GAFNet: A Global Fourier Self Attention Based Novel Network for multi-modal downstream tasks

Onkar Susladkar
Independent Researcher
onkarsus13@gmail.com

Gayatri Deshmukh
Independent Researcher
dgayatri9850@gmail.com

Dhruv Makwana
Independent Researcher
dmakwana503@gmail.com

Sparsh Mittal
IIT Roorkee, India
sparsh.mittal@ece.iitr.ac.in

R Sai Chandra Teja
Independent Researcher
saichandrateja@gmail.com

Rekha Singhal
TCS Research, India
rekha.singhal@tcs.com

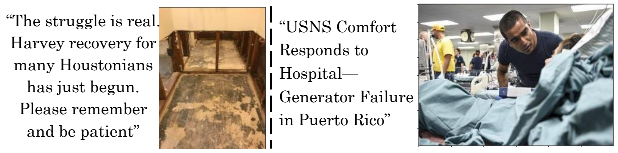
Abstract

In "vision and language" problems, multimodal inputs are simultaneously processed for combined visual and textual understanding for image-text embedding. In this paper, we discuss the necessity of considering the difference between the feature space and the distribution when performing multimodal learning. We deal with this problem through deep learning and a generative model approach. We introduce a novel network, GAFNet (Global Attention Fourier Net), which learns through large-scale pre-training over three image-text datasets (COCO, SBU, and CC-3M), for achieving high performance on downstream vision and language tasks. We propose a GAF (Global Attention Fourier) module, which integrates multiple modalities into one latent space. GAF module is independent of the type of modality, and it allows combining shared representations at each stage. Various ways of thinking about the relationships between different modalities directly affect the model's design. In contrast to previous research, our work considers visual grounding as a pretrainable and transferable quality instead of something that must be trained from scratch. We show that GAFNet is a versatile network that can be used for a wide range of downstream tasks. Experimental results demonstrate that our technique achieves state-of-the-art performance on multimodal classification on the CrisisMD dataset and image generation on the COCO dataset. For image-text retrieval, our technique achieves competitive performance.

1. Introduction

Multimodal information dramatically increases the effectiveness of communication. Whether a news article or a textbook, multiple modalities such as text, image, audio and video make the information easily accessible to a wide range of audiences. Further, in AI-based decision-making systems, multiple modalities can improve predictive performance [23, 18, 5].

In a multimodal learning problem setting, a network with each modality as input is prepared. A learner that connects the final layers of each network is trained to create a joint representation of all modalities. However, a challenge in multimodal learning is the heterogeneity between the modality, which refers to significant differences between modalities such as images and text. The datasets of different modalities may have different dimensions/structures, and distributions [33]. For example, in a tweet with both image and text, only the image may contain the complete information, or only the text may contain the full information. This is referred to as the deficiency in modality; Figure 1 shows some examples.


"The struggle is real. Harvey recovery for many Houstonians has just begun. Please remember and be patient"
"USNS Comfort Responds to Hospital—Generator Failure in Puerto Rico"

(a) GT = Other relevant information (b) GT = Affected individuals

Figure 1. Examples of multimodal tweets. In (a) and (b), ground truth (GT) is inferred primarily from text and image, respectively. This shows the challenges of multimodal learning.

Similarly, an image may have a sequence length of 256, whereas the text may have a length of four ("Wildfires strain California hospital"). Thus, their sequence lengths do not match. A single word in the text may correspond to a large portion of the image, or many words in the text may describe a small portion of the image. Conventional networks cannot describe explicit relationships between modalities.

This work was supported in part by IIT Roorkee, India under grant FIG-100874.

If the distributions between different datasets are different, one-to-one deterministic mapping between modalities cannot be designed. Therefore, we need to learn non-deterministic relationships by using approaches such as approximately selecting the closest one after creating a deterministic relationship (rule) [27] or using domain adaptation method [33].

Contributions: In this paper, we propose a novel and versatile network named GAFNet, which is useful for various problem settings in multimodal learning. GAFNet uses a novel module named GAF module, which uses a global attention mechanism in a multimodal scenario. The GAF module circumvents model collapse due to modality deficiency by using a shared representation for multiple modalities. GAFNet has three branches: image, text, and a middle branch that combines image and textual representations. GAFNet processes image and text inputs in separate branches using vision transformers (ViT) and BERT, respectively. These branches interact through GAF blocks.

GAF module can effectively handle the differences in dimensions, feature space and distributions of various modalities. GAF module works independently of the dimension or structure of the input modality. GAF module obtains an appropriate shared representation that integrates all modalities and has better accuracy than the conventional model. For capturing modality-invariant elements at the semantic level, we pre-train GAFNet on three datasets, viz., COCO, SBU and CC-3M.

While GAFNet can be deployed for a variety of multimodal applications, we showcase its efficacy by evaluating it on three tasks: multimodal tweet classification (Section 5.1), image generation under text guidance (Section 5.2) and image-text retrieval (Section 5.3). To use GAFNet for other tasks, a user needs to adjust the layers after the last GAF block. We showcase that multimodal learning can enhance generative models. For multimodal classification on the CrisisMMD dataset and image generation on the COCO dataset, GAFNet achieves SOTA (state-of-the-art) performance. For image-text retrieval, GAFNet achieves competitive performance on MSCOCO and Flickr30K datasets.

2. Related Work

In recent multimodal learning, the methods using deep neural networks have become the mainstream [23, 27, 4]. In multimodal learning, a network with each modality as input is prepared, and a learner that connects the final layers of each network is trained to create a joint representation of all the modalities. There is a growing interest in developing pre-training objectives for tasks with multiple modalities based on the success of self-supervised learning in intra-modal tasks (e.g., vision and language). For example, the CLIP [25] technique predicts the correspondence between the text and the image, resulting in a task agnostic model that is competitive with task-specific supervised models to leverage a much broader source of supervision from the text. Using a noisy dataset of over one billion image alt-text pairs, ALIGN [14] further scales up CLIP. The object detector (e.g., Faster R-CNN [26]) is used to capture vision features first; then a multi-layer transformer [28] is applied to the concatenation of the extracted vision and text features to learn joint embeddings.

SOHO [12] uses a visual dictionary to extract compact image characteristics from a complete picture. This leads to nearly 10× lower inference time than region-based approaches. For long-range dependencies spanning a series of fixed-size non-overlapping picture patches, ViLT [17] completely discards convolutional visual characteristics and utilizes vision transformers [8]. Osolo et al. [24] propose a transformer-based image captioning technique that uses Fourier transforms to improve efficiency while requiring fewer operations. While it provides high performance, it has a low inductive impact on the network since it combines data from the frequency and spatial domains.

3. Architecture of Proposed GAFNet network

Figure 2(a) shows the block diagram of our proposed network. The network has three branches: image, text, and a middle branch that combines image and textual representations. GAFNet processes image and text inputs in separate branches, and these branches interact through GAF blocks. For image modality, we use a pre-trained Vision Transformer (ViT) trained on the Imagenet21M dataset. For the text modality, we use a pre-trained Bert-base-uncased [1], or simply BERT. GAFNet uses an equal number (12) of ViT, BERT, and Global Attention Fourier (GAF) blocks to ensure that both modalities contribute equally. The model receives image and text representations as input. Then, for each representation, positional embeddings are computed. These image and text embeddings are supplied to the initial ViT and BERT blocks, respectively.

As in the transformer, intermediate representation is utilized to construct three matrices, Q, K, and V, which correspond to queries, keys, and values, respectively. This drives the transformer's multi-headed self-attention block. GAFNet exchanges the key and value between the BERT and ViT blocks at each step. ViT and BERT use the dot-product of the exchanged key vector and their own query vector to get attentional distributions across the value vector. This facilitates co-attention between textual and visual representations. This co-attention enables each ViT and BERT transformer block to exchange information. The outputs of textual and visual transformer blocks are fed into the GAF block, which reconciles the information present

The source code can be obtained from https://candlelabai.github.io/gaf

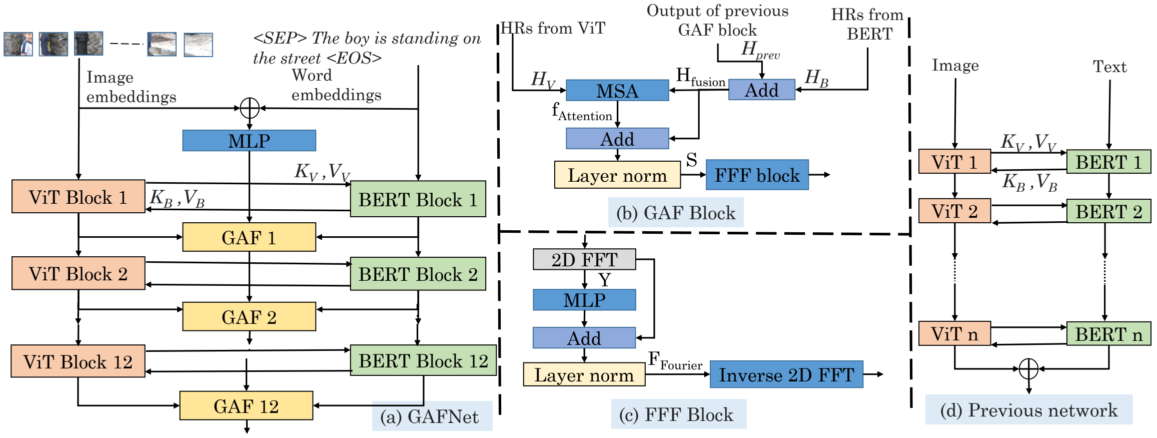


Figure 2. (a) GAFNet architecture (b) GAF Block (HR=hidden representation) (c) FFF Block (d) Block diagram of a previous network (ViLBERT) [22]

in language and vision to perform cross-modality learning. GAF block incorporates the FFF block, which uses feature vectors in frequency domains to analyze multi-modality frequency distribution.

3.1. Proposed "Global Attention Fourier" (GAF) block

We achieve cross-modality learning by integrating a global attention mechanism into the proposed network. For this purpose, we design a novel GAF block, which includes multi-headed self-attention (MSA), layer normalization, and Fourier feed-forward block. GAF considers the hidden states of ViT and BERT transformer blocks and creates the context vector containing the information of both image and text representations.

The first GAF Block accepts three inputs: (1) A concealed representation from the first ViT block, which serves as a key and value for the MSA. (2) Stacked representation of image and text positional embeddings, transmitted via the feed-forward layer. (3) Hidden representations of the first BERT block, given as query to the MSA. For the remaining GAF blocks, hidden representations from the corresponding ViT block serve as the MSA's key and value. In contrast, the output of the preceding GAF block and hidden representations from the corresponding BERT block serve as the MSA's query. This is shown in Figure 2(b).

Let H_V, H_B and H_{prev} be the hidden representations from ViT, BERT and previous GAF block, respectively. H_V acts as a key and value for the MSA, and the addition of H_B and H_{prev} acts as a query. The resultant vector obtained after addition is H_{fusion}, such that $H_{fusion} = H_B + H_{prev}$.

The MSA in the GAF block enables attention, which recognizes the agreement between the two modalities, viz., image and text. As described above, MSA obtains intermediary visual and linguistic depictions from the associated BERT, ViT and the previous GAF block. In order to use these intermediate representations, H_V is transferred through linear transformation function of key ($K(.)$) and value ($V(.)$) and H_{fusion} is passed through linear transformation function of query ($q(.)$). Later, they serve as key, query and value to the MSA. This linear transformation function is the feed-forward layer that gets trained during training to enable key, query and value vectors to learn and helps to provide attention to cross modalities.

The multi-headed self attention block (MSA) receives the keys, queries and values from each modality as input. As a result, the attention block generates attention-pooled characteristics for each modality (refer to equation 1 given below) that are conditioned on the other modality, i.e., executing image-conditioned language attention in the visual stream and language-conditioned image attention in the lingual stream. Let N be the number of attention heads in MSA (we use N=8 in this paper). Then,

$$f_{Attention} = Softmax(\frac{k(H_V) \times q(H_{fusion})}{\sqrt{N}}) \times v(H_V) \tag{1}$$

This aids in discovering the internal relationship between different modalities. Based on this, contextual clues can be transferred between modalities.

From the output of MSA, we obtain $S = LN(f_{Attention} + H_{fusion})$, where LN represents layer normalization. This is fed to the Fourier Feed Forward (FFF) block (Figure 2(c)). This block learns the diverse relationship between the text and image modalities, which further improves the learning capability of the network. In the FFF block, 2D Fourier transform is applied to S: 1D FFT along the sequence dimension and 1D FFT along the embedding dimension. This provides the frequency distribution across the fused sequence of vision and language.

The output of 2D FFT block is $Y = \Re(FFT(S))$, where $\Re$ indicates that only the real part of the result is kept and

FFT is the fast Fourier transform. We pass Y to MLP (multilayer perceptron), followed by residual connection and Layer Normalization (LN), to make these frequency features learnable. Note that we use MLP and the feed-forward layer synonymously.

$$F_{Fourier} = LN\left(MLP\left(Y\right) + Y\right) \tag{2}$$

Because the output features obtained after applying Eq. 2 are in the frequency domain, we need to transform them to the spatial domain before passing them to the next block. Hence, we perform Inverse 2D FFT on the calculated features, as shown in Figure 2(c).

$$F_{inverse} = \Re\left(IFFT\left(F_{Fourier}\right)\right) \tag{3}$$

Overall, in GAF, MSA pays attention to cross-modal features and the FFF block focuses on high-level characteristics. On comparing GAFNet (Figure 2(a)) with a recent work, named ViLBERT (Figure 2(d)), we can see that GAFNet concatenates image and text embeddings at the input side and also introduces novel GAF blocks for better shared-representation of both the modalities.

4. Pre-training

Humans have several ways of perceiving the world. Essential elements present in the surrounding can still be detected even if one channel is noisy or missing since they frequently occur in numerous channels. For instance, a horse can be described linguistically and visually. We use different pre-training techniques to enable GAFNet to capture modality-invariant elements at the semantic level. We now explain how we pre-train the GAFNet model using three pre-training techniques, viz., Masked Language Modeling (MLM), Image Text Matching (ITM) and Object Detection (OD). Figure 3 shows our pre-training approach. On 5 RTX 2080 GPUs, pre-training on these three datasets took nearly twenty days.

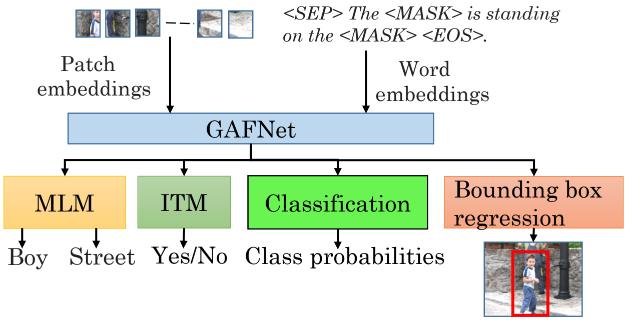


Figure 3. Pretraining approach

Each image region has been annotated with a region name like "girl in a pink dress". These annotations are utilized as pre-training data for this task. Table 1 shows the pre-training datasets used for GAFNet's pre-training[1].

Table 1. Datasets used for pre-training task

Dataset	Image	Caption	Annotations
COCO	0.11 M	0.55 M	0.45 M
SBU	0.86 M	0.86 M	-
CC-3M	2.9 M	2.9 M	-

Masked language modeling: The MLM approach is used to grasp the picture and text modalities by reconstructing the masked element using the unmasked remainders. It implies that MLM makes predictions about the text's masked words based on the visual idea. We chose a likelihood of 35% for the input tokens to be randomly masked. The model predicts these masked tokens using all the other unmasked tokens and visual cues. To complete this task, we take the output from the last GAF block of the GAFNet network and run it through an MLM with a linear layer, a SiLU activation layer, and another linear layer. This last linear layer generates an output vector by applying a softmax function.

Let $f(V,T)$ be the GAFNet, where V is the model's input for vision and T is the input for masked text. $M(f(V,T))$ is the MLM function which forecasts the probability of masked token t. Let y be a one-hot distribution where the probability of the ground-truth token for t is 1. We reduce the cross entropy loss by:

$$L_{MLM} = -\sum_{x \in M,V} y_x \times \log M(f(V_x, T_x))$$

Image-Text Matching (ITM): To create a representation of vision and language that is ubiquitous across languages, the ITM approach is modified. It foretells whether the text and image match. A statement and several picture areas serve as the inputs for ITM, and the output is a binary label that indicates whether or not the inputs were matched. During the training process, we select positive and negative pairs (V, T) from the dataset. The image and text in a paired sample are swapped out with another randomly chosen one from other samples to generate the negative pair. After giving this vision and text pair to it, we retrieve the output vector from GAFNet's last GAF block as the joint representation of the input image-text combination. Once the mean on the axis of sequence length has been calculated, the resultant vector is passed to the ITM branch. The ITM branch consists of an MLP layer and sigmoid function, which predict a score between 0 and 1. Let I(.) be the ITM classifier's function, and y be the ground truth vector. The loss L_{ITM} is calculated over the binary cross-entropy for respective labels as

[1] On SBU and CC3M datasets, only MLM and ITM are done since the bounding box information is not available for these two datasets. COCO dataset provides text, vision and bounding box information; hence, the model's pre-training is done using the three methods described here.

Table 2. Multimodal classification results on CrisisMMD dataset (acc =accuracy)

Model	Informative Task			Humanitarian Categorization Task		
	Acc	Macro F1	Weighted F1	Acc	Macro F1	Weighted F1
DenseNet [11]	81.57	79.12	81.22	83.44	60.45	86.96
BERT [7]	84.9	81.19	83.3	86.09	66.83	87.83
Compact Bilinear Pooling [10]	88.12	86.18	87.61	89.3	67.18	90.33
Compact Bilinear Gated Pooling [16]	88.76	87.5	88.8	85.34	65.95	89.42
MMBT [15]	82.48	81.27	82.15	85.82	64.78	88.66
Score Fusion	88.16	83.46	85.26	86.96	54.01	88.96
Feature Fusion	87.56	85.2	86.55	89.17	67.28	91.4
ViLBERT [22]	92.99	90.92	91.23	90.97	70.76	93.93
SSE-Cross-BERT-DenseNet [2]	89.33	88.09	89.35	91.14	68.41	91.82
GAFNet	**94.71**	**92**	**94.35**	**93.99**	**73.86**	**95.34**

$$L_{ITM} = -\sum_{x \epsilon T,V} y_x \times \log I(f(V_x, T_x)) \\ + (1 - y_x) \times \log(1 - I(f(V_x, T_x))) \tag{4}$$

Object Detection (OD): In this pre-training operation, we perform text-aware object identification to obtain the object's high-level semantic information. This enables the model to gain enough insights into the precise relationships between objects in various areas of the image and use textual information to distinguish between various things. Given the representation of a picture and text, we allow the model to estimate the bounding box (bbox) of the visual notion. We expect the model to learn fine-grained vision-language correlations more effectively by finding several visual ideas in the same image.

The object detection block contains two branches, namely the object classification branch and the bbox regressor branch (refer to Figure 3). GAFNet output $f(V, T)$ is fed to the classification and bbox prediction branches, which generate the class probabilities and anchor boxes, respectively. This enables GAFNet to identify the object's class and specific location in the image. As shown in Figure 9, this enables the model to learn the features for precisely locating a brown dog and distinguishing it from a black-and-white dog.

Let $O(.)$ be the regressor function that predicts bbox coordinates. Let Y_{BB} be the ground truth coordinates of bboxes. The regressor function uses IOU loss L_{IOU} defined as $L_{IOU} = \sum P(f(V, T)) \times IOU(O(f(V, T)), Y_{BB})$. Let $P(.)$ be the probability distribution function that predicts (1) confidence of the object being present in the anchor and (2) the classification probabilities of object classes. The corresponding losses are (1) $L_{Confidence} = \sum -p_c \times \log P(f(V, T))_{confidence}$ where p_c is ground truth confidence score. (2) $L_{Classification} = \sum -\log P(f(V, T))_{class}$. The total loss of object-detection pre-training (L_{OD}) is computed as a sum as follows:

$L_{OD} = L_{IOU} + L_{Confidence} + L_{Classification}$

Overall, the loss functions L_{MLM}, L_{ITM} and L_{OD} are added to obtain the final loss function, and the network is trained jointly for MLM, ITM and OD tasks. $TotalLoss_{pre-training} = L_{MLM} + L_{ITM} + L_{OD}$

5. Results

5.1. Multimodal classification

For multimodal classification, we evaluate GAFNet on CrisisMMD dataset [3]. This dataset contains multimodal tweets (image and text pairs) with corresponding annotations. We use binary_crossentropy loss for "informativeness classification" type in CrisisMMD dataset and categorical_crossentropy for "humanitarian classification" type in CrisisMMD dataset. The initial learning rate is 0.0001 with a cosine annealing scheduler. Training is done for 60 epochs, using a batch size of 24 and an AdamW optimizer. We use PyTorch and CUDA 11.2. We use several augmentations, including vertical flip, horizontal flip, random rotation, distortion, etc.

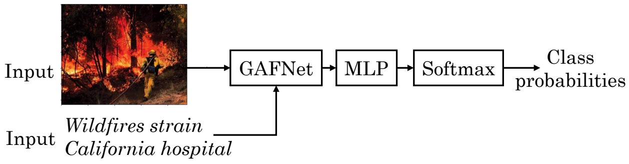


Figure 4. Architecture for multimodal classification

Figure 4 shows the changes to our model for performing multimodal classification. Here, the output of the last GAF block is sent to the feed-forward layer and then the softmax function is applied to get the class probabilities. In this dataset, the number of samples varies significantly across different categories. To evaluate this category imbalance problem, Weighted F1-score with accuracy and Macro F1-score are considered in Table 2. GAFNet significantly outperforms both baseline models and previously proposed techniques on all metrics. On both the tasks, GAFNet provides reasonably good Macro F1 and Weighted F1 scores. Note that the GAFNet results are obtained using both FFT and pretraining.

5.2. Image Generation

Figure 5 shows a high-level representation of the image generation architecture. The generator uses encoder-decoder architecture, with GAFNet as the encoder and pix2pix GAN [13] as the decoder. The inputs to the generator are text description t and segmentation mask s; and the inputs to the discriminator are s and an unknown image (which will be either a y or $g(s,t)$). Using the given parameters, a synthetic image conditioned on t and s is generated. By incorporating text information into the generation process and using GAFNet as the encoder, we create an exact connection between text and the relevant semantic regions of the mask and effectively control image creation. PatchGAN discriminator is employed as the discriminator.

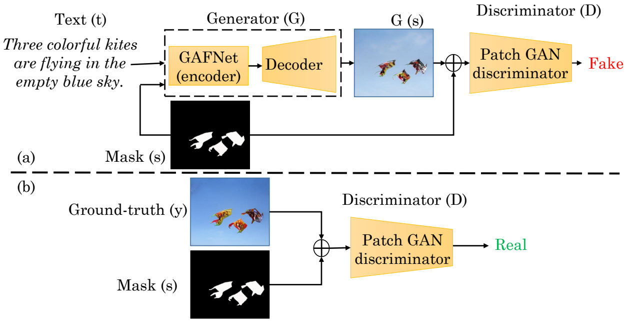


Figure 5. Proposed image generation approach

Instead of identifying the entire image as real or fake, this discriminator classifies specific (N×N) patches as real or fake. More restrictions are imposed as a result, which promotes finer high-frequency details. This model's training uses a mix of adversarial and generator pixel-to-pixel loss.

Let $M(.)$ be the encoder function (GAFNet) and $Dec(.)$ be the decoder function. Hence, generator ($G(.)$) is a function of M and Dec. Let D denote discriminator and y be ground truth image. Then, the generator loss vs. the discriminator loss is the conditional-adversarial loss ($L_{GAN}(G,D)$), shown as:

$$L_{GAN}(G,D) = \mathbb{E}_{s,y}[\log(D(s,y))] \\ + \mathbb{E}_{s,t}[\log(1 - D(s, G(M(s,t), Dec))] \tag{5}$$

Since generator is used for synthetic image generation, we use L1 loss for this task, defined as $L_{L1}(G) = \mathbb{E}_{s,t,y}[||y - G(M(s,t), Dec)||]$. Combining these two loss functions results in final loss, shown as:

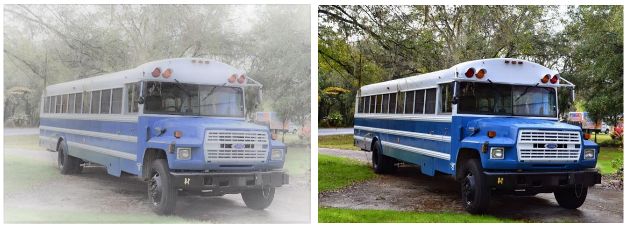
$$G^* = arg\ \min_G\ \max_D L_{GAN}(G,D) + \lambda L_{L1}(G)$$

The whole system gradually develops regional visual characteristics that are semantically matched with the provided description and generate high-quality photos that fit the global structure described by the segmentation mask.

Table 3. Inception Score (IS) and Rprecision (R-prcn) on COCO

Method	IS	R-prcn (%)
Real Images	27.41 ± 0.59	-
S-AttnGAN [29]	12.09 ± 0.28	75.24 ± 3.39
S-ControlGAN [19]	11.56 ± 0.16	80.43 ± 2.79
RefinedGAN [20]	15.96 ± 0.16	83.23 ± 1.37
RefinedGAN w/o POS	16.49 ± 0.18	84.01 ± 1.59
ViLBERT [22]	18.22 ± 0.02	85.99 ± 1.44
GAFNet	**19.04** ± **0.12**	**88.24** ± **1.01**

Results: We use L1 and SSIM loss for generator and binary cross-entropy loss for adversarial training. The initial learning rate is 0.0002 with a cosine annealing scheduler. Training is done for 150 epochs using a batch size of 12 and an AdamW optimizer. We have used inception score (IS) and Rprecision as the metrics. Table 3 shows the image-to-image translation results on COCO dataset. Clearly, GAFNet outperforms previous networks on both IS and R-prcn metrics. It shows that GAFNet can produce highly accurate images with high diversity. Also, note that RefinedGAN without POS (part of speech) performs better than RefinedGAN. This shows how futile bonds are formed between non-semantic words and visual features and how these worthless bonds degrade the quality of synthetic outputs.



(a) Output of ViLBERT (b) Output of GAFNet

Figure 6. Comparison of image generated from ViLBERT and GAFNet for the text "The blue-striped school bus is standing in the grass field."

Figure 6 compares GAFNet with ViLBERT for a sample generated image on COCO dataset. Both models can learn 100% semantic coherence with the text descriptions since both create blue buses standing on a green field. However, ViLBERT produces distorted textures while GAFNet produces more realistic and high-quality images. ViLBERT's co-attention between text and visual modalities helps it learn from both modalities. Yet, it cannot produce images with sharp details. Here lies the superiority of GAFNet. GAFNet can learn cross-modalities more effectively by using global attention in the GAF block. It can be seen that GAFNet can learn color and texture. GAFNet generates the blue color bus precisely, and the image texture and quality are better than that generated by ViLBERT.

As both ViLBERT and GAFNet were trained on the same distribution of dataset, they attempt to create an analogous image when the text and binary mask are provided as input. This is why Figure 6 shows comparable results. However, regarding image quality, GAFNet creates images with higher color and contrast than ViLBERT. This is because the GAF block draws attention to the visual and lingual stream and learns every aspect of an image, including color, contrast, and structure. As seen in Figure 7, both ViLBERT and GAFNet generate distinct images when black image (i.e., only background and no mask) and text are used as input. Here, GAFNet clearly outperforms ViLBERT.

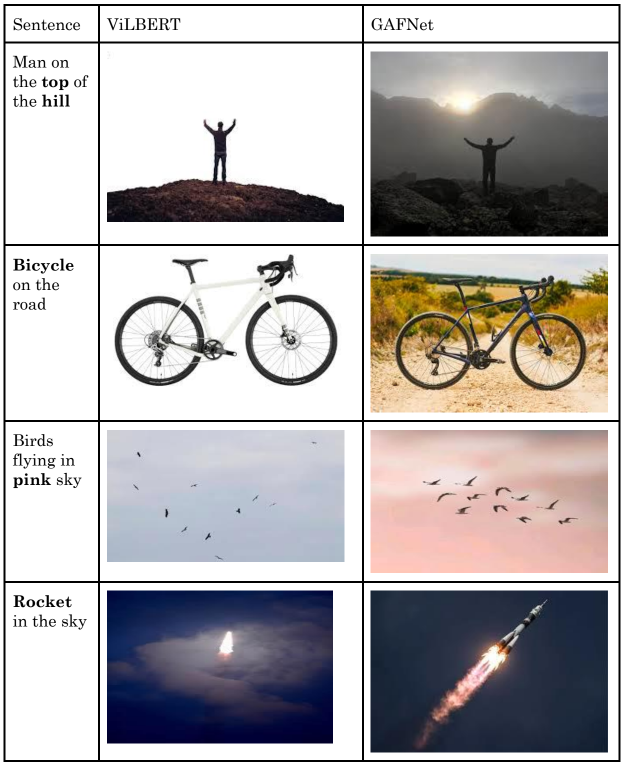


Figure 7. Comparison of images generated from ViLBERT and GAFNet for various texts

Figure 8 shows GAFNet output on the COCO dataset. GAFNet can generate high-resolution images with real-life objects. Also, the synthetic results generated from the model have a perfect semantic consistency with given text descriptions. From Figure 8, we can see that the image generated from the segmentation mask of kites matches the given text of "Three colorful kites are flying in the empty blue sky". On RTX 2080 GPU, for a batch size of six, image generation takes 2.8 seconds.

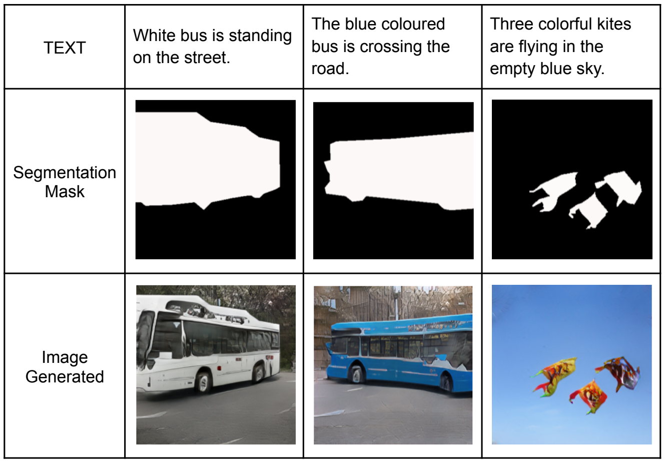

TEXT
White bus is standing on the street.
The blue coloured bus is crossing the road.
Three colorful kites are flying in the empty blue sky.
Segmentation Mask
Image Generated

Figure 8. GAFNet's image generation output on COCO dataset.

5.3. Image-text retrieval

Figure 9 shows examples of Grad-CAM visualization of images with related phrases. Clearly, GAFNet has a solid ability to understand. It can predict the correct regions in images even when the text descriptions differ by only one word. Additionally, GAFNet can align each word in the text with the matching image region. GAFNet can give attention to the region represented by the word (e.g., "black and white dog" vs. "brown dog").

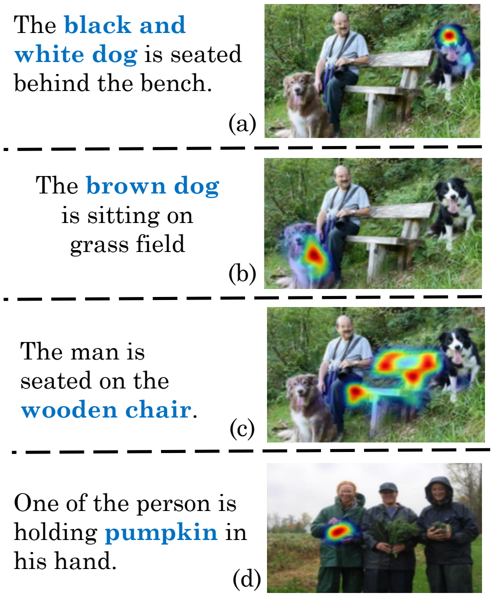

The **black and white dog** is seated behind the bench.
(a)
The **brown dog** is sitting on grass field
(b)
The man is seated on the **wooden chair**.
(c)
One of the person is holding **pumpkin** in his hand.
(d)

Figure 9. Sentences and specific words and corresponding Grad-CAM visualization (on unseen images)

Table 4 shows image-text retrieval results on MSCOCO and Flickr30K datasets. UNITER and VinVL are based on object-centric features, while ALIGN, METER, and ALBEF are based on overall image features. GAFNet performs

Table 4. Recall of different models (#PTI=pre-training images, TR/IR =text/image retrieval)

Method	# params	#PTI	MSCOCO (5K test set)						Flickr30K (1K test set)					
			TR			IR			TR			IR		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER [6]	300M	4M	65.7	88.6	93.8	52.9	79.9	88	87.3	98	99.2	75.6	94.1	96.8
METER-Swin [9]	380M	4M	73	92	96.3	54.9	81.4	89.3	92.4	99	99.5	79	95.6	98
ALBEF [21]	210M	4M	73.1	91.4	96	56.8	81.5	89.2	94.3	99.4	99.8	82.8	96.7	98.4
METER-CLIP	380M	4M	76.2	93.2	96.8	57.1	82.7	90.1	94.3	99.6	99.9	82.2	96.3	98.4
VinVL [32]	550M	5.6M	75.4	92.9	96.2	58.8	83.5	90.3		-			-	
ALIGN [14]	490M	1.8B	77	93.5	96.9	59.9	83.3	89.8	95.3	99.8	**100**	84.9	97.4	98.6
ALBEF [21]	210M	14M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	**100**	85.6	**97.5**	**98.9**
X-VLM [31]	216M	4M	80.4	95.5	98.2	63.1	85.7	91.6	96.8	99.8	**100**	86.1	97.4	98.7
X-VLM [31]	216M	16M	**81.2**	**95.6**	98.2	**63.4**	**85.8**	91.5	**97.1**	**100**	**100**	**86.9**	97.3	98.7
TCL [30]	-	4M	75.6	92.8	96.7	59.0	83.2	89.9	94.9	99.5	99.8	84.0	96.7	98.5
GAFNet	337M	4M	79.81	94.8	**98.65**	60.29	85.2	**92.7**	95.9	98.99	99.08	86.21	97.12	98.32

Table 5. Ablation study results on multimodal classification (acc =accuracy, M-F1=Macro F1, W-F1=Weighted F1)

Model	Informative Task			Humanitarian Categorization Task		
	ACC	M-F1	W-F1	ACC	M-F1	W-F1
GAFNet (no FFT, no pretraining)	90.23	89.3	90.01	88.7	70.77	91.91
GAFNet (no pretraining)	92.8	90.01	91.7	92.6	71.48	93.31

better than METER-Swin, which also employs the Swin Transformer as the image encoder. GAFNet performs significantly better than all previous techniques except XVLM in the 4M configuration, although previous approaches either have more parameters or more training data. METER also provides an empirical study demonstrating the vision backbone's importance in model performance. METER significantly improves retrieval task scores across the board, from Swin Transformer to CLIP-ViT.

5.4. Ablation Studies

5.4.1 Multimodal Classification

As shown in Table 5, on skipping FFT and pre-training, there is nearly 4% in accuracy and weighted F1 scores. Thus, both FFT and pre-training are vital for the model to distinguish the low-frequency and high-frequency features. Similarly, on using FFT but skipping pre-training, there is a 2% drop in accuracy and Macro F1 scores, which shows the importance of FFT in the proposed model.

5.4.2 Image Generation

As shown in Table 6, on not performing pre-training, there is a drop of 2% in inception score and R-prcn. Clearly, pre-training helps in learning feature importance in understanding the bonds between visual features and the semantic meaning of texts.

Table 6. Inception Score (IS) and Rprecision (R-prcn) on COCO

Method	IS	R-prcn (%)
GAFNet (no pretraining)	17.01 ± 0.28	85.61 ± 1.19

6. Conclusion

In this paper, we proposed GAFNet, a pre-trained network that simultaneously learns representations from both image and text modalities. By virtue of the global attentional GAF block, GAFNet accurately attends to both image and text modality information. Further, it uses Fourier transformation in the feed-forward layer to comprehensively capture the diverse semantic relationships of different modalities. GAF module offers a simple way to trade the shared representation in the conventional modality branch with a combined bi-directional shared representation. We also demonstrate that adding a few layers to GAFNet offers a simple way to extend the benefits of GAFNet to a wide range of downstream cross-modal tasks. GAFNet provides state-of-art or competitive results on a broad range of applications such as image generation under text guidance, multimodal classification and image-text retrieval. Our future work will focus on increasing the number of modalities (e.g., audio, category).

References

[1] https://huggingface.co/bert-base-uncased.

[2] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 14679–14689, 2020.

[3] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth international AAAI conference on web and social media*, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[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] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.

[10] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[12] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[15] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

[16] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[17] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[18] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[19] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] Bowen Li, Xiaojuan Qi, Philip HS Torr, and Thomas Lukasiewicz. Image-to-image translation with text guidance. *arXiv preprint arXiv:2002.05235*, 2020.

[21] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[22] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[23] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

[24] Raymond Ian Osolo, Zhan Yang, and Jun Long. An attentive fourier-augmented image-captioning transformer. *Applied Sciences*, 11(18):8354, 2021.

[25] 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.

[26] 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.

[27] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[29] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[30] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.

[31] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.

[32] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[33] Joey Zhou, Sinno Pan, Ivor Tsang, and Yan Yan. Hybrid heterogeneous transfer learning through deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.