# Hierarchical Graph Attention Network for Few-shot Visual-Semantic Learning

Chengxiang Yin[1], Kun Wu[1], Zhengping Che[2], Bo Jiang[2], Zhiyuan Xu[3], Jian Tang[3,*]
[1]Syracuse University, [2]Didi Chuxing, [3]Midea Group

## Abstract

*Deep learning has made tremendous success in computer vision, natural language processing and even visual-semantic learning, which requires a huge amount of labeled training data. Nevertheless, the goal of human-level intelligence is to enable a model to quickly obtain an in-depth understanding given a small number of samples, especially with heterogeneity in the multi-modal scenarios such as visual question answering and image captioning. In this paper, we study the few-shot visual-semantic learning and present the **H**ierarchical **G**raph **AT**tention network (HGAT). This two-stage network models the intra- and inter-modal relationships with limited image-text samples. The main contributions of HGAT can be summarized as follows: 1) it sheds light on tackling few-shot multi-modal learning problems, which focuses primarily, but not exclusively on visual and semantic modalities, through better exploitation of the intra-relationship of each modality and an attention-based co-learning framework between modalities using a hierarchical graph-based architecture; 2) it achieves superior performance on both visual question answering and image captioning in the few-shot setting; 3) it can be easily extended to the semi-supervised setting where image-text samples are partially unlabeled. We show via extensive experiments that HGAT delivers state-of-the-art performance on three widely-used benchmarks of two visual-semantic learning tasks.*

## 1. Introduction

Recently, significant progress has been made to various applications in single modality, such as object detection and machine translation, thanks to the advancing deep learning technologies [15] including convolutional [26] and recurrent [18] neural networks. However, in order for an artificial intelligence (AI) system to understand the real world around us, it requires not only the ability to memorize the rich information contained in a single modality, such as *visual* signals (i.e., images and videos) and natural *language* (i.e., captions and questions), but also a joint comprehension in multiple modalities. This, in general, is called multi-modal learning [5], and one typical example is the visual-semantic learning. For instance, a smart cooking robot in the kitchen is expected to make delicious dishes through understanding the recipe instructions as well as detecting and selecting the right ingredients on the table. The robot can hardly perform this task without the ability of either reading the texts or seeing the objects.

Lots of methods have attempted to address the multi-modal learning problems through dealing with visual-semantic learning tasks, such as visual question answering [32, 56, 54] and image captioning [10, 49, 12]. While these models are capable when massive human-annotated data and extensive training time are available, a real AI system should be able to quickly deliver an in-depth understanding using a small number of learning samples. The ability of solving visual-semantic tasks with limited samples, termed as few-shot visual-semantic learning, turns out to be challenging and critical for human-level intelligence.

Currently, for general few-shot learning problems, meta-learning [29, 36, 52] has become a standard methodology. Based on it, a few extensions have been recently made for few-shot visual-semantic learning. Fast Parameter Adaptation for Image-Text Modeling (FPAIT) [11] directly applied Model-Agnostic Meta-Learning (MAML) [13], a well-known meta-learning algorithm, to the few-shot visual question answering and image captioning. Analogously, another work [46] adopted a question answering model with two meta-learning techniques, prototypical networks [42] and meta networks [35]. Nonetheless, these attempts left much to be desired in terms of their scope and performance. Firstly and fundamentally, all these methods merely applied existing meta-learning algorithms without explicitly considering the multi-modal nature, to which we paid careful attention in this work. For example, Teney *et al*. [46] obtained their model input through a simple element-wise production between visual and semantic representations. Additionally, neither of them deal with the cases where labels are partially unlabeled, which is categorized as the semi-supervised learning setting. As labeling data can be expensive or even infeasible, semi-supervised learning is very

*Corresponding author

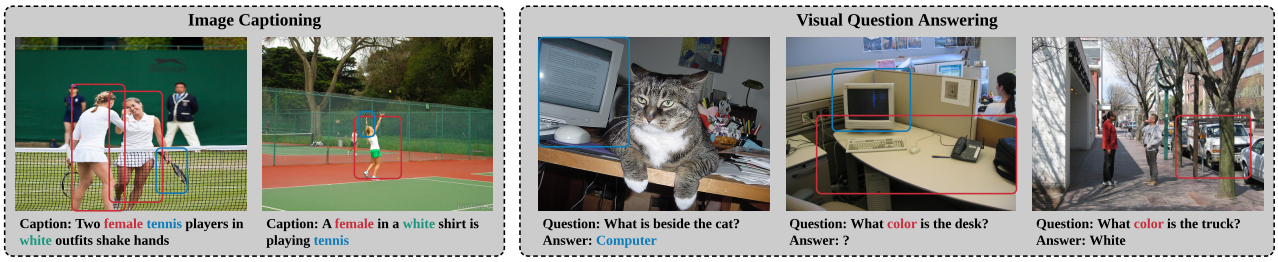


Figure 1. Image-text samples for image captioning (left) and visual question answering (right).

common and of great value in practice, and it becomes more severe when together with the few-shot setting.

For visual-semantic learning involving multiple modalities, especially the cases with limited and partially unlabeled image-text samples, it is vital to fully exploit the potential visual-semantic relationships, such as the intra-relationship of each modality (i.e., intra-modal relationships) and the inter-relationship between different modalities (i.e., inter-modal relationships). While the intra-modal learning have been examined meticulously such as Multimodal DBMs [43], the inter-modal learning leaves more space to explore and endows the model capacities to attentively capture complementary information.

We take two examples in Figure 1 for illustration. For the left-hand side image captioning example, both images contain female tennis players with white outfit. Correspondingly, the two ground-truth captions share the words "female", "white", and "tennis". In this example, even some words in the query caption are missing, the potential inter-modal relationship and the information captured from the visual modality can be used to supplement and strengthen the semantic modality and complete the caption. For the right-hand side visual question answering example, both the left two images contain a computer on the desk, quite different from the third image. Therefore, the predicted answer of the middle image is likely to be fooled (e.g., computer) by the visual similarity solely. Only if inter-relationship is captured through exploiting modal mutual interactions, the right visual clue can be distinguished from the distractors with the help of the semantic information and lead to the correct answer (i.e., white). While models based on Graph Neural Networks (GNNs) [16, 41] have been used to capture relational structures [48, 14, 22] in few-shot learning, they are incapable of jointly and subtly exploiting the intra- and inter-modal relationship for few-shot visual-semantic learning.

In this paper, to deal with the few-shot visual-semantic learning tasks, we propose the **H**ierarchical **G**raph **AT**tention network (HGAT). This two-stage network is able to model the intra-modal relationships and the inter-modal relationships with a few image-text samples and can be extended to a semi-supervised setting. In the first stage, visual-specific and semantic-specific GNNs are leveraged to model the intra-relationship of images and texts (i.e., visual-specific relationships and semantic-specific relationships), respectively. To model the inter-relationship between the visual and semantic modalities, an attention-based co-learning framework is presented to guide the node feature update of these GNNs. In the second stage, relation-aware GNNs are used to predict the result of the query sample by jointly learning visual representations, semantic representations, visual-specific relationships and semantic-specific relationships. We perform extensive experiments on three widely-used benchmarks, Toronto COCO-QA [40], Visual Genome-QA [25] and COCO-FITB [11], which showed that HGAT is a strong and effective model customized for few-shot visual-semantic learning.

The superiority of our proposed method can be summarized as follows: First, it sheds light on tackling few-shot multi-modal learning problems, especially for few-shot visual-semantic learning, a fairly new but critical setting for human-level intelligence, through taking advantage of the intra- and inter-modal relationships. Second, compared with FPAIT and several few-shot learning methods, it delivers state-of-the-art performance in terms of accuracy on both visual question answering and image captioning in the few-shot setting. In addition, several ablation experiments show the benefits of modeling of the visual-specific and semantic-specific relationships, the attention-based co-learning framework and the hierarchical graph-based architecture. Finally, it can be easily extended to the semi-supervised setting and delivers better performance compared with the other two graph-based methods.

## 2. Related Work

**Visual-Semantic Learning** Visual-semantic learning aims to build models that can process and relate the information for both visual and semantic modalities. Generally speaking, visual-semantic learning focuses on multimedia description tasks, such as visual question answering and image captioning. Various methods [1, 45, 55, 53] have been proposed for visual question answering. Recently, a multi-grained attention mechanism has been proposed [19] to address the failed cases on small objects or uncommon

concepts through learning word-object correspondence. Image captioning [3, 20, 17, 7, 51] aims to generate a natural language sentence to describe the image content. A recent work [47] introduces a hierarchical framework to explore both compositionality and sequentiality of natural language. Nevertheless, most of the existing visual-semantic learning works rely on a vast amount of human-annotated training data, which is very expensive. Conversely, the proposed model can deal with the cases with limited (or even partially unlabeled) data samples.

**Few-shot Learning** Few-shot learning has recently attracted extensive attention due to its superiority in learning with a few data samples. Mainstream approaches [44, 42, 50, 24, 28] on few-shot learning are based on similarity comparison among data samples using representation learning. Another popular approach [2, 39, 34, 4, 6] is to develop a meta-learner to optimize key hyper-parameters (e.g., initialization) of the learning model. A seminal work [13] presents a model-agnostic meta-learner to optimize the initialization of a learning model. However, all the works mentioned above focus only on the few-shot classification tasks, without careful consideration for the more complicated visual-semantic learning tasks, which involve multiple modalities. In this paper, we customize a model for few-shot visual-semantic learning.

**Graph Neural Networks** The Graph Neural Networks (GNNs) [16, 41] are used to deal with different types of graphs. Graph Attention Networks [48] can specify different weights to neighboring nodes by leveraging masked self-attentional layers. Additionally, GNNs can be employed for the few-shot classification problem. Garcia *et al*. [14] defines a node-labeling framework to cast few-shot learning as a supervised message passing task using GNNs. In contrast, EGNNs [22] learn to predict the edge-labels rather than the labels of nodes, and explicitly model the intra-cluster similarity and inter-cluster dissimilarity. Both GNNs and EGNNs can be extended to solve semi-supervised problems, while our model obtains a better performance on few-shot visual-semantic learning in the semi-supervised setting.

# 3. Methodology

In this section, we describe first the general definition and notations of visual-semantic learning, followed by its few-shot setting. Finally, we present the details about the proposed method.

## 3.1. Preliminaries

The general multi-modal problem aims to build models that process and relate information from multiple modalities [5]. We focus primarily, but not exclusively, on the visual and semantic modalities, and study the visual-semantic learning problem by tackling the *visual question answering (VQA)* and *image captioning (IC)* tasks. For VQA, given an image $\mathbf{I}$ and a related question $\mathbf{Q}$, we need to generate a corresponding answer $\mathbf{A}$. For IC, we follow the fill-in-the-blank setting [11], attempting to fill in the blank $\mathbf{A}$ of a given description $\mathbf{Q}$ for an image $\mathbf{I}$. Note that both the question/description $\mathbf{Q}$ and the answer/blank $\mathbf{A}$ are represented in a natural language format. Regularly, $\mathbf{A}$ is picked from a pre-defined set of different answers/labels. The traditional VQA and IC tasks seek a model $\mathbf{F}$, which can be a neural network, to map the observations $\mathbf{I}, \mathbf{Q}$ to the output $\mathbf{A}$.

## 3.2. Problem Statement

In *few-shot learning*, given only a few training samples, the model is expected to be able to adapt to a new task quickly. $N$-way $K$-shot problem settings are usually used to measure few-shot learning methods. Take an $N$-way $K$-shot VQA/IC task $\mathcal{T}$ with $M$ queries as an example: $\mathcal{T}$ consists of a support set $\mathcal{S}$ and a query set $\mathcal{Q}$, on which the model is learnt and evaluated respectively. $\mathcal{S}$ is a set of $N \times K$ samples, containing $K$ labeled image-text pairs for each of $N$ unique answers. $\mathcal{Q}$ contains another $M$ samples with the same answers as those in $\mathcal{S}$. Formally speaking, $\mathcal{T} = \mathcal{S} \cup \mathcal{Q}$, where $\mathcal{S} = \{(\mathbf{I}_i, \mathbf{Q}_i, \mathbf{A}_i)\}_{i=1}^{N\times K}$ and $\mathcal{Q} = \{(\mathbf{I}_i, \mathbf{Q}_i, \mathbf{A}_i)\}_{i=N\times K+1}^{N\times K+M}$; The label space of task $\mathcal{T}$ is defined as $\mathcal{C}_\mathcal{T} = \{\mathbf{A}_i\}_{i=1}^{N\times K}$. We have $\mathbf{A}_{(n-1)\times K+i} = \mathbf{A}_{(n-1)\times K+j}$ for $n = 1, \cdots, N$ and $1 \leq i, j \leq K$ (i.e., $|\mathcal{C}_\mathcal{T}| = N$), with $\{\mathbf{A}_i\}_{i=N\times K+1}^{N\times K+M} \subset \mathcal{C}_\mathcal{T}$.

In this work, we use *meta-learning* [13] to define few-shot visual-semantic learning problems, and it generally consists of two phases, meta-training and meta-testing. During the meta-training, a set of $T$ tasks $\{\mathcal{T}_t\}_{t=1}^{T}$ are generated from a meta-training dataset $\mathcal{D}_{\text{mtr}}$, and we develop a method that takes as input the support sets $\{\mathcal{S}_t\}_{t=1}^{T}$ and returns a model which minimizes the loss over the corresponding query sets $\{\mathcal{Q}_t\}_{t=1}^{T}$. During the meta-testing, another set of $T'$ tasks $\{\mathcal{T}_{T+t}\}_{t=1}^{T'}$ are generated from a meta-testing dataset $\mathcal{D}_{\text{mte}}$, and for $t = 1, \cdots, T'$, we expect the trained model can learn quickly from the $N \times K$ labeled image-text samples in the support set $\mathcal{S}_{T+t}$ and deliver highly-accurate labels for samples from the query set $\mathcal{Q}_{T+t}$. Note that the labels used in meta-training and meta-testing are mutually exclusive, i.e., $\mathcal{C}_{\text{mtr}} \cap \mathcal{C}_{\text{mte}} = \emptyset$ where $\mathcal{C}_{\text{mtr}} = \bigcup_{1\leq t\leq T} \mathcal{C}_{\mathcal{T}_t}$ and $\mathcal{C}_{\text{mte}} = \bigcup_{1\leq t\leq T'} \mathcal{C}_{\mathcal{T}_{T+t}}$. See details of the meta-training/testing in supplementary material.

Additionally, the problem can be extended to *semi-supervised learning* if a portion of labels in all support sets $\{\mathcal{S}_t\}_{t=1}^{T+T'}$ are unknown. In Section 4.3, the effectiveness of our model on semi-supervised setting will be presented.

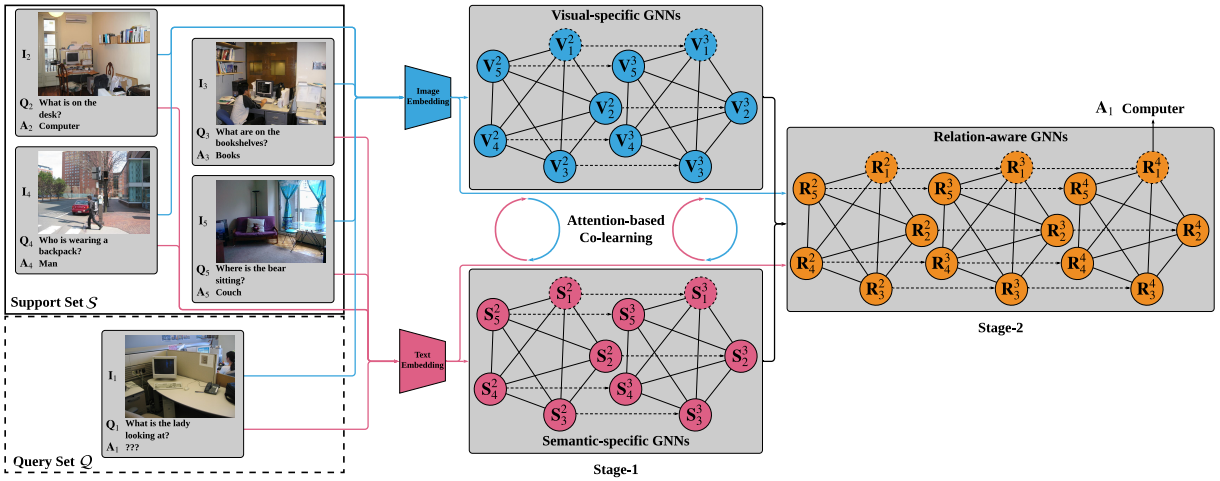


Figure 2. The architecture of Hierarchical Graph Attention Network. A 4-way 1-shot problem with one query sample ($N = 4$, $K = 1$, $M = 1$) is presented for simplicity. GNN nodes with solid line (Nodes 2-5) correspond to the samples from the support set $\mathcal{S}$, and nodes with dashed line (Node 1) correspond to the samples from the query set $\mathcal{Q}$. Dotted arrows between GNN layers represent node inheritances.

### 3.3. Hierarchical Graph Attention Network

This section describes the architecture of the Hierarchical Graph Attention Network (HGAT) as shown in Figure 2.

#### 3.3.1 Image and Text Embedding

To capture and preserve useful visual and semantic representations, we resort to modality-specific deep networks on image and text inputs. The neural networks for both image embedding and text embedding, i.e., $\phi(\cdot;\boldsymbol{\theta}_\phi)$ and $\psi(\cdot;\boldsymbol{\theta}_\psi)$, are jointly trained with other modules of HGAT. See more details of the model architecture for image embedding and text embedding in supplementary material.

#### 3.3.2 Graph Construction for Modal-specific GNNs

For each task $\mathcal{T}$, given the visual and semantic representations (extracted from the image and text embedding neural networks respectively) of all the image-text samples, we construct two graphs, the visual-specific GNNs (with blue nodes in Figures 2 and 3) and the semantic-specific GNNs (with red nodes), respectively. As shown in Figure 2, in Stage-1 of HGAT, both of the visual-specific and semantic-specific GNNs are two-layer GNNs. Each GNN layer contains $N \times K + M$ fully-connected nodes, and each node corresponds to an image-text sample from either the support set or the query set.

For each sample $(\mathbf{I}_i, \mathbf{Q}_i, \mathbf{A}_i)$, the feature vector of its corresponding node in the first layer of GNNs ($\mathbf{V}_i^1$ and $\mathbf{S}_i^1$) is initialized as the concatenation of its visual or semantic representation and the one-hot encoding of its label.

$$\mathbf{V}_i^1 = [\phi(\mathbf{I}_i;\boldsymbol{\theta}_\phi)||h(\mathbf{A}_i)] \quad (1)$$

$$\mathbf{S}_i^1 = [\psi(\mathbf{Q}_i;\boldsymbol{\theta}_\psi)||h(\mathbf{A}_i)] \quad (2)$$

where $||$ denotes vector concatenation operation, and $h(\mathbf{A}_i) \in [0,1]^N$ represents the one-hot encoding of the label $\mathbf{A}_i$. For any image-text sample $(\mathbf{I}_i, \mathbf{Q}_i, \mathbf{A}_i)$ from the query set $\mathcal{Q}$ or with unknown labels in the semi-supervised setting, we set $h(\mathbf{A}_i)$ to be a zero vector $\mathbf{0}^N$ instead.

For each node in the $l$th layer ($l > 1$), its feature vector is a concatenation of features inherited from its corresponding node in previous layer ($\mathbf{V}_i^{l-1}$ or $\mathbf{S}_i^{l-1}$) and an updated feature vector ($\mathbf{V}_i^l$ or $\mathbf{S}_i^l$) computed via the attention-based co-learning described in the following section.

#### 3.3.3 Attention-based Co-learning Framework

Each layer of the two modal-specific GNNs conducts associated node feature update in the proposed attention-based co-learning framework. For the node feature update in the $l$th layer ($l = 1, 2$), the inputs are two sets of nodes $\{\mathbf{V}_i^l\}_{i=1}^{N\times K+M}$, $\mathbf{V}_i^l \in \mathcal{R}^{F_V^l}$ and $\{\mathbf{S}_i^l\}_{i=1}^{N\times K+M}$, $\mathbf{S}_i^l \in \mathcal{R}^{F_S^l}$, and the outputs are two updated sets of nodes $\{\mathbf{V}_i^{l+1}\}_{i=1}^{N\times K+M}$, $\mathbf{V}_i^{l+1} \in \mathcal{R}^{2F_V^{l'}}$ and $\{\mathbf{S}_i^{l+1}\}_{i=1}^{N\times K+M}$, $\mathbf{S}_i^{l+1} \in \mathcal{R}^{2F_S^{l'}}$, where $F_V^l$, $2F_V^{l'}$, $F_S^l$, and $2F_S^{l'}$ represent the number of input and output feature channels of each node in the two modal-specific GNNs, respectively.

As an initial step, two shared learnable linear transformations, parametrized by $\mathbf{W}_V^l \in \mathcal{R}^{F_V^{l'}\times F_V^l}$ and $\mathbf{W}_S^l \in \mathcal{R}^{F_S^{l'}\times F_S^l}$, are applied to the two sets of nodes. Then, for each modal-specific GNN layer, a shared attentional mechanism $a$ is performed for each pair of nodes to compute the attention coefficients $e_{V_{ij}}^l \in \mathcal{R}$ and $e_{S_{ij}}^l \in \mathcal{R}$.

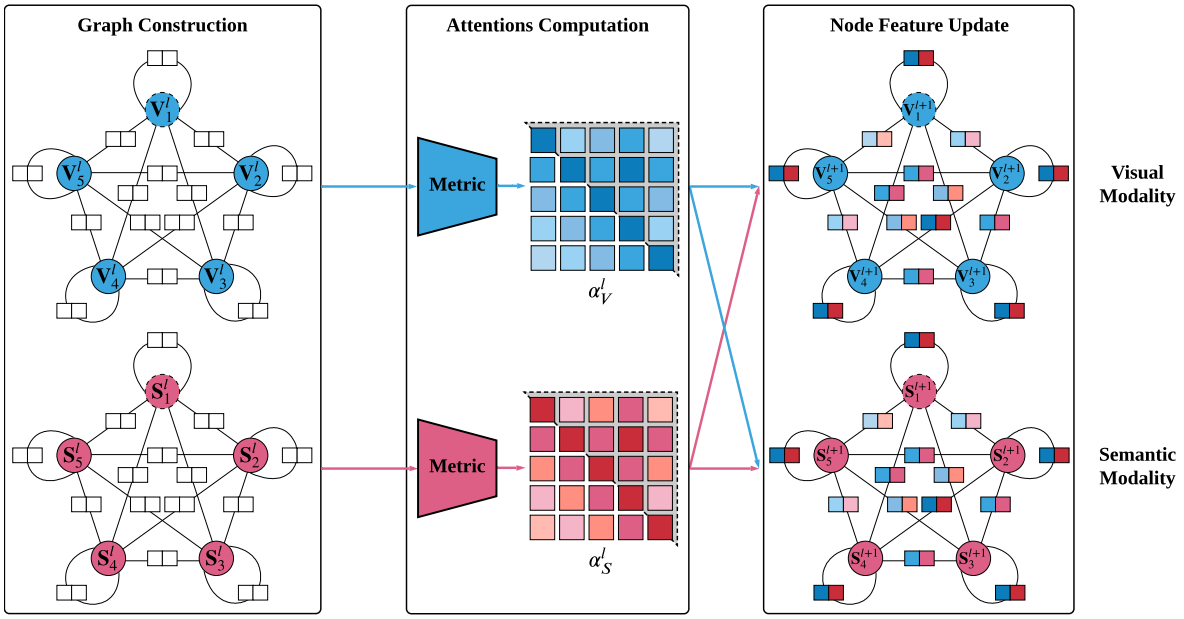


Figure 3. An illustration of the attention-based co-learning framework for one GNN layer. For simplicity, a 4-way 1-shot problem with one query sample ($N = 4$, $K = 1$, $M = 1$) is presented as an example. Nodes with solid line (Nodes 2-5) represent the samples from the support set $\mathcal{S}$, and nodes with dashed line (Node 1) represent the samples from the query set $\mathcal{Q}$. A two-dimensional attention is computed for each pair of nodes, to capture their relationship of the visual and semantic modalities, respectively. For simplicity, only half of the attentions (within the dotted triangles) are depicted in the node feature update.

$$e^l_{V_{ij}} = a(\mathbf{W}^l_V \mathbf{V}^l_i, \mathbf{W}^l_V \mathbf{V}^l_j) = \text{LReLU}\left({\mathbf{a}^l_V}^\top \left[\mathbf{W}^l_V \mathbf{V}^l_i || \mathbf{W}^l_V \mathbf{V}^l_j\right]\right) \tag{3}$$

$$e^l_{S_{ij}} = a(\mathbf{W}^l_S \mathbf{S}^l_i, \mathbf{W}^l_S \mathbf{S}^l_j) = \text{LReLU}\left({\mathbf{a}^l_S}^\top \left[\mathbf{W}^l_S \mathbf{S}^l_i || \mathbf{W}^l_S \mathbf{S}^l_j\right]\right) \tag{4}$$

where $e^l_{V_{ij}}$ and $e^l_{S_{ij}}$ indicate the importance of node $\mathbf{V}^l_j$ to node $\mathbf{V}^l_i$ in the visual-specific GNN and that of node $\mathbf{S}^l_j$ to node $\mathbf{S}^l_i$ in the semantic-specific GNN, respectively. LReLU denotes the Leaky Rectified Linear Unit [33] function. $\mathbf{a}^l_V \in \mathcal{R}^{2F^{l'}_V}$ and $\mathbf{a}^l_S \in \mathcal{R}^{2F^{l'}_S}$ both serve as learnable weight vectors, and $\mathbf{a}^\top$ represents the transpose of $\mathbf{a}$. The attentions $\alpha^l_{V_{ij}} \in \mathcal{R}$ and $\alpha^l_{S_{ij}} \in \mathcal{R}$ are obtained by normalizing the attention coefficients using the softmax function.

$$\alpha^l_{V_{ij}} = \text{softmax}(e^l_{V_{ij}}) = \frac{\exp(e^l_{V_{ij}})}{\sum_{k=1}^{N\times K+M} \exp(e^l_{V_{ik}})} \tag{5}$$

$$\alpha^l_{S_{ij}} = \text{softmax}(e^l_{S_{ij}}) = \frac{\exp(e^l_{S_{ij}})}{\sum_{k=1}^{N\times K+M} \exp(e^l_{S_{ik}})} \tag{6}$$

where $\alpha^l_{V_{ij}}$ and $\alpha^l_{S_{ij}}$ represent the attentions of the visual modality (blue-hue square matrix in Figure 3) and those of the semantic modality (red-hue square matrix), respectively. The values of $\alpha^l_{V_{ij}}$ and $\alpha^l_{S_{ij}}$ are expected to be positively correlated.

Once obtained, both attentions are shared by the associated node feature update of the two modal-specific GNNs. For example, the attentions of visual modality not only serve for the node feature update of the visual-specific GNNs but also utilize the relationship on visual modality to refine the semantic-specific GNNs. Analogously, the attentions of semantic modality are also helpful to both semantic-specific and visual-specific GNNs.

$$\mathbf{V}^{l+1}_i = \text{ELU}\left(\sum_{j=1}^{N\times K+M} \alpha^l_{V_{ij}} \mathbf{W}^l_V \mathbf{V}^l_j || \sum_{j=1}^{N\times K+M} \alpha^l_{S_{ij}} \mathbf{W}^l_V \mathbf{V}^l_j\right) \tag{7}$$

$$\mathbf{S}^{l+1}_i = \text{ELU}\left(\sum_{j=1}^{N\times K+M} \alpha^l_{V_{ij}} \mathbf{W}^l_S \mathbf{S}^l_j || \sum_{j=1}^{N\times K+M} \alpha^l_{S_{ij}} \mathbf{W}^l_S \mathbf{S}^l_j\right) \tag{8}$$

where $\alpha^l_{V_{ij}}$ and $\alpha^l_{S_{ij}}$ form the two-dimensional attention between each pair of nodes in both modal-specific GNNs and ELU denotes the Exponential Linear Unit [9] function.

Based on the attentions shared by the visual and semantic modalities, the associated node feature update is conducted, and the inter-modal relationships are modeled under the attention-based co-learning framework. Note that while the basic attention mechanism used here follows the Graph Attention Network [48], our proposed attention-based co-

learning framework is agnostic to the particular choice of attention mechanism.

#### 3.3.4 Relation-aware GNNs

The layerwise outputs of the modal-specific GNNs, i.e., $\{\mathbf{V}_i^{l+1}\}_{i=1}^{N\times K+M}$ and $\{\mathbf{S}_i^{l+1}\}_{i=1}^{N\times K+M}$ for $l = 1, 2$, extracting the hierarchical features with intra-relationship and inter-relationship from both the visual and semantic modalities, are further exploited by the relation-aware GNNs in Stage-2 of HGAT.

We construct the relation-aware GNNs with $N \times K + M$ nodes in each layer, which share similar structure with the modal-specific GNNs but also take the relationships obtained in Stage-1 for the feature initialization of each node. To be more specific, for the node feature update in the $l$th layer ($l = 1, 2, 3$), the inputs is a set of nodes $\{\mathbf{R}_i^l\}_{i=1}^{N\times K+M}$, $\mathbf{R}_i^l \in \mathcal{R}^{F_R^l}$ and the outputs are an updated set of nodes $\{\mathbf{R}_i^{l+1}\}_{i=1}^{N\times K+M}$, $\mathbf{R}_i^{l+1} \in \mathcal{R}^{F_R^{l'}}$ where $F_R^l$, $F_R^{l'}$ represent the number of input and output feature channels of each node in the relation-aware GNNs, respectively. The input to the first layer $\mathbf{R}_i^1$ is the concatenation of the visual and semantic embeddings, the one-hot encoding of the label, and the multi-modal features obtained in Stage-1.

$$\mathbf{R}_i^1 = \left[\phi(\mathbf{I}_i;\boldsymbol{\theta}_\phi)||\psi(\mathbf{Q}_i;\boldsymbol{\theta}_\psi)||h(\mathbf{A}_i)||\mathbf{V}_i^2||\mathbf{V}_i^3||\mathbf{S}_i^2||\mathbf{S}_i^3\right] \quad (9)$$

The input to the $l$th layer ($l > 1$) is a concatenation of features inherited from its corresponding node in previous layer $\mathbf{R}_i^{l-1}$ and an updated feature vector $\mathbf{R}_i^l$, which is computed in a similar way to the modal-specific GNNs. First, the attention coefficient $e_{R_{ij}}^l \in \mathcal{R}$ indicating the importance of node $\mathbf{R}_j^l$ to node $\mathbf{R}_i^l$ is calculated.

$$e_{R_{ij}^l} = a(\mathbf{W}_R^l\mathbf{R}_i^l, \mathbf{W}_R^l\mathbf{R}_j^l) = \text{LReLU}\left(\mathbf{a}_R^{l\top}\left[\mathbf{W}_R^l\mathbf{R}_i^l||\mathbf{W}_R^l\mathbf{R}_j^l\right]\right) \quad (10)$$

where $\mathbf{W}_R^l \in \mathcal{R}^{F_R^{l'}\times F_R^l}$ and $\mathbf{a}_R^l \in \mathcal{R}^{2F_R^{l'}}$ are learnable parameters. Then the attentions are computed by normalizing the attention coefficients using softmax function.

$$\alpha_{R_{ij}}^l = \text{softmax}(e_{R_{ij}}^l) = \frac{\exp(e_{R_{ij}}^l)}{\sum_{k=1}^{N\times K+M}\exp(e_{R_{ik}}^l)} \quad (11)$$

Afterwards, the attentions are used to compute the updated node features through a linear combination of the corresponding features, followed by a non-linearity activation.

$$\mathbf{R}_i^{l+1} = \text{ELU}\left(\sum_{j=1}^{N\times K+M}\alpha_{R_{ij}}^l\mathbf{W}_R^l\mathbf{R}_j^l\right) \quad (12)$$

Finally, to get the final prediction of the $i$th sample from HGAT, we set the last output dimension $F_R^{3'}$ to $N$, and use $\text{softmax}(\mathbf{R}_i^4) \in [0, 1]^N$ as the confidence score vector over the $N$ answers. The predicted label is $\hat{\mathbf{A}}_i = \text{argmax}_n\mathbf{R}_{i,n}^4$, where $\mathbf{R}_{i,n}^4$ is the $n$th element of $\mathbf{R}_i^4$ and $1 \le n \le N$.

### 3.4. Training HGAT

Given a set of $T$ tasks in the meta-training phase, the learnable parameters of the proposed HGAT, $\boldsymbol{\theta}_\phi \cup \boldsymbol{\theta}_\psi \cup \{\mathbf{W}_V^l, \mathbf{W}_S^l, \mathbf{a}_V^l, \mathbf{a}_S^l\}_{l=1}^2 \cup \{\mathbf{W}_R^l, \mathbf{a}_R^l\}_{l=1}^3$, are trained in an end-to-end manner by minimizing the following loss function over the task set.

$$\mathcal{L} = \frac{1}{TM}\sum_{\mathcal{T}\in\{\mathcal{T}_t\}_{t=1}^T}\sum_{i=N\times K+1}^{N\times K+M}\mathcal{L}_c(\mathbf{A}_i, \hat{\mathbf{A}}_i) \quad (13)$$

where $\mathcal{L}_c$ is defined as the cross-entropy loss, $\mathbf{A}_i$ and $\hat{\mathbf{A}}_i$ represent the ground truth answer and the predicted answer of the image-text samples from the query set $\mathcal{Q}$.

## 4. Performance Evaluation

We employed three compelling benchmarks, Toronto COCO-QA [40], Visual Genome-QA [25] and COCO-FITB [11] to evaluate the proposed HGAT on two typical visual-semantic learning tasks, visual question answering (VQA) and image captioning (IC).

### 4.1. Benchmark Datasets

| Benchmark | | TC-QA | VG-QA | COCO-FITB |
|---|---|---|---|---|
| Task | | VQA | VQA | IC |
| #Pair | Meta-training | 57,834 | 554,795 | 181,844 |
| | Meta-testing | 13,965 | 136,473 | 34,919 |
| #Class | Meta-training | 256 | 244 | 159 |
| | Meta-testing | 65 | 82 | 43 |

Table 1. Statistics on the three benchmark datasets. (TC-QA: Toronto COCO-QA; VG-QA: Visual Genome-QA.)

Table 1 shows the statistics on the three benchmark datasets for few-shot VQA and few-shot IC tasks. See more details of the three benchmark datasets, including preprocessing, in supplementary material.

### 4.2. Experimental Setup

**Few-shot setup** Following the common setup in few-shot learning [13, 42], for each task $\mathcal{T}$ of $N$-way $K$-shot learning, we set $N \in \{5, 10\}$, $K \in \{1, 5\}$ and $M = 1$. Take a 10-way 5-shot VQA task for example: given 10 different answers, each answer has 5 labeled image-question pairs, and these 50 samples serve as the support set to predict the result out of the 10 answers for the 1 unlabeled image-question pair from the query set. Therefore, we can evaluate both VQA and IC tasks in terms of the standard classification accuracy.

**Implementation details** In the meta-training phase, the proposed model was trained with Adam optimizer [23] with

| Method | Toronto COCO-QA | | | | Visual Genome-QA | | | | COCO-FITB | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 5-way accuracy | | 10-way accuracy | | 5-way accuracy | | 10-way accuracy | | 5-way accuracy | | 10-way accuracy | |
| | 1-shot | 5-shot | 1-shot | 5-shot | 1-shot | 5-shot | 1-shot | 5-shot | 1-shot | 5-shot | 1-shot | 5-shot |
| FPAIT | 59.38 | 71.92 | 45.11 | 60.20 | 75.49 | 79.12 | 61.66 | 67.62 | 60.13 | 70.88 | 47.10 | 59.31 |
| FPAIT+CLT | 60.61 | 72.17 | 46.37 | 60.92 | 75.05 | 79.28 | 60.82 | 67.48 | 61.01 | 71.13 | 47.79 | 60.91 |
| Prototypical Net | 60.12 | 71.72 | 45.31 | 59.67 | 75.43 | 80.33 | 62.32 | 67.23 | 60.56 | 71.16 | 47.52 | 59.38 |
| Relation Net | 61.75 | 71.89 | 45.60 | 60.13 | 77.21 | 80.72 | 63.14 | 68.10 | 61.35 | 71.68 | 47.92 | 59.55 |
| R2D2 | 61.83 | 72.60 | 47.13 | 59.36 | 77.44 | 81.08 | 64.71 | 71.55 | 60.87 | 71.60 | 47.73 | 59.33 |
| DN4 | 62.60 | 74.12 | 47.68 | 60.44 | 78.33 | 84.25 | 64.92 | 71.20 | 62.09 | 73.62 | 48.57 | 60.82 |
| GNN | 61.42 | 72.55 | 46.35 | 58.95 | 76.72 | 81.43 | 63.19 | 68.65 | 61.85 | 72.70 | 48.14 | 59.86 |
| EGNN | 62.21 | 73.41 | 46.99 | 60.01 | 77.67 | 83.26 | 64.07 | 70.87 | 62.67 | 72.98 | 48.22 | 60.13 |
| HGAT | **63.13** | **75.41** | **48.10** | **61.50** | **79.56** | **86.10** | **66.62** | **72.13** | **63.36** | **74.14** | **49.26** | **61.31** |

Table 2. Comparison of accuracy on Toronto COCO-QA, Visual Genome-QA, and COCO-FITB.

an initial learning rate of $1 \times 10^{-3}$ and weight decay of $1 \times 10^{-6}$. The task mini-batch sizes were set to 128, 32, 64, and 16 for 5-way 1-shot, 5-way 5-shot, 10-way 1-shot, and 10-way 5-shot, respectively. Our code was implemented in PyTorch [37] and run with NVIDIA Tesla P100 GPUs.

**Baselines** FPAIT [11] directly leverages MAML [13] to deal with few-shot VQA and IC tasks; Prototypical Net [42], Relation Net [44], R2D2 [6], and DN4 [28] focus on few-shot classification. GNN [14] and EGNN [22] are two GNN-based few-shot classification models. None of these algorithms, including MAML, has paid any attention to the few-shot visual-semantic learning, but it is noting that all of them can be extended to tackle few-shot VQA and IC as few-shot classification tasks. See implementation details of baseline methods in supplementary material.

### 4.3. Experimental Results

Results on the three benchmarks are shown in Table 2, and we can make the following observations:

1) HGAT outperforms all baselines in terms of classification accuracy in all settings. Concretely, in the case of 5-way 5-shot VQA on Toronto COCO-QA, HGAT gives a accuracy of 75.41%, excelling the second best by 1.29 percentage points, which indicates that the modeling of the intra- and inter-modal relationships using the hierarchical graph-based structure can lead to consistent advantages on few-shot visual-semantic learning. Similar trends can be observed for other test cases and benchmark datasets.

2) Among the graph-based methods, our HGAT brings noticeable improvements over GNN and EGNN. For example, when few-shot VQA is conducted on Visual Genome-QA, HGAT obtains classification accuracies of 79.56%, 86.10%, 66.62% and 72.13% in the four test cases, respectively, which are 1.89%, 2.84%, 2.55% and 1.26% higher than those of EGNN. Similar improvements can be observed regarding GNN. Although GNN and EGNN utilize the pairwise relationships of nodes, the intra-relationship of each modality, as well as the inter-relationship between different modalities, have not been fully exploited.

To further justify the superiority of HGAT for the few-shot visual-semantic learning tasks, experimental comparisons have been expanded to standard VQA and IC methods that are not specifically designed for few-shot learning, including HCA [31], SAAA [21], and CNN+TCN [11]. Please see the details in supplementary material.

### 4.4. Ablation Studies

| Case | S1 | S2 | VR | SR | AC | 5-way accuracy | | 10-way accuracy | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 1-shot | 5-shot | 1-shot | 5-shot |
| 1 | | ✓ | | | | 76.10 | 82.14 | 63.99 | 66.30 |
| 5 | ✓ | | ✓ | ✓ | ✓ | 77.47 | 83.26 | 64.90 | 70.03 |
| 6 | ✓ | ✓ | ✓ | | | 77.55 | 84.01 | 63.77 | 69.26 |
| 7 | ✓ | ✓ | | ✓ | | 78.14 | 83.88 | 64.23 | 68.61 |
| 8 | ✓ | ✓ | ✓ | ✓ | | 78.86 | 84.55 | 65.21 | 70.06 |
| 9 | ✓ | ✓ | ✓ | ✓ | ✓ | **79.56** | **86.10** | **66.62** | **72.13** |

Table 3. Ablation studies on Visual Genome-QA. (S1: Stage-1; S2: Stage-2; VR: Visual Relations; SR: Semantic Relations; AC: Attention-based Co-learning; Please see the full table in supplementary material.)

To validate the superiority of the proposed HGAT, several ablation experiments were conducted based on Visual Genome-QA for few-shot VQA. The following observations are made based on Table 3:

1) HGAT conducts separate exploitation of the intra-relationship of each modality, which can lead to better performance. Compared with Case-1, where no intra-modal relationships are exploited, there exists a jump on accuracy when the visual-specific relationships are modeled in Case-6. A similar improvement can be observed in Case-7, where semantic-specific relationships are modeled. Moreover, an additional gain can be noticed if both visual- and semantic-specific relationships are exploited in Case-8.

2) To validate the effectiveness of the attention-based co-learning framework, the experiment is conducted in Case-9, which achieves 0.70%, 1.55%, 1.41% and 2.07% im-

provements over Case-8. Note that the attention-based co-learning can only be achieved when both visual- and semantic-specific GNNs are leveraged in Stage-1.

3) The relation-aware GNNs in Stage-2 can deliver an additional performance gain on few-shot visual-semantic learning. For instance, compared with Case-5, where the relation-aware GNNs are replaced by fully-connected neural networks for label prediction, Case-9 brings up improvements of $2.09\%$, $2.84\%$, $1.72\%$ and $2.10\%$ on accuracies.

4) It should be noted that the Case-1 with only Stage-2 represents a 3-layer GNNs, and the initial feature of each node is the concatenation of the corresponding visual and semantic representations as well as the one-hot encoding of label. Case-1 performs comparably with the graph-based methods, GNN and EGNN as expected.

Moreover, for experimental analysis about the number of GNN layers, please refer to the supplementary material.

### 4.5. Semi-supervised Few-shot Learning

| Toronto COCO-QA | 5-way 5-shot accuracy | | | |
|---|---|---|---|---|
| | 40% | 60% | 80% | 100% |
| GNN-LabeledOnly | 64.62 | 67.30 | 70.31 | 72.55 |
| GNN-Semi | 66.04 | 68.44 | 71.48 | 72.55 |
| EGNN-LabeledOnly | 65.86 | 69.08 | 71.57 | 73.41 |
| EGNN-Semi | **67.18** | 69.92 | 72.61 | 73.41 |
| HGAT-LabeledOnly | 66.09 | 69.83 | 73.12 | **75.41** |
| HGAT-Distractor | 64.25 | 68.94 | 73.01 | **75.41** |
| HGAT-Semi | 67.16 | **70.78** | **73.95** | **75.41** |

Table 4. Comparison of semi-supervised learning results on Toronto COCO-QA for few-shot visual question answering.

Table 4 presents the comparisons of semi-supervised learning among HGAT, GNN, and EGNN. Experiments are conducted on the 5-way 5-shot VQA on Toronto COCO-QA, and results are presented when $40\%$, $60\%$, $80\%$ of the image-text samples are labeled. Note that the labeled samples are balanced among the $5$ classes. Take the $40\%$ case for example, for a task, each of the class contains 2 labeled samples and 3 unlabeled samples from the support set. 'LabeledOnly' is equivalent to the supervised few-shot setting, where only the labeled support samples are used. For instance, the 5-way 5-shot $40\%$ VQA with 'LabeledOnly' is equivalent to the 5-way 2-shot VQA. 'Semi' denotes the semi-supervised few-shot setting, where all the support samples are used, regardless of whether they are labeled. In addition, 'Distractor' means the unlabeled support samples are randomly sampled from other classes instead of the $5$ classes of the labeled support samples.

Besides, each of the three methods can acquire noticeable improvements when semi-supervised learning is performed compared with 'LabeledOnly' which demonstrates the unlabeled support samples can contribute to the learning in a few-shot setting. Moreover, for the proposed HGAT, the 'Distractor' leads to a minor performance degradation for each case compared with 'LabeledOnly'. This observation clearly shows that only the unlabeled samples from the classes of interest can contribute to the few-shot visual-semantic learning. Notably, for semi-supervised few-shot visual-semantic learning, the HGAT consistently outperforms the GNN and EGNN, except the $40\%$ case, where HGAT achieves a comparable accuracy given by EGNN.

### 4.6. Visualization

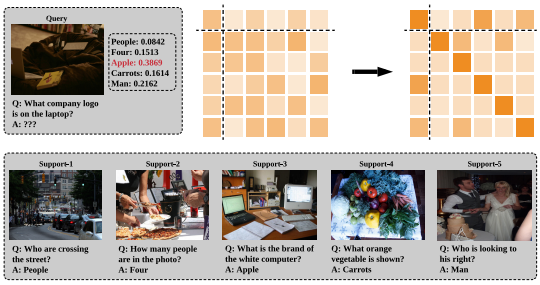


Figure 4. Attention visualizations of the 3rd layer in the relation-aware GNNs for 5-way 1-shot VQA on Visual Genome-QA. Dark/light color denotes higher/lower values. See more visualization samples in supplementary material.

Figure 4 shows the computed attentions for a 5-way 1-shot VQA task sampled from the meta-testing set. The left and right square matrix represent the attentions before and after the meta-training, respectively. We can notice that the attention between the query sample and the third support sample is larger than other off-diagonal attentions, which implies a stronger correlation between these two samples. Though the "apple" trademark, which is the decisive clue, occupies only a small portion of both images, HGAT can still associate the query to the support sample within the same class and give the correct answer.

## 5. Conclusions

This paper, through introducing **H**ierarchical **G**raph **AT**tention network (HGAT), presents a novel method for few-shot visual-semantic learning. Comprehensive experiments have been conducted on the widely-used Toronto COCO-QA, Visual Genome-QA and COCO-FITB benchmarks. The extensive experimental results have shown that 1) HGAT delivers the state-of-the-art performance in terms of accuracy on both few-shot VQA and IC tasks compared with few-shot learning and standard (non-few-shot) methods; 2) It sheds light on tackling the few-shot multi-modal learning problems, especially for the few-shot visual-semantic learning tasks, through hierarchical exploitation and co-learning of the multiple modalities. 3) It can be easily extended to the semi-supervised setting, outperforming other few-shot visual-semantic learning baselines in the semi-supervised setting.

## References


[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of IEEE CVPR*, 2018.

[2] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *NIPS*, pages 3981–3989, 2016.

[3] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Convolutional image captioning. In *Proceedings of IEEE CVPR*, pages 5561–5570, 2018.

[4] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. *ICLR*, 2019.

[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[6] Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *ICLR*, 2019.

[7] Fuhai Chen, Rongrong Ji, Jiayi Ji, Xiaoshuai Sun, Baochang Zhang, Xuri Ge, Yongjian Wu, Feiyue Huang, and Yan Wang. Variational structured semantic inference for diverse image captioning. In *NIPS*, pages 1929–1939, 2019.

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[9] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *ICLR*, 2016.

[10] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of IEEE CVPR*, pages 2625–2634, 2015.

[11] Xuanyi Dong, Linchao Zhu, De Zhang, Yi Yang, and Fei Wu. Fast parameter adaptation for few-shot image captioning and visual question answering. In *Proceedings of the ACMMM*, pages 54–62, 2018.

[12] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *Proceedings of IEEE CVPR*, pages 4125–4134, 2019.

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135. JMLR. org, 2017.

[14] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In *ICLR*, 2018.

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[16] Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings of IEEE IJCNN*, volume 2, pages 729–734. IEEE, 2005.

[17] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *NIPS*, pages 11135–11145, 2019.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[19] Pingping Huang, Jianhui Huang, Yuqing Guo, Min Qiao, and Yong Zhu. Multi-grained attention with object-level grounding for visual question answering. In *ACL*, 2019.

[20] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *ECCV*, pages 499–515, 2018.

[21] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017.

[22] Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-labeling graph neural network for few-shot learning. In *Proceedings of the IEEE CVPR*, pages 11–20, 2019.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[24] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML*, volume 2. Lille, 2015.

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[27] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of IEEE CVPR*, pages 156–165, 2017.

[28] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *Proceedings of the IEEE CVPR*, pages 7260–7268, 2019.

[29] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[31] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016.

[32] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In *Proceedings of the AAAI*, 2016.

[33] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, volume 30, page 3, 2013.

[34] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *ICLR*, 2018.

[35] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *ICML*, pages 2554–2563. JMLR. org, 2017.

[36] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[37] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *NIPS Autodiff Workshop*, 2017.

[38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543, 2014.

[39] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

[40] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *NIPS*, pages 2953–2961, 2015.

[41] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2008.

[42] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, 2017.

[43] Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, pages 2222–2230, 2012.

[44] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE CVPR*, pages 1199–1208, 2018.

[45] Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of IEEE CVPR*, pages 4223–4232, 2018.

[46] Damien Teney and Anton van den Hengel. Visual question answering as a meta learning task. In *ECCV*, 2018.

[47] Junjiao Tian and Jean Oh. Image captioning with compositional neural module networks. In *Proceedings of the AAAI*, pages 3576–3584, 2019.

[48] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[49] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of IEEE ICCV*, 2015.

[50] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016.

[51] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *Proceedings of IEEE ICCV*, pages 2621–2629, 2019.

[52] Chengxiang Yin, Jian Tang, Zhiyuan Xu, and Yanzhi Wang. Adversarial meta-learning. *arXiv preprint arXiv:1806.03316*, 2018.

[53] Chengxiang Yin, Jian Tang, Zhiyuan Xu, and Yanzhi Wang. Memory augmented deep recurrent neural network for video question answering. *IEEE transactions on neural networks and learning systems*, 2019.

[54] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of IEEE CVPR*, pages 6281–6290, 2019.

[55] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of IEEE CVPR*, pages 8807–8817, 2019.

[56] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Proceedings of the AAAI*, 2017.