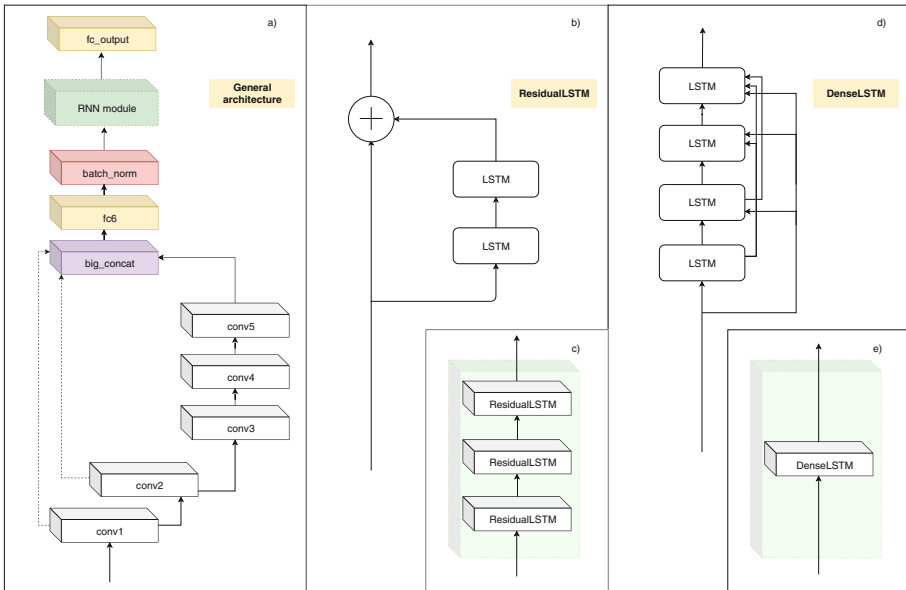


Fig. 1. A visual comparison of the proposed LSTM-based blocks. a) General structure of the Re^3 tracker; the main difference between Re^3 , ResidualRe^3 and DenseRe^3 is the RNN module (in green). b) Basic structure of the ResidualLSTM tracker, consisting of a series of two LSTMs; in the ResidualRe^3 alternative shown in c), a stack of three ResidualLSTM blocks has been deployed as the recurrent module. d) Basic structure of the DenseRe^3 tracker where four LSTMs have been densely connected through skip connections; resulting structure used as recurrent module is reported in e). (Color figure online)

3.4 Training and Implementation Details

The training procedure is the same as in the original Re^3 paper [1]. We here summarize the most important steps. Before starting the training, synthetic data are produced with several augmentation techniques such as horizontal flipping

and random noise generation, and weights from AlexNet are loaded in the CNN; LSTM states are initialized to zero, whereas other weights are set using MSRA initialization. The adopted optimizer is Adam with momentum and weight decay set to default values, and a learning rate decreasing from 10^{-5} to 10^{-6} after 10.000 iterations. Finally, the loss function is the Mean Absolute Error (MAE), and the number of iterations is 200.000. The training is initialized with 64 batch size, 2 unrolls and 1 probability of using the ground truth bounding boxes as a reference to crop the frame at the following time step; as soon as the loss plateaus, the batch size is halved and the unrolls are doubled (up to 32 unrolls). Moreover, the probability (initially 0) of mixing the predicted bounding boxes with the ground truth is increased using steps of 0.25; in this way, the network can learn from its errors during training thus being able to partly recover from errors at test time.

Since reproducibility of deep learning models is notoriously difficult to achieve, being the training procedure inherently random and affected by several factors including the training environment [14, 15], the original model has been retrained following the steps in the original publication on the ILSVRC2014 DET and ILSVRC2017 VID datasets, starting from the original code provided by the authors.

The training procedure for the modified networks followed the same used for training the Re^3 original version with some minor changes. For the Residual Re^3 tracker, a faster learning rate of 10^{-4} was initially set, then reduced to 10^{-5} and 10^{-6} when noticing that the loss function starts to plateau; moreover, a faster learning rate scaling of 10^{-1} was adopted for the finetuning of the CNN module weights. For the Dense Re^3 network, the procedure was similar, but we started to increase the probability of using the network prediction only after 32 unrolls and each time the loss function showed a plateau.

All the experiments were performed on a system configured with an i7 2600 CPU, 8 GB DDR3 1333 MHz RAM and an NVIDIA GTX 1060 3 GB GPU.

4 Experimental Results

First, we report the results of training and testing all the architectures on the ILSVRC2014 DET and ILSVRC2017 VID datasets, considering also the re-trained Re^3 model provided by [1]. Secondly, we compare results obtained by the original published Re^3 model on the challenging OTB50 and OTB100 benchmarks [16] along with several state-of-the-art trackers. The OTB100 benchmark consists of 100 different image sequences reporting objects from different classes and assignable to different attributes (occlusion, motion blur, out-of-view, etc.).

4.1 Training Residual Re^3 and Dense Re^3

The results for the architectures under test are reported in terms of two different metrics, namely, the number of targets lost by the tracker, and the Mean Intersection Over Union (IOU) between the predicted and the ground truth

bounding boxes. The parameters count is reported as well, to highlight how the new architectures remain comparable to the model from the original paper.

Our results show an improvement in the Mean IOU score and a lower number of lost targets compared to the retrained version of Re^3 , while keeping a comparable parameter count (Table 1). The evaluation of training results for DenseRe^3 showed significant improvements compared to the Retrained version of Re^3 and ResidualRe^3 (Table 1), thus demonstrating the advantages brought by the DenseLSTM blocks.

Table 1. Training results for different Re^3 architectures. It should be noticed how residual and dense LSTM improve performance with minimal increase in parameter count.

Tracker	Lost targets	Mean IOU	Parameter count
Re^3 (retrained)	350	0.64	85.699.686
ResidualRe^3	303	0.66	87.716.712
DenseRe^3	258	0.68	87.031.408

Concerning the original Re^3 architecture, it is worth noticing that we did not achieve the same performance of the model released by the authors even though, to the best of our knowledge, we followed the same training curriculum for Re^3 . Specifically, our retrained model achieves 350 lost targets with a Mean IOU of 0.64 versus the 243 lost targets and 0.72 Mean IOU for the weights provided by the authors.

This discrepancy is not entirely surprising, since reproducing results is a well-known issue of deep learning-related research, and maybe due to slight differences in implementation or training parameters. Recent research also highlighted the effect of random initialization on the estimated performance of image classification networks [14]; however, for complex deep learning architectures, such as object trackers, running multiple experiments per configuration requires substantial computational resources. In the future, we plan on exploring this issue in more detail. For the remaining experiments, we compare DenseRe^3 with the original model provided by the authors, which albeit less favourable allows an easier comparison with the previous literature.

4.2 Benchmarks Evaluation

Since we needed a state-of-the-art reference tracker to compare our results, we opted for the Recurrent Filter Learning tracker (RFL) [17]; besides its high performances, this model is based on a recurrent module thus representing an appropriate reference architecture for our experiments.

We report here the results of the One Pass Evaluation (OPE) protocol on the OTB50 and OTB100 benchmarks; while the results on OTB100 benchmark were computed at test time starting from the code as provided by the authors,

those for OTB50 were already provided for multiple trackers and thus have not been re-computed. Once the baseline was defined, we executed the OPE TB100 benchmark for the RFL tracker, the original version of Re³ and the DenseRe³ networks; the results (Fig. 2) show that our model can outperform the original version of Re³ in sequences characterized by low resolution, occlusions and out-of-view objects, while still performing similarly to the original architecture in other cases.

We then ran the OTB50 benchmark for the RFL, the original Re³, the DenseRe³ and other state-of-the-art trackers to evaluate the performances of our model with a larger pool of different architectures. The results (Fig. 3) are aligned with those obtained with OTB100; like with the other benchmark, the model performs particularly well with sequences characterized by the above attributes, reaching the top-4 positions for all attributes. In all the other sequences, it performs worse than Re³, even though it can reach the top-5 positions.

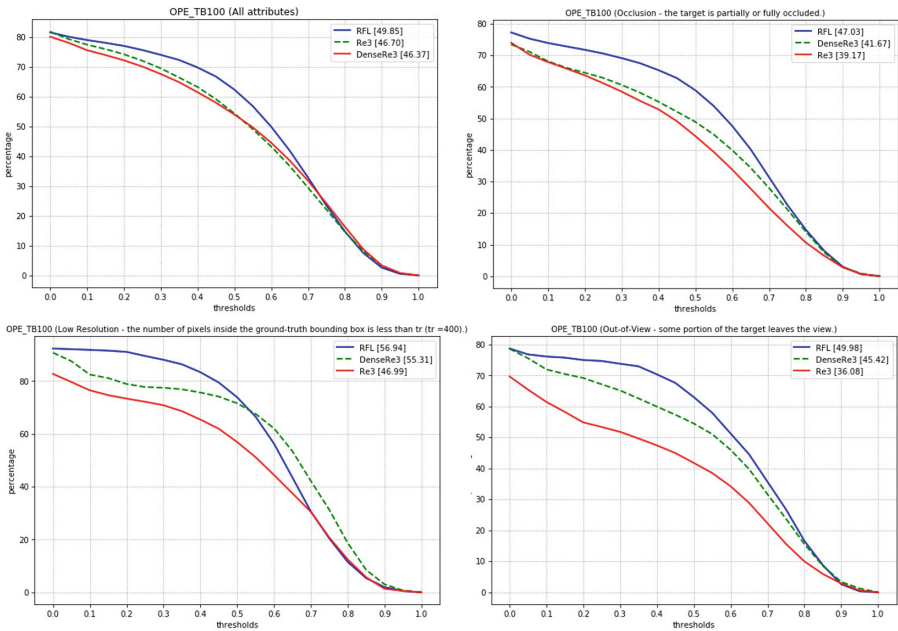


Fig. 2. Results on the OPE TB100 benchmark for RFL, the original version of Re³, and the proposed DenseRe³ architecture. The percentage of frames where the mean IoU is greater than a threshold (y axis) is plotted as a function of the threshold value (x axis). The success plots report the results for different sequence attributes (all attributes, occlusion, low resolution and out-of-view objects). Whilst in some cases the DenseRe³ model scores are similar to the z version, in other cases they show a better performance of our architecture.

Finally, two example sequences from the OTB100 benchmark are reported (see Fig. 4) to depict, in a visual fashion, the achieved improvement. The first

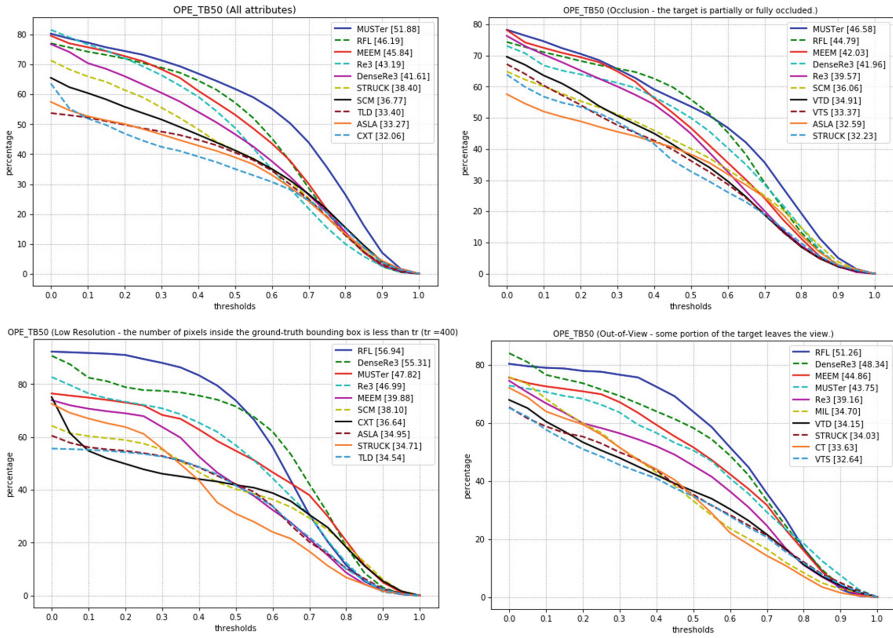


Fig. 3. Results on the OPE TB50 benchmark for the proposed DenseRe³ architecture and other state-of-the-art trackers. The percentage of frames where the mean IoU is greater than a threshold (y axis) is plotted as a function of the threshold value (x axis). The success plots reports the results for different sequence attributes (all attributes, occlusion, low resolution and out-of-view objects). In most of the subsets DenseRe³ scores are similar to those obtained by the original Re³ tracker. DenseRe³ achieves high performance in sequences with low resolution, occlusion and out-of-view target.

sequence, named “Matrix”, is characterized by multiple attributes like occlusion, fast motion and illumination variation. The second sequence, named “Ironman”, similarly presents multiple attributes as well, such as occlusion and out-of-view. Both the sequences have been annotated by Re³ and DenseRe³ with a red bounding-box representing the prediction of the tracker under test for each frame.

It’s evident how, in the sequence “Matrix”, Re³ (sub-sequence 1.a) loses the track of the object due to the fast motion of the body and the partial occlusion of the face features and consequently starts to track the hand of the second character. On the other side, DenseRe³ (sub-sequence 1.b) can keep track of the object also in presence of disturbances and it’s able to progressively recover from the error caused by the occluded frames. Moreover, a robust behavior can be appreciated in the fourth frame of the sequence where the model keeps track of the object in the presence of an important variation in the illumination.

In the second example, Re³ (sub-sequence 2.a) loses the track of the object when it goes temporary out-of-the-view and it’s not able to recover from its own

errors thus ending up losing track of the object. On the other hand, DenseRe³ (sub-sequence 2.b) shows again a robust behavior in case of occluded and out-of-view frames being able to keep track of the object even if with some difficulties due to the complexity of the frame.

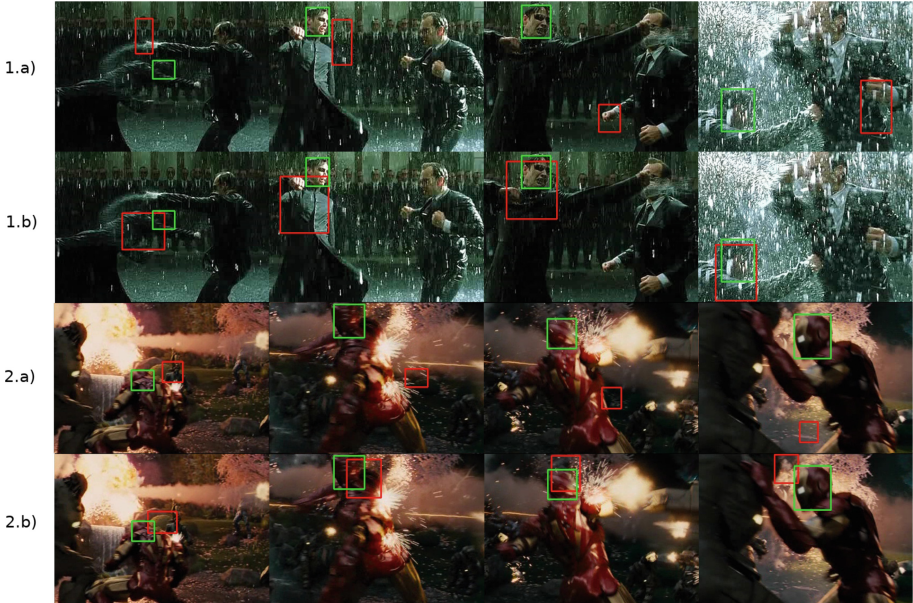


Fig. 4. Example sequences from the OPE TB100 benchmark evaluated on DenseRe³ and plain Re³. The first sequence, named “Matrix”, has been annotated with both Re³ (1.a) and DenseRe³ (1.b). Similarly the sequence named “Ironman” has been annotated by both Re³ (2.a) and DenseRe³ (2.b). The bounding boxes annotated by DenseRe³ intersect with the ground-truth (green bounding box) also in case of disturbances showing thus a robust behavior of the network if compared to plain Re³. (Color figure online)

5 Conclusions and Future Work

In this work, we explored the potential benefit of Residual and Dense LSTM in hybrid object tracking architectures. The idea of introducing residual and dense skip connections in LSTMs has been successfully explored in other applications, such as speech and text recognition.

We here investigate a case study in object tracking, in which we modified the architecture of an LSTM-based tracker, the Re³ architecture, using both ResidualLSTM and DenseLSTM modules. Our experiments showed that both ResidualLSTM and DenseLSTM modules can be successfully used to enhance

the robustness of the Re^3 tracker, as in low resolution or occlusion attributes, while keeping a parameter count comparable to the original version. In general the proposed architecture appears to be more robust to the presence of occlusions, low resolution and other disturbances. Residual and even more dense architecture allow to connect each layer not only with the previous layer, but also with previous ones. Skip connections are an essential component of deep convolutional neural networks allowing to increase the number of layers without incurring in vanishing gradients or other numerical instability. DenseRe^3 is characterized by four LSTM blocks, instead of the two blocks of the plain Re^3 tracker, thus effectively doubling the depth of the network. Nonetheless, the use of skip connections makes the information flow across the layers easier ensuring fast convergence, thus increasing performance in a comparable number of iterations. Previous works reported that increasing the number of layers in plain LSTM may lead to performance degradation, however, this phenomenon can be reduced or even reversed when residual connections are introduced [8]. We observed an even greater benefit from dense connections, but the relationship between performance and depth should be analyzed in a more systematic fashion.

We also hypothesize that in a densely connected structure, where the activations of each layer are fed to all subsequent ones, it is possible to more effectively “remember” the history of the object being tracked, thus improving the robustness in the presence of occlusions and background clutter, or when the object moves out-of-view.

We expect that similar improvements could be found on other architectures currently relying on plain LSTM modules. In the future, we plan to explore the advantages of ResidualLSTM and DenseLSTM blocks in other trackers or other visual tasks.

References

1. Gordon, D., Farhadi, A., Fox, D.: Re^3 : real-time recurrent regression networks for visual tracking of generic objects. *IEEE Robot. Autom. Lett.* **3**(2), 788–795 (2018)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
3. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks (2016)
4. Kim, H.I., Park, R.H.: Residual LSTM attention network for object tracking. *IEEE Signal Process. Lett.* **25**(7), 1029–1033 (2018)
5. Kim, J., El-Khamy, M., Lee, J.: Residual LSTM: design of a deep recurrent architecture for distant speech recognition. *arXiv preprint arXiv:1701.03360* (2017)
6. Ding, Z., Xia, R., Yu, J., Li, X., Yang, J.: Densely connected bidirectional LSTM with applications to sentence classification. In: Zhang, M., Ng, V., Zhao, D., Li, S., Zan, H. (eds.) *NLPCC 2018. LNCS (LNAI)*, vol. 11109, pp. 278–287. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99501-4_24
7. Gao, T., Du, J., Dai, L.R., Lee, C.H.: Densely connected progressive learning for LSTM-based speech enhancement. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5054–5058. IEEE (2018)

8. Wang, J., Peng, B., Zhang, X.: Using a stacked residual LSTM model for sentiment intensity prediction. *Neurocomputing* **322**, 93–101 (2018)
9. Ali, A., et al.: Visual object tracking—classical and contemporary approaches. *Front. Comput. Sci.* **10**(1), 167–188 (2016)
10. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S.: Fully-convolutional siamese networks for object tracking. In: Hua, G., Jégou, H. (eds.) *ECCV 2016*. LNCS, vol. 9914, pp. 850–865. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_56
11. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking (2015)
12. He, K., Sun, J.: Convolutional neural networks at constrained time cost. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5353–5360 (2015)
13. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016*. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
14. Bouthillier, X., Laurent, C., Vincent, P.: Unreproducible research is reproducible. In: *International Conference on Machine Learning*, pp. 725–734 (2019)
15. Marrone, S., Olivieri, S., Piantadosi, G., Sansone, C.: Reproducibility of deep CNN for biomedical image processing across frameworks and architectures. In: *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE (2019)
16. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: a benchmark. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2013)
17. Yang, T., Chan, A.B.: Recurrent filter learning for visual tracking. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 2010–2019 (2017)