An Attentive Fourier-Augmented Image-Captioning Transformer

Raymond Ian Osolo 1,2, Zhan Yang 2,3,* and Jun Long 2,3

1 School of Information Science and Engineering, Central South University, Changsha 410083, China; osoloian@csu.edu.cn
2 Network Resources Management and Trust Evaluation Key Laboratory of Hunan Province, Central South University, Changsha 410083, China; junlong@csu.edu.cn
3 Big Data Institute, Central South University, Changsha 410083, China
* Correspondence: zyang22@csu.edu.cn

Abstract: Many vision–language models that output natural language, such as image-captioning models, usually use image features merely for grounding the captions and most of the good performance of the model can be attributed to the language model, which does all the heavy lifting, a phenomenon that has persisted even with the emergence of transformer-based architectures as the preferred base architecture of recent state-of-the-art vision–language models. In this paper, we make the images matter more by using fast Fourier transforms to further breakdown the input features and extract more of their intrinsic salient information, resulting in more detailed yet concise captions. This is achieved by performing a 1D Fourier transformation on the image features first in the hidden dimension and then in the sequence dimension. These extracted features alongside the region proposal image features result in a richer image representation that can then be queried to produce the associated captions, which showcase a deeper understanding of image–object–location relationships than similar models. Extensive experiments performed on the MSCOCO dataset demonstrate a CIDER-D, BLEU-1, and BLEU-4 score of 130, 80.5, and 39, respectively, on the MSCOCO benchmark dataset.

Keywords: image-captioning; deep learning; transformers

1. Introduction

Although it is quite a trivial task for humans to describe an image, it was for many years, until recently, a very challenging task for a machine to perform at a level close to that of humans. The major breakthrough came with the introduction of deep learning-based image-captioning systems [1] that were inspired by the success of deep learning-based neural machine translation systems [2], which heralded probably the most significant age in the field of visual–language research. In automatic image caption generation systems, an image is presented to a machine which is then required to produce relevant and sensible human-readable descriptions (captions) that describe it. The challenge of accomplishing the task of automatic image-captioning stems from the fact that it is a vision-to-language task that requires a machine to both understand the semantic content in an image and have language understanding as well as generation capabilities. This is further complicated by the fact that it is an ambiguous task with many equally accurate descriptions possible for each image.

Automatic image-captioning systems have been mainly implemented using LSTMs (long short-term memory) [1,3–5] as the language model in an encoder–decoder architecture. In this architecture, a convolutional neural network (CNN) encoder pre-trained on a large image classification dataset such as ImageNet [6], is used to extract the image features into a fixed-length vector that is then fed to the decoder LSTM which learns to generate captions one word at a time by conditioning them over both the image features and the previously generated words. Even with the introduction of systems that employed...
CNN–CNN encoder–decoder models [7,8], the now classic method has still stood the test of time, especially with the introduction of attention-based variants [3] which significantly improved their performance.

In spite of their undeniable successes and impressive performances, LSTMs have several shortcomings. The inability to remember long sequences was a major factor in their development and advantage over vanilla RNNs, but the problem, which is caused by vanishing gradients, still exists. Their sequential nature also makes it difficult to parallelize parts of the training process as every subsequent caption token can only be dealt with after the previous one has been processed. As a consequence, they are incapable of taking full advantage of GPUs, which would make training them much faster, significantly impacting the training time. Furthermore, most of the previously mentioned models of image-captioning suffered from the issue of exposure bias. It is because of these reasons that more researchers have recently begun to leave previous methods and adopt transformer-based architectures. Transformer architectures, introduced by Vaswani et al. [9] for machine translation, have now achieved state-of-the-art results in many computer vision and natural language processing (NLP) tasks including visual–language tasks such as image-captioning [10–13].

Transformers for image-captioning in many of the proposed models are not without their shortcomings, though. Transformers are very data-hungry, requiring a lot of data to achieve decent results. This is because they are very unfocused at the beginning and require data to learn both where to focus, and what to focus on. Many of the transformer models in image-captioning do not adequately use the images to generate the captions, mostly relying on the training captions (which are fortunately already highly compressed and informative), thereby missing out on much of the fine-grained salient semantic information available in images that could improve the quality of generated captions. This problem was much worse in non-attention-based captioning models that would only look at the image once and still produce decent captions; thus, the practice to give little attention to the images in image-captioning persisted. Another challenge experienced in image-captioning concerns increasing the level of detail in the captions generated, specifically increasing the number of objects and attributes mentioned in captions generated. Often, captions only mention a single object, including probably its location and the action being performed. As a result, there is a need to find more ways to highlight and extract more of the latent features contained in images so that they can provide additional information which can be exploited by intra-attention mechanisms.

It is with the above in mind that we propose a transformer-based image-captioning method that purposefully uses fast Fourier transforms to extract additional salient information from image features in order to create a richer representation of input images and lead to the generation of more detailed representative captions, enabling the image to contribute more towards the generation of captions without specifically grounding caption objects to regions in the image, as shown summarily in Figure 1 and in the full framework diagram in Figure 2. The ability of Fourier transforms to break down a function into its individual frequencies, as well as the mixing mechanisms they provide [14] without using learnable parameters, offer a way to extract and highlight some existing discriminatory image information that may remain unused if self-attention is only applied directly to the input features. We (1) leverage the power of Fourier signal decomposition and mixing to extract otherwise ignored salient image information, and (2) use a simple yet effective concatenation of features to create a richer image feature representation to foster better quality and more detailed caption generation. The additional image feature information allows for the self-attention mechanism to learn more relationships among features and the associated captions, leading to demonstratively more detailed captions. Our model was tested on the MSCOCO benchmark dataset and our contributions can be summed up as follows:

- We propose AFCT, a fully attentive transformer-based captioning model that uses a multi-layer encoder–decoder architecture to extract fine-grained salient image features that are used in caption generation.
• The model proposed leverages the ability of Fourier transforms to separate input signals into their constituent frequencies to extract and augment the object proposal features, creating a richer and more detailed image feature representation that learns a deeper relationship between objects and their locations in images.

• We perform a thorough analysis of the effect of the use of Fourier transforms and a fully connected layer for image-captioning. Furthermore, we demonstrate why Fourier transforms work much better with image features than with captions by performing several experiments on variants of the proposed model.

• We demonstrate the superior performance of the proposed method qualitatively through the quality of captions generated, as well as quantitatively through tests performed on the MSCOCO benchmark dataset, showcasing competitive scores when compared to similar and more complex models without using additional training or pre-training methods.

![Diagram of Fast Fourier Transform and Regional Proposal Features](image1)

Figure 1. The fast Fourier transform extracted features fused with regional proposal features.

![Diagram of Fourier-enhanced Image-captioning Framework](image2)

Figure 2. Fourier-enhanced image-captioning framework diagram.

2. Related Work

Automatic image-captioning is not a new field of study and a lot of research has already gone into developing systems that tackle it. As in [15], machine learning-based image-captioning methods can be divided into two broad categories: traditional and deep learning-based image-captioning methods. To extract image features, traditional
machine learning methods [16–22] employed tedious and cumbersome manual feature engineering, and relied on sentence templates, sentence-retrieval methods, and ranking-based techniques to create the descriptions, which inevitably were generally very rigid in nature and not novel.

The field of image-captioning remained more or less a fringe research area until, inspired by the successes of deep neural networks applied to neural machine translation [2], deep learning-based methods thrust automatic image-captioning into the limelight. Most of these deep learning-based methods [1,3,23–25] involve the use of neural networks such as, in particular, recurrent neural networks (RNNs) and, specifically, LSTMs [26] or gated recurrent units (GRUs) [27], in conjunction with convolutional neural networks (CNN) [28], in an encoder–decoder architecture in which a state-of-the-art CNN architecture such as VGGNet [29] and ResNet [30] pre-trained on the ImageNet dataset, used as the encoder for image feature extraction. These features, together with the associated image captions, are then used to train an LSTM or GRU decoder (language model) to generate relevant captions based on the input image; this is achieved by training to generate captions using the image features and the previously generated captions to predict the next word in the sequence. Attention mechanisms that allow the decoder to focus on different parts of the image during generation boosted the performance of these methods [3]. In an attempt to further boost performance and increase the diversity of the captions generated, some works [7,8] have successfully utilized CNNs as decoders, employed generative adversarial neural networks and reinforcement learning [31,32], and recently, more commonly, employed transformers. This was done to alleviate RNN issues including being too complex and slow to train, as well as their sequential nature in making it difficult for them to take full advantage of the parallelization advantages offered by the use of GPUs.

Transformers, first proposed in the paper “attention is all you need” [9], have greatly changed the vision, language, and the vision–language research communities. Methods involving transformers such as ViLBERT [33] have achieved state-of-the-art results in vision, language, and vision–language tasks. Originally proposed following encoder–decoder architectures, there are now variants that exist using only one of the modules. In image-captioning, transformers, with their attention-only architectures (without recurrent units), may follow the encoder–decoder architecture [10,13], wherein the encoder receives and processes image features, and sends them to the decoder module (language model) in which captions are also input and output. Pre-trained unified transformer architectures for visual–language tasks have been proposed, wherein the model is pre-trained on a large corpus and then fine-tuned for downstream tasks such as image-captioning and vision-question-answering [11,12]. The meshed-memory transformer [10] uses a multilevel representation of the encoder output and a meshed cross-attention network at the decoder. AoANet [13] proposes an alternative to the scaled dot product attention wherein the initial query is concatenated with the output of the dot product attention, the result of which is then transformed by two linear layers to produce an information signal and a sigmoid attention gate which, when multiplied together, can be used as a substitute for the normal scaled dot product attention. DLCT [34] uses both region proposal and grid features to generate captions. This requires a specially designed complex attention mechanism they refer to as a dual-way self-attention mechanism and a locality-constrained cross-attention mechanism, alongside an alignment graph to combine the features and eliminate any resulting noise. CPTR [35] uses a pre-trained vision transformer as the encoder, completely removing convolutional features. Due to the fact that a transformer is initially very unfocused, a lot of data is required to train it. As a result, in order to use a visual transformer on the captioning dataset, the authors had to use an encoder pre-trained on a much larger dataset than their captioning dataset.

In order to add more details to the captions, some works such as [4] used templates which are later filled using associated external news articles. Wu et al. [36] used an intermediate attribute prediction layer and injected external knowledge mined from a knowledge-base into the model to improve the captions generated. There is a need for methods that
can make such detailed captions without resorting to external knowledge. The current language models are quite powerful and the datasets contain enough textual information that it is still possible to improve the quality of captions and generate more detailed captions without the need for external knowledge-bases. This can be done by extracting more salient information from the image features as well as learning better associations and relationships during training both between the features and with the captions that describe the images.

Inspired by the work explored above, we introduce the Attentive Fourier-Augmented Image-Captioning Transformer, AFCT, a transformer-based image-captioning model that fully explores the input image features to make the image in image-captioning matter to a greater degree. We achieve this by using a very simple yet effective fast Fourier transform algorithm to augment the regional proposal features from a faster R-CNN network. The model learns a deeper and more complex relationship between the input image features (and subsequently the associated captions) by setting a self-attention layer to create relevancy scores over the input regional proposal features supplemented by the Fourier extracted features. We demonstrate a model that makes better use of the fine-grained salient semantic information contained in the images to generate accurate, fluent, relevant, and more detailed captions that show a better understanding of the object, its location, and the surroundings when generating captions. Fourier transforms have been more commonly used in theoretical work [14] such as for solving partial differential equations [37], reducing the exploding and vanishing problems in RNNs [38], speeding up CNN computations [39,40], and the development of Fourier convolutional neural networks [41].

Fourier transforms have also been extensively utilized in many image-processing applications. For example, it can be used for image transformation and compression [42,43]. Shi et al. [44] used 2D Fourier transforms that also employed a hash function to take advantage of the sparsity in the frequency domain to estimate the largest k coefficients. The authors expressed that their method is faster than the Fastest Fourier Transform in the West (FFTW) [45], which is generally taken as the fastest fast Fourier transform implementation. Fast Fourier transforms have also been used to fuse satellite images as in [46]. In the image classification domain, Fourier transforms were explored by [47] for the extraction of useful information in the frequency domain using the Fourier transform of input images and deep neural networks. In contrast, we work with feature vectors that have been pre-processed by a faster R-CNN network because we need more fine-grained information than is required for classification. A more recent proposal and most similar to our work is the study by Lee-Thorp et al. [14] who propose replacing the self-attention layer with a Fourier transform or fully connected layer. In contrast, we maintain the self-attention layer but use the Fourier transforms to mine more salient image information and employ it to an image-captioning task rather than a translation or text-retrieval task. Similar to [10], we also use a multilevel encoding of image regions, but in contrast, we do not use a mesh connectivity.

3. Material and Methods

Overview: The ultimate objective of the method that this paper proposes is to design a transformer-based image-captioning system that is able to intelligently leverage knowledge contained in both images and text, with special attention given to the image features, in order to generate grammatically, semantically, and syntactically correct captions. As part of the explanation of our proposed model, a general explanation of how a visual–language transformer operates will also be given, loosely based on the popular “The Illustrated-transformer” article [48] in order to provide a better understanding of the proposed model. In a transformer network, following the encoder–decoder architecture, the input sequence tokens to the encoder and decoder are passed in parallel. Positional encodings are used to keep track of the position of each embedding in the input sequence. The positional embeddings are concatenated to the input embeddings, providing them some context as to their position in the input sequence, and fed to the transformer.
3.1. Method

In our proposed model, AFCT, the input features at the decoder are 2048-dimensional image feature vectors from a faster R-CNN model trained on the visual genome dataset. These region proposal features undergo a linear transformation that projects them into a lower dimensional encoder representational space, reducing their dimension from 2048 to 512. From the 512-dimensional image features, a set of Fourier transform features are extracted. These features are concatenated to the original image regional proposal features, forming a richer representation of the input image features over which a multi-head self-attention network is applied to generate relevance scores.

Each subsequent encoder or decoder uses the output of the previous layer as its input. No positional embeddings are used for the encoder but they are retained for the decoder. In the encoder section, Vaswani et al. [9] described the task as a query (Q), key (K), and value (V) problem in which for every input token, a Q, K, and V vector is generated by applying three matrices that are learned during the training process. For each item in the input sequence, an attention or relevance score is calculated by doing a dot product of the query vector with the key vector. The output will be the product of the softmax of the relevance scores (attention vectors) and the weighted average of the values. These attention vectors describe how important each part of the input sequence is to the other parts and to itself. The output of this sub-module is then passed through a position-wise feed-forward network, producing a feature vector representation of the inputs that is then sent to the decoder. The decoder receives this encoded vector through an encoder–decoder cross-attention sub-module. The caption embeddings, concatenated to the positional embeddings, are fed to the auto-regressive decoder module first via a masked multi-head self-attention sub-module. The output of the section is then sent to the encoder–decoder sub-module from which the relationships between the image representations and captions are determined, matched, and mapped. Then, through a position-wise feed-forward network, the output of the layers are sent to a feed-forward layer, onwards to a classification linear layer, and finally to a softmax layer in which the probability distribution over the input vocabulary is the output. A normalization layer is applied after every sub-module and residual connections between the layers reduce the amount of information lost as it propagates through the layers. This is formally described below.

Given an input sequence, \( x_n \), where \( n \in [0, N - 1] \), the features first undergo a linear transformation into:

\[
X_i = \phi(x_n W) + b, \tag{1}
\]

where \( \phi \), \( W \), and \( b \) correspond to the ReLu activation function, weight, and bias of a fully connected layer.

Fourier transforms are then used to extract more of the less obvious fine-grained salient information. A discrete Fourier transform (DFT) can be used when you have signals only known at specific \( N \) instants, as is our case. Therefore, for the input sequence, the DFT can be computed by:

\[
X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i nk}{N}}, \quad 0 \leq k \leq N - 1, \tag{2}
\]

where \( X_k \) is the new representation generated by the Fourier transform from the input sequence tokens \( x_n \).

The DFT in Equation (2) can be computed using a fast Fourier transform (FFT), which is an algorithm for computing the DFT of a sequence or its inverse. Inspired by the techniques proposed in [14], we performed a 1D Fourier transform on the input image sequence in the hidden dimension and a 1D Fourier transform in the sequence dimension, and then passed the output through two linear transforms as shown below and further demonstrated in Figure 3:
\[ X_{\text{fft}} = \text{Real}(F_{\text{seq}}(F_{\text{hidden}}(x))) \]
\[ X'_{\text{fft}} = \phi((\phi(X_{\text{fft}}W_f) + b_1)W_f^2) + b_2, \]  

where \( F_{\text{seq}} \) and \( F_{\text{hidden}} \) refer to conducting a fast Fourier transformation on the input \( x \) along the sequence and hidden dimension, respectively. \( \phi \) is the GeLU activation function used. \( W_f^1, W_f^2, \) and \( b_1, b_2 \) are the weights and biases of a fully connected layer. We compute the discrete Fourier transform by using the fast Fourier transform algorithm as it was found by the authors to be faster on GPUs. The two sets of features, namely \( X_i \) and \( X'_{\text{fft}} \), are then concatenated to derive a more detailed representation of the input image features. The Fourier transforms highlight more of the intrinsic less obvious salient information that may otherwise not be apparent if only the features were used, allowing the model to easily identify the more important features and their relations in context.

\[ X_i \in \mathbb{R}^{N_i \times D}, \quad X'_{\text{fft}} \in \mathbb{R}^{N_f \times D}, \]
\[ X = \text{concat}(X_i, X'_{\text{fft}}). \]

The multi-head self-attention layer consists of 8 identical heads. During training, each head learns a different aspect of the attention mechanism. A single head could cause a token to emphasize too much attention on itself rather than on its relevance to the other tokens in the sequence. The values of the queries (Qs), keys (Ks), and values (Vs) are calculated for each of the input tokens as shown below:

\[ Q = X_e \ast W_Q, \quad K = X_e \ast W_K, \quad V = X_e \ast W_V, \]

where \( X_e \) represents the encoder input features \( x_1 \ldots x_n \), while \( W_Q, W_K, \) and \( W_V \) are learned projection matrices for the queries, keys, and values, respectively. The output of each head is independently calculated, concatenated, and then multiplied by another learned weight matrix (\( W_o \)) to fit within the expected dimensions for the position-wise feed-forward network.

\[ \text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V, \]
\[ \text{Multihead}(Q, K, V) = \text{Concat}(h_1, \ldots, h_n)W_o \]
\[ h_i = \text{Attention}(X_eW_Q, X_eW_K, X_eW_V), \]

where \( h_{i:n} \) are the attention heads. Residual connections are used to send information from the input of one layer to the next in order to reduce information loss due to the many transformations undergone by the data and also to help the gradients propagate through the model more easily. Firstly, the multi-head attention output is added to the input and normalized

\[ X_m = \text{Norm}(X_e + \text{Multihead}(Q, K, V)). \]

Secondly, the output of the multi-headed attention goes through two transformations in the position-wise feed-forward layer.

\[ X_f = \phi(X_mW_f) + b, \]
where $\phi, W_f, and b$ correspond to the ReLu activation function, weight, and bias of a fully connected layer. Finally, the output of the feed-forward is added to its input and the result is normalized:

$$X' = \text{Norm}(X_m + X_f),$$  \hspace{1cm} (10)

where $X_m, X_f \in \mathbb{R}^{N \times D}$. $X'$ is the final output vector from the encoder. As we do not just use the final output of the last decoder but rather the final outputs of the three encoders (multi-level encoder–decoder network), all three values are saved for interfacing with the corresponding decoder.

The decoder section mostly remained the same as the transformer proposed by [9], consisting of a masked self-attention sub-module, a cross-attention module, and a position-wise feed-forward network. The captions are both the inputs into the decoder and the targets. The caption token embeddings, concatenated to sinusoidal positional embeddings, and the image representation interact in the encoder–decoder cross-attention layer. Queries in this layer are from the decoder, while the keys and values are from the encoder output, which is why it is referred to as cross-attention. The $i$-th encoder layer is connected to the $i$-th decoder layer, unlike in the original source [9], in which only the last encoder layer is connected to the decoder. This is similar to [10] who demonstrated the benefits of multilevel encoding of image regions, but in contrast, we do not use a mesh connectivity, which we believe adds much more complexity and requires more computational power for a meager improvement in performance. Formally, the above can be described as:

$$\text{Multihead}(Q_d, K_e, V_e) = \text{Concat}(h_1, \ldots, h_n)W_o$$

$$h_i = \text{Attention}(X_dW_{Q_i}, X_eW_{K_i}, X_eW_{V_i}),$$  \hspace{1cm} (11)

where $Q_d$ and $X_d$ are the decoder queries and inputs, respectively, while $K_e, V_e$, and $X_e$ are the encoder keys, values, and inputs, respectively.

Figure 3. The fast Fourier transformation.
3.2. Training and Loss Function

During training, we first optimized the cross entropy, as is commonly done for word sequence generation, followed by CIDEr-D metric optimization as introduced by Rennie et al. [49] in their “Self-critical Sequence Training for Image Captioning” (SCST) paper, in which they minimized the negative expected value of a caption’s CIDEr-D metric. This uses policy gradient methods from reinforcement learning, which allows for non-differentiable metrics to be directly optimized.

Given \( (y^*_1:T) \) as a sequence of ground truth word tokens, where \( T \) is the sequence length of the sequence \( y_1, \ldots y_T \), the cross-entropy loss is given by:

\[
L(\theta) = -\sum_{t=1}^{T} \log(p\theta(y^*_t | y^*_1:t-1)),
\]

where \( \theta \) represents the model parameters. The CIDEr score is optimized by SCST:

\[
L_{RL}(\theta) = -E_{(y_1:T \sim p\theta)}[r(y_1:T)],
\]

where \( r \) is the score function. The gradient is approximated by:

\[
\nabla_{\theta} L_{RL}(\theta) \approx -(r(y^*_1:T) - r(\hat{y}_1:T)) \nabla_{\theta} \log p\theta(y^*_1:T)
\]

3.3. Materials: Datasets and Evaluation Metrics

In order to evaluate the effectiveness of the proposed methods in tackling the issues explained in this paper, tests were performed on the MSCOCO dataset [50] and all the evaluation metrics commonly used in image-captioning were used to quantitatively evaluate the captions generated by the model. The Microsoft COCO dataset provides five caption descriptions for each of its images that are split between training, validation, and test sets containing 113k, 5k, and 5k images, respectively. We used features generated from a faster R-CNN with a ResNet-101 model trained on the visual genome dataset using scripts provided by Anderson et al. [5].

The evaluation metrics that were used include BLEU-1 to BLEU-4 [51], CIDEr [52], METEOR [53], and SPICE [54]. BLEU is an n-gram precision-based metric, METEOR performs unigram matching, CIDEr uses TF-IDF to give more weight to important n-grams, and SPICE computes an F1-score over caption scene-graph tuples, i.e., the balance between the precision and the recall. CIDEr and SPICE metrics correlate best with human evaluators. We used the predefined splits described in [23] for the results in Table 1 as many works also use the same splits.

| Method       | B1    | B4    | M    | R    | C    | S    |
|--------------|-------|-------|------|------|------|------|
| SCST [49]    | -     | 34.2  | 26.7 | 57.7 | 114.0| -    |
| Up-Down [5]  | 79.8  | 36.3  | 27.7 | 56.9 | 120.1| 21.4 |
| ETA [55]     | 81.5  | 39.3  | 28.8 | 58.9 | 126.6| -    |
| SGAE [56]    | 80.8  | 38.4  | 28.4 | 58.6 | 127.8| -    |
| AFCT-F0      | 80.1  | 38.56 | 29.0 | 58.4 | 127.9| 22.1 |
| VRAtt-soft-trans [57] | 80.5  | 38.5  | 28.9 | 61.8 | 129.2| 22.8 |
| AoA [13]     | 80.2  | 38.9  | 29.2 | 58.8 | 129.8| 22.4 |
| AFCT         | 80.5  | 38.7  | 29.2 | 58.4 | 130.1| 22.5 |

4. Experiments and Results

4.1. Settings Implementation

As in [10], we used the Karpathy splits of the MSCOCO dataset for the training, evaluating, and offline testing of the model, which included a batch size of fifty, eight attention heads, and three encoder and three decoder layers. We also use a model dimension of 512.
Scripts made available by [5] were used for a faster R-CNN network with a ResNet-101 trained using attribute annotations from the visual genome dataset for the object detections. As is the norm, the model was first trained optimizing the cross-entropy loss, followed by CIDEr-D optimization. The Adams optimization algorithm and the beam search algorithm with the beam width set to five were used. We also employed dropout and early stopping as a regularization strategy. A fixed vocabulary size of 10,000 was used for our experiment. The experiments were implemented using PyTorch [58] run on a system that had four 2080Ti GPUs with a 64-bit 128 GB RAM Ubuntu System with Intel E5-2600 CPUs. The Fourier transforms were implemented using PyTorch’s Accelerated Fast Fourier Transforms module. All reported evaluation metrics were done with the model that had the best CIDEr-D validation score and CIDEr-D optimization started with the model that had the best CIDEr-D validation score during the cross-entropy optimization section.

4.2. Ablation Study

We present five variants of our model. First, we implemented a variant that uses only Fourier extracted features (AFCT-F1) for the encoder, while the decoder remains unchanged. AFCT-F1 is used to investigate whether there are any qualitative benefits to adding the Fourier extracted features to the total image feature representation. Second, we implemented a variant with six layers instead of the standard three (AFCT-F2). This was chosen to ascertain whether there were any additional benefits to having more layers. Third, we implemented a variant similar to AFCT-F1 but with only one attention head (AFCT-F3). This variant examines the general impact of having just a single head. Fourth, we implemented a variant without the Fourier or the multi-head attention (AFCT-F4). AFCT-F4 is used to investigate the impact of having neither the Fourier attention, nor the self-attention layer when extracting the image features, with the input sent straight to the position-wise feed-forward network. Fifth, we implemented AFCT-F5, a variant that replaces both the encoder self-attention layer and the decoder masked self-attention layer with a Fourier transform layer and a masked Fourier transform layer, respectively. This was created to demonstrate the difference between Fourier transforms applied to images vs. text in an image-captioning scenario. Assume there were eight attention heads unless explicitly stated otherwise. See Table 2 for the model configuration. The results of these studies are shown in Table 3 and discussed in the Discussion section (Section 5).

| Method   | # Layers | # Heads | Model-Size (MB) | Encoder | Decoder      |
|----------|----------|---------|-----------------|---------|--------------|
| AFCT-F1  | 3        | 8       | 485             | FFT     | CA, masked SA|
| AFCT-F2  | 6        | 8       | 836             | FFT     | CA, masked SA|
| AFCT-F3  | 3        | 1       | 485             | FFT     | CA, masked SA|
| AFCT-F4  | 3        | 8       | 411             | FC      | CA, masked SA|
| AFCT-F5  | 3        | 8       | 522             | FFT     | CA, masked FFT|

Table 2. Model configuration for the variants used in the ablation studies. Abbreviations: CA = cross-attention; SA = self-attention; FFT = fast Fourier transform; and FC = fully connected.

| Method   | B1   | B4   | M   | R   | C   | S   |
|----------|------|------|-----|-----|-----|-----|
| AFCT-F1  | 80.3 | 38.5 | 28.8 | 58.2 | 127.7 | 22.5 |
| AFCT-F2  | 80.2 | 38.0 | 28.5 | 57.7 | 126.7 | 22.2 |
| AFCT-F3  | 80.3 | 38.5 | 28.8 | 58.2 | 127.7 | 22.5 |
| AFCT-F4  | 80.2 | 38.1 | 28.8 | 58.0 | 128.0 | 22.4 |
| AFCT-F5  | 77.2 | 30.2 | 24.8 | 52.3 | 104.9 | 17.7 |
| AFCT     | 80.5 | 38.7 | 29.2 | 58.4 | 130.1 | 22.5 |

Table 3. Results from the different variants. All variants’ results reported after CIDEr-D optimization. Abbreviations: B1 = BLEU-1; B4 = BLEU4; M = METEOR; R = ROUGE-L; C = CIDEr; and S = SPICE.
4.3. Comparison with Other Models

It is difficult to make accurate or fair comparisons between different models. There are usually differences in the choice of datasets, hardware capabilities, implementation settings, evaluation methods, and even regarding for how long models are trained or optimized. This does have a direct impact on the results researchers report. Nevertheless, we do our best to select the most appropriate models for comparison while also stating any significant differences between our model and the compared models when we deem it necessary. The baseline model that we refer to as AFCT-F0 is a captioning transformer similar to [10] but with the mesh part removed, leaving behind a multi-level cross-attention model for which the adjacent encoders and decoders are connected to each other but not in a meshed network. The original has three encoders and three decoders that interact in a multi-level cross-attention model in which the output of the $i$-th encoder is sent to the input of the $i$-th decoder and all other decoders, an arrangement they refer to as a meshed decoder. The model presented by the authors clearly demonstrated the merits of using the results of all the encoder stages and not just the output of the final encoder stage as input to the decoder. Other compared models include the attention on attention [13]; self-critical sequence training for image-captioning [49]; bottom-up and top-down attention for image-captioning and visual-question-answering [5]; SGAE [56]; Entangled Transformer [55]; and VRAtt model [57].

5. Discussion

The results from our model are competitive in all evaluation categories and validated by the results of the variants used in the ablation studies to explore the effect of each of the design decisions we made. This is especially apparent regarding the CIDEr score in Table 1 and the quality of the captions as attested to by human evaluators. Automatic evaluation of language generation results is still an ongoing area of research. No single method is good enough to adequately evaluate the generated text with respect to the ground truth but the metrics are provided nevertheless to make it possible to compare this manuscript to other papers. We used the full suite of methods, alongside human evaluators, and provide graphs that we hope together demonstrate a reasonably accurate analysis both quantitatively through the metrics and graphs (Figure 4), and qualitatively through the sample captions shown in Tables A1–A4 in Appendix A.1. We intentionally do not highlight any captions sections in order to not bias anyone towards what to notice so as to let reader devise their own deductions about the differences between the captions generated by the different models. The optimum number of encoders and decoders for this set up is three encoders and three decoders. This finding is not only proven by the stellar results achieved by our model but further solidified by the better performance of the AFCT-1 and AFCT-3 variants, which had three encoders and three decoders, when compared to the AFCT-2 variant, which had six layers.

The feature representation of an image contains much of the latent information that can be exploited to identify objects and their relationships to each other. Fourier transforms provide a way to accomplish such a task because they decompose input features into their constituent elements in the frequency domain. In this domain, many features that were not apparent in the original domain become highlighted as the transform improves the signal-to-noise ratio. In our image-captioning application, this additional object and relation information improved the quality of the captions generated. Alternative ways to improve the quality of the image feature information, such as using external knowledge sources or pre-training, are thus avoided. In addition, the Fourier transforms themselves are just a linear transformation and thus hardly add any new parameters that need to be learned. The advantage of this is that there will only be an insignificant increase in the training time required, while attaining a significant improvement in performance.

The AFCT-1 variant that used a Fourier transform to replace the self-attention model, which performed second best and was only beaten by the proposed model. This shows FFTs are capable of extracting and highlighting many of the underlying features, thus
providing the cross-attention sub-module with highly discriminating and information-rich features on which to model the captions. The AFCT-3 variant used only one attention head instead of eight but this did not seem to affect its performance. This is because no self-attention was used, thus the outputs of the heads were usually the same and there was no measurable advantage of using one or eight heads. In the self-attention technique, every input feature receives a score of its relevance to itself and to other features. This may result in each feature giving itself a higher score than all the other features. This situation is mitigated by having multiple attention heads, which by focusing on different attention areas, reduces this bias. The result of the Fourier transformation over the features is deterministic and always the same no matter how many heads observe it. The decoder section of the network used cross-attention and masked self-attention, greatly benefiting from the additional attention heads. The main model and most of the variants (except AFCT-3) used the same number of attention heads in the encoder and decoder sections for symmetrical reasons, which we believe may have also helped with the training process. The model that contained only the fully connected layer (AFCT-4) had higher first epoch scores in all the metrics when compared to the other variants. This can be attributed to the fact that the other variants contained extra learnable parameters (courtesy of the Fourier section) that were randomly initialized and still needed to be learned. It still had lower metrics than the proposed model (AFCT) in spite of the fact that it also contained more learnable parameters (FFT linear layers and the self-attention weight matrices). This shows the agility and robustness of the proposed model as, despite containing these extra parameters, AFCT was still able to beat the baseline and variants by capturing fine-grained salient image information right from the start.

One reason why Fourier transforms worked so well on the images is because an image is organized in a grid on which every pixel is related to the pixels around it. This is a phenomenon that can be well exploited by both self-attention and Fourier transforms to produce information-rich image representations. For text, while a word may derive meaning and context from the words around it, more often than not, the prediction of a new word may depend on another word that was output several tokens before, in addition to the words around it. Self-attention, which creates a relevance score of not only the words around it but also of every other word in an input word sequence, has a much higher chance of capturing this dependence than a Fourier transform. We demonstrate this by constructing a variant (AFCT-F5) that not only uses a Fourier transform layer at the encoder but also replaces the decoder masked self-attention with a masked Fourier layer. Early epochs can be deceiving but they do provide a lot of valuable information. It takes a while for the linear layers to start converging, but as the parameters are learned, the performance of this variant also becomes competitive, especially after the commencement of the CIDER-D optimization. It still, however, exhibits the lowest score in every metric and by a wide margin. The only difference between AFCT-5 and AFCT-1 is the masked Fourier layer used in the decoder in place of the masked self-attention. It is evident, as described above, that while FFTs are capable of extracting relevant latent information in images, they perform quite poorly with long-term dependencies which are quite common in textual inputs, in contrast to image inputs which mainly depend on the features around themselves. The associated graphs are shown in Figure 4 and the scores are presented in Table 3.

To add to the points above, image-captioning, unlike some other tasks such as object detection and/or classification, requires exploring more fine-grained information because it not only encompasses the latter two but also requires associations and relationships to be detected in the image. This is where a Fourier transform can help by decomposing the feature vectors further to offer the self-attention mechanism more information to use when detecting relationships between each feature and other features for the purpose of determining the degree to which each feature is related to other features, i.e., the relevance of each feature to the others. This has the added advantage of detecting and learning relationships that may have been lost during the transformation from the spatial domain. The mixing of the input tokens, unlike when using convolutional kernels which limit interactions to a
neighborhood, provides context for more long-term dependencies, allowing for the model, in some instances, to, e.g., generate a caption that captures more than one object, their colors, associations, and general location. Thus, alongside the intra-attention mechanism, the fast Fourier transformed tokens promote the transfer of salient information.

Another way of looking at the performance of the model and the different variants is to consider what exactly each evaluation metric takes into consideration. As described in Section 3.3, the different evaluation metrics can help explain different aspects about the captions generated, giving a quantifiable insight into their quality. They are not perfect but they are things that can be measured and comparisons can be made. Considering CIDEr and SPICE metrics correlate best with human evaluators, they are usually considered to be the most important metrics. AFCT-F4, the variant with only fully connected layers, performed quite well in the BLEU scores but its captions are not as detailed as the ones produced by the other models. This is accurately reflected by the lower METEOR, CIDEr, and Spice score it received, which runs in contrast to its good BLEU scores. This shows that the fully connected layer was not able to capture as much fine-grained salient information as the other models that employed a combination of one or both of the Fourier and self-attention layer. In addition, the fully connected variant, based on the captions generated, had the most inconsistent performance. For instance, sometimes it generated very good captions and at other times very poor captions. The captions show a model that can identify the items in the picture but...
struggles with dependencies, capturing relevance and associations. These are some of the strengths of self-attention layers and, to a lower extent, FFT layers. The higher CIDEr and SPICE scores correlate well with the good captions generated by AFCT and AFCT-F1.

Another major takeaway from this experiment was the fact that even if we were able to mine more of the latent information from the image features, the language model would still be dominant and able to create meaningful sentences even with the least information provided, for example, with the fully connected layer variant. This shows that the information is discriminatory and adequate enough for the decoder language model to make a correct prediction both greedily or through a beam search. This is usually a merit considering as a captioning model, it has to output natural language, but this ability also causes it to hallucinate at times, as can be seen in the sample captions in Tables A1–A4. Considering for every input there will be a legitimate resulting output probability distribution, the model will always produce a caption no matter its level of confidence. In reality, these guesses are usually right most of the time, but this highlights the limitations of many of the current captioning systems; they are not production-ready. At the level of the current models, if a confidentiality variable is to be introduced, it can only be used to rank sentences, a task that beam search is already doing. If it is used by the model to judge whether or not to output a sentence, a large number of captions for which the model has a low confidence for, but are actually right, will not be output. This is also compounded by the fact that the current evaluation metrics used still do not correlate well enough with human judgment and, in a situation like ours in which we are comparing captions from the variants, fail to notice quality sentences.

In our case, we did not notice any significant decrease in the time required to train a model that uses self-attention when compared to one that used a Fourier transform. This could be attributed to the fact that the decoder section still contains two sets of eight attention head self-attention sub-modules each, while the Fourier transform section also contains linear layers (with weights that need to be learned) after the fast Fourier transform, as shown in Equation (3). If there is a time difference, it probably only becomes obvious when one replaces all sections with a Fourier transform or when conducted on much larger models. Actually, the AFCT-F4 (fully connected) variant did take about 5% less time to train but this supposed time gain was too short to be independently verified due to the varying GPU and CPU loads on the shared GPU server. The model sizes do vary as expected, with the six-layer AFCT-F2 variant being the biggest one, at twice the size of the smallest one, the AFCT-F4 variant, which only contained the position-wise feed-forward layer in the encoder. That being said, the time required for each epoch varied greatly depending on the complexity and size of the model, and especially concerning the hardware used. While we used the same GPUs, data was stored on different SSDs in order to perform simultaneous tests. During cross-entropy, it required about 40 min to 90 min per epoch, and 3 to 7 h per epoch for CIDEr optimization, varying largely depending on the complexity or size of the variant.

A lot more can be deduced from and discussed regarding the major and subtle differences between the captions generated by the different variants of the model. This will be left up to the reader to further scrutinize the nuances. The performances of the models, in which the self-attention layer was replaced by either a Fourier or a full connected layer, performed much better than expected and warrants further research. An important question, which is not a subject of this paper’s discussion but that we encountered during our experiments, concerns determining at what point to perform the CIDEr-D optimization. Due to a finite amount of computing power, it is not possible to try all possible combinations but the decision does have a significant impact on the final results obtained. We utilized CIDEr-D optimization and thus used the model with the best validation CIDEr-D score from the cross-entropy training stage as the starting point, another hyperparameter. Regardless, how many cross-entropy epochs that have a non-improving CIDEr-D score should one wait for before making the decision to switch to CIDEr-D optimization? This does add to the difficulty in comparing the various models. One may set it at six, while another may wait for forty epochs. A helpful indicator is whether the validation score is still dropping,
but at some point, one will need to make a decision based on some logic, intuition, and experience. This is some food for thought.

6. Conclusions

In this paper, we presented AFCT, a simple transformer-based captioning system that achieves competitive results comparable to more highly complex transformer-based models for image-captioning without the need to employ pre-training or complex alignment techniques. We showed that Fourier transforms can mine and highlight latent image information that would otherwise be ignored when creating captions, which thus can be leveraged to augment existing feature vectors in order to increase the amount of semantic image information in the encoder output vector. Furthermore, we analyzed and provided insight into the use of fast Fourier transform features as alternatives or supplements to regional features for self-attention in image-captioning applications. The experiments we performed highlighted both the amazing performance of current methods as well as their shortcomings, which make them not yet production-ready. In our future research, we hope to build on this and even go deeper into researching how to better exploit the potential exhibited by Fourier transforms and fully connected layers as replacements for the computationally expensive self-attention layers in transformers.

Author Contributions: Conceptualization, R.I.O.; methodology, R.I.O.; software, R.I.O.; validation, R.I.O., Z.Y., and J.L.; formal analysis, R.I.O., Z.Y. and J.L.; investigation, R.I.O.; resources, R.I.O. and J.L.; data curation, R.I.O.; writing—original draft preparation, R.I.O.; writing—review and editing, R.I.O., Z.Y. and J.L.; visualization, R.I.O.; supervision, J.L.; project administration, J.L.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under grant U2003208; in part by the Science and Technology Plan of Hunan under grant number 2016TP1003; and in part by the Key Technology R&D Program of Hunan Province under grant number 2018GK2052.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://cocodataset.org/ (accessed on 15 July 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Additional Material

Appendix A.1. Sample of Captions Generated

Table A1. A table showing a sample of the captions generated by our model (AFCT) and some of its variants (AFCT-F1, AFCT-F5, and AFCT-F4). Abbreviation: GT = ground truth. Details about each variant can be found in the ablation studies in Table 2. Sample captions part 1/4.

| GT: a market features a variety of fruits and vegetables. | AFCT-F1: boxes of fruit on display in a market. | AFCT-F5: a fruit stand at a market with bananas. | AFCT-F4: a market with a variety of fruit on display. | AFCT: a fruit stand with a variety of fruits and vegetables. |
| GT: a man jumping a skateboard at a skate park. | AFCT-F1: a person doing a trick on a skateboard on a wall. | AFCT-F5: a man doing a skateboard on a air. | AFCT-F4: a black and white photo of a person on a skateboard. | AFCT: a black and white photo of a person jumping on a wall. |
Table A1. Cont.

GT: a person is holding a cup filled with tiny doughnuts. AFCT-F1: a hand holding a bowl of donuts on a table. AFCT-F5: a person is holding a plate with. AFCT-F4: a person holding a bowl of donuts on a table. AFCT: a person holding a glass bowl filled with donuts.

GT: a train ridding on a rail with a large view of a mountain range. AFCT-F1: a train traveling down the tracks in a mountain. AFCT-F5: a train traveling down the train tracks with. AFCT-F4: a train traveling down the tracks near a mountain. AFCT: a train on the tracks with mountains in the background.

GT: a telephone booth next to a yellow fire hydrant. AFCT-F1: a yellow and blue fire hydrant on a sidewalk. AFCT-F5: a red fire hydrant sitting on the sidewalk. AFCT-F4: a yellow and blue fire hydrant on the side of a. AFCT: a yellow and blue fire hydrant sitting on the sidewalk.

GT: a horse buggy and driver sitting in front of a building. AFCT-F1: a white horse pulling a carriage down a street. AFCT-F5: a man riding a horse in a street. AFCT-F4: a man riding a horse drawn carriage on a street. AFCT: a man is riding a horse drawn carriage down the street.

GT: a bedroom with twin beds and matching linen. AFCT-F1: two beds in a hotel room with a lamp. AFCT-F5: a hotel room with a bed and beds. AFCT-F4: two beds in a hotel room with a lamp on. AFCT: two beds with colorful pillows in a hotel room.

GT: a small kitchen with a stove and counters. AFCT-F1: a kitchen with stainless steel appliances and white cabinets. AFCT-F5: a white kitchen with a sink and a stove. AFCT-F4: a modern kitchen with stainless steel appliances and. AFCT: a kitchen with white cabinets and a stainless steel appliances.

Table A2. Sample captions part 2/4.

GT: two people sitting at a bench on a city street. AFCT-F1: a man and a woman sitting on a bench looking at their cell phones. AFCT-F5: two women sitting on a bench with a park. AFCT-F4: a man and a woman sitting on a bench looking. AFCT: a man and a woman sitting on a bench looking at a cell phone.

GT: young students sitting in a classroom watching a person on a television screen. AFCT-F1: a group of people sitting in a classroom watching a screen. AFCT-F5: several people sitting at a video game with. AFCT-F4: a group of people sitting in a room watching television. AFCT: a group of people sitting in a classroom watching a screen.
Table A2. Cont.

GT: a moped is parked next to a fire hydrant on the sidewalk. AFCT-F1: a yellow motorcycle parked next to a yellow fire hydrant. AFCT-F5: a yellow fire hydrant sitting on the street. AFCT-F4: a yellow scooter parked next to a yellow scooter on. AFCT: a yellow scooter parked next to a yellow fire hydrant.

GT: 25 images of dogs are formed into a square grid. AFCT-F1: four pictures of a dog laying on a bed. AFCT-F5: a cat is standing in a dog and. AFCT-F4: a collage of pictures of a dog and a dog. AFCT: a collage of different pictures of different pictures.

GT: a small walkway with a microwave and fridge and a door. AFCT-F1: a kitchen with a microwave and a refrigerator. AFCT-F5: a white kitchen with a microwave and a kitchen. AFCT-F4: a microwave oven sitting on top of a refrigerator. AFCT: a microwave sitting on top of a refrigerator in a kitchen.

GT: a woman in a yellow raincoat consults her cell phone. AFCT-F1: a woman in a yellow jacket looking at her cell phone. AFCT-F5: a woman standing on a cell phone in. AFCT-F4: a woman standing on the street looking at her cell phone. AFCT: a woman standing on the street looking at her cell phone.

GT: a man standing next to a table filled with bottles of wine. AFCT-F1: a man in a suit and tie standing in front of wine bottles. AFCT-F5: a man standing at a table with wine. AFCT-F4: a man in a suit standing in front of a table with. AFCT: a man in a suit standing at a table with wine bottles.

GT: a collage of white squares with pictures of different meals. AFCT-F1: a series of pictures of different types of food. AFCT-F5: a display bookstore with different kinds of food. AFCT-F4: a series of pictures of different types of food. AFCT: a bunch of different pictures of different food items.

GT: people are walking by a truck with an advertisement against the former president of iran. AFCT-F1: a little girl standing in front of a food truck. AFCT-F5: two people standing at front of a truck. AFCT-F4: a little girl standing in front of a food truck. AFCT: a woman standing in front of a truck with people.

Table A3. Sample captions part 3/4.

GT: a laptop is on a table with a frosty beverage nearby. AFCT-F1: a laptop computer sitting on a table with a glass of orange juice. AFCT-F5: a laptop computer sitting on a table with. AFCT-F4: a laptop computer sitting on a table next to a glass. AFCT: a laptop computer sitting on a table with a glass of orange juice.
Table A3. Cont.

| GT                                                                 | AFCT-F1                                                                 | AFCT-F5                                                                 | AFCT-F4                                                                 | AFCT                                                                 |
|-------------------------------------------------------------------|------------------------------------------------------------------------|------------------------------------------------------------------------|------------------------------------------------------------------------|-----------------------------------------------------------------------|
| a car parked next to a stop sign                                   | a car is parked next to a stop sign                                     | a stop sign sitting on a street                                         | a green car parked next to a stop sign on                               | a green car parked next to a stop sign on a street                     |
| huge airplane with propellers flies across the blue sky            | red and white plane flying in the sky                                   | blue plane flying in the blue sky with                                 | red and white plane flying in the sky                                   | red and white blue airplane flying in the sky                           |
| large birthday cake decorated with cereal and a caterpillar sits on | birthday cake with candles on top of a table                           | birthday cake sitting on a plate with                                 | a cake with white frosting on a white plate with                       | birthday cake with white frosting and blue writing on a plate           |
| outside photo of parked cars with a blackbird sitting on a parking | car parked next to a bird on a parking meter                            | a bird sitting on a parking meter                                      | black bird sitting on top of a parking meter                            | black bird sitting on top of a parking meter                             |
| man in armor is seated on a wood frame by a horse                  | woman sitting on a bench next to a horse                                | woman riding a horse on a horse                                        | black and white photo of a woman with a horse                           | black and white photo of a woman with a horse                           |
| kitchen with a refrigerator and sink with bucket in it             | empty refrigerator and a sink in a kitchen                              | dog sitting in a kitchen in                                            | bathroom with a toilet and a roll of toilet paper                      | toilet in a bathroom with a wooden wall                                  |
| stop sign at an intersection with trees and a building behind it   | two street signs on top of a stop sign                                  | a stop sign with a street with a building                              | a stop sign with two street signs on top of                             | a stop sign with two street signs on top of                              |
| a cat with a lethargic look sitting inside a suitcase              | a black and white cat laying in a red suitcase                          | a dog is laying in a suitcase in                                       | black and white cat laying in a suitcase                               | black and white cat laying in a red suitcase                             |
| small bathroom with the seat up on the toilet                     | white toilet in a bathroom with a toilet                               | white toilet with a toilet and a bathroom                              | bathroom with a toilet and a roll of toilet paper                      | toilet in a bathroom with a roll of toilet paper                        |
Table A4. Sample captions part 4/4.

GT: there is a lady checking another lady ’s teeth. **AFCT-F1**: two women are brushing their teeth in a bathroom. **AFCT-F5**: two women holding a woman in a toothbrush. **AFCT-F4**: two women are brushing their teeth in a bathroom. **AFCT**: two women standing next to each other brushing their teeth.

GT: a big group of people posing for a picture. **AFCT-F1**: a black and white photo of a group of children. **AFCT-F5**: several people and white photo of people in. **AFCT-F4**: a black and white photo of a group of children. **AFCT**: a black and white photo of a group of children.

GT: a television and some books in a room. **AFCT-F1**: a living room with a television and a table. **AFCT-F5**: a living room with a television and a tv. **AFCT-F4**: a living room with a television and a table. **AFCT**: a living room with pictures on the wall and a television.

GT: the bike is coming down the street with his lights on. **AFCT-F1**: a person riding a motorcycle down a city street at night. **AFCT-F5**: a street driving down a street on night. **AFCT-F4**: a car driving down a city street at night. **AFCT**: a person riding a motorcycle down a city street at night.

GT: a child skier is headed down a small slope on their skies. **AFCT-F1**: a person riding a snowboard down a snow covered slope. **AFCT-F5**: a child riding a snowboard down a snow. **AFCT-F4**: a person and a child on skis in the snow. **AFCT**: a young boy riding a snowboard down a snow covered slope.

GT: there is a old photo of the public market center sign. **AFCT-F1**: a black and white photo of a city street with cars. **AFCT-F5**: a clock walking on a street with cars. **AFCT-F4**: a black and white photo of a busy city street with. **AFCT**: a black and white photo of a busy city street with a clock.

GT: a family sitting in the living room playing video games. **AFCT-F1**: a group of people in a living room playing a video game. **AFCT-F5**: two people playing a video game with a wii. **AFCT-F4**: a group of people sitting in a living room watching television. **AFCT**: a group of people sitting in a living room playing a video game.

GT: lovely hand-made pots line the stone sill of a window. **AFCT-F1**: a stone wall with various vases on it. **AFCT-F5**: a large vat on a wall with plants. **AFCT-F4**: a stone wall with vases and a stone wall. **AFCT**: a stone wall with plants and a brick wall.

GT: man standing at edge of busy road in metropolitan street. **AFCT-F1**: a busy city street with cars and a bus. **AFCT-F5**: a city street on a street with cars. **AFCT-F4**: a man walking down a city street with a bus. **AFCT**: a busy city street with cars and a man walking down.
References

1. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.

2. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. arXiv 2014, arXiv:1409.0473.

3. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image captioning with generalization view. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.

4. Biten, A.F.; Gomez, L.; Rusinol, M.; Karatzas, D. Good News, Everyone! Context driven entity-aware captioning for news images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 12466–12475.

5. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.

6. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 2015, 115, 211–252. [CrossRef]

7. Wang, Q.; Chan, A.B. Cnn++ cnn: Convolutional decoders for image captioning. arXiv 2018, arXiv:1805.09019.

8. Anje, J.; Deshpande, A.; Schwing, A.G. Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5561–5570.

9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.

10. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-Memory Transformer for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10578–10587.

11. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-semantic alignments pre-trained for vision-language tasks. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 121–137.

12. Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; Gao, J. Unified vision-language pre-training for image captioning and vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13041–13049.

13. Huang, L.; Wang, W.; Chen, J.; Wei, X.Y. Attention on attention for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 4634–4643.

14. Lee-Thorp, J.; Ainslie, J.; Eckstein, I.; Ontanon, S. FNet: Mixing Tokens with Fourier Transforms. arXiv 2021, arXiv:2105.03824.

15. Hossain, M.; Vinyals, O.; Bengio, S.; Erhan, D. Show and tell: A comprehensive survey of deep learning for image captioning. ACM Comput. Surv. 2019, 51, 118. [CrossRef]

16. Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A.C.; Berg, T.L. Babytalk: Understanding and generating simple image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. 2013, 35, 2891–2903. [CrossRef] [PubMed]

17. Li, S.; Kulkarni, G.; Berg, T.L.; Berg, A.C.; Choi, Y. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, Portland, OR, USA, 23–26 June 2011; pp. 220–228.

18. Mitchell, M.; Han, X.; Dodge, J.; Mensch, A.; Goyal, A.; Berg, A.; Yamaguchi, K.; Berg, T.; Stratos, K.; Daumé, H., III. Midge: Generating image descriptions from computer vision detections. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Avignon, France, 23–27 April 2012; pp. 747–756.

19. Kuznetsova, P.; Ordonez, V.; Berg, A.C.; Berg, T.L.; Choi, Y. Collective generation of natural image descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers–Volume 1, Association for Computational Linguistics, Jeju Island, Korea, 8–14 July 2012; pp. 359–368.

20. Kuznetsova, P.; Ordonez, V.; Berg, T.L.; Choi, Y. Treetalk: Composition and compression of trees for image descriptions. Trans. Assoc. Comput. Linguist. 2014, 2, 351–362. [CrossRef]

21. Ordonez, V.; Kulkarni, G.; Berg, T.L. Im2txt: Describing images using 1 million captioned photographs. In Proceedings of the Advances in Neural Information Processing Systems, Granada, Spain, 12–15 December 2011; pp. 1143–1151.

22. Hodosh, M.; Young, P.; Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. J. Artif. Intell. Res. 2013, 47, 853–899. [CrossRef]

23. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.

24. Karpathy, A.; Joulin, A.; Fei-Fei, L.F. Deep fragment embeddings for bidirectional image sentence mapping. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 1889–1897.

25. Kiros, R.; Salakhutdinov, R.; Zemel, R.S. Unifying visual-semantic embeddings with multimodal neural language models. arXiv 2014, arXiv:1411.2339.

26. Hochreiter, S.; Schmidhuber, J. Long short-term memory. Neural Comput. 1997, 9, 1735–1780. [CrossRef]
55. Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled transformer for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 8928–8937.

56. Yang, X.; Tang, K.; Zhang, H.; Cai, J. Auto-Encoding Scene Graphs for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019.

57. Zhang, Z.; Wu, Q.; Wang, Y.; Chen, F. Exploring region relationships implicitly: Image captioning with visual relationship attention. *Image Vis. Comput.* **2021**, *109*, 104146. [CrossRef]

58. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8026–8037.