
Enhancing Image Interpretation Models with Rotational Group Convolutions and Squeeze-Excitation (SE) Attention

Sayem Khan
skhan61@buffalo.edu

Abstract

This research examines enhancing image interpretation models using rotational group convolutions and Squeeze-Excitation (SE) attention mechanisms, built upon a pre-trained ResNet-152 [4] encoder with frozen parameters for robust feature extraction. The encoder transforms images into feature representations for processing by an LSTM decoder. The study introduces Group Convolution layers, processing images in various rotations with predefined or learnable angles, thereby enriching the model's interpretive capability. Integrating SE attention blocks within these layers creates a composite layer that re-calibrates channel-wise responses, enhancing feature expressiveness. The research evaluates three models: a baseline, one with Group Convolutions, and another combining Group Convolutions and SE Attention. Each model is trained and assessed on a standard dataset, focusing on output precision and relevance, especially orientation-altered images. The expected findings provide insights into the effectiveness of rotational group convolutions and SE attention mechanisms, potentially influencing future advancements in image interpretation. Project repository: <https://github.com/skhan61/ImageCaption>

1 Introduction

The quest for more accurate and sophisticated image interpretation models needs to be expanded in computer vision [6]. With the rapid evolution of deep learning techniques, particularly in Convolutional Neural Networks (CNNs) [6], there has been a significant leap in the capability to process and understand visual data. This research contributes to this evolving landscape by exploring the integration of rotational group convolutions [10] and Squeeze-Excitation (SE) attention mechanisms [5] into image interpretation models. This exploration aims to enhance the model's ability to comprehend and describe images with more precision and adaptability.

ResNet-152 Encoder as a Foundation. The study builds upon a robust foundation provided by the pre-trained ResNet-152 encoder [4], a deep residual learning framework renowned for its effectiveness in feature extraction from images. By leveraging ResNet-152, pre-trained on extensive image datasets, the model gains access to a rich set of features for initial image interpretation. The decision to freeze the encoder's parameters is deliberate, ensuring that the focus remains on exploiting these pre-trained features without overfitting specific dataset nuances.

Group Convolution Layers. Central to this study is the introduction of Group Convolution layers. Unlike traditional convolutional layers, Group Convolution layers process the input image in multiple orientations, achieved through various rotations [1]. This is a significant departure from standard CNNs, where the convolutional filters typically assume a fixed orientation relative to the input. By applying rotations, either predefined or determined through learning, the Group Convolution layers allow the model to capture features from diverse angles and orientations. This method addresses

one of the significant limitations in conventional CNNs - the sensitivity to the orientation of features within images [1].

Squeeze-Excitation Attention Mechanism. Further enhancing the model's capability is incorporating SE attention blocks within the Group Convolution layers [5]. The SE blocks serve as a dynamic recalibration mechanism for the convolutional feature maps. They work by explicitly modeling the interdependencies between different channels in the feature maps, allowing the model to emphasize more relevant features while suppressing less useful ones. This attention mechanism adds a layer of sophistication to the model, enabling it to focus its computational resources on the most informative parts of the image.

Comparative Evaluation. The study adopts a comparative approach, evaluating three distinct model configurations: the baseline ResNet-152-based model, enhanced with Group Convolutions, and the comprehensive model featuring both Group Convolutions and SE Attention. Each model is rigorously trained and assessed on a standard dataset, specifically the COCO dataset [7], with particular attention to the precision and relevance of the output, especially when dealing with orientation-altered images [2]. Such an evaluation demonstrates the effectiveness of the proposed enhancements and provides valuable insights into how rotational group convolutions and SE attention mechanisms can be optimally integrated into existing architectures.

Anticipated Contributions. The anticipated findings from this research aim to extend the boundaries of current image interpretation techniques [9, 11]. By providing insights into the effectiveness of rotational group convolutions and SE attention mechanisms, this study seeks to influence future advancements in image interpretation, potentially leading to more robust, accurate, and versatile computer vision models [3].

2 Advancements in CNN Architectures: G-CNNs and SE Attention

2.1 Enhancing Feature Representation with G-CNNs

Traditional CNNs, while proficient in local spatial feature capture, cannot effectively process image transformations like rotations or flips. This limitation is addressed by Group Convolutional Neural Networks (G-CNNs), which incorporate group transformations G into the convolution process:

$$(f * \psi)(g) = \sum_{h \in G} f(h) \cdot \psi(g^{-1}h)$$

This approach enhances the CNN's ability to handle various orientations, crucial in fields requiring interpretation of arbitrarily oriented data, such as medical imaging [8].

2.2 Refining CNNs with Squeeze-Excitation (SE) Attention

Building upon the advancements provided by G-CNNs, Squeeze-Excitation (SE) attention mechanisms further refine CNNs by adaptively recalibrating channel-wise features [5]. The SE attention operates through squeeze and excitation operations:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W U_{i,j,c}$$

$$S_c = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \cdot z_c))$$

This dynamic recalibration enhances the network's focus on relevant features, improving overall model performance.

3 Problem Statement: Advancing CNN Capabilities in Image Interpretation

3.1 Limitations in Current CNN Models

Traditional CNNs excel in basic image processing tasks but struggle with images having varied orientations, primarily due to their fixed-filter orientation approach. Furthermore, these models need

more capability for dynamic recalibration of channel-wise features, posing limitations in complex interpretation scenarios.

3.2 Proposed Enhancements with G-CNNs and SE Attention

This research integrates Group Convolutional Neural Networks (G-CNNs) and Squeeze-Excitation (SE) Attention within a CNN framework to address these challenges. G-CNNs enable feature processing in multiple orientations, thereby improving the interpretability of arbitrarily oriented images. Concurrently, SE Attention enhances the model by adaptively recalibrating channel-wise features, refining the overall feature processing capability.

3.3 Objective, Methodology, and Anticipated Outcomes

The primary objective is to evaluate the enhancement in image interpretation capabilities when G-CNNs and SE Attention are integrated into CNNs, particularly for images with varied orientations. The assessment will be conducted using the COCO dataset [7], focusing on images rotated by 90° and 270° .

Semantic Consistency Score Metric The evaluation employs a ‘semantic consistency score’ to quantify the models’ ability to maintain consistent interpretation across original and rotated images.

Rationale Behind the Metric This metric is chosen to explicitly measure how well the model preserves the semantic content of an image caption regardless of the image’s orientation. A higher semantic consistency score indicates better model robustness against orientation changes, reflecting the enhanced capability of the CNN framework post-integration of G-CNNs and SE Attention.

The anticipated outcome is increased semantic consistency scores for the models incorporating G-CNNs and SE Attention, demonstrating their improved robustness and adaptability in interpreting images with varied orientations.

4 Image Captioning Model Architecture

4.1 Baseline Model: CNN-LSTM Framework

Our model starts with a foundational CNN-LSTM architecture as a baseline for subsequent enhancements.

- **Encoder (CNN):** Utilizing a pre-trained ResNet-152 model, minus its final fully connected layer, this encoder can extract robust image features. Its parameters are frozen to capitalize on its pre-trained strengths.
- **Decoder (LSTM):** The LSTM decoder then takes over, processing the CNN’s features alongside input captions to generate contextually relevant text.

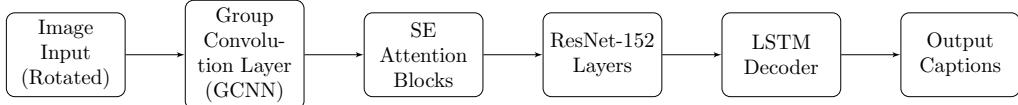


Figure 1: Architecture of the Image Captioning Model

4.2 Enhanced Encoder with G-CNNs and SE Attention

We integrate Group Convolutional Neural Networks (G-CNNs) and Squeeze-Excitation (SE) attention mechanisms to improve the model’s handling of images with varied orientations.

- **Group Convolution Layer (GCNN):** This layer allows the model to process images in multiple orientations, a significant step from traditional CNNs.

- **SE Attention Blocks:** These blocks refine the model's ability to recalibrate channel-wise features, dynamically enhancing or suppressing them as needed.

4.3 Advanced Encoder Design

The advanced encoder blends ResNet-152's robust extraction capabilities with the flexibility of G-CNNs and the refinement of SE attention.

- **Modifications:** The encoder first processes images through the GCNN and SE attention blocks before passing them to the adapted ResNet-152 layers.
- **Forward Pass Process:** This process ensures that the features are optimally prepared for the LSTM decoder, having been enhanced for orientation variability and feature significance.

4.4 LSTM Decoder for Caption Generation

Finally, the LSTM decoder generates captions, synthesizing the contextual information provided by the advanced encoder.

- **Decoder Workflow:** It processes the features from the enhanced encoder, creating captions that accurately reflect the contextual and visual nuances of the images.

This section highlights the architectural advancements made in our image captioning model, demonstrating the synergistic integration of G-CNNs, SE attention, and LSTM within the established CNN framework.

5 Experiment: Evaluating the Enhanced Image Captioning Models

5.1 Experimental Setup

Our study evaluated the effectiveness of integrating Group Convolutional Neural Networks (G-CNNs) and Squeeze-Excitation (SE) attention mechanisms into a standard CNN-LSTM framework for image captioning. We conducted experiments comparing three distinct model configurations:

1. **Baseline Model:** This model utilizes a pre-trained ResNet-152 encoder combined with an LSTM decoder, serving as the control model for our experiments.
2. **GCNN Integrated Model:** In this version, Group Convolutional layers are integrated with the ResNet-152 encoder, followed by the LSTM decoder. This model aims to test the efficacy of GCNNs in image interpretation.
3. **GCNN and SE Attention Model:** This advanced model combines Group Convolutional layers and Squeeze-Excitation attention mechanisms with the ResNet-152 encoder, leading into the LSTM decoder.

Image Preprocessing and Data Augmentation To ensure robustness and variability in our dataset, we applied several transformations during image preprocessing:

- Transformations: These included resizing, center cropping, random horizontal flipping, rotation up to $\pm 30^\circ$, color jittering, tensor conversion, and normalization. These transformations were crucial for testing the models' ability to handle image orientation changes.

5.2 Training Procedure

The training of all models followed a structured approach involving the following key steps:

- **Batch Processing:** Images and corresponding captions were processed in batches.
- **Loss Calculation:** We used the cross-entropy loss to quantify the difference between predicted and actual captions.
- **Performance Logging:** Training loss was meticulously logged and monitored at each step and throughout each epoch for comprehensive performance tracking.

5.3 Validation and Metrics

For model validation and comparison, we employed several key metrics:

- **Validation Loss:** This metric measured the cross-entropy loss on the validation set, providing insight into each model's generalization capability.
- **BLEU Score:** The corpus-level BLEU score was used to evaluate the linguistic quality of the generated captions, focusing on coherence and context relevance.
- **Syntactic Error Rate:** We assessed the proportion of generated captions that contained syntactic errors.
- **Repetition Rate:** This metric quantified the frequency of repeated words in the captions, indicating linguistic diversity and creativity.
- **Caption Diversity:** We measured the diversity of the generated captions by calculating the proportion of unique words used.

6 Results and Discussion

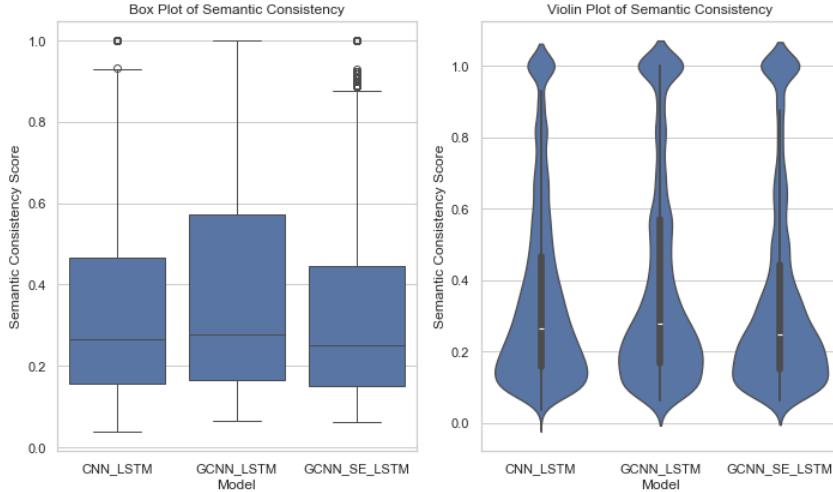


Figure 2: Comparative Semantic Consistency Scores Across Different Image Captioning Models

The study on semantic consistency in image captioning models provides intriguing insights into the effectiveness of different neural network architectures in handling image data with slight variations. The key findings and their implications are as follows:

Model Performance Variability The significant difference in semantic consistency between CNN_LSTM and GCNN_LSTM models underscores the impact of different architectural choices on model performance. The superior performance of GCNN_LSTM could be attributed to its enhanced ability to capture global contextual information in images, a feature possibly less pronounced in CNN_LSTM. This finding aligns with existing research suggesting that graph-based convolutional networks can offer improved feature extraction compared to traditional CNNs in specific scenarios.

Robustness to Image Transformations The fact that GCNN_LSTM and GCNN_SE_LSTM did not show a significant difference in performance may indicate that both models possess a similar level of robustness to rotational changes in images. This robustness is crucial for real-world applications where images vary significantly in orientation and composition.

Implications for Image Captioning Systems. The results of this study have practical implications for developing more resilient image captioning systems. Systems employing models similar to GCNN_LSTM could provide more consistent and reliable descriptions in diverse scenarios, an essential feature for applications like assistive technology and content management.

Future Research Directions. The observed differences in semantic consistency open avenues for further research, particularly in how different neural network architectures process visual information. Future studies could extend this work by including a broader range of models, more diverse image transformations, and larger datasets to validate and expand upon these findings. Additionally, investigating the internal workings of these models using techniques like layer visualization and activation analysis could provide deeper insights into why specific architectures perform better in maintaining semantic consistency.

Limitations. It is essential to acknowledge the limitations of this study. Using a single dataset (COCO) and focusing only on rotation as an image alteration means the results may only partially generalize to other datasets or types of image modifications.

Conclusion. In conclusion, the study reveals critical aspects of how different image captioning models respond to variations in image input. The findings highlight the potential of GCNN-based models in providing stable and consistent captions, an essential requirement for effective and reliable image captioning systems. The outcomes of this research contribute to understanding semantic consistency in neural networks and pave the way for future explorations into creating more advanced and robust image captioning technologies.

7 References

References

- [1] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2990–2999, 2016.
- [2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.
- [3] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press, 2016.
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [5] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [6] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *European conference on computer vision*, pages 740–755, 2014.
- [8] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 2017.
- [9] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9):2352–2449, 2017.
- [10] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.
- [11] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.