KU LEUVEN

FACULTY OF
ENGINEERING SCIENCE

Energy-Based Learning in Weakly Supervised Scene Graph Generation

Berkay Güler

Thesis submitted for the degree of
Master of Science in
Electrical Engineering, option
Information Systems and Signal
Processing

Supervisor:
Prof. dr. ir. Tinne Tuytelaars

Assessors:
Prof. dr. ir. Marian Verhelst
Prof. dr. Matthew B. Blaschko

Assistant-supervisor:
Ir. Bo Wan

Academic year 2022 – 2023

Preface

In 2021, I obtained a Electrical Engineering BSc degree in Middle East Technical University in Turkiye and I decided to come KU Leuven to pursue a MSc degree in the same field. This thesis work is one of the milestones in my life which shows a significant growth in my specialization field signal processing and computer vision.

First of all, I would like to thank my thesis advisor Professor Tinne Tuytelaars for showing guidance along this long-lasting process. My daily supervisor Bo Wan also played an quite important role in my procedure. He was always there to enlighten me about the concepts of this thesis and helped me to stay on the right track for the methods I wanted to follow. I gained a considerable amount of knowledge under his daily supervision.

Lastly, allow me to present my appreciations to my family and friends since they showed their support whenever I needed. Without them, it would be impossible for me to complete this extensive project.

Berkay Güler

Contents

Abstract

In scene graph generation (SGG), many works in literature have shifted towards weak supervision instead of guiding the training in fully supervised settings for two main reasons. Fully supervised training demands substantial labor for annotating the dataset leading to human label errors and bias in training. Furthermore, weak supervision is cost-effective and more scalable since it allows gathering image-label pairs more effortlessly, eliminating the need for annotations. Besides the training scheme, SGG works in weak supervision traditionally predicts the $\langle subject, predicate, object \rangle$ triplets by the sum of cross-entropy losses. Such a loss formulation leads to interpreting every object pair independently; however, every image possesses an inherent structure making every object in the image related. Hence, this thesis suggests exploiting the benefits of energy-based frameworks for the SGG tasks under weak supervision to consider the underlying structure in the image. This energy framework maximizes the joint likelihood of objects and relations. That is why it differs from the studies that use the sum of cross-entropy losses. The main contributions of this thesis start from a conventional relation model and modify this model for the weak training scheme to come up with the baseline results. Then, an energy-based approach from the literature is adapted into the same weakly trained setup and integrated into the modified relation model. The integration contains newly designed filtering and sampling techniques by this thesis to assist the energy framework. It is assumed that the energy methods would solve the common challenges in weakly supervised SGG such as the long-tailed dataset, and the biased training due to image-level relation labels for the relation model. The experimental results demonstrated that the energy methods contribute to the overall SGG performance of the baseline model even under weak supervision, proving the main suggestion in this thesis.

List of Figures and Tables

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

BG	Background
biLSTM	Bidirectional Long-term short memory
CNN	Convolutional Neural Network
EGNN	Edged Graph Neural Network
EBM	Energy-Based Method
EBM	Energy Based Methods
FPN	Feature Pyramid Network
GNN	Graph Neural Network
GRU	Gated Recurrent Unit
GT	Ground Truth
HOI	Human-Object-Interaction
LSTM	Long-term short memory
mAP	Mean Average Precision
MLP	Multi Layer Perceptron
PhrDet	Phrase Detection
PredCls	Predicate Classification
R@K	Recall-at-K
RNN	Recurrent Neural Network
RPN	Region Proposal Network
SGCls	Scene Graph Classification
SGD	Stochastic Gradient Descent
SGDet	Scene Graph Detection
SGG	Scene Graph Generation
SGLD	Stochastic Gradient Langevine Dynamics
VG	Visual Genome
Vtrans	Visual Translation
mR@K	Mean-Recall-at-K
MCMC	Monte Carlo Markov Chain

Symbols

N	Number of region proposals
N_r	Number of relation classes
N_o	Number of object classes
O	Detected objects matrix
R	Relation score matrix
o_i	i-th proposal's object dist.
r_{ij}	Relation dist. between i-th & j-th
f_i	Feature vector for i-th proposal
l_i	GloVe embedding for i-th proposal
b_i	Bounding box embedding for i-th proposal
c_i	Obj. context biLSTM hidden dim.
d_i	Edge context biLSTM hidden dim.
$f_{i,j}$	Union feature vector for (i,j) proposal pair
$w_{i,j}$	Frequency bias vector for (i,j) proposal pair
W	Kernel layer to map dimensions
n_k	Node embedding for kth proposal
$r_{j \to i}$	Edge embedding between i-th & j-th proposals
m_i	Incoming message vector for the i-th proposal
$p_{j \to i}$	Incoming message vector for the relation vector of (i,j) proposal pair
N	Node vector representation after pooling
E	Edge vector representation after pooling
e^-	Scalar negative energy value for prediction
e^+	Scalar positive energy value for ground truth
$E_\theta()$	Energy model function
G_{SG}	Predicted scene graph
G^+_{SG}	Ground truth scene graph
f_{gate}	Function that finds attention score for input node
g_{gate}	Function that finds attention score for input edge
R^{scores}_{ij}	Relation probabilities found applying softmax on R
O^{scores}	Object probabilities found applying softmax on O
o^{score}_i	The object probability for the i-th proposal
r^{score}_{ij}	The relation probability for the (i,j) proposal pair
Tri^{score}	Triplet scores

Chapter 1

Introduction

Over the last decade, deep learning contributed to the area of computer vision drastically by improving most of the results of numerous types of research. Thanks to these enhancements, tasks such as image classification and object detection are no longer regarded as intriguing; therefore, people initiated to employ more demanding and higher-level visual understanding tasks like scene graph generation, which possesses a comprehensive image representation by considering the relationships between the detected objects.

The scene graph generation (SGG) [4] is to recognize the visual semantics of a given image by detecting the triplets of $\langle subject, predicate, object \rangle$. The objects detector provides locations and categories for an object pair, namely *subject* and *object*, and the predicate is the relation word between this $\langle subject, object \rangle$. These predicted triplets are sorted depending on their confidence scores and provide a meaningful and structured representation of the input image. SGG can be realized mainly in two methods: **i)** One-staged methods predict these triplets with transformer-like structures in one evaluation. **ii)** Two-staged methods first detect object locations and labels; then, they determine the relation label using these proposals. Since SGG generates this instructive representation for the given image, it can serve as a processing block in different vision application areas such as image generation [5], image/video captioning [6], image retrieval [7], visual question answering [8], or image understanding [9].

This thesis mainly focuses on two-staged SGG methods, and the training of these two-stage scene graph models could be implemented primarily in two ways. The two standard schemes are fully supervised and weakly supervised training. In a fully-supervised setting, the bounding box annotations for the objects are present. For the relation classification stage, the fully supervised setting provides ground truth relation categories for each subject-object pair at the instance level in the ground truth. On the contrary, the weakly supervised setting does not provide any of those bounding boxes for the objects in training, and it only has image-level object labels and relations or image-level graphs in the ground truth. That means in the weakly settings, one only knows which objects and relation labels are present in the image or a graph representation that lacks box annotations. Figure 1.1 illustrates

the main differences between fully and weakly supervised settings. Although SGG is initially designed for the fully supervised setting, this approach requires a ton of human effort on thousands of images resulting in quite time-consuming tasks and lots of annotation errors such as mislabeling or extensive labeling. The main advantage of weak supervision over the fully supervised setting is that it is easier to collect these ground truth image captions sharing information about the image. Thus, the recent works shift their attention towards weak supervision.

In recent years, a method called energy-based learning has begun to gain popularity in several areas. The studies such as [10, 11] utilize energy framework to enhance their image generation model accuracy, whereas the papers like [3, 12] find the energy-based method beneficial in discriminative tasks. According to LeCun *et al.* [13], the energy-based method assigns a scalar value for particular input-output pair. For instance, the input could be an object pair, and the output is the relation between the pair. This scalar value is called the energy, and the model learns to minimize it for the correct input pairs and maximizes it for the incorrect input configurations. Generally, scene graph tasks contain a summation of cross-entropy losses for the relation classification of each triplet. Adding an energy loss in the back-propagation guides this learning process thanks to energy methods. Suhail *et al.* [3] proves that energy-based approaches are effective for scene graph tasks in fully supervised settings. This energy loss is calculated by using graph-based ground truths of the images. Therefore, energy methods in the weakly supervised scene graphs represent a new domain worth exploring.

This thesis explores energy-based methods for SGG tasks under weak supervision. It starts by modifying a state-of-the-art relation model to obtain weakly supervised baseline results. Then, it continues by revising a novel energy model architecture to improve the previous baseline results by adding it to the general pipeline. The required modifications for each model will be made to convert the models suitable for weak supervision. Section 5 conducts a series of experiments to answer the research questions in Section 1.2.

1.1 Challenges

As mentioned earlier, most of the existing work [2, 3, 4, 8, 14] on scene graph generation treats the problem as a fully supervised one and follow the strategy of using bounding box annotations for each object and their interactions. However, composing such a dataset needs vastly annotated images done by humans which is expensive to collect and brings its own bias and labeling errors into the dataset. Thus, a more convenient approach to scale the number of images in the dataset is to apply weakly supervised settings instead of fully supervised, where these weak supervisions only need the image-level information.

Even though weak supervision is more scalable and easier to collect, it does come with its own set of challenges and limitations. Firstly, such weak supervision does not offer detailed information about the image and usually provides shallow visual semantics. Secondly, employing an off-the-shelves object detector brings its noisy

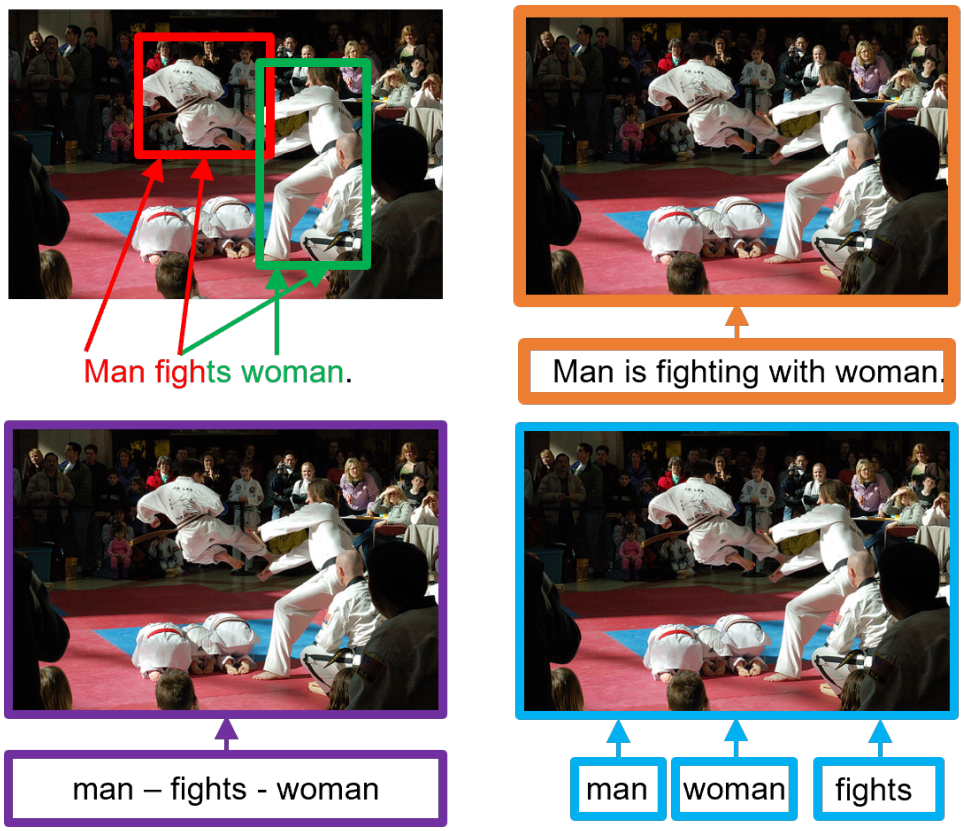


FIGURE 1.1: Comparison of scene graph training schemes under fully supervised settings (top-left), weakly supervised with image caption (top-right), weakly supervised with image-level labels (down-right), and weakly supervised with image-level graph (down-left).

predictions. Generally, these noisy predictions count as background objects and are eliminated in full supervision since the bounding box annotations are provided for the foreground objects. However, one has to trust the proposals obtained by the off-the-shelf object detector or train their object detector under weak supervision.

Figure 1.2 illustrates an example of scene graph generation and another challenge in the scene graph generation task. Typically, most existing works [2, 4, 8, 14, 15, 16] about SGG, whether they are weakly supervised or fully supervised, utilize sum of cross entropy losses for the predicted $\langle subject, predicate, object \rangle$ triplets. Such a loss calculation ignores the underlying structure in the given image and treats every triplet as an independent one. However, an image usually contains triplets that are highly related to each other so a new method like energy-based learning can incorporate this structure in the image with an additional energy loss formulation and convert the problem from the sum of likelihood terms into a maximization for the joint density function of the image. Addition of this new method results in better predictions as it can be seen in Figure 1.2 where the predicate word `riding`

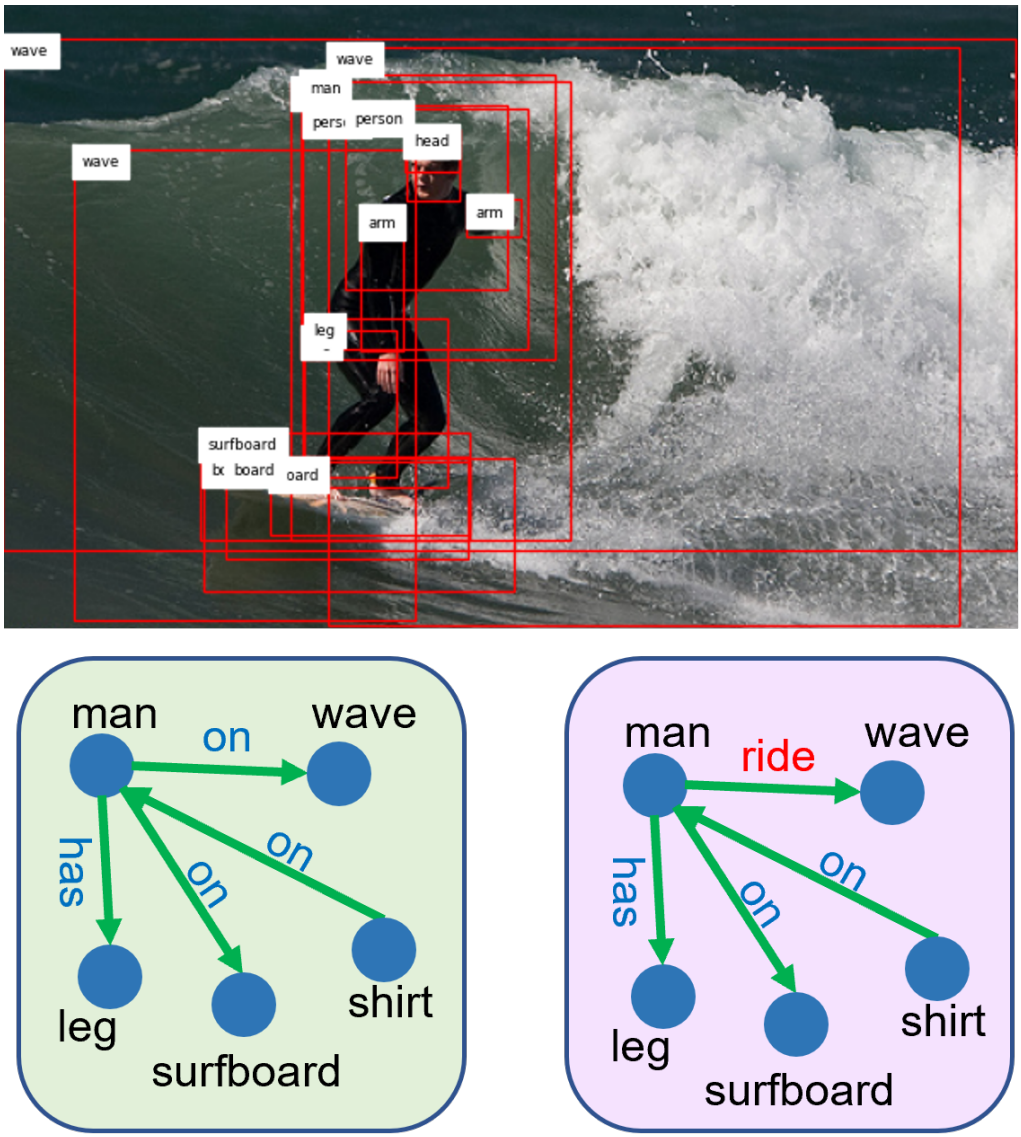


Figure 1.2: A generated scene graph by the baseline only cross entropy model (green) and the energy model (purple).

is replaced with `on` since the latter presents a better understanding of the image and tackles down the biased dataset issue. The other shared problem of the SGG is biased dataset training. All of the stated previous studies perform their tests on Visual Genome [17] dataset which has a long-tailed distribution. Hence, some predicate words such as `eating` are only seen a small fraction of the time whereas geometric predicates like `on` is present in most of the images. Therefore, the method of one's model should try to overcome this issue. Energy-based learning SGG provided in [3] proves to achieve success when it comes to dealing with this long-tailed distribution problem.

1.2 Research Questions

The main objective questions for the thesis to solve the issues mentioned above can list as follow: *Is it possible for energy-based methods to improve SGG model results under weak supervision as in full supervision?*

The thesis work will search for an answer to the objective question, and this question can be divided into sub-questions to answer the challenges of the SGG tasks. ***i)*** *What results can be obtained by only considering the image-level relation labels?*, ***ii)*** *To what extent do energy models mitigate the effect of biased training from the image-level relations?* and ***iii)*** *Can energy models work under weak supervision even if the foreground/background objects are absent in the ground truth?*

1.3 Main Contributions

As briefly discussed before, this thesis aims to explore whether the energy-based learning method also works for weakly supervised training for scene graph generation tasks. This study starts by conducting a reproduction study of the results presented by Suhail *et al.* [3] to have a solid starting point. The contributions built on this starting point can be summarized as follows: Firstly, the scene graph model designed by Motif [2] will be modified to establish a weakly supervised baseline. Secondly, the energy-based method designed by Suhail *et al.* [3] will be adjusted to make it suitable for weak supervision and integrated into the modified Motif to improve the baseline model results. This integration utilizes a proposed sampling and filtering mechanism that functions under weak supervision. The source code for the work is available in `https://github.com/gulerrberkay/scene-graph-ebm`.

1.4 Thesis Summary

This section states a general overview of all upcoming chapters in this thesis. **Chapter 2** discusses the literature works regarding SGG tasks in different branches. The analysis for the selected papers is provided to compare it with the energy models. **Chapter 3** provides general background information about scene graph generation methods. Even though many studies have been discussed in Related Works section 2 previously, this thesis will mainly focus on the Motif model [2]. The working

mechanism of this study will be shared in detail to provide readers with how a generic relation head model functions and identifies relations between subject-object pairs in a given image. In addition, since the aim of this thesis is to show that energy methods also work for weak settings, graph-based architecture and energy loss formulation designed by [3] will be discussed. Lastly, performance evaluation metrics for scene graph models and reasons for using such metrics will be familiar to the readers. **Chapter 4** begins with the problem setup for this thesis work. Assumptions and limitations about the topic will be introduced to readers. Then, all of the added code features and how they are implemented will be presented thoroughly. **Chapter 5** investigates both quantitative and qualitative results for weakly settings. All experiments and ablation studies for energy-based weakly settings will be examined to understand which hyperparameters and design choices have a direct effect on getting better results. **Chapter 6** summarizes all the contributions that have been done in this thesis and further discusses what can be implemented to extend this thesis in the future.

Chapter 2

Related Works

In this chapter, the related works in the literature for different object detectors and scene graph models and their approaches are listed and analyzed. For SGG, the methods divided into groups by looking at their strategy to tackle the SGG problem. Then, these method is briefly summarized so that one can understand better what they try to achieve. Lastly, their methods and results are discussed with this thesis work method to analyze the pros and cons of the proposed method over the related works in the literature.

2.1 Object Detector

Object detection can be considered one of the fundamental blocks for any vision task. Thus, one should consider the available architectures for object detection when initiating an SGG task. Here, this subsection presents various studies for object detection.

Faster R-CNN [1] is one of the most popular object detectors in the literature, and it has been used extensively in different vision tasks. Faster R-CNN still offers one of the best performances for object detection when it is compared with other detectors thanks to its fine-grained regression and classification. This fine-grained alignment serves better features and boxes for the detected objects. That is why it improves the detection performance and almost every SGG paper utilizes this architecture for detecting objects. Section 3 explains the method of Faster R-CNN in more detail.

YOLO [18] is also a popular object detector. It follows a one-staged strategy than two-staged methods since it divides the image into grids and predicts class probabilities and bounding box regions directly on grids. The main advantage of YOLO is that the speed performance of the architecture outperforms many of the detectors and can perform in real time. Despite the speed improvement, YOLO suffers from localization errors for small objects due to its grid method.

DETR [19] utilizes the merits of transformers for detecting objects in a single pass. According to [19], it also drops the need for the anchors by using the attention strategy of transformers, and the CNN operation is no longer needed because the

attention provides global scale information about the image. Even though this model possesses a simpler approach than Faster R-CNN but it still lacks the performance of Faster R-CNN for small objects.

2.2 Scene Graph Generation

The scene graph community has extensively surveyed two different training schemes, namely fully and weakly supervised settings. Fully supervised SGG usually has various approaches depending on the main block in the architecture. If the relation predictor block is an RNN-based layer, then the study automatically becomes an RNN-based solution. Each solution has its advantages and disadvantages. On the contrary, weak supervision studies follow unique ways of solving the problems; thus, they cannot be categorized as RNN-based or GNN-based. This subsection provides a summary of all studies in two different parts.

2.2.1 Fully Supervised Scene Graph Generation

Fully supervised SGG has different branches to solve the same problem of visual relation detection. The selection of the architecture determined the differences between these branches. This subsection examines some of these different approaches in the papers.

Graph Neural Network (GNN) based SGG

Some works like Graph R-CNN [20] try to address this issue with Graph Neural Network (GNN) structures since the problem of SGG has a graph-like structure if one sees objects as the nodes and relations as the edges of this graph. Graph R-CNN [20] starts from the detected objects and calculates the relatedness of these objects using MLP layers. To refine the graph, and to incorporate global context, it adds an attention layer where messages between neighboring nodes are calculated to update each node's states.

Recurrent Neural Network (RNN/LSTM) based SGG

Other branches of fully supervised SGG find Recurrent Neural Network structures like LSTMs quite valuable while predicting the relations in the image. That is the natural advantage of RNNs when it comes to discovering the relationships of object pairs from the structured input sequences. Most of the SGG studies [2, 4, 8, 14] tries to learn the relations in the given image in a fully supervised manner and evaluate their results on Visual Genome [17] dataset. Initial works like [4] utilizes RNN structures to aggregate messages between detected object features and predict a scene graph. Their approaches primarily focus on local features rather than global union features. On the other hand, Motif [2] collects visual features from every proposal and union region to combine them in LSTM layers called object and edge contexts to improve performance. VCTree [8] takes the hierarchical structure present

in the image by constructing a tree-based LSTM called BiTreeLSTM. Hierarchical entities are fed into this layer to understand the context of the given image. However, previous models do not solve the problem of biased training caused by the dataset. Thus, Tang *et al.* [14] formulates a different kind of loss formulation to resolve the issue of the biased dataset. Tang *et al.* [14] uses the same structures in the previous works, but with a novel hand-crafted loss formulation prevents the drawbacks of using a biased dataset in training. Even though these models offer a solid starting point, they treat every predicted object triplet independently. Such an approach ignores the underlying structure of the image.

Translation Embedding based SGG

According to Zhang *et al.* [21], a visual translation method tries to represent the objects and subjects in the image in lower dimension and it tries to satisfy a straightforward equation $s + p \approx o$ where the translated vectors are s, p, and o. This summation can be also considered as a vector translation between object and subject embeddings low dimension representations. Thus, the TransE-based models learns three projection kernel matrices $W_s x_s + t_p \approx W_o x_o$.

One of the early works, in this domain is VTransE [21]. Their approach contains an object detection module to find classes of objects and a bilinear interpolation module where the visual features of these proposals are extracted in a differentiable way. They feed their relation module with every possible relation triplet to find a relation translation vector at the training. However, this method ignores the long-tailed distribution of the Visual Genome, so it is not successful for the triplets to occur less in training data.

CNN-based SGG

CNN-based approaches are also part of the SGG tasks. In this method, the goal is to extract local and global features via CNNs, and then predict both objects and relations by classification. According to Zhu *et al.* [22] CNNs architectures are successful in extracting global visual features; hence, it makes sense to utilize their benefits in the related visual task SGG. However, Zhu *et al.* [22] also states that CNNs based-approaches offer deep interaction features between objects which makes them computationally expensive. Therefore, while keeping the deep features without disruption, lowering the computational needs of CNN-based SGG seems to be a challenge for this branch.

LinkNet [23] provides one of the first studies regarding CNNs-based SGG. They introduce a global context encoding module to capture spatial-context information between objects. Their method consists of three stages where the first stage is to find proposals. This stage is followed by the object classification stage where object distributions, features, and context features are then fed into an object classifier to find the object labels. Lastly, the relation classifier finds out all the relations in the image. The drawback of their method is that detecting relations between all related objects is computationally expensive.

2.2.2 Weakly Supervised Scene Graph Generation

Besides the full supervision, there are also notable studies under weak supervision. Zhang *et al.* [24] proposed a weakly supervised object detection and relation classifier learning strategy where the model is trained with the image-level object and relation labels. Their relation branch consisted of two sub-branches where pair and relation scores are calculated from a score map pairwise pooling mechanism. These score maps were computed via CNN; thus, spatial context may miss minor details in the image, and may cause problems in relation prediction.

VSPNet [15] is one of the earliest works, and the authors decided to use ground truth graphs of the images and use graph alignment algorithms between ground truth and their predicted parsed graph. Their hypothesis was to incorporate the context in the image by using a graph as a ground truth and not treating the objects in the image independently, which ignores the structure of the input image. Also, they want to solve the common problem of quadratic growth of relations in SGG. If one treat object-subject pairs as the nodes and relations as the edges, one should see that the complexity of relation calculation grows quadratically.

Ye *et al.*[25] tackles the problem of weak SGG by trying to include image captions in their method setup. They start from a VSPNet[15] like graph setup where they try to find attention scores between object proposal regions and ground truth text graph. Text graphs can be either parsed from an image caption with an existing language parser or VSPNet-like clean ground truth graph. These attention scores will be trained by image-level relation labels and iteratively updated to find a better representation of the image. Their method takes the underlying structure in the image as it learns with a text graph. Since they use captions for the supervision, their problem setup is weaker than both [24] and [15]. The calculated text graph benefits from the caption description to enrich the attention scores, and these scores will distinguish small details better. Learning from a graph again makes this method better than [24] in terms of understanding the structure in the given image.

Baldasarre *et al.* [26] approaches the weakly supervision problem with image level relations like previous works. Their difference is that they utilize a Graph Neural Network structure whose nodes are initialized by the proposal features. The edges between these nodes are only the spatial coordinates and angles. All nodes and edges are aggregated with straightforward MLP layers to find the present relations in the image. The authors then find the corresponding nodes that contribute to the highest-scored relations with a method called sensitivity analysis.

A similar task of Human-Object-Interaction (HOI) studies is also related to scene graph generation due to similarities. HOI investigates the relations between subjects and objects where the subjects are always human. Bo *et al.* [27] addresses the problem of weakly HOI tasks by applying a multi-task loss on their visually and textually enhanced feature vectors. Their enhancement architecture contains CLIP [28] encoded features and attention mechanism between visual and textual features. One of the loss formulations that they proposed is taking the maximum score for each relation class from the predicted relation score matrix to find the image-level relation scores. The maximization is then followed by a binary cross-entropy loss

which guarantees that at least one pair for a particular class has that relation. This method is also suitable for this thesis work; thus, Section 3 analyzes this weak loss formulation in-depth.

Some of the newer works Shi *et al.*[29] and Li *et al.* [16] engaged the weakly supervision with a different approach where they aimed to convert the weakly supervised problem into a fully supervised problem. To perform such a transformation, Shi *et al.*[29] utilizes a first-order graph alignment algorithm, which is simpler than VSPNet[15], to find a pseudo ground truth scene graph. They align their unlocalized scene graphs by calculating attention scores between unlocalized text graph entities and object features from proposals. On the other hand, Li *et al.* [16] proposes to train a grounding module instead of using a graph alignment. This grounding module is trained with positive-negative images where images in the batch are paired with text graphs. Even though [16, 29] both acquire state-of-the-art results, their method does not solve the problems of existing fully supervised SGG problems like unbiased dataset issues.

2.3 Energy-Based Methods

The energy-based approaches are mainly popular for the image generation tasks. In studies like [10, 11] increases their generative model accuracy thanks to energy models. Suhail *et al.* [3] is one the significant important papers for the implementation in discriminative tasks. In future the applications of the energy model may grow due to their flexibility to be using any arbitrary function. LeCunn *et al.* [13] states that energy-based models try to learn the underlying data distribution of input space while they evade calculating the partition function used for normalizing which is computationally untraceable. This thesis mainly focuses on the study that has been carried out by Suhail *et al.* [3] where the authors' experiments with energy methods about classification tasks. The researchers formulate an energy loss formulation that will be beneficial for scene graph learning. This scalar energy value will represent the whole structure in the image. The energy loss will be computed from the novel graph architecture and it will serve as a guide for actual scene graph learning.

2.4 Conclusion

In this chapter, various studies from different branches of SGG are examined to compare them with this thesis work. The advantages and disadvantages of these papers are listed and a brief analysis is provided. The next chapter will go more in-depth into the papers that will be mainly utilized for this thesis.

Chapter 3

Background

In this chapter, building block elements for weakly supervised scene graph training are discussed thoroughly. Object detection section 3.1 presents a well-known object detection algorithm used by most recent studies about computer vision. Scene Graph Generation 3.2 part investigates a particular scene graph generation model to understand deeply how it works. Lastly, the Energy Based Learning 3.4 part shows how energy-based learning can be utilized to improve the results for a scene graph model.

3.1 Object Detection

The scene graph task requires object pairs to find relations between them. Thus, it is necessary to start with an object detection block. Since object detection itself is a complicated task, there are different approaches to finding objects in the given image and some of these methods are already introduced in Section 2. Over these detectors, Faster R-CNN is the one usually utilized in most of the SGG studies [2, 3, 4, 8, 15, 30] in the literature due to its success in detection performance. Since the speed of the detector is not important for the task ahead, the one-staged methods are not used or examined in this thesis.

3.1.1 Faster R-CNN

The important blocks of Faster R-CNN are depicted in Figure 3.1. The process of Faster R-CNN can be summarized as follows: Firstly, a feature map for the given image is calculated from the backbone convolutional layers. These feature maps are the representations of the image in a different form and they are used in the training of the Region Proposal Network (RPN). According to [1] the most significant part of this structure is the RPN, as it finds the most probable spatial locations in the image while training. To train the RPN, the authors in [1] proposed using sliding windows over the feature map. These sliding windows are also called anchors, and they predict object locations depending on their score while sliding over the feature map. Two different lower dimensional layers called class and regression are utilized

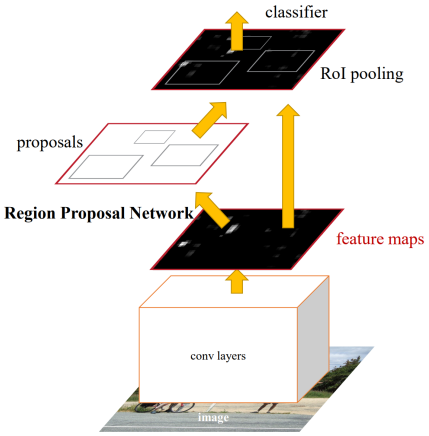


Figure 3.1: Faster R-CNN model blocks [1].

for mapping these sliding windows. These lower dimensional layers then predict a new class and bounding box location for the particular window and which returns as a loss for the training.

Faster R-CNN is quite beneficial for the scene graph and it will be the first block for the processing flow of the scene graph detection task. Faster R-CNN will provide a fixed amount of proposals and predicted object classes for the given image. The features in these proposals will be also extracted and used along with the box object labels in the next stage of the SGG task.

3.2 Scene Graph Generation

After detecting the objects in a given image, the next goal is to identify relations between the detected objects. This section investigates how an example SGG model designed by Motif [2] operates in fully supervised settings. In addition, all the experiments, and adjustments in the following chapters will be performed on this Motif model, so it is crucial to understand this relation model architecture.

Figure 3.2 displays the general diagram of the Motif model. Proposals coming from the object detector, which is pre-trained on the Visual Genome dataset, are aggregated in two different biLSTM structures. First, biLSTM is called *object context* where the inputs of the layer are the features and the labels of proposals. The aggregation formula is provided in equation 3.1 where f_i is the feature vector for i-th proposal, and $\mathbf{l}_i$ is the label for the ith proposal.

$$C = biLSTM\left([f_i\,;\mathbf{W_1}\mathbf{l}_i]_{i=1,\ldots,n}\right) \tag{3.1}$$

The *object context* $C = [c_1, ..., c_n]$ is the final form of hidden states of the biLSTM. Every proposal and its label contributes to the resulting hidden state representation. $\mathbf{W_1}$ is a learnable mapping matrix and updates its parameters during training.

According to Figure 3.2, the next part is called **Decoding** part where proposal labels are refined. Object detector outputs a fixed amount of proposals for every

image. Some of those proposals possess low confidence scores meaning that those proposals could be noisy predictions. Motif addresses this issue by adding an LSTM layer between *object context* and *edge context*, and it deals with noisy proposals by making their label equal to the background. *This is only possible in full supervision when one has object labels at the instance level.* **Decoding** part refines input proposal labels which makes the model determine which objects are background or foreground objects.

$$D = biLSTM\left([\mathbf{c}_i\,;\,\mathbf{W_2}\hat{\mathbf{o}}_i]_{i=1,\ldots,n}\right) \tag{3.2}$$

Equation 3.2 formulates how hidden states for the last biLSTM are calculated where $\hat{\mathbf{o}}_i$ is the output commitments refined by the previous **Decoding** part. $D = [d_1, ..., d_n]$ contains edge features for every proposal and these features will perform in relation calculations where every relation class probabilities will be computed. Since there are n different proposals in Figure 3.2, $n(n-1)$ relations should be computed from the scene graph model which prevents finding object relation between itself.

$$Pr(x_{i\to j}) = softmax(\mathbf{W_r}(d_i \circ d_j)) \circ f_{i,j}) \tag{3.3}$$

Finally, equation 3.3 presents the relation score calculation for each object pair. $f_{i,j}$ is the feature vector of the union box that contains the object pairs. Including union box features in relation calculation also takes spatial context around the objects into account. The matrix D calculated in *edge context* will be used for finding the relation scores for each object pair. Using $softmax$ to find the best possible predicate will provide probabilities for each relation class. That means at the output of the SGG module, a relation score matrix $S \in \mathbb{R}^{MxA}$ where M is the number of object pairs and A is the number of relation classes.

Loss Formulation. The loss for Motif has two parts. The first part is computing cross-entropy losses for the LSTM object decoder layer. Since each proposal needs to be refined as a background and foreground object, the instance-level ground truths and predictions from output return a refine object loss for LSTM by applying a simple cross-entropy loss. After sampling foreground objects through full supervision, one also needs foreground and background relation sampling between the refined foreground objects. The ground truth in full supervision allows this sampling, and it finds all the ground truth relations between detected foreground objects in the image. Hence, one has relation labels for each object pair, and the number of these object pairs does not necessarily have to be equal to $N(N-1)$ where N is the number of proposals. As these relations are discovered through the ground truth sampling whereas in weak supervision it is common to see the number of object pairs grow quadratically to calculate a loss. Finally, one computes again a cross-entropy loss for each object pair by using predicted relation and ground truth relation for that object pair.

Inference. In inference, the aforementioned formulas for the Motif compute the relation score matrix for every $N(N-1)$ object pair whereas in training makes use of only foreground objects and relations. This relation matrix contains scores for every relation class of each pair and it needs to be sorted to provide Top-K triplets

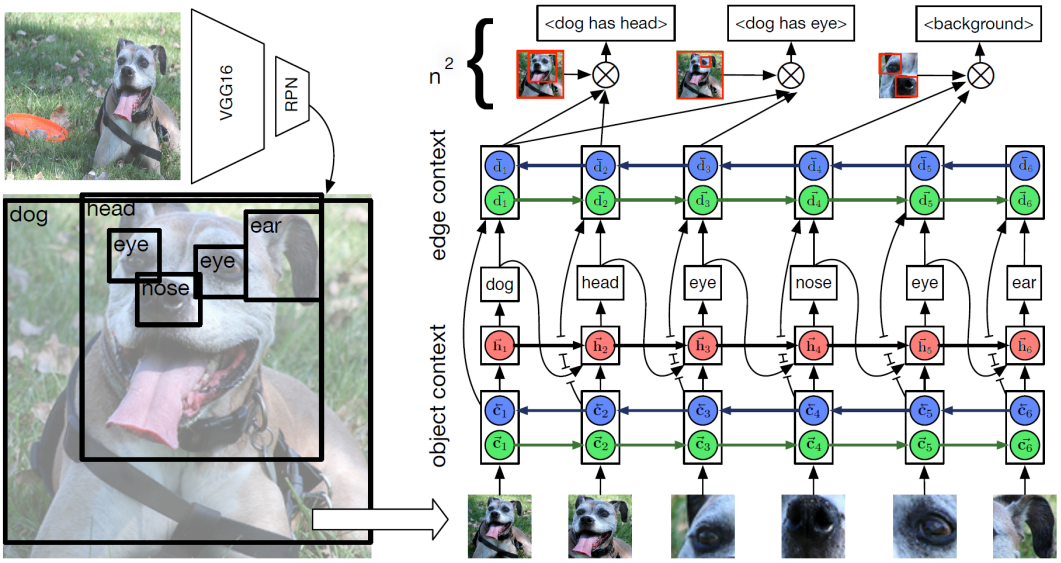


Figure 3.2: Motif model diagram [2].

for the image. The score of a triplet is calculated from a very simple equation: $Score_{triplet} = Obj_i * Obj_j * Rel_{i-j}$. The Top-K scored triplets are then evaluated through the evaluation metrics in Section 5.2.

Usually, LSTM structures require sequence inputs where previous samples are connected with future samples so that LSTM can find the patterns in the sequence training data. Figure 3.2 shows that the Motif model utilizes a sequence of proposals. Rowanz *et al.* [2] proposes to sort these proposals by confidence first, and apply them to biLSTM afterward to create a proposal sequence. The authors also experimented with different sequence orders such as proposals box sizes, random order, and left-to-right using the x-coordinate. They report all of these different sorting results; however, from the results, it is conspicuous that sorting the proposals with different settings does not have a significant impact on the results.

3.3 Weakly Supervised Setting

To turn the problem into a weakly supervised one, one needs to compute a different loss formulation different than before. Wan *et al.* [27] offers a bag-of-words approach to calculate a weakly loss in an HOI task. Being inspired by this idea, one can convert fully supervised SGG loss into a weakly one by calculating image-level relation scores after taking maximization over object pairs.

As mentioned earlier, SGG module Motif will provide a relation score matrix $R \in \mathbb{R}^{N' x N_r}$ where $N' = N * N - 1$ is the number of object pairs, N_r is the number of relation classes, and N is the number of proposals. This matrix R can be reduced into a vector $\tilde{r}$ by taking the maximum values of scores over object pairs. This results in a vector $\tilde{r} \in \mathbb{R}^{N_r}$ that presents image-level relation scores. Equation 3.4 shows

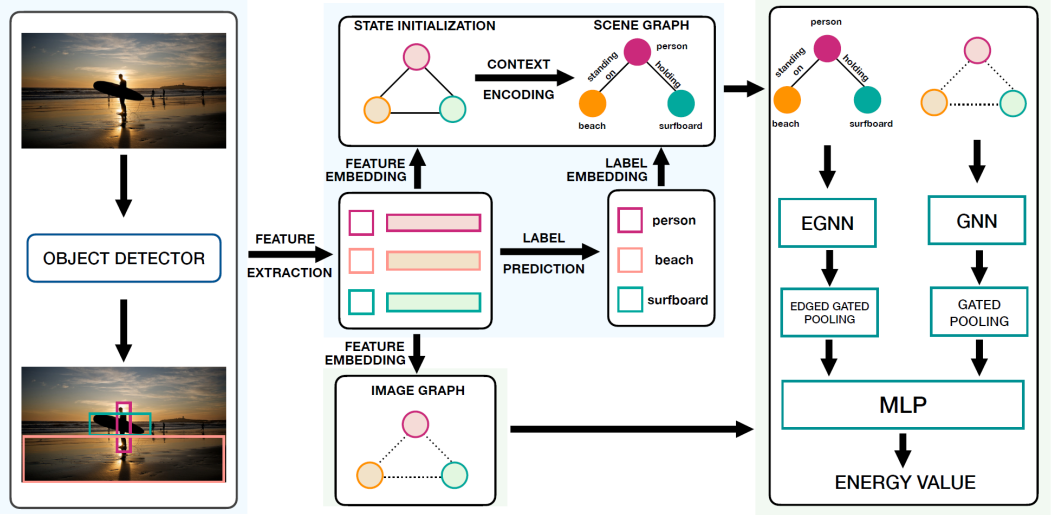


FIGURE 3.3: Energy model overview [3].

this maximization operation over object pairs where k is the k-th object pair.

$$\tilde{r} = \max_{k} R \tag{3.4}$$

3.4 Energy Based Learning

LeCun *et al.* [13] states that energy-based learning is mainly about detecting the dependencies in the input domain. One can assure this capture by using an energy model that outputs low energy values for correct inputs. Since energy-based learning learns a joint probability density function for the input space, the structure in the input domain can be modeled. LeCun *et al.* [13] emphasizes that learning a probability distribution $p(x)$ needs to fulfill two fundamental constraints: **i)** the probability distribution needs to assign a non-negative value for every input value x i.e. $p(x) \geq 1$. **ii)** the integral of probability distribution needs to be equal to 1 for each input i.e. $\int_a^b x^2\,dx = 1$. These two constraints can be satisfied for any chosen function $E_\theta(x)$ where θ is the model parameters if one chooses a Boltzmann distribution. Thus, equation 3.5 defines the new distribution for the energy model.

$$q_\theta(x) = \frac{\exp(-E_\theta(x))}{Z_\theta} \tag{3.5}$$

It is clear that the new distribution $q_\theta(x)$ satisfies previous conditions with any arbitrary function of choice. The goal is to make this learnable $q_\theta(x)$ distribution closer to *true* unknown distribution $p(x)$ by training. Although, the Boltzmann distribution provides flexible choices for function selection, the denominator in equation 3.5 is usually intractable so the training of this model seems cannot rely on

determining exact likelihoods. One popular method to train an energy model under these conditions contrastive divergence as LeCun *et al.* [13] suggests.

Contrastive Divergence. The training aims to minimize a defined loss function to teach the model about the task. According to [3, 13], in energy models, one cannot simply try to maximize the likelihoods of the input data energies i.e. maximization of $\exp(-E_\theta(x))$ cannot work since one does not know anything about the normalization function Z_θ. It is not certain that Z_θ stays constant while training; thus, minimization is not guaranteed. Most of the works address this issue by reformulating likelihood maximation with a derivative operation as it can be shown in equation 3.6.

$$\begin{aligned}\nabla\mathcal{L}_{MLE}(\theta; p) &= -\mathbb{E}_{p(x)}[\nabla_\theta \log q_\theta(x)] \\ &= \mathbb{E}_{p(x)}[\nabla_\theta E_\theta(x)] - \mathbb{E}_{q_\theta(x)}[\nabla_\theta E_\theta(x)]\end{aligned} \tag{3.6}$$

[13] provide equation 3.6 which requires sampling from the distribution of energy model. Usually, Monte Carlo Markov Chain (MCMC) methods solve this sampling issue. Suhail *et al.* [3] points out that only relative input configurations and energies are needed to find a better-predicted scene graph as input. That is why equations 3.6 and 3.7 have the same structure where minimization in equation 3.7 corresponds to MCMC sampling.

EBM Overview. Suhail *et al.* [3] proposes using energy-based approaches mainly for solving two problems about the conventional scene graph model training. The first issue is that a scene graph such as Motif sums all cross-entropy losses. This approach treats every relation triplet as a distinct entity and neglects the underlying structure within the image. Designing a graph-based energy learning strategy should incorporate prior structure in the input image and it should improve the results. The authors also state that this additional energy loss allows the model to learn from fewer data making it practical for guessing rare training words.

EBM Method. The model pipeline for Suhail *et al.* [3] is depicted in Figure 3.3. Since the authors use energy-based learning as an additional loss source, any scene graph model such as Motif or VCtree [8] can be selected by their approach.

Figure 3.3 starts from the object detector where the detector finds proposal regions and predicted labels. Suhail *et al.* [3] extracts these predicted labels and encodes them into a scene graph. Thus, the scene graph in Figure 3.3 does only contain predicted label embeddings. The scene graph also has edge label embeddings predicted from the scene graph model such as Motif. The image graph, on the other hand, is the object feature vectors extracted from the proposal regions.

The RoI features extracted from the proposal regions initialize the image graph to enhance the energy value computed for the image. Figure 3.3 continues with the novel graph-based approach for calculating the energy of the image. To incorporate the visual structure in the image, the scene graph and the image graph are sent to Edged Graph Neural Network (EGNN) and Graph Neural Network (GNN) blocks in the model where they will undergo message-passing algorithms to enhance the representation.

$$\mathcal{L} = E_\theta(G^+_\mathcal{I}, G^+_{SG}) - \min_{G_{SG} \in SG} E_\theta(G_\mathcal{I}, G_{SG}) \tag{3.7}$$

Suhail *et al.* provides [3] equation 3.7 that shows how the energy value for a given image is calculated. The ground truth scene graph label embeddings and the image graph are compared with the predicted scene graph and image graph. That means the energy-based method requires a ground truth of the image-level graph. The aim of this equation 3.7 is to bring the predicted scene graph's energy to lower values by moving towards the negative direction of the steepest descent. Suhail *et al.* states that solving such an equation requires taking derivates concerning predicted graphs; thus, Stochastic Gradient Langevine Dynamics (SGLD) methods [31] are needed to find the solution of the optimization problem of equation 3.7. SGLD is a method for sampling from a distribution; therefore, it requires some optimization steps to get closer to the answer. These steps are shown in equation 3.8 where O is the scores of the predicted objects and R is the relation score matrix for each object pair. These matrices are updated by taking a step towards a lower energy configuration while increasing the energy for random samples in the distribution. Increasing the number of iterations may be beneficial for obtaining a more accurate scene graph node and edges states; however, it also increases computational load and time.

$$\begin{aligned} O^{t+1} &= O^t - \frac{\lambda}{2}\nabla_O E_\theta(G^t_{SG}) + \epsilon \\ R^{t+1} &= R^t - \frac{\lambda}{2}\nabla_R E_\theta(G^t_{SG}) + \epsilon \end{aligned} \tag{3.8}$$

The most effective parts of the energy graph model proposed by Suhail *et al.* are the EGNN and GNN parts where the messages between nodes and edges are calculated. These messages enrich the structure predicted by the model and later they will be fed into MLP layers to come up with one scalar energy value for the predicted scene graph. One also needs to calculate an energy value for the ground truth configuration as well. The method for these blocks is investigated in Section 4 in detail but it is important to note that EGNN and GNN have the same message-passing algorithms which will be explained in Section 4

Even though Suhail *et al.* proposes using the image graph as an additional source of information. Since the RoI features for the ground truth are also required for the image graph, it is not suitable for weak supervision. Therefore, it is removed from the proposed methods. The removal of the image graph also makes the new energy model lacking the visual features information. The modified energy model needs to trust only object distributions.

3.4.1 Language Priors & Dataset Analysis

Despite the achievements of previously discussed methods for scene graph generation, it is still possible to enhance the results even more by adding prior information on predictions. For instance, one can consider the predicate word `riding`. It becomes abundantly clear that the subject of this predicate is presumably `human` or the object is probably `horse`. If one knows the predicate, subject, or object, the chances of detecting the other two words increase considerably. Another example would be the

Relations			
Type	Examples	Classses	Instances
Geometric	above, behind, under	15	228k (50.0%)
Possessive	has, part of, wearing	8	186k (40.9%)
Semantic	carrying, eating, wearing	24	39k (8.7%)
Misc.	far, from, made of	3	2k (0.3%)

Table 3.1: Relation types in Visual Genome [2].

subject-object pair of `man, shirt`. Perhaps the relation word is `wearing`, and the model should put bias towards this relation.

Language Priors. Language priors are utilized in some of the fully supervised scene graph literature [8, 14, 2] stated earlier. It is first introduced in Motif [2] where they calculated the statistics of Visual Genome. This is called Frequency biasing, and they assign a prior probability for each triplet in the ground truth. That means given the subject-object pair probability of category for the relation word is not uniform. These prior probabilities are calculated by counting the frequency of occurrences. In weak supervision, the works [16, 25, 26, 29] employ a similar strategy. Ye *et al.* [25] tries to capture these language priors from the image captions. Baldassarre *et al.* [26] utilizes the same frequency bias approach under weak supervision.

Visual Genome. Visual Genome is a large-scale, well-known dataset consisting of 108k images and 600k relations. Before going into the proposed methods, an analysis provided by the Motif [2] is worth mentioning here. As mentioned earlier, one of the main challenges in scene graph generation studies is handling the bias present in the dataset. Table 3.1 provides the relation label amount per type. Geometric and possessive relation words are the most dominant predicates in the dataset. In contrast, meaningful semantic relations such as `carrying, eating` seem to have only a minor portion of 8.7%. As mentioned earlier, one essential challenge is emphasizing the importance of semantic type predicates in training so that the generated scene graph in Figure 1.2 has a better representative of the given image.

3.5 Conclusion

This chapter provided beneficial strategies for scene graph generation tasks in detail. It also has provided a brief introduction to the Faster R-CNN object detector. A popular method Motif has been shown for the scene graph task, and how it works is provided layer by layer. How a weak supervision loss can be computed for the scene graph task is given. Lastly, how energy loss can guide cross-entropy loss from the predicted scene graph is understood thoroughly. The methods delivered in this chapter are functional to formulating the method used in this thesis in the next chapter.

Chapter 4

Method

This chapter explains the followed methodology in this thesis work. In Section 4.1, why these model selections are made is explained briefly. In Section 4.2, the problem parameters like inputs, and outputs will be defined and why this setup is prepared will be explained in detail. Section 4.3 is dedicated to discussing each processing block in the problem solution. The designed model and each block selection are further analyzed. Section 4.4 describes the loss functions used in training for the model.

4.1 Methodology Principles

While selecting the previously discussed model designs, important principles such as state-of-the-art models, popular message-passing methods, and structures in SGG are taken into consideration. These principles lead to the design choices made in the thesis work.

Detector. Faster R-CNN provides a two-stage detection task where the proposal regions are first predicted; then, they are extracted from these regions. Having a two-stage detector allows one to produce more accurate object regions both for small and large objects; hence, this architecture is chosen to find the proposal regions in this thesis.

Relation Model. In Section 2, many works in literature have been stated and compared with our selection. Since Motif is one of the state-of-the-art models and it has been used in these works, it provides a solid starting point in relation models. One can also select a different relation model since the energy model is also compatible with other models like [4, 8] but Motif is more lightweight than these models so the training needs less time for Motif than the other models.

Energy Model. For energy models, any selection for the architecture is compatible with the contrastive divergence loss due to the Boltzmann distribution; however, the message passing between objects and relations is not present in Motif-like structures. Thus, having this message passed in your energy model would be beneficial in training. Many works of literature work about the weakly supervised SGG utilizes

these message-passing algorithms in various ways. Thus, the model provided by Suhail *et al.* [3] is the best option to incorporate message passing.

4.2 Problem Setup and Method Overview

In this subsection, the proposed problem setup will be introduced. Given image I, the scene graph model $\mathcal{M}$ generates a tuple of (O, R) i.e. $\mathcal{M}(I) = (O, R)$ where $O \in \mathbb{R}^{NxN_o}$ represents the detected objects in image I. The relations between the detected objects are shown in $R \in \mathbb{R}^{(N*N-1)xN_r}$. The number of proposals, object classes, and relation classes are notated as N, N_o, N_r respectively in the previous formulas.

For weak supervision, only image-level ground truths are available. That means for a given image I, the ground truth object labels for this image $\mathcal{O}_I$ is a subset of $\mathcal{O} \in \{0, 1, ...N_o\}$. For the same image, image-level ground truth labels $\mathcal{R}_I$ is a subset of $\mathcal{R} \in \{0, 1, ...N_r\}$. In addition, the relation score matrix R has a quadratically increased dimension by the number of proposals due to weak supervision.

The designed scene graph model starts from the object detector to generate a set of object proposals with their predicted labels and RoI features. This object detector should be an off-the-shelf detector trained on a different dataset like Bo *et al.* [27] used in their work, or it should be a weakly supervised like the works in Zhang *et al.* [15, 24]. The combination of the predicted proposals is fed into the next part of the model to find out the relations. However, in this thesis, a detector pre-trained on Visual Genome is utilized to prove the concept rather than following the methods mentioned.

The next part in the training pipeline is a Motif [2] like biLSTM structure which is also called *base model*. *Base model* is a modified Motif that lacks the decoder part of the Motif. Removing the decoder part means that one has to trust the proposals and labels predicted by the object detector. Since the energy part is compatible with other relation prediction models, it is possible to replace the *base model*; however, only the modified Motif is investigated in the experiments. Motif presents a biLSTM-based structure for relation predictions. In the works [2, 8, 22], it has been shown that biLSTM-based structures are successful when it comes to understanding the context in the given image.

After the Motif, the framework introduces *energy model* which is only available in the training and compatible with any *base model* architecture. This framework enables to use *energy model* in training to guide the relation predictions arising from *base model*. *Energy model* computes and scalar energy value which represents the whole image. This thesis hypothesizes that this energy value should help the training of *object, edge contexts* of modified Motif to find better relation prediction results under weak supervision. Due to the training bias that emerged from the biased dataset and image-level training, it is also predicted that the energy model should help to mitigate the drawbacks of these biases. **The energy values for the predicted scene graph and the ground truth scene graph are used to incorporate this inherent structure.**

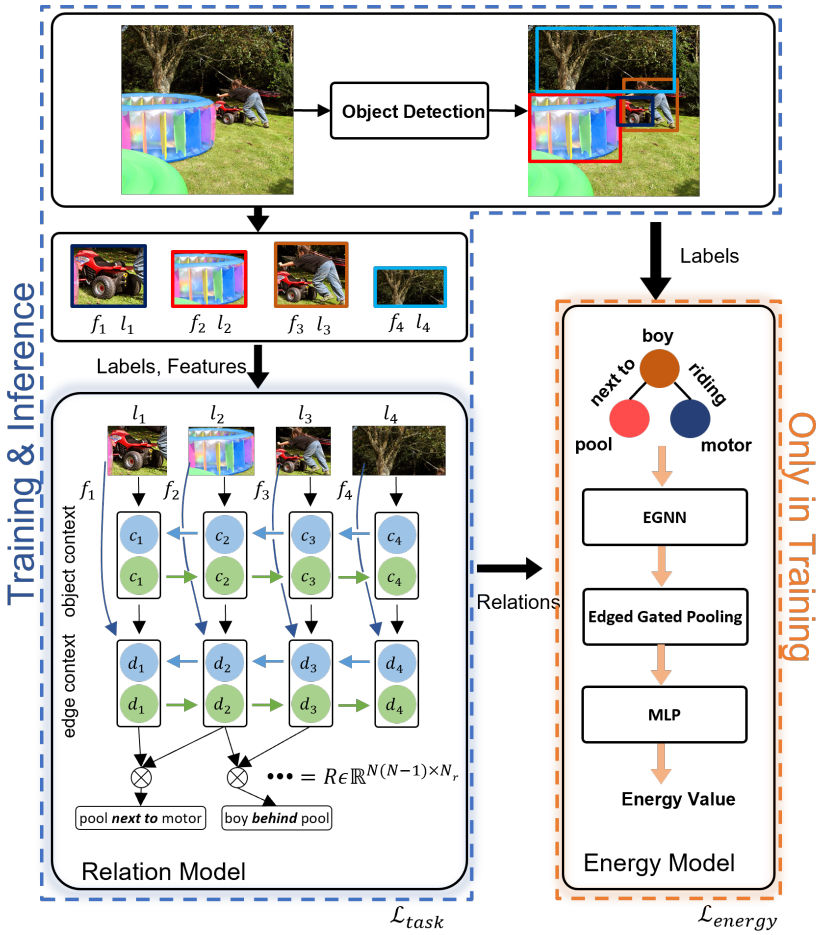


FIGURE 4.1: Method overview.

To train the model, two different losses will be defined in Section 4.4. The first of these losses is the task loss found from image-level relation labels. The other one is the energy loss calculated with a contrastive divergence-based approach.

4.3 Model Design

This subsection reviews the model overview depicted in Figure 4.1 and provides the modified versions of Motif [2] and energy-based model [3] architectures. The overall pipeline contains these two modules. Each of these modules will be examined closely to understand how they work in the next two subsections.

4.3.1 Relation Model

The first operation applied on given image I is the object detection to extract object proposal regions. This can be seen in the upper part of Figure 4.1. The predicted

bounding boxes $B = \{b_1, ..., b_N\}$ of these proposals also assigned with a feature vector $f_i \in \mathbb{R}^{4096}$ and a label $l_i \in \mathbb{R}^{200}$ for the ith proposal.[1] Since these proposals are non-contextualized and extracted from different parts of the image, they need to be mixed into each other to get a better understanding of the image according to authors of Motif [2]. They will be sent to *object context* layer in the next step.

Object Context

The proposals should create a sequence by ordering them before feeding them into the *object context* biLSTM. The Default option of left-to-right ordering is utilized for ordering the proposals. That means by using the bounding boxes B, the proposals on the left will be the first ones entering the object context. There are other options in [2] to order the proposals; however, they do not have a significant effect on the results. According to Motif [2], this first biLSTM layer makes every proposal to contribute to the relation results in the next layer.

$$C = biLSTM\left([f_i\,;\mathbf{W_1}\mathbf{l}_i]_{i=1,...,n}\right) \tag{4.1}$$

The formulation of the first layer is delivered in equation 4.1 where $C = [c_1, ..., c_N]$ are hidden states for each proposal. $\mathbf{W_1}$ is a mapping kernel for arranging the dimension of l_i.

Edge Context

The next step is to compute a relation matrix $R \in \mathbb{R}^{(N*N-1)xN_r}$. This relation matrix conveys the result of the relation scores of object pairs and its dimension increases quadratically with N. The first step is to start from contextualized object vectors in C, and use them to calculate contextualized edge matrix D as given in equation 4.2.

$$D = biLSTM\left([\mathbf{c}_i\,;\mathbf{W_2}\mathbf{l}_i]_{i=1,...,n}\right) \tag{4.2}$$

$D = [d_1, ..., d_n]$ represents the edge vectors for the proposals. However, these representation lacks global knowledge about the object pair. Therefore, the union box feature $f_{i,j}$ for ith and jth proposals needs to be extracted and integrated with the next step. In addition, the language priors discussed in Background Section 3 are added to enhance the results even further. $w_{i,j}$ is the bias vector representing the language priors.

$$R = (\mathbf{W_h}d_i \circ \mathbf{W_t}d_j) \circ f_{i,j} + w_{i,j} \in \mathbb{R}^{(N*N-1)xN_r} \tag{4.3}$$

Equation 4.3 shows the calculation of relation score matrix R. $\mathbf{W_2}, \mathbf{W_h}, \mathbf{W_t}$ are dimension mapping matrices for d_i. R matrix represents the relation distribution of the image for each pair, and it is beneficial in the next step of the pipeline.

[1] Note that l_i has a dimension of $\mathbb{R}^{200}$ since the loaded GloVe embeddings [32] has this dimension.

Inference

In the previous section, the relation score matrix R is derived from various operations on *object* and *edge contexts*. This matrix represents relation scores for each triplet and it needs to be sorted to find the best-predicted triplets for the given image I. One can use $softmax$ on R and O to calculate probabilities of relations and objects from logits. The score for the triplet is then shown in equation 4.4 where o_i^{score} is a row from O^{scores} and it denotes the object score for the ith proposal. r_{ij}^{score} is a row from R_{ij}^{scores} which provides a score for the relation between the ith and jth proposals.

$$\begin{aligned} R_{ij}^{scores} &= softmax(R) \\ O^{scores} &= softmax(O) \\ Tri^{score} &= o_i^{score} * o_j^{score} * r_{ij}^{score} \end{aligned} \tag{4.4}$$

4.3.2 Energy Model Architecture

This subsection analyzes energy model architecture presented in Figure 4.1. One should note that **the energy model is only activated in the training phase** so it is not loaded in the inference phase.

The energy model in Figure 4.1 starts from a scene graph encoding scheme where the nodes of this encoded graph are initialized from the detected proposals O, and the edges are set up from relation matrix R. Therefore, the problem setup for the energy model can be summarized as follows: Given a scene graph generation model $\mathcal{M}$, and an image I, the predicted scene graph can be denoted as $G_{SG} = \mathcal{M}(I)$. Due to the energy loss formulation discussed in subsection 4.4, one also needs ground truth scene graph configuration. Hence, the ground truth scene graph is denoted as G_{SG}^{+}, and it is initialized from **the ground truth of image-level graph** for the object labels and relations. That means the proposed energy model requires a graph-based ground truth instead of using only image-level labels. See Figure 1.1 to see the difference. The most important modification in this part is the removal of the image graph mentioned in Section 3.4.

Finally, the general formulation of the energy-based method in Figure 4.1 can be written down. For a given energy model E_θ, the scalar energy value for the predicted scene graph G_{SG} can be obtained from the equations 4.5 and 4.6.

$$E_\theta(G_{SG}) = MLP[f(EGNN(G_{SG}))] = e^{-} \tag{4.5}$$

$$E_\theta(G_{SG}^{+}) = MLP[f(EGNN(G_{SG}^{+}))] = e^{+} \tag{4.6}$$

EGNN and pooling layer $f()$ in equations 4.5 and 4.6 will be explained in detail in the next two subsections. One should also note that the energy value for the predicted scene graph is named as negative energy whereas the energy of the ground truth scene graph is denoted as positive energy.

Edged Graph Neural Network (EGNN)

Suhail *et al.* [3] proposed this novel edged graph neural architecture for their energy computations. The nodes of the graph are the detected object proposals and the relations are called edges between the nodes. These nodes and edges are predicted from the modified Motif structure that we have previously; thus, $\mathcal{M}(I) = (O, R)$. Every row of this matrix $O \in \mathbb{R}^{NxN_o}$ holds an object proposal n_j where $j \in 1, ..., N$. On the other hand, each row in relation matrix $R \in \mathbb{R}^{(N*N-1)xN_r}$ holds a relation distribution $r_{j\to i}^{t-1}$ between the node n_i and n_j. The object and edge embedding n_i, $r_{j\to i}^{t-1}$ are produced by applying the rows of O and R i.e. o_i & r_{ij} to a simple MLP layer.

To aggregate all information present in the image, a message-passing algorithm is formulated in equation 4.7.

$$m_i^t = \alpha \underbrace{W_{nn}\left(\sum_{j\in N} n_j^{t-1}\right)}_{\text{node to node message}} + (1-\alpha) \underbrace{W_{en}\left(\sum_{j\in N'} r_{j\to i}^{t-1}\right)}_{\text{edge to node message}} \tag{4.7}$$

Equation 4.7 emphasizes that each node embedding should be enriched by the neighbor nodes and edges by utilizing a weighted sum where the weight $0 < \alpha < 1$ is a trade-off hyper-parameter between node and edge messages. The reader may also realize that there is a notation t on the message m_i^t. After collecting the m_i^t, it will be applied on the node and edge states with the help of two GRUCells both for nodes and edges. The hidden state of these GRUCells are initialized from the predicted nodes and edges of G_{SG} and these hidden states are updated τ times. Thus, τ is also another hyper-parameter for message iterations.

$$\underbrace{p_{j\to i}^t = W_{ne}[n_i^{t-1}||n_j^{t-1}]}_{\text{node to edge message}} \tag{4.8}$$

Besides the node updates, there are also edge updates at the same time with a similar approach shown in equation 4.8. This equation states that the edge update message $p_{j\to i}^t$ is computed from the concatenation of nodes n_i and n_j. As explained in Section 3.4.1, the direction of the subject and object nodes is the fundamental information to predict the relation word. Hence, the node-to-edge update message should contain this direction by combining them in subject-object order where n_i is the subject and n_j is the object for this triplet.

$$\begin{aligned} n_i^t &= GRU_n(m_i^t, n_i^{t-1}) \\ r_{j\to i}^t &= GRU_e(p_i^t, r_{j\to i}^{t-1}) \end{aligned} \tag{4.9}$$

To capture dependencies and the structure in the image two different Gated Recurrent Units are utilized in equation 4.9. The intuition behind this cell is that node and edge states are updated with their neighbors and now each embedding contains information about nearby nodes and images. However, this graph neural

network representation has still a high dimension so it needs to be reduced by the next pooling layer.

Pooling Layer

To obtain scalar energy, a pooling operation should be employed on the updated node and edge states. This pooling aggregates all the information of message-updated nodes and edges into one scalar value. Suhail *et al.* [3] suggests performing a gating operation on each node and edge before adding them to each other. The function f_{gate} is a linear layer that outputs the attention score for a given node embedding n_k. The attention score is multiplied by the related node. This operation is performed on all the nodes n_k and $r_{j\to i}^{t-1}$ as it is seen in equation 4.10. Lastly, all nodes and edges are summed up to obtain two vectors namely $N \in \mathbb{R}^{N_o}$ for the nodes, and $E \in \mathbb{R}^{N_r}$ for the edges. The scalar energy e^- is obtained by the output of an MLP layer after concatenating N, E.

$$\begin{aligned} N &= \sum_k f_{gate}(n_k) \odot n_k \\ E &= \sum_{ij} g_{gate}(r_{i\to j}) \odot r_{i\to j} \\ e^- &= MLP(N; E) \end{aligned} \tag{4.10}$$

Sampling in Weak Supervision

The energy model incorporates a message-passing structure, as explained earlier. However, these messages should be only sent between the foreground objects (nodes). Full supervision enables the usage of ground truth boxes to select the correct nodes; hence, the messages between the neighboring nodes and edges are meaningful. On the other hand, the weak supervision ground truth does not possess foreground object locations, so one should find a way to sample these foreground nodes and edges.

This thesis proposes a filtering operation on ground truth image-level objects and edge labels to find out which nodes are foreground objects and which edges are foreground relations. The relation model computes $N(N-1)$ relations in the previous part, but most of these relations are unrelated and should not be utilized in the energy model. To sample these correct edges, one should first find the foreground objects. The object detector outputs a fixed amount of proposals for every given image. One can filter the background proposals by looking at their label predicted by the detector, and compare it with image-level object labels. If the proposal's label is present in the ground truth, this node should be considered the foreground node. This method eliminates some of the noisy predictions made by the detector.

The foreground relations are also sampled with a similar strategy. If an object pair label is present as a relation pair in the image-level ground truth, that pair should be considered a foreground relation pair. Thus, only these nodes require the message-passing algorithm mentioned previously.

The handcrafted Background Score & Softmax Input

In Section 4.3.2, the filtering mechanism is proposed to eliminate some of the background objects and unnecessary messages passing between these nodes. However, the energy model still needs one additional tweak to become functional.

The baseline relation model predicts a relation matrix that includes background scores. Section 4.4 introduces the binary cross-entropy loss for image-level supervision. This supervision rewards only the image-level relations but it does not guide the background score for a particular object pair. Hence, the background scores predicted by the modified Motif are not valuable. That is why another technique introduced in this thesis work is to assign these background scores for the object pair manually. This assignment is called the handcrafted background score, and equation 4.11 shows the formulation for this approach.

$$\begin{aligned} BG &= 1 - \underset{r}{\operatorname{argmax}}(softmax(R')) \in \mathbb{R}^{N(N-1)*1} \\ R &= [BG; R'] \end{aligned} \tag{4.11}$$

R' in equation 4.11 is the sliced version of the matrix R, and it does not contain the meaningless background scores predicted by the relation model. A softmax function converts the predicted relation scores into probabilities. The background score BG is assigned as the subtraction between the maximum probability for a relation category, and the maximum probability of 1. After obtaining the handcrafted background score, it is concatenated to the relation matrix R to get a better relation matrix for the energy model. Section 5.4.3 also discusses the selection of sigmoid instead of softmax here.

4.4 Weak Supervision Loss Formulations

To train the model, a multi-loss function that has three distinct terms has been proposed. Each of these loss functions aims to improve scene graph generation performance: **i)** an image-level binary cross entropy loss $\mathcal{L}_{base}$ that guides the learning towards image-level labels; **ii)** an energy loss $\mathcal{L}_e$ that concentrates on graph-level properties and underlying structure in the image; **iii)** a regularization loss for energy values to prevent gradient explosion for energy values. The formulation is given in equation 4.12 with relative weights $\lambda_e, \lambda_{reg}, \lambda_{base}$ that will be selected empirically.

$$\mathcal{L}_{total} = \lambda_e \mathcal{L}_e + \lambda_{reg} \mathcal{L}_{reg} + \lambda_{base} \mathcal{L}_{base} \tag{4.12}$$

The first term in the total loss is the energy loss calculated from the nodes and edge graph representations discussed previously. To minimize the function given in equation 4.13, one has to find a solve an optimization problem that requires sampling operation on the input data distribution. As mentioned earlier in Section 3.4, the energy of the predicted scene graph can be minimized by solving this function iteratively using Stochastic Gradient Langevine Dynamics [31].

$$\mathcal{L}_e = E_\theta(G^+_{SG}) - \min_{G_{SG} \in SG} E_\theta(G_{SG}) \tag{4.13}$$

The intuition behind the minimization is that one starts from an energy of a predicted scene graph configuration and makes this energy lower along the gradient path to find guidance for the base model predictions. The most important difference between the image-level binary cross-entropy loss and the energy loss is that the ground truth for the energy loss contains an image-level graph whereas the ground truth for the binary cross-entropy only contains relation labels. This graph-based loss supports training toward learning the joint likelihood. Also, one should note that the image-level graph is still considered to be weak supervision since no bounding box information is not used for the training.

$$\mathcal{L}_{reg} = E^2_\theta(G^+_{SG}) + E^2_\theta(G_{SG}) \tag{4.14}$$

Equation 4.14 prevents gradient overflow for the predicted energy values. Suhail *et al.* [3] investigated that learning under equation 4.13 makes energy values become very large which is usually undesired due to computational stability.

$$\mathcal{L}_{base} = -\sum_{c=1}^{C} (\mathbb{1}_{[c \in \mathcal{R}]} \log a + \mathbb{1}_{[c \notin \mathcal{R}]} \log 1 - a) \tag{4.15}$$

Finally, base model loss also called task loss is derived from a binary cross entropy loss on image-level relation class scores found by max operation in Section 4.3.1. If a relation class is present in the image, $\mathbb{1} = 1$. Otherwise, $\mathbb{1} = 0$. The base loss also requires the predicted image-level scores which are computed by the relation score matrix R via a maximization operation mentioned in Section 3.3.

4.5 Conclusion

In this chapter, the proposed method is discussed in detail starting from the problem setup, and the required modifications are listed to convert a problem into weak supervision. How the relation and energy models work with each other is also provided. Lastly, the weak supervision loss formulations are written down to understand what the model is learning. In the next chapter, the experiments performed on this proposed method will be investigated.

Chapter 5

Experimental Results

5.1 Experimental Setup

The experiment results are presented after performing tests on the Visual Genome dataset [17]. A pre-processed version of the Visual Genome by Xu et al. [4] has been employed in all experiments. Visual Genome contains 108k images with 150 object categories and 50 relation categories. However, 57,723 images are utilized in training, while 5000 images are set aside for the validation set. The test set consists of 26,446 images meaning that the split on the whole set corresponds to a 70%-30% split for the training and test phases.

5.2 Evaluation Metrics

After training scene graph models, they have to be evaluated under some settings. Most of the literature work [2, 3, 4, 8, 14] about fully supervised scene graphs employ similar test strategies to compare their results with each other. However, some of these settings are not suitable for weak supervision. The reader may find extra information about all of these settings in Appendix B. The only related evaluation setting for the weak supervision is **SGDet**.

Scene Graph Detection (SGDet): Predicting relation labels, object labels, and bounding boxes given the image.

Table 5.1 summarizes what is given and what is expected at the output for the **SGDet** setup.

Setting	Given	Output
SGDet	Image	Rel. labels, Obj. Labels, Bound. Boxes

TABLE 5.1: Inputs and outputs for test settings.

5.2.1 Relationship Recall

For the selected **SGDet** setting, the most common metrics in the scene graph community are the variations of well-known metric Recall@K. For scene graphs, the mean average precision (mAP) metric is not quite enlightening since it falsely penalizes wrong predictions. Annotating ground truth for every possible relation is impossible and some other relations which may have a close meaning should be also considered as true and they should not be penalized by the metric. Thus, next a few subsections introduce the most common metrics found by scene graph studies which possesses these properties.

To classify a predicted triplet as a correct triplet in **SGDet** setting, three conditions has to be satisfied: **i)** The object and subject is considered to be correctly predicted if their predicted bounding boxes has ≥ 0.5 IoU with the ground truth bounding boxes. **ii)** The predicted relation for this object-subject pair must be correct.

Recall @K

Rowanz et al. [2] reports their results with a conventional metric **Recall @K(R@K)** for scene graph evaluation. Zhu et al. [22] states that **R@K** computes the fraction of times correct relationship is predicted in top the most confident K predictions. K is usually equal to three numbers: 20, 50, and 100. Selecting K a lower value forces the results to be in a stricter limit than higher values like 100.

$$R@K = \frac{|TopK \cap GT|}{|GT|} \tag{5.1}$$

Equation 5.1 illustrates for a given image how to calculate a **R@K** value for only one image. To collect the results for whole test set, all recall values in every image are averaged out. One should also remember that the most confident triplets are collected via the equation 4.4.

Mean Recall @K

An enhanced version of **R@K** is the metric **mean Recall @K(mR@K)** proposed by Tang et al. [8] and Chen et al. [33]. **R@K** is mostly a biased metric because Visual Genome is dominated by trivial relations predicate words such as 'on,' 'has.' To prevent the adverse effects of such bias caused by the dataset, **mR@K** retrieves each predicate separately, and calculates a **R@K** for every one of them and takes the average. Such a new metric results in more reliable results a scene graph models since it achieves to get contribution from every predicate word.

The calculation of mR@K is pretty straightforward. Firstly, one has to iterate over all images, and collect all R@K values for the predicates using the formula in 5.1 where GT is equal to a list of triplets containing a particular predicate such as `on`, and Top-K will be equal to the predicted triplets that contain `on`.

$$\left[\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix}_{on} \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix}_{has} \cdots \begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \end{bmatrix}_{above} \right]$$

The collection matrix $\in \mathbb{R}^{N_r}$ above is written to display the procedure for mR@K. Each column of this matrix contains a column vector inside. These column vectors represents the appended image R@K score for each predicate separately. Thus, a_{ij} shows the R@K score from i-th image for the j-th predicate word. Taking the mean of the column vectors delivers the R@K result for the predicate in that column. After taking the means for every column vector, one can take the average of these mean scores to calculate mR@K for the test result. One should note that some images contains only a couple of predicate words, so most of the time R@K for other predicates are equal to zero. If a calculated R@K value a_{ij} is zero, it is not appended to the column vectors above to have a meaningful value after averaging.

5.3 Implementation Details

Detector. A pre-trained Faster R-CNN [1] with ResNeXt-101-FPN backbone [34, 35] is utilized for obtaining the proposal regions and the weights of these layers are freezed during the scene graph model training. In addition, these weights are acquired by the study of Tang et al. [36]. It should be noted that using a pre-trained detector on Visual Genome under weakly settings is illegal because the pre-training for the detector is done with fully supervision. However, in this work the aim was not acquiring a state-of-the-art result but it is prove the concept of effectiveness of energy-based method in weakly settings. Due to pre-training on objects, the experiment results will be a bit boosted comparing with existing literature works. The number of proposals is selected as 30. The detector outputs 151 object classes (including background class) for 30 proposal regions.

Scene Graph Model. The scene graph model with only cross-entropy loss is trained with a Stochastic Gradient Descent(SGD) optimizer. The learning schedule of the training follows a warm-up phase where the learning rate is increased to 0.01 linearly until 400 iterations. The learning rate is reduced by a factor of 0.1 at every plateau region. The plateau region means that the last 3 validations did not produce any significant improvement in the results. The validations are only applied at every 2000 iterations. SGD also has a momentum of 0.9 and a batch size of 4 in every experiment. The weights of the model are randomly initialized and they have a weight decay of 0.0001. The frequency bias information is incorporated in the training phase to enhance performance. The hidden state dimensions of biLSTMs in modified Motif are selected as 512. The output for the baseline model predicts 51 relation classes for each object pair.

Energy Model. The energy model itself also has some specific hyper-parameters. Firstly, the number of iterations in the message passing step is set to 3. In the SGLD optimization step, the gradient of the nodes and the edges are taken and added back in order to update them. At each optimization update, the node and edge states

Notation	Explanation	Dimension
$N = 30$	Number of region proposals	
$N_r = 50$	Number of relation classes	
$N_o = 150$	Number of object classes	
O	Detected objects matrix	$50 * 151$
R	Relation score matrix	$(50 * 49) * 51$
W	Kernel layer to map dimensions	Depends on usage.
o_i	i-th proposal's object dist.	$1 * 151$
r_{ij}	Relation dist. between i-th & j-th	$1 * 51$
f_i	Features for i-th proposal	$1 * 4096$
l_i	GloVe embedding for i-th proposal	$1 * 200$
c_i	Obj. context biLSTM hidden dim.	$1 * 512$
d_i	Edge context biLSTM hidden dim.	$1 * 512$
$f_{i,j}$	Union feature vector for (i,j) proposal pair	$1 * 4096$
$w_{i,j}$	Frequency bias vector for (i,j) proposal pair	$1 * 51$
n_k	Node embedding for kth proposal	$1 * 512$
$r_{j \to i}$	Edge embedding between i-th & j-th proposals	$1 * 512$
m_i	Incoming msg vector for the i-th proposal	$1 * 512$
$p_{j \to i}$	Incoming msg vector for the rel. vector of (i,j) pair	$1 * 512$
N	Node vector representation after pooling	$1 * 512$
E	Edge vector representation after pooling	$1 * 512$
e^-, e^+	Scalar energy values	$1 * 1$

Table 5.2: Hyper-parameter selection & Dimensions for the notations.

are scaled back to the range of [0, 1], and the gradients of these nodes and edges are clipped between [-0.01, 0.01]. The node and edge embeddings has a size of 512. The NVIDIA RTX 3090 is mainly used for training both baseline and energy models. The relative weights $\lambda_e, \lambda_{reg}, \lambda_{base}$ are selected as 1 empirically.

5.4 Experiments

In this section, the test results of the proposed method are shared and explained. The experiments start from the reproduction results to have a solid starting point for the thesis. For the weak supervision test, two models namely only the relation model and the relation+energy model are compared along the experiments where the relation model is the modified Motif model trained with only image-level binary cross-entropy loss. The second model is the activated energy part for the modified Motif and its results show the contribution of energy-based learning for training.

5.4.1 Experiment 1 - Reproducing Literature Results

This experiment illustrates the results of the reproduction of the previous works under some slight modifications for the energy model. Under full supervision, how much increase is obtained if the image graph of the energy model part is removed will be explored. This means that in this experiment, the decoder part of Motif is not removed but only the image graph is deactivated for the experiment. The reproduction under this modification establishes a solid fundamental for weak supervision tests since the image graph cannot be utilized in weak supervision.

Table 5.3 illustrates the retrieved results from the reproduction both for the energy-based and cross-entropy models under full supervision. The first and second rows refer to the results of previous works, and one can easily find them in the works [14] and [3]. The third row shows the results of the reproduction, which has comparable results with the first and second rows. The energy model in the reproduction improves the cross-entropy baseline model in mR@20-50-100 metrics smaller than the previous works. mR@100 metric is increased by 0.52 points in this thesis whereas Suhail *et al.*[3] has a 1.15 points growth in their work. On the other hand, a stricter metric mR@20 has 0.31 points rise in this thesis, whereas Suhail *et al.* [3] has a boost of around 0.6 points. This result indicates that the reproduction seems to have similar behavior to the previous work. The hyper-parameter selections and the image graph removal caused a less increase for the energy model results rather than a sharp one. The reproduction does not have the same hyper-parameters entirely as in [3]. For instance, the effective batch size is 4 in this thesis, but setting this parameter as 16 would be a better choice as they do in [3].

R@20-50-100 results have a trend of decreasing for the original work in the second row in Table 5.3. The energy model is designed to get rid of the bias of the most common predicates such as 'on,' 'has.'. If one gets rid of the bias in training, this will reduce the predictions with these frequent words. That leads to a decrease in R@20-50-100 when the original work activates the energy model since R@K metric depends on these frequent predicates more. On the other hand, the reproduction does not seem to affect R@20-50-100 values when switching energy model on. This result is analyzed further by the next Figure 5.1.

Figure 5.1 illustrates a thorough analysis made for predicates in the R@100 metric. It displays the difference between the R@100 results of the energy model and only the cross-entropy model. Thus, the green bars indicate an increase for R@100

Figure 5.1: Analysis of each predicate word separately for no image graph energy model under full supervision.

Model	Method	Scene Graph Detection					
		R@20	R@50	R@100	mR@20	mR@50	mR@100
Motif [14]	Cross Entropy	25.48	32.78	37.16	4.98	6.75	7.9
	EBM-Loss	-	-	-	-	-	-
Motif [3]	Cross Entropy	25.62	32.97	37.41	5.07	6.91	8.12
	EBM-Loss	24.39	31.74	36.29	5.67	7.71	9.27
Motif(Ours)	Cross Entropy	25.13	32.01	35.9	4.91	6.73	7.89
	EBM-Loss	25.13	32.05	35.86	5.22	7.12	8.41

TABLE 5.3: Paper reproduction results.

in the energy model, while the red bars point out a diminish for the energy model. It turns out that the reproduction caused some decreases for the common words such as 'near,' 'with,' 'above, and 'holding' though it did not affect the most recurring ones such as 'on,' 'has.' This explains the values of the third row in Table 5.3 for R@20-50-100. Figure 5.1 also shows that a sharp increase for the semantic predicates like 'standing on,' 'carrying,' 'eating,' and 'riding.' This sharp increase causes us to obtain better mR@20-50-100 results than only the cross-entropy model.

A slightly lower result in R@K for the reproduction of the energy model was expected but R@K values stayed quite close to only cross-entropy model results. These results are assumed to be caused by not having a perfect reproduction due to the hyper-parameter selection and the removal of the image graph.

Besides the red and green bars in Figure 5.1, there are predicates with no bars. These predicates did not have any successful prediction in evaluation; hence, they do not have a difference indicating bar for R@100. These predicates usually belong to the semantic type of relation classes. When they are detected, they provide a highly meaningful scene graph; however, due to the sampling frequency of these predictions, they are the most challenging ones to detect. The removal of the image graph affected mostly these predicates in the evaluation.

Another significant result to mention is that R@K and mR@K are seen to be in a balance which means preventing the bias causes you to get lower R@K values as R@K is mostly dependent on the most common predicates. Having a method like the energy method rises your mR@K values as in the second row of Table 5.3 but this causes a decrease in R@K values. The third-row reproduction did not get this decrease for R@K; however, selecting the right parameters should boost the performance of mR@K which should directly affect R@K results too. Since reproduction seems to have a less increase for mR@K, R@K values are not affected too much.

The frequencies of predicates could be seen from Figure A.1.

5.4.2 Experiment 2 - The weak supervision test results

Experiment 2 is the main experiment of the thesis that displays the performances of the only cross-entropy model, and the energy method-activated model under weak supervision. It also compares these results with other state-of-the-art studies; however, the comparison is unfair for the literature studies due to the pre-trained object detector used in this thesis.

Table 5.4 summarizes the overall results and comparisons with other studies. The first row shows a fully supervised model Motif results from [3]. The second row consists of state-of-the-art models from literature that shows the current best results for weak supervision. The third row contains the baseline cross-entropy, and the energy model to prove that the energy-based methods are helpful in SGG tasks.

One should start from the third row to examine Table 5.4. The pre-trained object detector achieves 0.27 mAP on the test set, and this leads to the values seen in the third row of Table 5.4. The cross-entropy model achieves to detect the relations in the given image concretely when one compares it with the first row. That means the image-level relation loss presents a high baseline for the relation model-modified Motif. The energy model improves these mR@20-50-100 values marginally. However, R@20-50-100 values seem to get a drop of 0.89, 0.84, and 0.39 respectively while activating the energy model. This drop in R@K is expected because the energy model should put more emphasis on non-frequent predicates but it was assumed to get more performance for the mR@K values in the energy model.

To gain deeper insights into the results mentioned in the third row of Table 5.4, R@100 values for each predicate separately are depicted in Figure 5.2. Figure 5.2 illustrates a trend having a decrease for the frequent predicates, and a boost for the uncommon predicates. This behavior is similar to the full supervision results in Figure 5.1 as it also improves the predicates on the tail, and slightly reduces the ones around the head of the distribution. For instance, the semantic predicates such as 'standing on,' 'carrying,' 'eating,' and 'riding' have an increasing trend when the energy model is activated whereas the common ones like 'on,' 'has' have diminished.

In addition, the maximum improvement for a predicate in Figure 5.1 corresponds to 'eating,' and this growth is around 0.08 points in R@100. The maximum difference in Figure 5.2 belongs to again 'eating' while this time having a difference of around 0.04 points in R@100. This result states that even under weak supervision, the improvements are still similar to full supervision in terms of values. Even though the values in Table 5.4 seem to only possess a marginal increase for mR@20-50-100 in energy models, examining the predicates in detail shows that the energy method also tackles down the long distribution bias of Visual Genome.

Even though it has been proved that the energy-based method helps to cross-entropy model by tackling the problem of bias in the dataset, there may be some solutions to enhance the values in mR@20-50-100 more. The marginal increases are assumed to be caused by several reasons:

i) The model cannot be trained to provide meaningful background class scores for relations with image-level labels. Section 4 suggests solutions for this issue but these methods may not be the optimal solutions for assigning a background class.

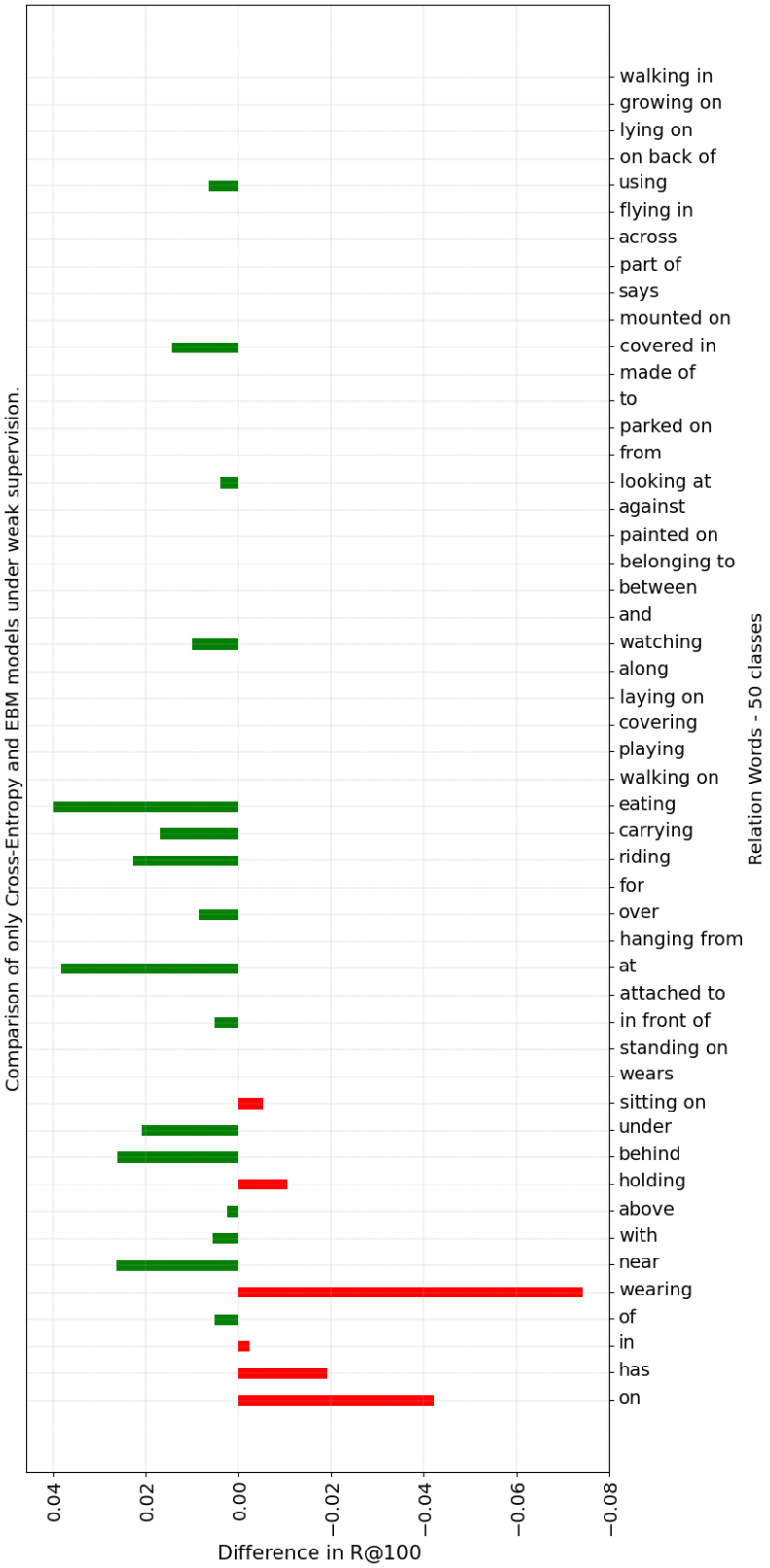


FIGURE 5.2: Analysis of each predicate word separately for energy model under weak supervision.

Model	Supervision	Scene Graph Detection					
		R@20	R@50	R@100	mR@20	mR@50	mR@100
Motif [3]	Full	25.62	32.97	37.41	5.07	6.91	8.12
VSPNet [15]	Weak	-	4.70	5.40	-	-	-
LSWS [25]	Weak	-	7.30	8.73	-	-	-
WSGM [29]	Weak	4.12	5.59	6.45	-	-	-
Motif(Ours)	Weak-CE	22.25	28.51	31.75	3.14	4.58	5.59
	Weak-EBM	21.36	27.67	31.36	**3.22**	**4.70**	**5.74**

Table 5.4: **Quantitative results.** The weak supervision test results for CE & EBM models, and comparison with other studies in literature.

Moreover, the binary cross-entropy weakly loss causes to have a smoother predicted relation class distribution than full supervision. To prevent this smoothness and solve background class issues, some solutions are offered in Section 4. Despite these solutions, in debug sessions it is observed that the inputs of the energy model after training seem to have still high values for more than two label categories in some object pairs which is undesirable for the energy model as it needs sharp inputs. Thus, the proposed solutions achieve minor successes for mR@K values while still proving the concept by having the same trend as in full supervision.

ii) The noisy proposals introduced by the object detector disturb the energy model training process. In full-supervision, the inputs of the energy model for object scores contain a learnable distribution. That means along the training, the object scores are also updated and classified as background or foreground. However, this study does not refine the predicted proposals instead it uses them directly. Along with training, one common behavior is that the node-to-node message kernel usually slowed down early than any other layer of the model due to using the object score from the detector directly.

The second row in Table 5.4 displays some weak supervision SGG results in the literature. Even though these studies utilize similar image-level weak supervision losses in training, they use off-the-shelves detectors pre-trained on different datasets than Visual Genome. Hence, it would be unfair to compare the performance of this thesis with these studies. Moreover, this thesis considers using the frequency baseline prior information calculated for Visual Genome. This prior information puts more bias towards particular relations for the given object pair labels. This statistical bias can explain the enormous R@K gap between the state-of-the-art model and our results. Most weakly supervised papers do not consider using the direct statistical information of Visual Genome. For instance, [25] learns this prior knowledge via captions.

Model	Method	Scene Graph Detection					
		R@20	R@50	R@100	mR@20	mR@50	mR@100
Motif(Baseline)	CE	22.25	28.51	31.75	3.14	4.58	5.59
Motif (No sampling)	EBM	22.22	28.43	31.67	3.02	4.40	5.41
Motif (+fg/bg)	EBM	21.90	28.17	31.44	3.02	4.51	5.50
Motif (Sigmoid)	EBM	21.43	27.94	31.43	3.10	4.56	5.58
Motif (+Softmax)	EBM	21.36	27.67	31.36	**3.22**	**4.70**	**5.74**

TABLE 5.5: **The ablation studies.** The summary of all contributions for the proposed methods.

5.4.3 Ablation Studies

In this section, the selected ablations have been applied to the proposed method to show their contribution to overall results. Motif (+Softmax) results in Table 5.5 show the best energy model results discussed earlier, and Motif (Baseline) is the best baseline results for the only cross-entropy model.

Foreground/Background Sampling

The most significant difference between full and weak supervision is that full supervision offers foreground objects and relations; thus, the loss formulation simply becomes a cross-entropy rather than an image-level binary cross-entropy. This thesis proposes a sampling technique under weak supervision using the image-level object and relation labels as mentioned in Section 4. The effectiveness of this method is shown in Table 5.5. This Table contains the results for the baseline and final energy models at the first and last rows respectively. Motif (No sampling) displays the results without applying any relation sampling. As you may recall, the baseline model predicts relations for every possible object pair in the image. However, using the same strategy in the energy model leads to a decrease in performance as mR@100=5.41 is obtained. Therefore, only meaningful object pairs' are utilized as the inputs of the energy model instead of using every relation. The addition of fg/bg sampling is shown in the row with Motif (+fg/bg) model name. Motif (+fg/bg) improved the results of Motif (No sampling) for every mR@20-50-100 metric. The diminish in R@K values is also expected since the energy model starts to predict the predicates from the tail of the distribution instead of the biased ones.

Softmax or Sigmoid Inputs

Table 5.5 also consider two additional settings called Motif (Sigmoid), and Motif (+Softmax). This ablation is performed to see how the inputs of the energy model should be selected. During experiments, it is noticed that the energy model is functional only if the inputs possess a sharp distribution. That means the relation distribution computed by the baseline part and the object distribution provided by the detector should contain distinct values for a particular class. For instance, an

object pair should not have two large scores for two different relation classes. Since the relation model is trained with binary cross-entropy losses, the obtained relation distributions for every object pair do not have a sharp distribution and do not have an instructive background score.

Section 4 suggests a handcrafted background score for each object pair but that method does not guarantee a sharp distribution for the background class too. The final method Motif (+Softmax) utilizes a softmax function to modify the inputs provided from the baseline model. Softmax function rewards the highest relation class for the energy model part. Motif (Sigmoid) is a similar approach that considers replacing the softmax with the sigmoid function while keeping all the other methods the same. In the experiments, the Motif (Sigmoid) achieved some improvements on Motif (+fg/bg) model in Table 5.5; however, it is still a sub-optimal result compared to Motif (+Softmax) as it does not provide better results for mR@20-50-100 metrics.

Another significant result from Table 5.5 is that every additional method suggestion improved the overall mR@20-50-100 metrics. Motif (+fg/bg) offered 0.11 points of advancement over Motif (No sampling) in mR@100. The addition of sigmoid and softmax activations for the energy model inputs also contributed to mR@100 0.08 and 0.28 points. Thus, there is a gradual but marginal increase in mR@20-50-100 with every addition of layers which proves these methods are beneficial for the overall model. On the other hand, R@20-50-100 metrics seem to get worse at every additional method but this result is also expected since the balance between mR@K and R@K metrics. The energy model puts more emphasis on semantic predicate thanks to the proposed methods which decreases the biased metric R@K.

The comparison of softmax and sigmoid functions above seems a bit counter-intuitive at the beginning because the baseline model is trained with a binary cross-entropy loss. If the outputs of the baseline part calculate the scores with a sigmoid function why the energy model seems to work better with the inputs from a softmax function? The answer to this question reveals itself in the energy model loss. The energy loss is not calculated from the image-level labels but they are found from image-level graphs. Hence, a competition between the relation classes for an object pair is required for the energy loss but rather rewarding the most confident relation label for this pair is more important.

The ablation for the handcrafted background score is not mentioned separately in this subsection because it is the main part of Motif (+Softmax) and Motif (Sigmoid). See Appendix C to examine the in-depth analysis of Motif (sigmoid).

5.4.4 Qualitative Results

The trained models' performances should be compared by checking the results for the example images in the test set. Figure 5.3 illustrates the differences between the cross-entropy and the energy model in four example images. The energy model depicted in the purple box generates more instructive scene graphs for given images as opposed to the baseline model. Although both models generate similar relationships for most object pairs, the energy model identified some critical object pairs and found a better predicate. For instance, the energy model determined `<man eats pizza>` triplet instead of putting a more biased triplet such as `<man has pizza>`.

An intriguing finding from the qualitative results is that the images with obvious large objects lead to better performance for the energy model. For example, the relation `ride` is detected in the third image in Figure 5.3. This third image contains two large objects: 'man' and 'horse.' The energy model seems to achieve finding a better predicate when the detected objects are confident by the detector. Thus, not having a confident object distribution for proposals seems to be one of the bottlenecks for the current setup.

5.5 Conclusion

In this chapter, the experiment results for baseline and energy models are conducted under full and weak supervision. The full supervision experiment is performed by removing the image graph part of the energy model and it has been observed that the removal of the image graph affects the prediction of instructive relation predicates but the energy model has improved the overall performance of the baseline model in terms of mR@20-50-100. Then, the same experiments display a similar trend for the energy model under weak supervision. The success of the same semantic predicates has increased even under weak supervision which also proves one of the research questions of this thesis. Finally, the ablation study results proved the effectiveness of the proposed methods. The marginal improvement for the energy model is only gathered after combining every proposed method according to ablations.

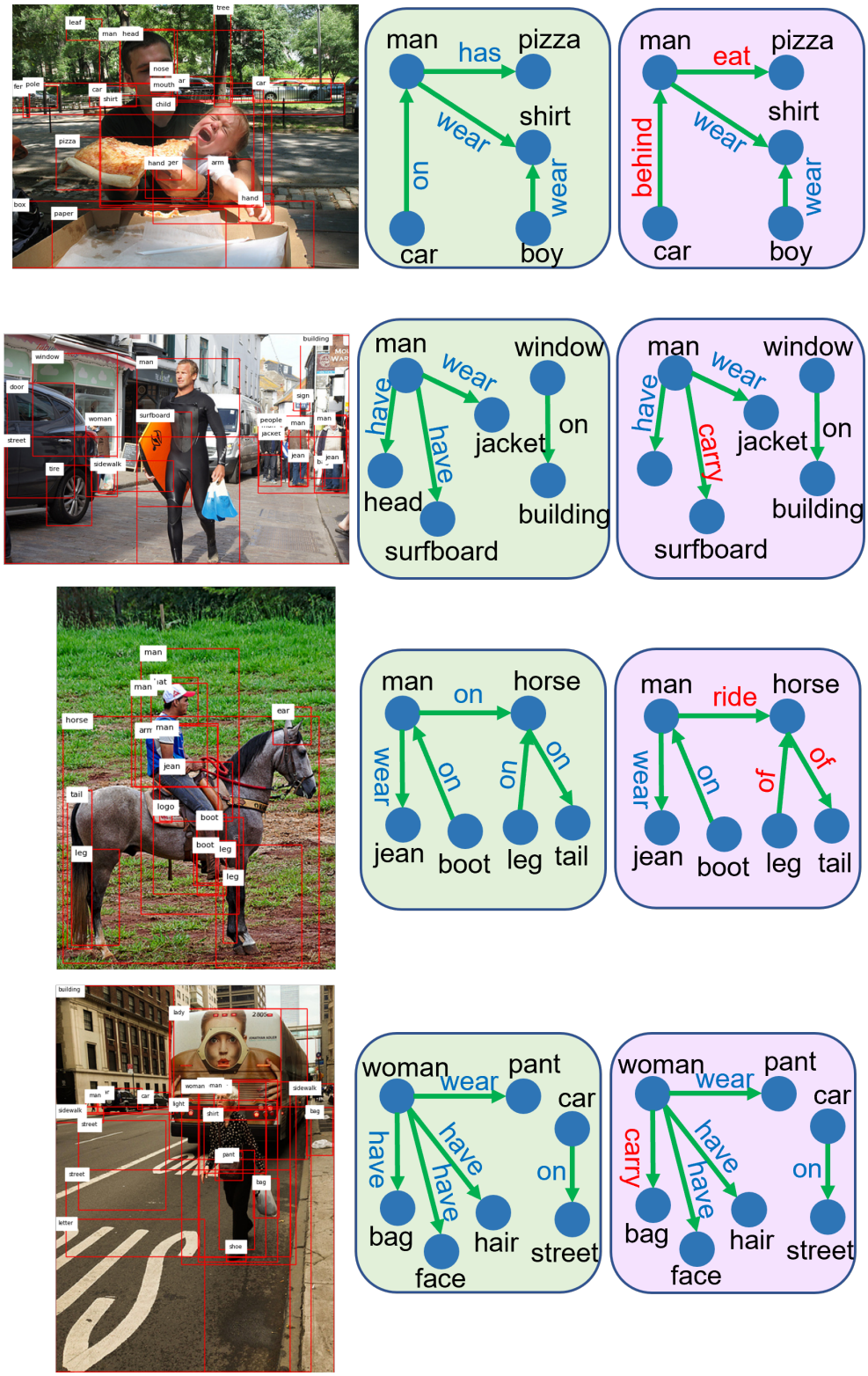


Figure 5.3: **Qualitative results.** Visualizations from scene graph detection both for cross-entropy (in green) and energy model (in purple)

Chapter 6

Conclusion

This section concludes the thesis work with the contributions presented, the confronted limitations, the answers to the initial research questions, and how the direction for future work should take shape.

6.1 Summary of Contributions

The thesis study aimed to explore the effectiveness of energy-based learning in weakly supervised SGG tasks and proposed several contributions to reach this primary goal. The following list summarizes all the efforts committed in the thesis:

i Some modifications on a state-of-the-art relation model have been applied to convert it compatible with weakly supervised training. This modified Motif provided concrete results despite being trained only on image-level labels.

ii Similar architecture alterations are also utilized in the energy model to adapt it to the same weak setting.

iii The designed filtering and sampling mechanism in the energy model improved the results of the baseline model marginally while having the same improvement trend for the rare predicates as in full supervision.

6.2 Limitations

During the experiments, the trained baseline and energy models faced some limitations which diminishes their overall performances. As mentioned in the Challenges section 1.1, trusting an off-the-shelves object detector affects the overall performance of the energy model adversely. The baseline model performs still concretely despite using noisy proposals; however, the energy model is observed to be very sensitive to its inputs. To mitigate the effect of the noisy proposals, the softmax activation is proposed to obtain a sharper object and relation distributions.

Secondly, the background score produced by the baseline is useless since binary cross-entropy training only offers supervision for 50 predicate categories. The hand-crafted background score is suggested to solve this issue but it seems to be still a suboptimal method to find out the logit for this category.

6.3 Revisit Research Questions

At the beginning of the thesis several research questions has been asked to provide a general guideline for the research. This section tries to answer these questions depending on the observations of the model performance.

i *What results can be obtained by only considering the image-level relation labels?*

Even with only image-level relation labels, the baseline model seems to produce reasonable R@K and mR@K results. The pre-trained detector also has benefits to getting this result; however, it provides a reasonable starting point for the rest of the work.

ii *To what extent do energy models mitigate the effect of biased training from the image-level relations?*

According to the experimental results, the modified energy model mitigates the drawbacks of biased training and dataset marginally due to the limitations discussed in the previous subsection. However, it also possessed the same improvement trend as in the fully supervised reproduction results. That means the energy models can deliver their benefits even under weakly supervised training.

iii *Can energy models work under weak supervision even if the foreground or background objects are absent in the ground truth?*

Yes, they can perform under these settings thanks to the proposed filtering & sampling techniques in Section 4. Even if the foreground boxes are absent, one can use the image-level labels to enhance the baseline results marginally. In addition, the Ablation Study section 5.4.3 displayed the effect of sampling.

6.4 Future Works

This thesis aimed to prove that the energy-based methods contribute to the cross-entropy-based models in SGG tasks even if the model is trained under weak supervision. The results have shown that the improvement with the proposed methods follows the same trend as in full supervision for the rare predicates with some marginal advancements. Hence, this research still seems to be worth exploring.

Caption-based supervision

As stated before, weak settings are preferred because full supervision needs extensive human effort, and it possesses a considerable amount of annotator errors. The weak supervision allows the designer to collect a dataset easier leading to more data in training. The image-level labels and graphs are still considered the easy part of weak supervision. A more realistic and scalable approach is to convert the training into image-caption-based training. This caption training should be the main direction of this thesis, and to show that energy-based methods are beneficial even in weaker settings.

Revisit Background Score

The final experiment results show that the handcrafted background score seems to be suboptimal still because the improvements are minor after the training. The performance of the energy model seems to be highly related to meaningful background scores. Therefore, one may need to find a better solution for this part.

Object Detection

Another hypothesis that may help is to supervise the object detections too. Even though the detector provides 0.27 mAP performance for the test set, the object distributions are not helpful for the energy model. A refinement layer for the detections may be helpful along the process by removing the background objects completely. Since the energy model utilizes the graph-based supervision that contains objects, it will also help the refinement of these foreground objects thanks to energy loss. However, this supervision for object refinement should be done in a weak setting without breaking any rules.

Comparison with Literature

Besides training a detector, an off-the-shelves detector is needed to compare the actual performance of the weakly supervised baseline. Since the current object detector is pre-trained on Visual Genome, comparing this model with state-of-the-art models is unfair in weak supervision. This addition should be one of the priorities for future work to see how the modified Motif is performing with other models under weakly supervised training.

Appendices

Appendix A

The Sampling Frequencies in Visual Genome

This Appendix holds the data distribution of Visual Genome dataset. This distribution possess a long-tailed behaviour as it can be seen in Figure A.1. The predicates like 'on,''has,''in' almost occurs most of the time. This also shows the noise of annotator too since it only has small amount of semantic type predicate such as 'flying in.'

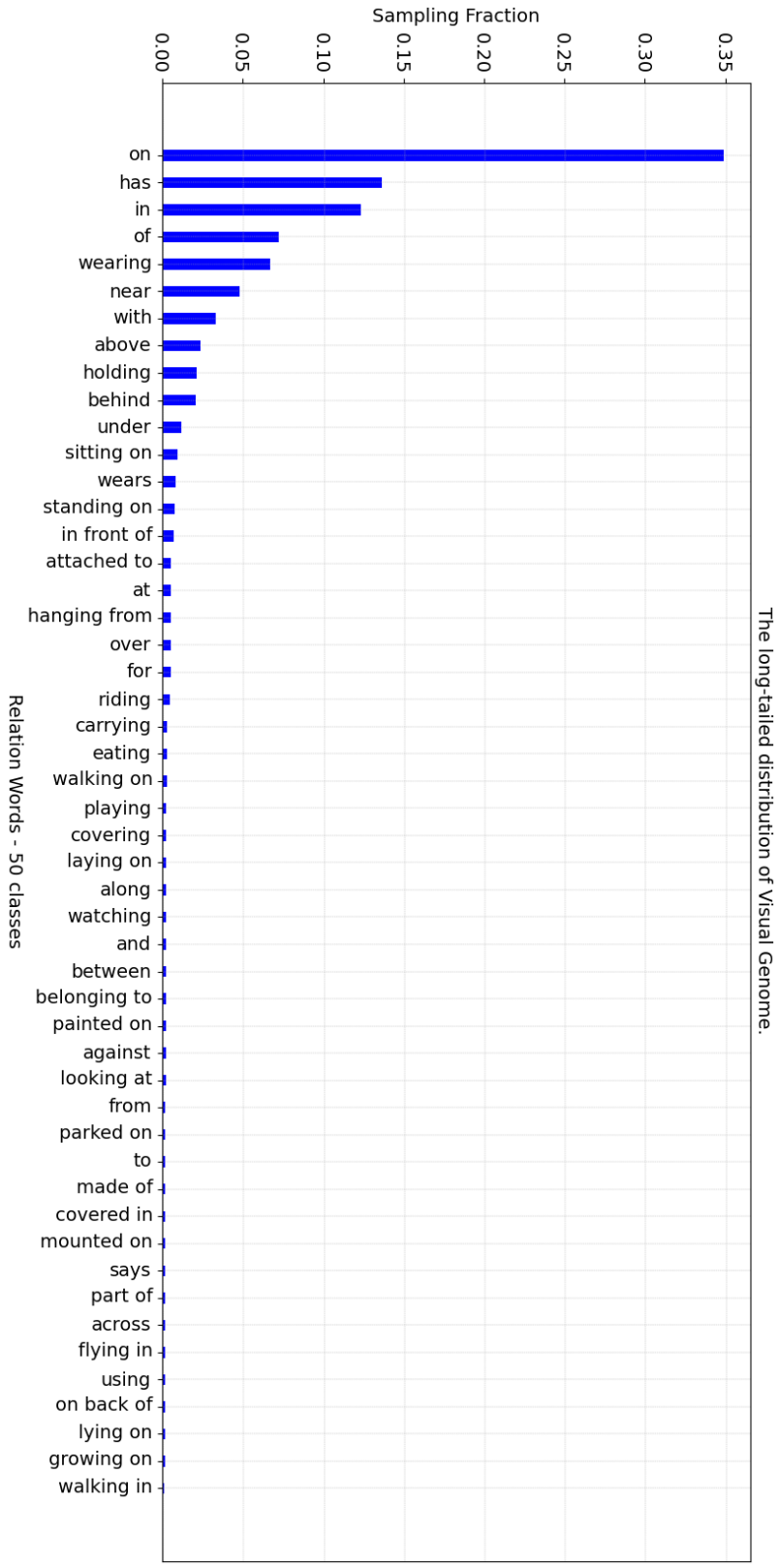


Figure A.1: Analysis for the total frequencies of each predicate in Visual Genome dataset.

Appendix B

Evaluation Metrics in Literature

The scene graph detection task contains different evaluation metrics in literature. In this work, only **SGDet** is reported in Section 5 since **SGDet** is the only suitable metric for weak supervision. However, there are also some other evaluation metrics for SGG in full supervision. This appendix introduces them in detail to show their differences.

B.1 Evaluation Metrics

This section provides the most popular evaluation settings for the SGG task in the studies like [3, 2, 4, 8, 14]. Since full supervision allows object label, and bounding boxes usage, many different settings can be arranged to observe the performance.

Predicate Classification (PredCls): Predicting relation labels, given the image, bounding boxes, and object labels.

Scene Graph Classification (SGCls): Predicting relation and object labels, given the image and bounding boxes.

Phrase Detection (PhrDet): Predicting relation labels, object labels, given the image.

Scene Graph Detection (SGDet): Predicting relation labels, object labels, and bounding boxes, given the image.

Table B.1 summarizes what is given and what is expected at the output for different setups. For instance, **PredCls** utilizes bounding boxes and object labels in training and enhancing their proposal features by mixing them with ground truth

Setting	Given	Output
PredCls	Img., Bound. Boxes, Obj. Labels	Rel. labels
SGCls	Img., Bound. Boxes	Rel. labels, Obj. Labels
PhrDet	Img.	Rel. labels, Obj. Labels
SGDet	Img.	Rel. labels, Obj. Labels, Bound. boxes

TABLE B.1: Inputs and outputs for test settings.

label and bounding box embeddings. Thus, results for this setting should clearly be drastically higher than the other settings.

Even though fully supervision provides three different settings for evaluation, in weakly supervision first two settings namely **PredCls** and **SGCls** are usually discarded. That is because weakly supervision avoids using bounding box information both in prediction and ground truth. Therefore, only setting that is significant for weakly supervised scene graphs is **SGDet**.

Appendix C

The Ablation Study - Sigmoid Activation

This Appendix delivers a detailed analysis of the Motif (sigmoid) introduced in Section 5.4.3. The same trend also occurs for this activation function. The overall results for the mR@K metric are very close to the baseline model results.

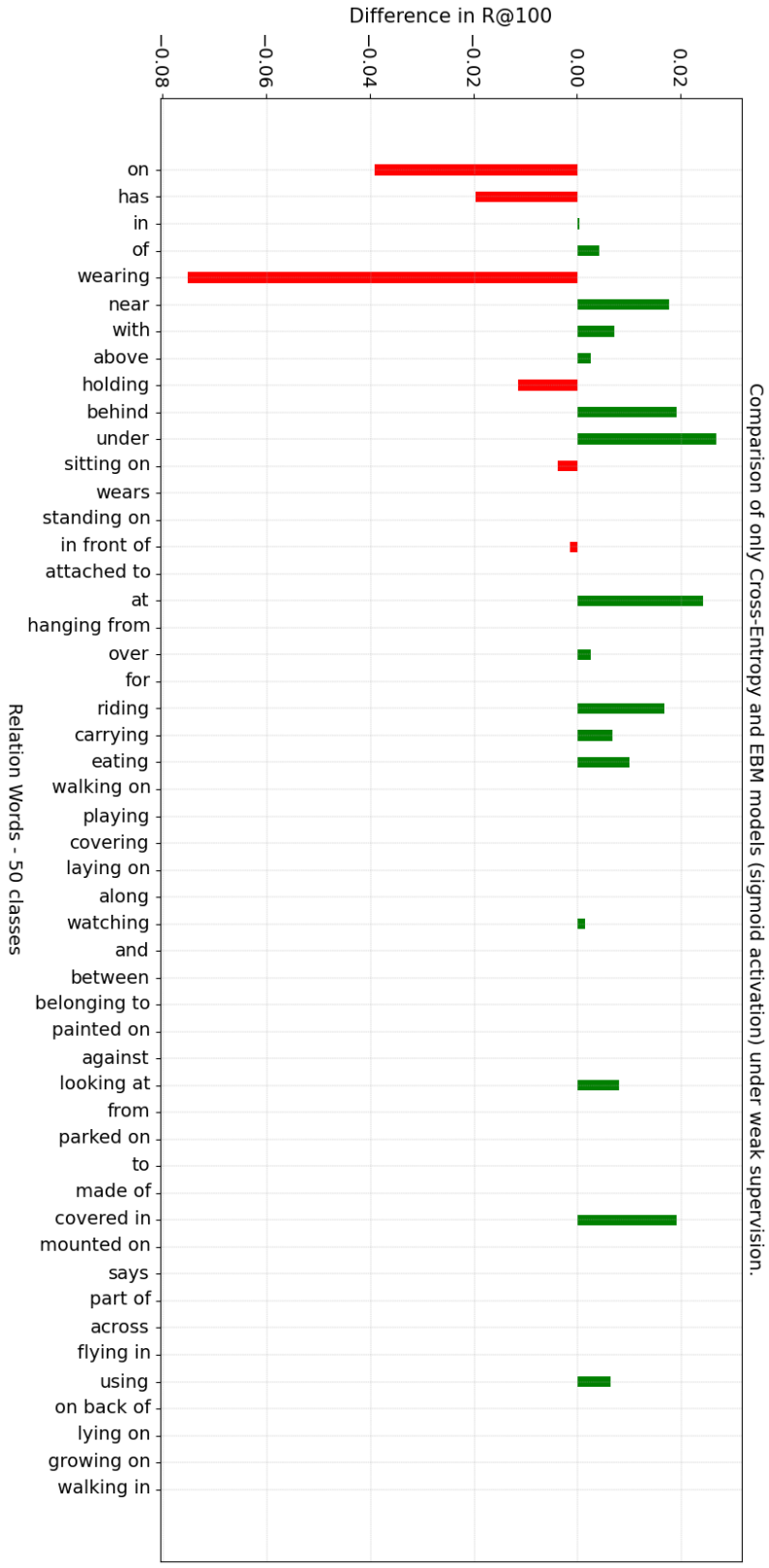


Figure C.1: In-depth analysis of predicates for the Motif (sigmoid).

Bibliography

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[2] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018.

[3] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13936–13945, 2021.

[4] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

[5] Gaurav Mittal, Shubham Agrawal, Anuva Agarwal, Sushant Mehta, and Tanya Marwah. Interactive image generation using scene graphs. *arXiv preprint arXiv:1905.03743*, 2019.

[6] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments, 2019.

[7] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[8] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019.

[9] Somak Aditya, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermüller. Image understanding using vision and reasoning through scene description graph. *Computer Vision and Image Understanding*, 173:33–45, 2018.

[10] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

[11] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

[12] Bo Pang and Ying Nian Wu. Latent space energy-based model of symbol-vector coupling for text generation and classification. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8359–8370. PMLR, 18–24 Jul 2021.

[13] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[14] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020.

[15] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[16] Xingchen Li, Long Chen, Wenbo Ma, Yi Yang, and Jun Xiao. Integrating object-aware and interaction-aware knowledge for weakly supervised scene graph generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4204–4213, 2022.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.

[18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.

[19] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[20] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018.

[21] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017.

[22] Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, et al. Scene graph generation: A comprehensive survey. *arXiv preprint arXiv:2201.00443*, 2022.

[23] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. *Advances in neural information processing systems*, 31, 2018.

[24] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *Proceedings of the IEEE international conference on computer vision*, pages 4233–4241, 2017.

[25] Keren Ye and Adriana Kovashka. Linguistic structures as weak supervision for visual scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8289–8299, 2021.

[26] Federico Baldassarre, Kevin Smith, Josephine Sullivan, and Hossein Azizpour. Explanation-based weakly-supervised learning of visual relations with graph networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 612–630. Springer, 2020.

[27] Bo Wan, Yongfei Liu, Desen Zhou, Tinne Tuytelaars, and Xuming He. Weakly-supervised hoi detection via prior-guided bi-level representation learning. *arXiv preprint arXiv:2303.01313*, 2023.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[29] Jing Shi, Yiwu Zhong, Ning Xu, Yin Li, and Chenliang Xu. A simple baseline for weakly-supervised scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16393–16402, 2021.

[30] Xingchen Li, Long Chen, Wenbo Ma, Yi Yang, and Jun Xiao. Integrating object-aware and interaction-aware knowledge for weakly supervised scene graph generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4204–4213, 2022.

[31] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[32] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[33] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019.

[34] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2017.

[35] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark, 2018. Accessed: [24/05/2023].

[36] Kaihua Tang. A scene graph generation codebase in pytorch, 2020. https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch.

Energy-Based Learning in Weakly Supervised Scene Graph Generation

Berkay Güler
KU Leuven
Leuven, Belgium
berkay.guler1@gmail.com

***Abstract*—Nowadays, the literature about scene graph generation (SGG) focuses more on weak supervision in training. That is mainly caused by the fact that full supervision possesses demanding human effort, and inherent bias and errors due to the annotators. The weakly supervised setting overcomes these issues since there is no need for annotations for the images, thus making it also more scalable and cheaper. Besides the problems in training scheme, the traditional approaches for the SGG task utilize a sum of cross-entropy losses for the detected object pairs ignoring the composition present in the image. This thesis aims to utilize a proposed energy-based framework to understand the structure in the image, and to improve the results of a baseline relation model under weak supervision. To achieve this goal, a state-of-the-art relation model is modified to make it compatible with weakly supervised setting. Similar modifications are applied to the energy model and these two models are integrated into each other. According to the experimental results, the modified energy framework[1] enhances the performance of the baseline model for the rare predicate words marginally, and it follows the same improvement trend for these rare relation categories even under weak supervision.**

***Index Terms*—SGG, Energy Learning, Weak Supervision**

I. Introduction

As the area of computer vision grows each year, traditional tasks such as object detection are not considered to be exciting anymore. Therefore, the researchers initiated more instructive challenges like scene graph generation (SGG) to get a better representation of the images. This graph-based representation is also beneficial to other several related applications, including visual question answering [1], image captioning [2], and image retrieval [3].

A typical scene graph model tries to recognize the visual semantics in the image and detects the triplets of $\langle subject, predicate, object \rangle$. The detector identifies the locations and categories of objects, and the overall model finds the predicate word meaning the relationship between this pair. The detection of these triplets could be done in two main ways namely one-staged or two-staged methods. This thesis mainly focuses on the two-staged methods. In this method, the proposal regions and features are extracted by an object detector. Then, these proposals are utilized to find the relation category between the proposals.

The training scheme of a typical SGG model primarily considers two ways. The first one is full supervision, where the bounding box annotations are present the in the ground truth. For the classification stage, these box annotations provide detailed information about the relationships between each object pair. In contrast, weak supervision does not possess any bounding box annotations. The literature for the weak supervision [4]–[9] utilizes different weak supervision schemes. The most popular ones are the image-level relation labels and the ground truth image-level graphs. These popular schemes are easier cases than employing image captions and are mainly used to provide an upper limit before applying image caption training. Although SGG is initially designed for full supervision, weak supervision studies started to gain attention in recent years. This is mainly caused by the fact that full supervision annotations are expensive and prone to annotator errors and biases. Also, caption-based supervision is easier to collect making the weak supervision more scalable. Therefore,

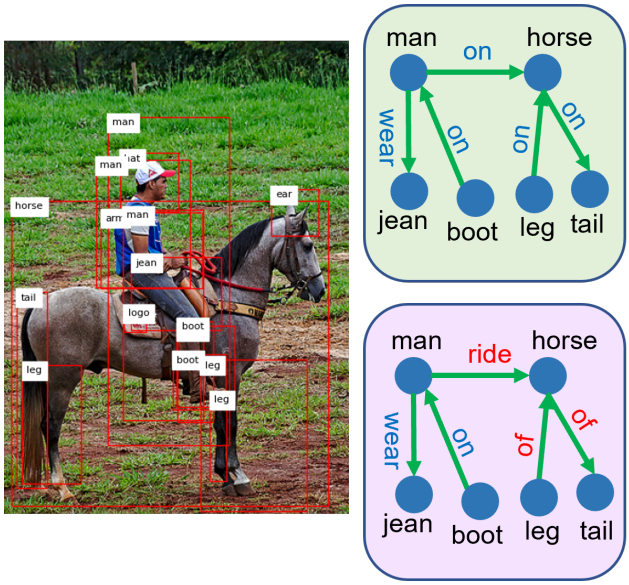


Fig. 1: **Scene Graph Generation:** Figure illustrates an example scene graph generation both from the baseline cross-entropy model (green) and energy model (purple). It is clearly observed that the baseline model is biased towards to the common predicates like `on, wear`; however, the energy model incorporated the composition of the image and identified the rare predicates `ride, of`.

[1]Code: https://github.com/gulerrberkay/scene-graph-ebm

this paper mostly focuses on weak supervision thanks to these benefits.

The weakly supervised SGG contains fundamental challenges that have to be confronted. Firstly, such weak supervision does not provide as detailed information as full supervision during training. Secondly, most of the studies in weak supervision benefits from an off-the-shelves object detector to identify the proposed regions for the object. However, these detectors usually produce some noisy object predictions along with confident ones. One has to trust these predictions in the training phase to train their model. Thirdly, a typical two-stage scene graph structure aggregates the context information using several methods. These models are usually trained with the sum of cross-entropy losses in weak and full supervision cases. However, applying such a loss technique lacks information on the composition of the image, and it treats every object pair independently. By commonsense, it is known that most objects in the image are somehow related, so this underlying structure has to be considered in the loss function. Fourthly, the well-known dataset Visual Genome [10] has a long-tailed distribution for the predicate words, so the problem of SGG has a biased training process. This bias causes the conventional models to predict the frequent predicate more than the rare ones.

In recent years, energy-based learning strategies start to get promising results for the image generation studies like [11], [12]. Also, they have an increasing usage in discriminative studies such as [13], [14]. These energy frameworks assign scalar energy values for the input-output configurations. Given the input x with label y, the joint energy model can be denoted as $E_\theta(x, y)$. [15] states that any function could be selected if one satisfies the two main criteria of probability distributions: **i)** the probability distribution needs to assign a non-negative value for every input value x i.e. $p(x) \geq 1$. **ii)** the integral of probability distribution needs to be equal to 1 for each input i.e. $\int_a^b x^2\, dx = 1$. This could be achieved by a Boltzmann distribution as in $q_\theta(x) = \frac{\exp(-E_\theta(x))}{Z_\theta}$. [15] also mentions that the denominator of this crafted distribution is usually untraceable. Thus, it is not trivial to find the best parameters that maximize the likelihood. According to [15], most methods address this issue by writing the derivate of log-likelihood as $\nabla \mathcal{L}_{MLE}(\theta; p) = \mathbb{E}_{p(x)}[\nabla_\theta E_\theta(x)] - \mathbb{E}_{q_\theta(x)}[\nabla_\theta E_\theta(x)]$. This new representation requires Monte Carlo Markov Chain (MCMC) methods to sample from the data distributions to compute the expectation.

This paper proposes using energy-based frameworks in weak supervision to address all aforementioned problems. It has been proved that the energy-based approach improves baseline cross-entropy models under full supervision thanks to the study of [13]. Therefore, exploiting such energy formulation should guide the learning process of the baseline model leading the better results. According to [13], the main reason behind the success of the energy models is that they convert the problem of maximization of the sum of likelihoods into the maximization of joint likelihood problem by taking the structure in the input image into account. Thus, the main research question in this paper is how much the energy models improve the baseline relation model results even under weak supervision.

Contributions. The summary of contributions are as follows in this paper:

- A conventional relation model is adjusted and made compatible with the current weak supervision setup. The baseline results established solid fundamentals for the next part.
- The energy framework is also modified for the weakly training. The integration of the energy model into the relation model is done to see any improvements in the results for the weak supervision.
- To realize the stated integration, a filtering and sampling mechanisms under weak supervision are proposed. The ablation studies for each additional method are provided to illustrate their effectiveness.

II. Related Works

Weakly Supervised SGG: Scene graph generation is one of the popular tasks in vision-related research, and in recent years various papers have been shared mostly under full supervision. The studies like [1], [16]–[20] utilizes the full supervision training scheme to tackle the SGG task. However, as mentioned earlier weak supervision gained popularity against full supervision because of the inherent problems of fully supervised training. The weakly supervised SGG papers like [4]–[9] utilize similar strategies while training their model. VSPNet [4] utilizes the image-level relation labels, and the ground truth image graph as the ground truth to align their predictions with this ground truth graph. The study of [5] extends to problems from ground truth graphs to image captions to make the problem setup even weaker. Their method wants to benefit from the linguistic structure present the captions to assist the training under weak supervision. [6] employs a simpler approach to find the relations between the objects. Their method contains a sensitivity analysis that detects the effect of each object on the predicted relations to find out which objects produced the high-scored relations. Both studies of [8], [9] consider turning the weak supervision problem into a fully supervised one by generating pseudo ground truths graphs. [8] achieves the goal by applying graph alignment algorithms on the detected proposals whereas [9] utilizes a previously trained grounding module to localize each detected proposal. These localized graphs become the pseudo-ground truths, transforming the problem into a fully supervised one.

Energy Based Modeling: In recent years, the energy-based methods started to gain attention in image generative studies like [11], [12]. These papers exploit the energy model architecture to increase the image generation accuracies. Even though the energy models are often popular in generative tasks, the studies [13], [14] focusing on discriminative tasks also started to use the energy models to get improvements in their results. Since [13] proved that the energy framework is highly

useful for the SGG task, applying this methodology in the weakly supervised SGG training seems to be promising.

III. Method

This section introduces the proposed methodology in the paper. Section III-A introduces the main variables in the proposed method. Section III-B discusses each layer in the architecture along with the formulations of the structure. Section III-C investigates the overall loss functions for the training of the designed architecture.

A. Problem Setup

This subsection introduces the proposed problem setup for weakly supervised SGG. Given image I, the scene graph model $\mathcal{M}$ calculates a tuple of (O, R) i.e. $\mathcal{M}(I) = (O, R)$ where $O \in \mathbb{R}^{NxN_o}$ represents the detected objects in image I. The relations between the detected objects are shown in $R \in \mathbb{R}^{(N*N-1)xN_r}$. The number of proposals, object classes, and relation classes are notated as N, N_o, N_r respectively in the previous formulas.

For weak supervision, only image-level ground truths are available for the baseline model. That means for a given image I, the ground truth object labels for this image $\mathcal{O}_I$ is a subset of $\mathcal{O} \in \{0, 1, ...N_o\}$. For the same image, image-level ground truth relation labels $\mathcal{R}_I$ is a subset of $\mathcal{R} \in \{0, 1, ...N_r\}$. In addition, the dimension of the relation score matrix R increases quadratically by the number of proposals in weak supervision because it needs to calculate the relation scores for all pairs.

B. Model Design

This subsection investigates the model overview depicted in Figure 2 layer by layer and provides how the adjusted Motif [16] and energy framework [13] functions.

1) Relation Model: This paper chooses the architecture of Motif [16] which is considered one of the traditional LSTM-based relation models. It utilizes the natural ability of Recurrent Neural Network (RNN) based architectures to detect the relationships between the objects. Although using proposals as a sequence input for the LSTMs seems counter-intuitive, the LSTM structures are highly utilized in the SGG community due to their success. The contribution that converts Motif suitable in weakly supervised training is the elimination of the Decoder part in the original Motif [16]. This layer was decoding the proposals and assigns refine object labels for them. However, it cannot be utilized in weak supervision since the ground truth labels for the proposal boxes are not available.

The relation model partitions the problem of relation detection with equation 1.

$$Pr(G_{SG}|I) = Pr(B|I)Pr(O|B, I)Pr(R|B, O, I) \quad (1)$$

The first and second terms $Pr(B|I), Pr(O|B, I)$ in equation 1 denotes the problem of object detection. An off-the-shelf object detector detects proposal regions for the input images and provides the predicted object distributions. The labels of these proposals can be calculated from the maximum-scored object category. The last term $Pr(R|B, O, I)$ illustrates the relation identification part where the detected object labels, and locations of these objects are all being used to find the scene graph.

The relation model first extracts the features and labels thanks to the detector as mentioned earlier. The predicted bounding boxes $B = \{b_1, ..., b_N\}$, feature vectors $f_i \in \mathbb{R}^{4096}$ and labels $l_i \in \mathbb{R}^{200}$ are detected for the i-th proposal.[2]. [16] argues that these detected proposals need to be contextualized by using them in biLSTM layers. Hence, the next step is called the *object context* where these objects produce a better representation of the given image I.

Object Context. A sequence of inputs is created by the proposals by feeding them into the first biLSTM layer called object context. The default option from [16] is selected for ordering the proposals. This default option orders the proposals from left to right by looking at the detected bounding box.

$$C = biLSTM\left([f_i\,;\mathbf{W_1}\mathbf{l}_i]_{i=1,...,n}\right) \quad (2)$$

Equation 2 provides the formulation of the first layer where $C = [c_1, ..., c_N]$ are hidden states for each proposal. $\mathbf{W_1}$ is a mapping kernel for arranging the dimension of l_i.

Edge Context. The next step is to compute a relation matrix $R \in \mathbb{R}^{(N*N-1)xN_r}$. This matrix holds all relation scores for every object pair. That is why its dimension rises rapidly if one chooses to increase the number of proposals. The contextualized proposals in C serve as inputs in this layer to apply a similar procedure on edge vectors.

$$D = biLSTM\left([\mathbf{c}_i\,;\mathbf{W_2}\mathbf{l}_i]_{i=1,...,n}\right) \quad (3)$$

$D = [d_1, ..., d_n]$ represents the edge vectors for the proposals. However, these representation lacks global knowledge about the object pair. To incorporate this additional information, the union's visual features need to be mixed into the solution. Moreover, most of the studies [1], [6], [16], [17], [19] mentioned in Section II activates the frequency baseline which puts more bias towards some relations if a particular object pair is detected. After adding this information, equation 4 displays how to compute the matrix R for relation scores.

$$R = (\mathbf{W_h}d_i \circ \mathbf{W_t}d_j) \circ f_{i,j} + w_{i,j} \in \mathbb{R}^{(N*N-1)xN_r} \quad (4)$$

$\mathbf{W_2}, \mathbf{W_h}, \mathbf{W_t}$ are dimension mapping matrices for d_i. R matrix holds every relation score for each object pair; thus, this graph representation can be used in the energy model.

Inference. The evaluation of the relation model needs sorting operation on the triplets. Each predicted triplet has three components $\langle subject, predicate, object\rangle$. The scores of each component are calculated by performing a softmax operation on matrices R, O. Three scores will be multiplied to obtain the score for the corresponding triplet. The most confident triplets will be considered in the evaluation part.

[2]Note that l_i has a dimension of $\mathbb{R}^{200}$ since the loaded GloVe embeddings [21] has a vector of this dimension.

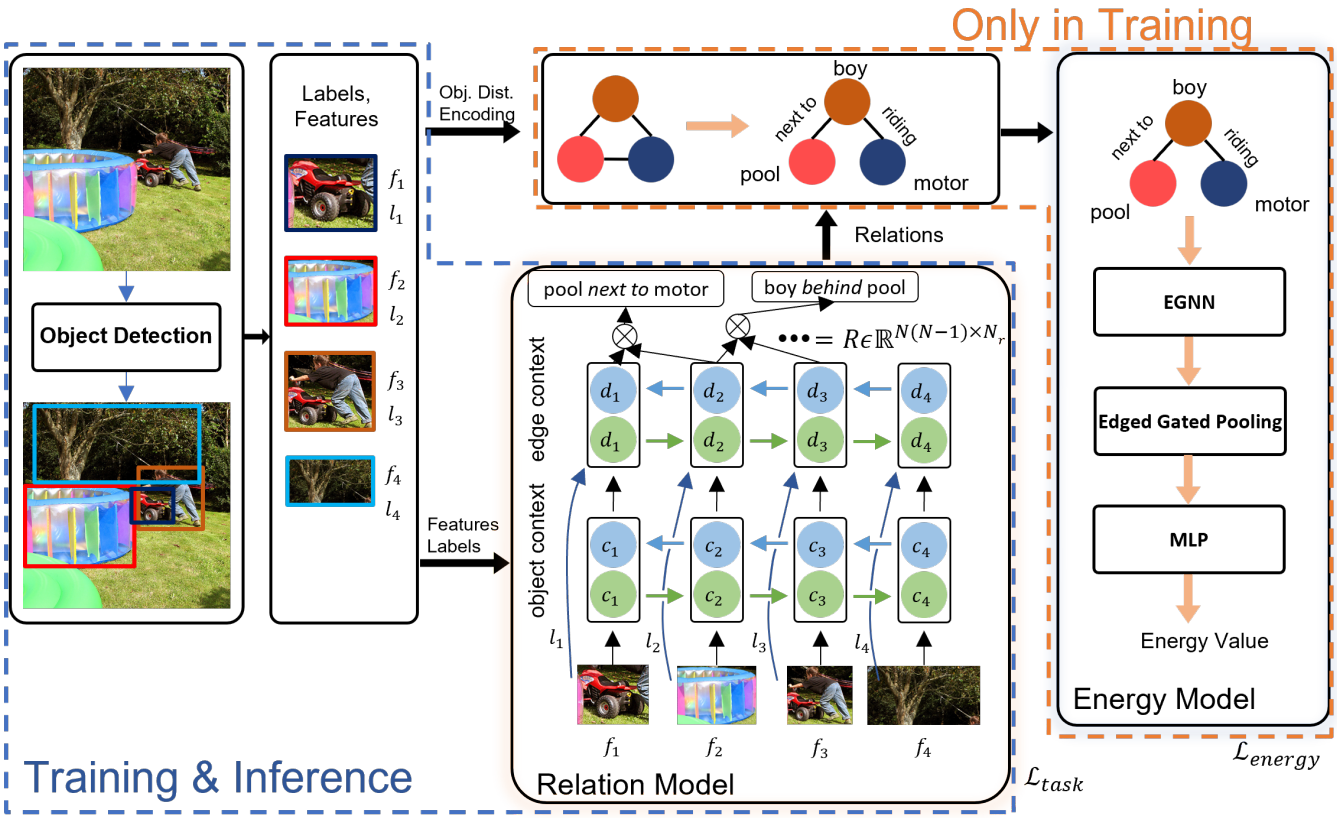


Fig. 2: **Model Overview.** The overall pipeline starts from the object detector on the left to find the labels, features, the bounding boxes. These features and labels are fed into the relation model to make them contextualized. The edge context utilizes the inputs from labels and object context to calculate the relation score matrix R. The relation model is always activated both for training and inference. The energy model, on the other hand, is only activated in training. It guides the training procedure for the relation model by calculating the energy loss from the predicted and ground truth scene graphs.

2) Energy Model: This subsection analyzes the energy model architecture depicted in Figure 2. An important remark about the energy model is that it is only activated in the training phase. Thus, it is not employed in the inference phase. The most important contribution is the removal of the image graph from the original work [13]. The image graph contains the RoI features from the proposals and ground truth object boxes; hence, it incorporates the visual features in the energy loss. Since it is not suitable for weak supervision due to the absence of box annotations, this image graph is removed from the formulations. The effects of this removal can be seen in Section IV.

Overview. The energy model in Figure 2 starts from the encoded scene graph predicted by the relation model. Therefore, the problem setup for the energy model can be summarized as follows: Given a scene graph generation model $\mathcal{M}$, and an image I, the predicted scene graph can be denoted as $G_{SG} = \mathcal{M}(I) = (R, O)$. The tuple of (O, R) is two main inputs of the energy model as they represent the scene graph. This tuple undergoes Edged Gated Neural Network (EGNN) and pooling layers proposed by the authors of [13]. These layers essentially apply message passing between the detected objects and relations to get a meaningful representation. The pooling functions decrease the dimension of the hidden embeddings into one scalar value called the energy.

$$E_\theta(G_{SG}) = MLP[f(EGNN(G_{SG}))] = e^- \quad (5)$$

$$E_\theta(G^+_{SG}) = MLP[f(EGNN(G^+_{SG}))] = e^+ \quad (6)$$

Equations 5 and 6 denote the general formulation of the energy model. e^- is called the negative energy and holds the information for the predicted graph configuration. e^+ is referred to as the positive energy and it is the scalar representation of the ground truth configuration. Therefore, **the energy model needs the ground truth image-level graph to calculate the energy loss.** The image-level graph should guide the training by considering the structural information present in the image. The graph-level representation in the ground truth is still considered weak supervision as it does not utilize box annotations. For instance, an image caption can be turned into an image-level graph by a language parser [22].

EGNN Layer. According to [13], this layer aims to enrich each node and edge state by applying message passing algorithms between them. One should note that in energy model,

the nodes represent object proposal, and the edges represent the relations between the object pairs.

These nodes and edges are predicted from the modified Motif structure; thus, $\mathcal{M}(I) = (O, R)$. Every row in this matrix $O \in \mathbb{R}^{NxN_o}$ presents an object proposal embedding n_j where $j \in 1, ..., N$. At the same time, each row in relation matrix $R \in \mathbb{R}^{(N*N-1)xN_r}$ refers to a relation distribution $r_{j\to i}^{t-1}$ between the node embeddings n_i and n_j. The object and edge embedding n_i, $r_{j\to i}^{t-1}$ are produced by applying the rows of O and R i.e. o_i & r_{ij} to a kernel MLP layer.

$$m_i^t = \alpha \underbrace{W_{nn}\left(\sum_{j\in N} n_j^{t-1}\right)}_{\text{node to node message}} + (1-\alpha) \underbrace{W_{en}\left(\sum_{j\in N'} r_{j\to i}^{t-1}\right)}_{\text{edge to node message}} \quad (7)$$

Equation 7 shows the message embedding vector for the i-th proposal. These messages are aggregated from the neighboring nodes and edges for each node. α denotes the weighting for the node and edge messages. The message passing helps the model to aggregate all relevant information treating every node related. Even if there is no information between the nodes, this information must be shared for the foreground objects.

$$\underbrace{p_{j\to i}^t = W_{ne}[n_i^{t-1}||n_j^{t-1}]}_{\text{node to edge message}} \quad (8)$$

A similar edge update message should be also computed from the neighboring nodes. However, the order of the nodes affects the resulting predicate so before sending the edge message, the nodes should be concatenated properly as it is shown in equation 8.

$$\begin{aligned} n_i^t &= GRU_n(m_i^t, n_i^{t-1}) \\ r_{j\to i}^t &= GRU_e(p_i^t, r_{j\to i}^{t-1}) \end{aligned} \quad (9)$$

To capture the composition in the image, the proposed energy model architecture leverages Gated Recurrent Units (GRUs). As in the modified Motif, GRUs cells aggregate the messages from neighboring nodes and edges. The hidden states of the GRUs are initialized from the node and edge embeddings. The reader may notice that there is a superscript t on the messages. The embeddings need to be updated in several iterations to get better node and edge states. That means the initial states are provided by the relation model.

Pooling Functions. Obtaining a scalar energy value requires a pooling operation on the updated node embeddings. Since one scalar value holds all the information present in the image, the authors in [13] suggests a clever way to aggregate all the nodes and edges. Equation 10 illustrates the attention-based pooling operation. Each node and edge is fed into a linear layer to get a particular attention score between 0-1. This attention score is multiplied with the embeddings to apply a weighted aggregation. The resulting vectors N, E represent the clear edge and node vector representations as it gives low attention to unrelated nodes and edges. Lastly, the concatenation of the N, E retrieves the scalar energy value for the input scene graph.

$$\begin{aligned} N &= \sum_k f_{gate}(n_k) \odot n_k \\ E &= \sum_{ij} g_{gate}(r_{i\to j}) \odot r_{i\to j} \\ e^- &= MLP(N; E) \end{aligned} \quad (10)$$

Sampling in Weak Supervision. As explained earlier, the message-passing algorithm is needed in the energy model, to refine the node and edge embeddings. However, this message passing should be only applied between the foreground objects to obtain valuable messages. Full supervision contains bounding box information so it knows foreground object locations. However, weak supervision does not allow you to know which proposals are the foreground proposals. This paper suggests that only the proposals that are in the image-level ground truth object labels should be considered in message-passing algorithms. Even though the relation model considers $N(N-1)$ object pairs, most of these pairs are unnecessary and create nothing but noisy messages. Therefore, removing some proposals in the energy model improves the model's performance. The detailed results are provided in the Section Ablation Studies IV-E.

The Handcrafted Background Score. Another problem that needs to be resolved is the background score. Section III-C provides the loss formulations for the combined model. The relation model is mainly trained with binary cross-entropy loss that uses image-level relation labels. However, this loss formulation does not train the model for producing a valuable background score since weak supervision does not know which pairs are unrelated. At the same time, the model still predicts meaningless scores for the background relation category. This paper recommends a handcrafted background score formulation as shown in equation 11.

$$\begin{aligned} BG &= 1 - \underset{r\in\mathbb{R}^{N_r}}{\text{argmax}}(softmax(R')) \in \mathbb{R}^{N(N-1)*1} \\ R &= [BG; R'] \end{aligned} \quad (11)$$

This formulation assigns a background score for the predicted relation score matrix R. The scores turned into probabilities using a softmax function. The background score BG for each object pair is then the subtraction of the maximum relation category probability for that pair from one. R' is the sliced version of the relation score matrix that does not contain meaningless background scores. Section Ablation Studies IV-E also considers the utilization of sigmoid instead of softmax.

C. Weakly Loss Formulations

To train the model a multi-loss function with three effective terms is formulated. Each loss function contributes to the overall performance of the model as follows: **i)** an image-level binary cross entropy loss $\mathcal{L}_{base}$ supervising the training regarding the image-level labels; **ii)** an energy loss that compares the predicted scene graph energy with the ground truth

image-level graph to incorporate the structure in the image; **iii)** a regularization loss for the calculated energy values. The formulation is the sum of all loss functions as shown in equation 12. The relative weights $\lambda_e, \lambda_{reg}, \lambda_{base}$ will be selected empirically.

$$\mathcal{L}_{total} = \lambda_e \mathcal{L}_e + \lambda_{reg} \mathcal{L}_{reg} + \lambda_{base} \mathcal{L}_{base} \tag{12}$$

The first term in the multi-loss function is named energy loss. The derivation of the energy loss is displayed in equation 13. The computation of the energy loss requires solving an optimization problem that minimizes the equation 13. This is often dealt with by solving this function iteratively using Stochastic Gradient Langevine Dynamics (SGLD) [23].

$$\mathcal{L}_e = E_\theta(G^+_{SG}) - \min_{G_{SG} \in SG} E_\theta(G_{SG}) \tag{13}$$

The intuition behind the energy loss is that the optimization updates the predicted scene graph nodes and edges while minimizing the energy of the prediction. These updates guide the training by assigning smaller energy values for the correct predictions. For the predictions that are not in the dataset, the energy values should be maximized to have a balance. This is also called contrastive divergence-based learning. Please see the Background Section of the thesis material to get more detailed information about these updates.

$$\mathcal{L}_{reg} = E^2_\theta(G^+_{SG}) + E^2_\theta(G_{SG}) \tag{14}$$

The writers of [13] argues that the scalar energy values become too large while training. Hence, they address this issue by adding an L2 regularization loss.

$$\mathcal{L}_{base} = -\sum_{c=1}^{C} \left(\mathbb{1}_{[c \in \mathcal{R}]} \log a + \mathbb{1}_{[c \notin \mathcal{R}]} \log 1 - a \right) \tag{15}$$

Finally, the base model loss is calculated by using a binary cross-entropy for the image-level relation labels. If a relation category is present in the ground truth image-level labels, the target probability is assigned as 1. If it is not present, the target should be equal to zero. The predicted image-level relation scores are computed by operating a maximization function on the relation score matrix R as shown in equation 16. The max operation on rows should produce predicted relation scores for the image.

$$\tilde{r} = \max_{rows} R \in \mathbb{R}^{N_r} \tag{16}$$

IV. Experimental Results

This section gives information about the experimental setup and results of the proposed method.

A. Experimental Setup

This section mentions the dataset split for the tests. The popular image dataset that contains relationship labels Visual Genome [10] is employed in the experiments. A pre-processed version of the Visual Genome by Xu et al. [17] is popular among the SGG community so it is utilized in all experiments. Visual Genome contains 108k images and the utilized split considers 150 object categories and 50 relation categories. In addition, this split leverages only 57,723 images in training, while 5000 images are kept separately for the validation set. The test set consists of 26,446 images meaning that the split on the whole set corresponds to a 70%-30% split for the training and test phases respectively.

B. Evaluation Metrics

This paper finds useful two conventional evaluation metrics for the weakly supervised SGG task. Recall@K (R@K) metric [16] is one of the oldest metrics for SGG. Although R@K is a biased metric due to the long-tailed distribution of the Visual Genome, it is still beneficial in weakly supervised SGG tasks. Another metric is the mean Recall@K (mR@K) introduced by [1]. mR@K eliminates the bias in R@K by taking the average of all predicates R@Ks. Hence, it shows the effect of the energy model if it improves the results for the rare predicates.

The evaluation metrics have to be calculated under a certain setting. **SGDet**: Scene Graph Detection is the scheme that is going to be shown in the results. SGDet expects to predict the scene graph in the given image I. The scene graph means that the bounding boxes of the object should be detected with IoU ≥ 0.5, and the labels of the objects must be correct. Moreover, the relations between these correct object pairs need to be identified correctly.

C. Implementation Details

Detector. A pre-trained Faster R-CNN [24] with ResNeXt-101-FPN backbone [25], [26] is utilized for obtaining the proposal regions, and the weights of these layers are frozen during the scene graph model training. The pre-training on Visual Genome is a violation for the weak settings but this paper aims to prove the concept of the effectiveness of the energy-based method in weak settings. The detector outputs 30 proposals with 151 object categories including the background object category.

Relation Model. The relation model is trained with a Stochastic Gradient Descent(SGD) optimizer that has a learning rate of 0.1. The learning rate increases linearly until 400 iterations so it has a warm-up stage. The learning reduces at every plateau region according to the validation results. The decay happens only if 3 consecutive validation did not improve the validation results (plateau region.) SGD also has a momentum of 0.9 and an effective batch size of 4 in every experiment. The model's weights are initialized randomly, and a weight decay of 0.0001 is applied. During the training phase, the incorporation of frequency bias information is utilized to improve performance. The context biLSTMs have a hidden dimension of 512. The main GPU in training was The NVIDIA RTX 3090.

Energy Model. The α in message passing is set as 0.5 to keep a balance between messages. The message update iterations only applied 3 times. In SGLD optimization steps, the node and edge states are scaled back to the range of [0, 1]

when each update is made for the node and edge states. The node and edge embeddings have a vector size of 512. The relative weights $\lambda_e, \lambda_{reg}, \lambda_{base}$ are selected as 1.

D. Quantitative Results

This subsection presents the quantitative results from the experiments.

Model	Method	Scene Graph Detection	
		R@20/50/100	mR@20/50/100
Motif [19]	CE	25.48/32.78/37.16	4.98/6.75/7.90
	EBM	- / - / -	- / - / -
Motif [13]	CE	25.62/32.97/37.41	5.07/6.91/8.12
	EBM	24.39/31.74/36.29	5.67/7.71/9.27
Motif(Ours)	CE	25.13/32.01/35.90	4.91/6.73/7.89
	EBM	25.13/32.05/35.86	5.22/7.12/8.41

TABLE I
Paper reproduction results for full supervision.

The first experiment contains full supervision results for the modified relation and modified energy models. Table I illustrates these modification results. See the Thesis writing for the detailed modification lists for the models. This test is mainly done to establish solid ground for the other experiments in the paper. In addition, it displays how the modified energy model improves the baseline relation model. The first row of Table I states the original Motif model results for full supervision. The second row shows the re-implementation results from [13]. Finally, the third row in Table I mentions our full supervision results. Our modified energy model still enhances the baseline results for the mR@20-50-100 metrics. For instance, in mR@100, our baseline predicts 7.89 points whereas the energy model improves this value by 0.52 despite removing one of the important branches in the original work of [13] namely the image graph. Our results also display that R@20-50-100 values did not seem to change for the modified energy model. When one compares the third row with the second row, one sees that reproduction provided similar improvements for the mR@20-50-100 values but R@20-50-100 values did not really been affected. The expectation was to get some decreases in the R@K values for the energy model as it diminishes the bias in training. To understand the behavior of the experiment, Figure 4a displays an in-depth analysis of the R@100 values for each predicate separately for full supervision. As it can be seen from Figure 4a, the energy model supported the prediction for the rare predicates such as 'eat,''carry,' and 'riding.' but it did not really decrease the most biased predicates like 'on' or 'has.' The averaged biased predicates like 'holding,''above,' and 'near.' seemed to be diminished from R@100 by a couple of points. The reproduction does not seem to be perfect as in the original work of [13] but it seems to be succeeding for the uncommon predicates. The reason behind not getting the perfect reproduction is that the effective batch size was selected 16 in the original work whereas this paper selects 4 due to computational availability and the removal of an important branch in the original work. Another important result to mention is that mR@K and R@K are in balance. If one uses a method to boost mR@K, the R@K metric should decrease automatically since it is a biased metric. Putting more emphasis on unbiased predicates should take the accuracy of biased predicates away. Since reproduction does not increase mR@K values as significantly as in the original work(0.52 points for mR@100.), R@K metric can stay stable. Same Figure 4a also shows that some predicates do not even have a bar which means they did not have any successful predictions. The removal of the image graph in the original work caused a decrease in the performance for these semantically meaningful predicates.

The second experiment reveals the main results of this paper since it shows the effectiveness of the energy model under weak supervision. It also consists of a row for the results of the state-of-the-art studies. Table II summarizes all these experiments. The first row reveals fully supervised training results of the original Motif; hence, it possesses the highest values in every metric. The second row displays the state-of-the-art in weakly supervised SGG. Lastly, the third row projects our modified Motif baseline and the energy model results. One should start from the third row to examine the results. The weakly supervised baseline model has comparable results with the first row. Even though only image-level labels are used in the training, the weakly supervised baseline presented concrete results. The addition of the energy model improves these baseline mR@20-50-100 values marginally. However, R@20-50-100 values seem to get a drop of 0.89, 0.84, and 0.39 respectively when activating the energy model. This drop in R@K is expected because the energy model should put more emphasis on non-frequent predicates. However, it was assumed to get more performance for the mR@K values in the energy model.

To understand why the effectiveness of the energy model is marginal, a detailed predicate analysis is employed in Figure 4b similarly. Figure 4b illustrates a trend having a decrease for the frequent predicates, and a boost for the uncommon predicates. This behavior is similar to the full supervision results in Figure 4a as it also improves the predicates on the tail, and slightly reduces the ones around the head of the distribution. For example, the semantic predicates such as 'standing on,' 'carrying,' 'eating,' and 'riding' have an increasing trend when the energy model is activated whereas the common ones like 'on,' 'has' have diminished in Figure 4b. Furthermore, the values for these trends are also quite similar to the ones in full supervision. For instance, the maximum improvement for the predicate 'eating' in full supervision corresponds to 0.08 points in R@100 when the energy model is activated. The enhancement in weak supervision is around 0.04 points in R@100 making the improvements very close in terms of values. Although Table II illustrates minor advancements with the energy model, the detailed analysis reveals that the energy models tackle down to the biased nature of SGG training

Model	Supervision	Scene Graph Detection					
		R@20	R@50	R@100	mR@20	mR@50	mR@100
Motif [13]	Full	25.62	32.97	37.41	5.07	6.91	8.12
VSPNet [4]	Weak	-	4.70	5.40	-	-	-
LSWS [5]	Weak	-	7.30	8.73	-	-	-
WSGM [8]	Weak	4.12	5.59	6.45	-	-	-
Motif(Ours)	Weak-CE	22.25	28.51	31.75	3.14	4.58	5.59
	Weak-EBM	21.36	27.67	31.36	**3.22**	**4.70**	**5.74**

TABLE II

Quantitative Results. The weak supervision test results for CE & EBM models, and comparison with other studies in literature.

which also proves the main goal of this paper.

The minor improvements in the energy model are assumed to be caused by several reasons:

i) Section III proposes a handcrafted background probability calculation but this approach seems to be sub-optimal for assigning a background score. This score along with softmax activation is proposed to prevent smooth inputs for the energy model because the energy model requires sharp object and edge distributions for its inputs. However, the inputs are still not optimal as they contain high scores for 2 relation categories sometimes, as it is observed from the trained model debug sessions afterward. That is why the energy model achieves only minor successes for mR@K results.

ii) Another debug observation for the energy model is that the usage of the off-the-shelf object detector causes node-to-node messages to stop early as the nodes are not updated during training. Typically, the energy model also helps the object detection process in full supervision; however, the object detector outputs constant node states to be used directly without refinement in weak supervision. In addition, some of these constant object detections are noisy and not eliminated in the training process. Since full supervision allows a refinement stage for the detected object, the noisy proposals are removed during training. The suggestion for future work here is to add a refinement stage for object detection by still supervising detection results weakly. This should allow for the energy model to help the node states too, and not to stop learning early for the node-to-node kernels.

Furthermore, the second row in Table II displays some weak supervision SGG results in the literature. Even though these studies utilize similar image-level weak supervision losses in training, they use off-the-shelves detectors pre-trained on different datasets than Visual Genome. Therefore, it would be unfair to compare the performance of this thesis with these studies. Another idea that explains the huge gap in Table II between the weakly supervised baseline model result and the state-of-the-art model is the utilization of frequency baseline mentioned in [16]. These baseline assigns calculated statistics from Visual Genome on predictions to put more bias towards some relations depending on the object pair labels. This is not used in weakly SGG studies.

E. Ablation Studies

In this section, the selected ablations have been applied to the proposed method to show their contribution to overall results. Motif (+Softmax) results in Table III shows the best energy model results discussed earlier, and Motif (Baseline) is the baseline results for the only cross-entropy model.

Model	Method	Scene Graph Detection	
		R@20/50/100	mR@20/50/100
Motif(Baseline)	CE	22.25/28.51/31.75	3.14/4.58/5.59
Motif (No sampling)	EBM	22.22/28.43/31.67	3.02/4.40/5.41
Motif (+fg/bg)	EBM	21.90/28.17/31.44	3.02/4.51/5.50
Motif (Sigmoid)	EBM	21.43/27.94/31.43	3.10/4.56/5.58
Motif (+Softmax)	EBM	21.36/27.67/31.36	**3.22/4.70/5.74**

TABLE III

The ablation studies. The summary of all contributions for the proposed methods.

Foreground & Background Sampling. This paper proposed a sampling technique for weak supervision to make the energy model functional. As one may recall that full supervision offers the foreground objects and relations. Thus, the message passing between these foreground objects is useful in the energy model. However, weak supervision does not allow you to know which proposals are foreground or background. If you apply your message passing between all the pairs, you get the result for Motif (No sampling) in Table III. As one can see the energy model seems harmful when you allow message passing between unnecessary nodes. The addition of weak sampling leads to the results for the Motif (+fg/bg) in Table III. In that case, the energy model results improved compared to the previous case but it still did not improve the Motif (Baseline).

Softmax or Sigmoid Inputs. Preventing smooth distributions for the inputs of the energy model is provided with this proposed method. Softmax inputs also introduce the handcrafted background score discussed previously. As you can see from Table III, Motif (+Softmax) achieves the best results for the mR@20-50-100 metrics. Having the best results for the mR@K also causes to get the lowest values for R@20-50-100 as expected since these two metrics are in balance. Another suggestion was to use the sigmoid function instead

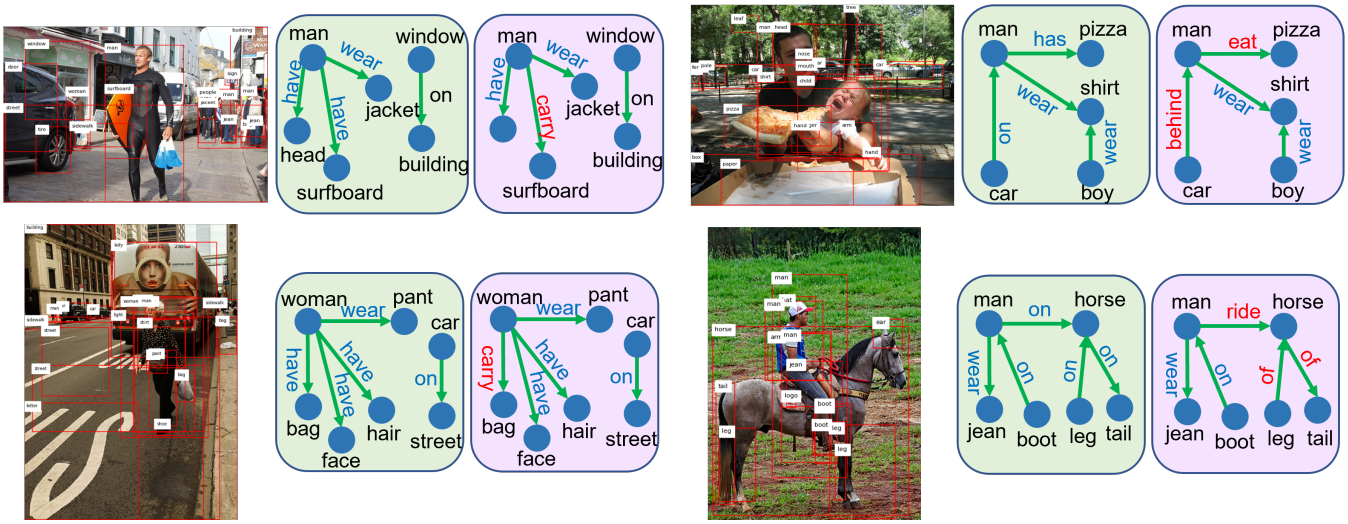


Fig. 3: **Qualitative Results.** Visualizations from the scene graph detection both for only cross-entropy (in green), and energy model (in purple)

of softmax. Even though it improved the results for the Motif (+fg/bg sampling), it is not as successful as softmax activation. The explanation of softmax is assumed to be caused by the sharp distribution requirements of the energy model. Sigmoid activation rewards every high relation score whereas softmax only rewards the highest relation category.

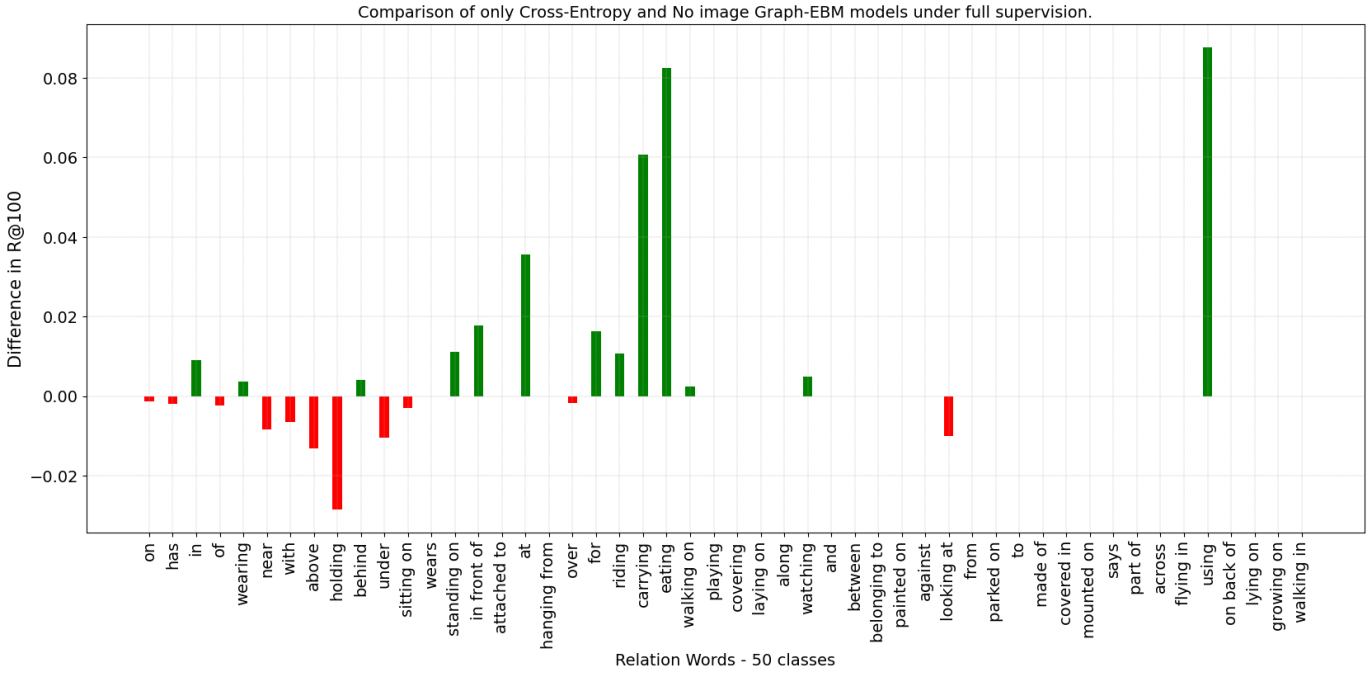
F. Qualitative Results

The performances of the models should be also compared to some example images. Figure 3 illustrates the best triplets for the given four example images. The results point out that the energy model is more successful in generating more instructive scene graph representations. One may recall the improvement trend in Section IV-A for the semantic predicates. These predicates could be observed easily in these example images. For instance, the energy model determined `<man eats pizza>` triplet instead of putting a more biased triplet such as `<man has pizza>`.

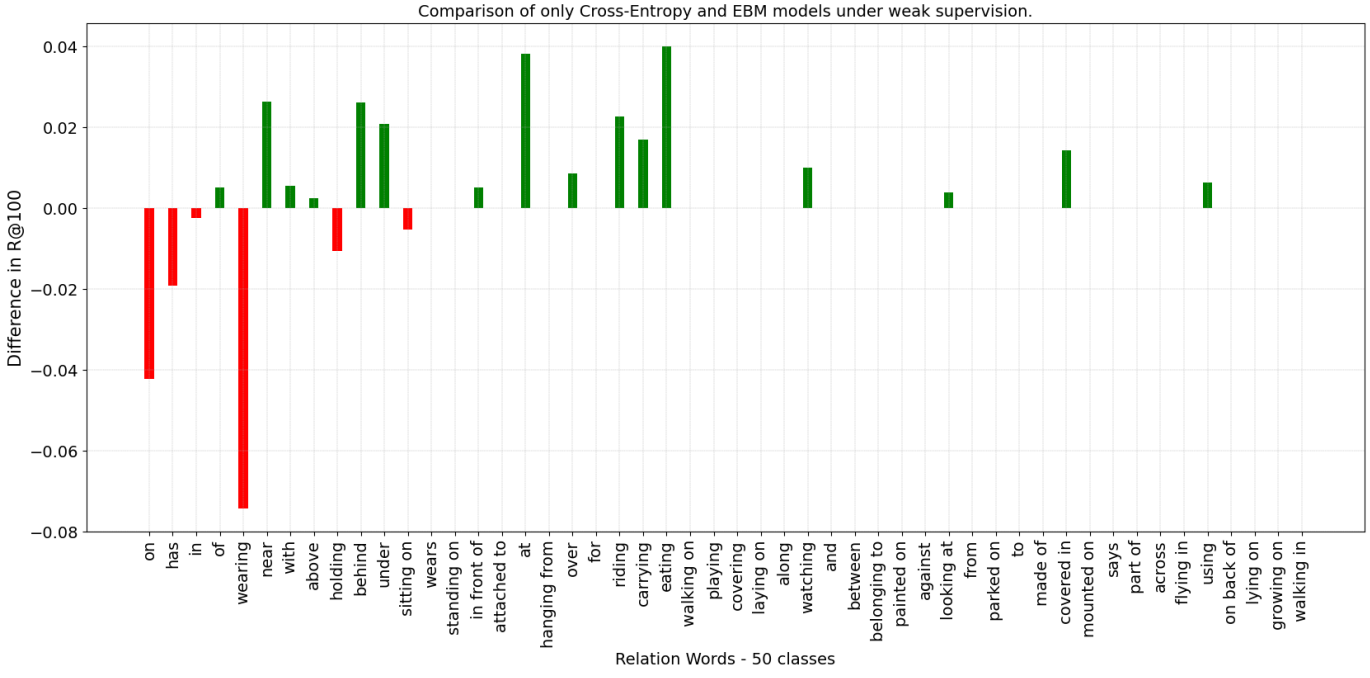
An intriguing finding from the qualitative results is that the images with obvious large objects lead to better performance for the energy model. For example, the relation `ride` is detected in the third image in Figure 3. This third image contains two large objects: 'man' and 'horse.' The energy model seems to achieve finding a better predicate when the detected objects are confident by the detector.

V. Conclusion

This paper suggested to benefits of energy-based approaches in weakly supervised SGG tasks. The experiment results showed that the modified relation model produces concrete results for SGG, and the energy approach improves the baseline model marginally. However, a detailed predicate analysis revealed that the improvements of the rare predicate in the dataset follow the same trend as in full supervision, proving the main goal of this paper. This work can be enhanced if one finds suggestions for the problematic parts of the method such as proposing a better background scoring method, and aggregation of additional information for the energy model instead of leveraging only from the object and relation distributions. In addition, the problem setup can be also extended to caption-based supervision to employ a weaker but more general strategy.



(a)



(b)

Fig. 4: In-depth analysis of each predicate word separately for the modified energy model under full supervision (a), weak supervision (b).

REFERENCES

[1] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019.

[2] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments, 2019.

[3] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[4] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] Keren Ye and Adriana Kovashka. Linguistic structures as weak supervision for visual scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8289–8299, 2021.

[6] Federico Baldassarre, Kevin Smith, Josephine Sullivan, and Hossein Azizpour. Explanation-based weakly-supervised learning of visual relations with graph networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 612–630. Springer, 2020.

[7] Bo Wan, Yongfei Liu, Desen Zhou, Tinne Tuytelaars, and Xuming He. Weakly-supervised hoi detection via prior-guided bi-level representation learning. *arXiv preprint arXiv:2303.01313*, 2023.

[8] Jing Shi, Yiwu Zhong, Ning Xu, Yin Li, and Chenliang Xu. A simple baseline for weakly-supervised scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16393–16402, 2021.

[9] Xingchen Li, Long Chen, Wenbo Ma, Yi Yang, and Jun Xiao. Integrating object-aware and interaction-aware knowledge for weakly supervised scene graph generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4204–4213, 2022.

[10] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.

[11] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

[12] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

[13] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13936–13945, 2021.

[14] Bo Pang and Ying Nian Wu. Latent space energy-based model of symbol-vector coupling for text generation and classification. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8359–8370. PMLR, 18–24 Jul 2021.

[15] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[16] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018.

[17] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

[18] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018.

[19] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020.

[20] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017.

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[22] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.

[23] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[25] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2017.

[26] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark, 2018. Accessed: [24/05/2023].