Deep Vision with Recurrent Neural Networks

Davit Buniatyan

Research Proposal for PhD

Contents

1. Introduction

Deep learning techniques such as Convolutional Neural Networks (CNN) achieved state-of-the-art results in Computer Vision and formulated branch in it Deep Vision in order to solve problems such as image classification, object detection and tracking, and semantic segmentation. However, these techniques often lack accuracy in low-level computer vision problems such as depth reconstruction that requires sequence-to-sequence pixel level labeling. RNNs are primarily used in modeling one-dimensional sequential inputs, however with the introduction of multi-dimensional RNNs and grid structure, which is a recent improvement, RNNs show state-of-the-art results in low-level Computer Vision tasks such as segmentation. The evaluation of recurrent architectures on spatiotemporal data is problematic because of limited training data.

The aim of this research is to evaluate the capabilities of Multidimensional Recurrent Neural Networks in depth reconstruction and propose modernized architecture that will take the advantage of both CNNs and RNNs and achieve real-time processing on spatiotemporal data, for instance, on image sequences (Video) or 3D environments (Virtual Reality). We will compare the model with widely used techniques, which have reached top-level results. Furthermore, in order to train and evaluate the algorithm due to the lack of efficient data we will generate synthetic realistic and non-realistic visual data using computer generated graphics and estimate the generalization feature of the model on real-life input. The goal of the research is to develop architecture that could be applied in robotics, for example, personal assistants or self-driving cars and enhance the performance of real-time computer vision by capturing spatiotemporal context.

2. Background

There has been extensive research in Computer Vision for generating depth map from single or more images. A. Saxena introduced a state-of-the-art technique Make3D, which uses Markov Random Field to infer a set of "plane parameters" that capture both the 3D location and 3D orientation of a patch [1,2]. B. Liu attempted using different approach by considering the semantic labels. [3] There have been several manual or interactive solutions; however, an automatic depth extraction from videos and single images has been proposed that achieved a significant result [4].

Deep learning techniques, specifically Convolutional Neural Networks pioneered by Yan LeCun in 1989 [5], only recently reached top results almost in every problem in computer vision by taking the advantage of modern GPUs and parallel processing. As a result, a multi-scale deep neural network has been proposed to predict a depth map using a single image [6]. A CNN has been trained to extract depth from stereo images with the matching error 2.61% on KITTI dataset [7]. Recently, Google researchers published paper called DeepStereo, which is based on dual convolutional neural network architecture that also reconstructs the depth map for directly generating new views from imagery [8]. Using unsupervised learning network architecture called auto-encoders, depth and motion extraction from a set of images either varied in time, in location (different cameras) or both was successfully trained [9].

The architecture of Recurrent Neural Networks (RNNs) introduces temporal behavior on non-restricted sequential input compared to other neural networks and are widely used in Natural Language Processing (NLP). More abstract vision problems such as Image/Video Captioning have been achieved with the sequential combination of CNN and RNN, where the former extracts specific features of visual data and the latter express those features in human readable language [10].

Fully connected RNNs suffer from the vanishing gradient problem and it makes their application on long sequences such as images (per pixel input) useless. To solve this problem Hochreiter and Schmidhuber introduced Long-Short-Term Memory (LSTM) cells that allow to preserve long-term connection on top of recurrent structure [11]. They outperformed other methods on problems such as handwriting recognition. As RNNs get sequential input, the

applications on low-level computer vision tasks are inappropriate. A. Grave proposed Multidimensional RNN that allows the network to learn spatiotemporal context by getting input from every dimension at each time step [12]. For example, at each pixel in the image (2D) the hidden layer will get a memory from (x-1) and (y-1) pixels. Also another limitation of RNN is that it only keeps past memories at each input. Bi-directional training (multi-directional in case of more than one dimension) was proposed to have past and future context while processing given input. Scene labeling using 2D LSTM outperformed other approaches not only in performance but also in efficiency [13]. Grave's multidimensional multidirectional recurrent neural network has been extended to grid-like LSTM that achieved near state-of-the-art result in digit labeling based on MNIST dataset [14]. Successful integration of CNN and Conditional Random Field, as a recurrent layer, achieved top results in image segmentation problem on Pascal VOC 2012 segmentation benchmark [15].

Taking into consideration the previous research and the correlation between segment labeling, motion and depth reconstruction, further advancement in this problem using combination of convolutional-recurrent neural network is considered to be potential.

3. Research Objectives and Approach

The aim of this research is to architect end-to-end novel Recurrent Neural Network for reconstructing the depth from sequence of images based on spatiotemporal context and achieve state-of-the-art results and real time computational complexity to provide industry ready solution. Furthermore, introduced architecture could be applied in other Computer Vision problems and introduce new approach in Deep Vision. This method is going to be unconstrained by the type of input.

3.1 Objectives

The following objectives are going to be addressed.

1. How well the proposed architecture is able to extract long-term spatiotemporal context?

   We are going to estimate empirically the importance of spatiotemporal context in pixel level depth estimation and understand the significance of short (neighborhood pixels) and long (segment/object) term context.

2. To what extent do Multidimensional RNNs, trained on synthetic data, generalize on real-life input?

   As we are going to heavily base our training on synthesized data and validate the performance on real dataset, it is of vital importance to estimate the effectiveness of training recurrent neural networks on simulated environments and how well they generalize in real-life situations.

3. Complexity analyses of visual RNNs for parallel processing (Graphics Processing Unit, CUDA) on real-time applications.

   The significance of the research is to design an algorithm that could be applied in real life problems. Our objective is to keep the constraint on computational complexity as real-time as possible in order providing high applicability.

3.2 Theoretical Framework

In contrast with Long-term Recurrent Convolutional Network (LTRC) that was proposed for image captioning [16], our model is not going to take a single frame as a time step input. LTRC model extracts features of a single frame and passes to the LSTM layer for text generation. The following model is reasonable for high-level problems such as image captioning. A. Karpathy also used this approach for an image captioning system trained on ImageNet [10]. Similar to CRF-RNN model [15], our system is going to take pixel input; however, it will take into consideration not only neighbor pixels but also temporal neighborhood, thus achieving spatiotemporal pixel-level context while processing the depth of the pixel.

From theoretical point of view, one could notice a correlation between Integral Images and Recurrent Neural Networks. Integral Images have been introduced for efficient evaluation of the sums of image values aligned over rectangular region. Furthermore, it has been successful in real-time face detection and other computer vision tasks including depth estimation [17, 18]. RNNs in computer vision might have profound explanation of how spatiotemporal context affects efficiency and effectiveness of the model. For example, in the scene-labeling task, 2D LSTM achieved a top result with running time on a single Central Processing Unit (CPU) compared to other counterparts trained on Graphical Processing Unit (GPU) [13]. Thus thorough examination of the correlation will provide significant background for RNNs' further potential in low-level and high-level computer vision problems.

3.3 Applications

Primary applications of the solution are in the following fields,

- Robotics

  Even though RGBD cameras are expanding, they are still not cheap enough for embedding them into robots. Depth reconstruction could be used as an alternative for environment reconstruction. They could be potentially used for enhancing gesture recognition and object detection in robots. Emotech, London-based startup, which develops personal assistant robot Olly, is interested in the proposed technology for enhancing computer vision abilities of their product [19].

- Visual FX

  In the movie production industry, software such as Boujou or Nuke have tools for cg artists to reconstruct scenes for post-processing, for instance, adding 3D characters into the scene or modifying objects. The limitation of the current software is that the primary usage is in outdoor static scenes. Real-time depth reconstruction will enhance artists' workflow not only in static but also in dynamic scenes.

- RGBD cameras

  Cameras such as Kinect lack accuracy in estimating the depth of far objects. The solution could be an additional online processing layer to improve the accuracy.

- Virtual Reality

  In order to enhance the virtual reality presence, one might propose 3D reconstruction of video content. The algorithm could be used to generate point cloud of movies.

3.4 Key Resources

Primary resources required for the research:

- Deep Learning Frameworks

  There is no up-to-date implementation of Multidimensional Recurrent Neural Networks using current deep learning frameworks and as a result one of them should be extended. As seen so far, Caffe is considered to be most applicable. On the other hand, several tweaks to Theano one-dimensional iterator make MDRNNs quickly implementable.

- Hardware

  As our implementation concentrates on real-time processing and high applicability, we are not looking for cloud clusters for training and processing; however, a workstation with dual Tesla K80 GPU will significantly decrease the time for experiments with different architectures due to parallel processing. In addition, the effectiveness of rendering realistic image sequences will increase.

- Data

  Kinect camera will be required for collecting ground truth real data. For synthesising realistic data, we shall use Autodesk 3Ds Max with Vray Renderer and for real-time graphics we shall use either Unity3D or Unreal Engine.

3.5 Methods

3.5.1 Data Collection

Given the need to synthesize data, it will be collected with the reference to generating Flying-Chairs in FlowNet paper (2015). During data collection, a testing set will be separated to avoid overfitting the model [20]. The method of data generation may vary such that it might be possible to conduct real-life data collection. As current datasets are limited for training multidimensional RNNs, I propose several approaches to collect:

- Scraping Stereo Videos from Youtube and reconstructing depth map via computing stereo matching costs with CNNs described in LeCun's paper (2015): The advantage of this method is that we have vast amount of available resources. However, given the error of 2.61% our results will be artificially lowered.
- Capturing ground truth depth and RGB from high graphics video games (such as GTA 5, Far Cry 3, Battlefield 4): The advantage of this method is that we will have a vast amount of ready-to-use simulated environments for training that will supposably generalize, (for example, generalization of FlowNet from Sintel Dataset); however, capturing the depth map will need additional activity and time to spend on playing games.
- Rendering custom 3D image sequences using either game engine or Visual FX tools. This might take time and computational resources for generating high quality dataset; however, depth variations of the same images will be feasible to generate in order to enhance training significantly.
- Real ground truth data collection with Kinect camera. Time consuming and will not be accurate in outdoor scenes.

3.5.2 Model & Training

Variations of end-to-end Multidimensional Recurrent Neural Networks are going to be considered for experimentation including Convolutional Neural Network or Conditional Random Field integration. Furthermore, theoretical foundation of integral images might introduce novel architecture for pixel-level depth estimation. The model is going to be trained with back-propagation and with online stochastic gradient descent. Other optimization methods will be observed.

3.5.3 Benchmarking

The trained algorithm will be tested on KITTI and NYU Depth dataset to benchmark results with Depth MRF, Make3D, Feedback Cascades, Semantic Labels and DepthTransfer for single images. Furthermore, their implementations, if available, are going to be used for benchmarking the computation complexity of methods. The comparison of video depth reconstructed sequences will be held with DepthTransfer. To estimate synthetic-to-real-life generalization feature of the model, we are going to reconstruct real life input using 3D modeling software and train/test.

3.5.4 Scopes & Limitations

We are going to only concentrate on RNN Depth Reconstruction from spatiotemporal data, thus it might not work well for single images. Furthermore, extensive dynamic scenes, atmospheric effects and non-rigid animations will introduce additional layer of inaccuracy. In case the synthetic data generalization feature is low or the generated/collected data is not suitable, the model might not fit well for real input.

4. Research Plan

Academic Year 1	**Quarter**
Research goal refinements & Literature Review	Q4
Implementation of existing methods	Q1
Prototype I and initial benchmarking	Q2
Data Collection	Q3
Year 2	
Prototype II trained on our data	Q4
Theoretical foundation of RNNs in Deep Vision I	Q1
Experiments with various models	Q2
Initial draft of thesis	Q3
Year 3	
Theoretical foundation of RNNs in Deep Vision II	Q4
Further experimentation with profound models	Q1
Second draft & final draft	Q2
Final preparations & submission	Q3

Q4: Oct 1 - December 24, Q1: January 8 - March-31, Q2: April 1 - June 30, Q3: July 1 - September 31,

5. Current Work

My final year project in my UCL bachelor's degree under the supervision of Dr. Lourdes Agapito is depth reconstruction from image sequences using Recurrent Neural Networks. Currently working on implementation of A. Graves Multidimensional Recurrent Neural Networks using Theano and deep learning framework Keras. Planning to have initial network by the end this term. I am interested in continuing this research while acquiring PhD and will concentrate all my second term on achieving initial results.

6. Conclusion

Deep Vision achieved a new wave of state-of-the-art solutions for computer vision problems, mainly due to CNNs. However, they often lack accuracy in pixel-level labeling tasks. Proposed Recurrent Neural Network is going to consider spatiotemporal context of the pixel and estimate depth reconstruction with real-time processing in 'mind'. Based on the integral image concept and its usage, one might think of reversing CNN-RNN usual integration so that RNN produces integral image first and only then, on top, CNN will extract features. Andrej Karpathy called the unreasonable effectiveness of recurrent architectures magic, explained and visualized RNNs [21]. My goal during this research is to unveil this magic by providing profound explanation under deep vision context. Furthermore, using Depth Reconstruction as a specific use case, I shall introduce methods that could be applied to other computer vision problems by expanding the borders of Deep Vision and provide industry-ready solutions for robotics and other domains with real-time processing capabilities.

7. Bibliography

[1] Saxena, Ashutosh, Min Sun, and Andrew Y. Ng. "Make3d: Learning 3d scene structure from a single still image." Pattern Analysis and Machine Intelligence, IEEE Transactions on 31.5 (2009): 824-840.

[2] Saxena, Ashutosh, Sung H. Chung, and Andrew Y. Ng. "Learning depth from single monocular images." Advances in Neural Information Processing Systems. 2005.

[3] Liu, Beyang, Stephen Gould, and Daphne Koller. "Single image depth estimation from predicted semantic labels." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[4] Karsch, Kevin, Ce Liu, and Sing Bing Kang. "Depth Transfer: Depth Extraction from Video Using Non-Parametric Sampling." Pattern Analysis and Machine Intelligence, IEEE Transactions on 36.11 (2014): 2144-2158.

[5] LeCun, Yann, and Yoshua Bengio. "Convolutional networks for images, speech, and time series." The handbook of brain theory and neural networks 3361.10 (1995).

[6] Eigen, David, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network." Advances in Neural Information Processing Systems. 2014.

[7] Žbontar, Jure, and Yann LeCun. "Computing the stereo matching cost with a convolutional neural network." arXiv preprint arXiv:1409.4326 (2014).

[8] Flynn, John, et al. "DeepStereo: Learning to Predict New Views from the World's Imagery." arXiv preprint arXiv:1506.06825 (2015).

[9] Konda, Kishore, and Roland Memisevic. "Unsupervised learning of depth and motion." arXiv preprint arXiv:1312.3429 (2013).

[10] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." arXiv preprint arXiv:1412.2306 (2014).

[11] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[12] Graves, Alex. Supervised sequence labelling with recurrent neural networks. Vol. 385. Heidelberg: Springer, 2012.

[13] Byeon, Wonmin, et al. "Scene Labeling with LSTM Recurrent Neural Networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[14] Kalchbrenner, Nal, Ivo Danihelka, and Alex Graves. "Grid long short-term memory." arXiv preprint arXiv:1507.01526 (2015).

[15] Zheng, Shuai, et al. "Conditional random fields as recurrent neural networks." arXiv preprint arXiv:1502.03240 (2015).

[16] Donahue, Jeff, et al. "Long-term recurrent convolutional networks for visual recognition and description." arXiv preprint arXiv:1411.4389 (2014).

[17] Facciolo, Gabriele, Nicolas Limare, and Enric Meinhardt-Llopis. "Integral images for block matching." Image Processing On Line 4 (2014): 344-369.

[18] Viola, Paul, and Michael J. Jones. "Robust real-time face detection." International journal of computer vision 57.2 (2004): 137-154.

[19] Emotech LTD, UCL Startup, Emotech.co

[20] Sutskever, Ilya. Training recurrent neural networks. Diss. University of Toronto, 2013.

[21] Karpathy, Andrej, Justin Johnson, and Fei-Fei Li. "Visualizing and understanding recurrent networks." arXiv preprint arXiv:1506.02078 (2015