# Paper review of "Parsing Natural Scenes and Natural Language with Recursive Neural Networks"

**Alexandra Ioana Neagu**
TU Delft
neagu.alexandra0206@gmail.com

## Abstract

In my review of the (Socher et al., 2011) paper, I found it to be a significant contribution to the fields of computer vision and natural language processing. The authors propose a novel approach utilizing Recursive Neural Networks (RNNs) for scene understanding and parsing tasks, showcasing its effectiveness through rigorous experimental evaluations. Despite minor weaknesses, such as parsing accuracy compared to established parsers, the paper offers valuable insights and innovations. Overall, this work demonstrates the potential for advancing the state-of-the-art in both domains. My assessment, though confident, acknowledges the possibility of missed details or misunderstood points, ensuring a fair evaluation.

## 1 What is this paper about, and what contributions to NLP does it make?

This paper introduces recursive neural networks (RNNs) as a tool for predicting recursive structures in multiple modalities, primarily focusing on scene understanding in computer vision and parsing natural language sentences. The authors highlight the prevalence of recursive structures in both natural language syntax and scene images, emphasizing their importance in understanding and classifying these modalities.

The proposed RNN architecture operates by oversegmenting images into small regions representing parts of objects or background, extracting vision features from these regions, and mapping them into a semantic space using a neural network. Subsequently, the RNN computes scores for merging neighbouring regions into larger ones, generates new semantic feature representations for these merged regions, and assigns class labels to them. These merging decisions implicitly define a tree structure, where each node corresponds to a merged region and is associated with the RNN outputs.

Similarly, in natural language parsing, words are mapped into a semantic space and merged into phrases in a syntactically and semantically meaningful order, with the RNN computing the same outputs for each node in the parse tree.

The main contributions of the paper lie in several areas. Firstly, it presents one of the first deep learning methods to achieve state-of-the-art results in segmenting and annotating complex scenes, surpassing other methods based on conditional random fields or combinations thereof. Secondly, the learned features from the proposed RNN architecture outperform existing methods such as Gist descriptors for scene classification. Additionally, what I believe is a great contribution of the paper is how it demonstrates the generalizability of the algorithm by applying it to natural language parsing, achieving competitive performance on the Wall Street Journal dataset.

### 1.1 Paper summary

In summary, the paper contributes a novel RNN architecture for predicting hierarchical tree structures in scene images and parsing natural language sentences, showcasing its effectiveness and versatility across different modalities.

*The related work section* outlines the key research areas that inform and motivate the method proposed in the paper. It highlights the significance of scene understanding in computer vision, which involves tasks such as annotation, segmentation, and classification of objects within scenes. While some existing methods rely on global descriptors for scene classification, they often lack a deeper understanding of scene content. The paper introduces an algorithm that recursively merges image segments into super segments in a semantically and structurally coherent manner, outperforming other approaches by learning representations for super segments. Moreover, it discusses the importance of syntactic parsing in natural language processing and how the proposed RNN architecture simultaneously learns parsing and phrase representation in a

continuous vector space. Additionally, the section discusses the connection between NLP techniques and computer vision, emphasizing the generality of the proposed approach compared to previous methods. Finally, it contrasts the paper's approach with existing deep learning algorithms, highlighting its novelty in handling variable-sized inputs and capturing the recursive nature of natural language.

*Section 3* discusses the process of mapping segments from scene images and words from natural language sentences into a syntactico-semantic space for operation within the RNN. For scene images, the features of image segments are computed, including colour, texture, appearance, and shape features, and then mapped into a semantic space using a neural network layer. Similarly, for natural language sentences, words are mapped to vector representations using a word embedding matrix, capturing cooccurrence statistics. Both image segments and words are then represented in the semantic space of the RNN for further processing.

*Section* 4 elaborates on the use of RNNs for structure prediction, focusing on the discriminative parsing architecture. The goal is to learn a function that maps inputs to binary parse trees. Inputs consist of activation vectors representing image segments or words in sentences, along with a symmetric adjacency matrix indicating segment relationships. The section discusses the formulation of structured margin loss for proposing parses, emphasizing correct merging sequences for image segments and the single correct tree for language parsing. Max-margin estimation is employed to ensure correct tree scoring and margin separation. The RNN model operates through a greedy structure prediction approach, recursively merging segments based on local scores until a single parent activation represents the entire image or sentence. Category classifiers are integrated into the tree nodes with a softmax layer and cross-entropy loss, enabling the prediction of class labels. The section concludes with improvements for language parsing, including the use of beam search algorithms to handle sentence structure prediction efficiently (since words only have at most 2 neighbours, the greedy approach used in the natural scene context is somewhat overkill for the natural language context).

*Section 5* discusses the learning process for the proposed model. It addresses the non-differentiability of the objective function due to the hinge loss and introduces the subgradient method to generalize gradient descent. The gradient-like direction, called the subgradient, is computed for the model parameters. Backpropagation through structure is utilized to compute the derivative of the objective function, enabling the minimization of the objective using L-BFGS over the complete training data. Despite the potential non-differentiability issues, the authors state that practical observations suggest no significant problems.

Lastly, *Section 6* details the experiments conducted to evaluate the proposed RNN architecture in both vision and NLP tasks. The parameters for tuning the model are discussed, with emphasis on robustness to parameter variations. The experiments cover scene understanding, segmentation, and annotation tasks, showcasing significant improvements over previous methods in pixel-level accuracy and scene classification. The RNN model's effectiveness is demonstrated through experiments on the Stanford background dataset. Additionally, supervised parsing experiments on the Wall Street Journal dataset exhibit competitive performance compared to existing parsers. Nearest neighbour analyses illustrate the capability of learned feature representations to capture semantic and syntactic similarities in both images and phrases. The paper concludes by highlighting the contributions of the RNN architecture in advancing segmentation, annotation, and parsing tasks.

## 2 What strengths does this paper have?

This paper showcases several strengths. It introduces a novel approach using RNNs for predicting hierarchical tree structures in both scene images and natural language sentences. This innovative method represents a significant departure from traditional techniques and contributes to advancing the fields of computer vision and natural language processing. The experiments conducted demonstrate superior performance compared to existing methods in tasks such as segmentation, annotation, scene classification, and parsing. The paper provides compelling evidence of the effectiveness of the proposed RNN architecture in achieving state-of-the-art results across multiple domains. The robustness of the proposed method to parameter variations is highlighted, indicating its potential applicability across different scenarios. Moreover, the ability of the RNN model to handle both vision and NLP tasks underscores its versatility and generalizability.

In my opinion, the paper is well-written and effectively communicates the methodology, experimental setup, and results. The authors provide clear explanations of the inputs, processing steps, and evaluation metrics used in their approach, enhancing the reproducibility and understanding of the work. They also give access to their RNN model, which further enhances the reproducibility of the paper.

## 3 What weaknesses does this paper have?

While the paper demonstrates several strengths, there are also some potential weaknesses. Although the paper compares the proposed method with existing approaches in scene understanding and natural language parsing, the comparison may be limited in scope. The evaluation focuses primarily on performance metrics such as accuracy, but a more comprehensive comparison with a broader range of methods, including recent deep learning approaches, could provide a clearer understanding of the proposed method's strengths and limitations. Furthermore, while the experiments demonstrate promising results on benchmark datasets, the scalability of the proposed approach to larger and more diverse datasets remains unclear. Scaling up the model to handle larger images or more complex natural language sentences may pose challenges in terms of computational resources and training time.

In terms of interpretability and explainable AI (XAI), the paper lacks an in-depth analysis of the interpretability of the learned representations and decision-making processes of the RNN model. Understanding how the model arrives at its predictions and interpreting the learned features could provide valuable insights into its functioning and limitations. Lastly, the paper does not touch at all upon potential limitations or constraints in real-world deployment scenarios.

## 4 What are the research questions that are answered in the experimental evaluation?

In the experimental evaluation, the paper addresses several research questions related to the performance and capabilities of the proposed RNN architecture. Two main research questions evaluated in the experimental section are:

1. **RQ1: Performance in Scene Understanding**: The key results are that the RNN architecture is evaluated for multiclass segmentation and annotation tasks on the Stanford background dataset. It outperforms previous methods in pixel-level multiclass segmentation accuracy, achieving a higher accuracy rate compared to methods such as Pixel CRF and logistic regression on superpixel features. This demonstrates the effectiveness of the RNN in scene understanding tasks, providing more accurate segmentation and annotation of scene images.

2. **RQ2: Scene Classification Accuracy**: The key results are that the RNN architecture is evaluated for scene classification tasks on the Stanford background dataset. It achieves a higher accuracy rate compared to the state-of-the-art features (Gist descriptors) for scene categorization. By leveraging the entire parse tree and learned feature representations of the RNN, the accuracy of scene classification is improved, indicating the effectiveness of the proposed method in accurately categorizing scenes into predefined types.

As a research question of secondary importance, the paper also aims to evaluate the parsing accuracy and efficiency of the proposed RNN architecture compared to the widely used Berkeley parser. The RNN achieves a final unlabeled bracketing F-measure of 90.29%, slightly lower than the Berkeley parser's score of 91.63%. However, the development F1 scores are virtually identical, with the RNN achieving 92.06% and the Berkeley parser 92.08%. This suggests that while the RNN's parsing accuracy is slightly lower than the Berkeley parser, it still performs admirably well. Additionally, the RNN's continuous representations of words and phrases allow it to make good parsing decisions without providing a parent with information about the syntactic categories of its children.

## 5 How could the work presented in this paper be improved?

A potential research question that I would investigate in order to improve the work presented in this paper would be *How does incorporating contextual semantic information from surrounding segments or words impact the performance of the RNN architecture in scene understanding and language parsing tasks?*. While the RNN architecture shows promising results in segmenting scenes and parsing

language, incorporating contextual semantic information could further enhance its performance. By considering the semantic context of neighbouring segments in scene understanding or neighbouring words in language parsing, the RNN may better capture the relationships and dependencies within the data. This could lead to more accurate segmentation of complex scenes or more precise parsing of syntactically intricate sentences. Additionally, leveraging contextual semantic information could improve the robustness of the RNN architecture across different datasets and tasks, making it a more versatile and effective tool in various applications.

## 6 How reproducible is the work?

Given the detailed description of the methodology, parameter settings, datasets used in the paper, and the entire code of the RNN model, researchers and practitioners would likely rate the reproducibility of this work as a **4 out of 4**. They could mostly reproduce the results described here, potentially substituting public data for any proprietary data used. I say mostly here because some results with precise decimals, such as the natural language parsing accuracy, may be hard to reproduce up to those exact decimals. While the parameter settings are specified, there may be some subjectivity in determining certain parameters. However, with effort and access to the necessary resources, reproducing the results should be quite feasible (Raff, 2019).

## 7 How confident am I in my assessment of this paper?

I would rate my confidence in the assessment of this paper as a **4 out of 5**. I feel quite sure about the evaluation and tried to check the important points carefully. While it's unlikely that I missed something significant, there's always a chance that some details might have been overlooked which could affect my ratings. However, I'm willing to defend my evaluation, and overall, I believe I have a good understanding of the paper and its contributions.

## References

Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. *Advances in Neural Information Processing Systems*, 32.

Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 129–136.