VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning

Jun Chen¹, Han Guo², Kai Yi¹, Boyang Li³, Mohamed Elhoseiny¹
¹ King Abdullah University of Science and Technology (KAUST), ² Carnegie Mellon University, ³ Nanyang Technological University
{jun.chen,kai.yi,mohamed.elhoseiny}@kaust.edu.sa
hanguo@cs.cmu.edu, boyang.li@ntu.edu.sg

Abstract

The limited availability of annotated data often hinders real-world applications of machine learning. To efficiently learn from small quantities of multimodal data, we leverage the linguistic knowledge from a large pre-trained language model (PLM) and quickly adapt it to new domains of image captioning. To effectively utilize a pretrained model, it is critical to balance the visual input and prior linguistic knowledge from pretraining. We propose VisualGPT, which employs a novel self-resurrecting encoder-decoder attention mechanism to quickly adapt the PLM with a small amount of in-domain image-text data. The proposed self-resurrecting activation unit produces sparse activations that prevent accidental overwriting of linguistic knowledge. When trained on 0.1%, 0.5% and 1% of the respective training sets, VisualGPT surpasses the best baseline by up to 10.0% CIDEr on MS COCO [43] and 17.9% CIDEr on Conceptual Captions [63]. Furthermore, VisualGPT achieves the state-of-the-art result on IU X-ray [15], a medical report generation dataset. Our code is available at https://github.com/Vision-CAIR/VisualGPT.

1. Introduction

Recent performance gains in image captioning [13, 24, 28, 31, 75] are achieved on top of large-scale data corpora such as MS COCO [43] or Conceptual Captions [63], each containing hundreds of thousands of captions. Manual annotation of captions requires considerable time and effort. On the other hand, semi-automatic collection of image-caption pairs from the Internet, as used by Conceptual Captions [63], may generate incorrect or undesirable training data even after multiple rounds of cleaning. Data for specialized domains like medical report generation [15, 40] and low-resource language captioning [18, 74] cannot be easily scaled. Improving the data efficiency of image captioning networks would enable quick data curation, description of rare objects, and applications in specialized domains.

In this paper, we investigate the data efficiency problem for image captioning. This problem is distinct from the novel object captioning problem [1, 23], which relies on abundant in-domain data but zero out-of-domain data. Instead, we aim to improve the performance of image captioning systems trained on a small subset of in-domain data.

We propose to improve data efficiency by leveraging pretrained language models (PLMs) [17, 34, 46, 60], such as BERT [16], XLNet [77], and GPT [6, 58, 59]. Via self-supervised learning, these models acquire rich linguistic and semantic knowledge, which has been shown to inform downstream tasks in NLP [7, 21]. However, the adaptation of PLMs pretrained on unimodal textual data for multimodal tasks remains under-investigated.
A key challenge in utilizing PLMs is to bridge the domain gap between multi-modal data and the unimodal textual data the PLMs are pre-trained on. In Figure 2, we compare the part-of-speech distributions of MS COCO and WikiText-2 [51]. MS COCO employs 75% more nouns but 14% fewer verbs, which indicates a bias toward descriptions of static objects rather than actions. This suggests that, in order to effectively utilize PLMs in image captioning, we must balance prior linguistic knowledge acquired from pre-training and visual input information.

Figure 1 depicts the overall architecture of our proposed model, dubbed as VisualGPT. In the commonly used encoder-decoder architecture for image captioning, we initialize the parameters of the decoder from PLMs such as GPT-2 [59], whereas the encoder layers are randomly initialized. In addition, we propose an attention mechanism with self-resurrecting activation units (SRAUs), which balances the input from the visual encoder and the linguistic input from the previous decoder layer. The proposed mechanism can produce sparse activations while not being as vulnerable to the zero-gradient problem as regular gates; the self-resurrecting gates can be “turned on” again after being zeroed out.

Empirical results demonstrate that, when trained on 0.1%, 0.5%, and 1% of the MS COCO and Conceptual Captions data, VisualGPT outperforms several strong baseline models. We achieve the state-of-the-art result on IU X-ray [15], a medical report generation dataset. With several ablation experiments, we verify the effectiveness of PLMs and the proposed self-resurrecting attention mechanism.

Contributions. We make the following contributions:

- We explore the data efficiency problem for image captioning by utilizing pretrained language models (PLMs) as the caption decoder. With only a small amount of in-domain training data, the proposed technique quickly adapts PLMs to the cross-modal task of image captioning. To our knowledge, this is the first work that focuses on efficiently adapting large pretrained language models for image captioning.
- We propose a novel encoder-decoder attention with self-resurrecting activation units (SRAUs), which can balance features from the visual and textual modalities. SRAU produces sparse activations that reduce accidental overwriting of pretrained weights.

2. Related Work

Image Captioning. Image captioning has been extensively studied in computer vision research. Early methods [19, 31, 37, 65, 79] focus on filling templates with extracted objects, attributes, and relationships. With the advent of deep learning, researchers proposed end-to-end neural networks that encode an image into vector representations and decode a caption word by word [27, 71]. Many improvements to the encoder [11, 38, 49, 75, 76, 80, 81], the decoder [72, 73, 78], and the attention mechanism [8, 13, 24, 33, 36] has since been proposed. Encoding the image using object regions has proven beneficial [2]. Reinforcement learning enables model optimization with non-differentiable evaluation metrics [14, 45, 61, 64]. [9, 12] investigate fine-grained control of caption generation. [14, 64] adopt GAN-like architectures that encourage human-like captions.

A few formulations of the image captioning problem deviate from the traditional supervised learning paradigm. Novel object captioning aims to describe objects that do not exist in the training data [1, 23, 41, 50, 70]. Feng et al. [20] propose unsupervised captioning without using paired image-caption supervision. Kim et al. [29] focus on learning efficiency and improve the data efficiency by learning from auxiliary unpaired image-caption data.

Self-supervised NLP Models. Self-supervised training of large neural networks on textual data proves to be an important technique in the creation of high-performance NLP models. Several self-supervision signals have been proposed, such as autoregressive language modeling [5, 52], which includes the GPT series of models [6, 58, 59], and masked language modeling, which includes ELMo [56] and BERT-related methods [16, 32, 47].

In this paper, we propose a quick adaptation technique for network weights obtained using the language modeling (LM) objective. However, the proposed technique can easily be applied to other models, as the masked language modeling objective can be converted to the LM objective by masking only the last word in the textual sequence. Unlike neural networks pretrained on multimodal data (e.g., [39, 48, 57, 66, 67, 82, 83]), our method only requires a small amount of multimodal training data and focuses on adapting linguistic knowledge learned from the textual modality.
3. Preliminaries: Transformer for Captioning

The Transformer [68] has become one of the standard models for image captioning. At its core lies the multi-head dot-product attention mechanism. Taking three input matrices, query \( Q \), key \( K \), and value \( V \), the attention function can be written as

\[
\text{Attn}(Q, K, V) = \text{softmax}\left( \frac{(W^q Q)(W^k K)^T}{\sqrt{D}} \right) W^v V, \tag{1}
\]

where \( W^q, W^k, \) and \( W^v \) are trainable parameters and \( D \) is a scaling factor. Intuitively, the attention operation can be seen as encoding \( W^q Q \) as convex combination of the row vectors of \( W^v V \). The multi-head attention repeats the process with multiple sets of \( W^q, W^k, \) and \( W^v \); the results are concatenated and linearly projected back to the same dimensionality.

In visual captioning tasks, we apply a visual encoder whose output is \( I \in \mathbb{R}^{O \times S} \). \( O \) is the length of the input sequence, which in this work is a sequence of objects in the image. \( S \) is the hidden dimension size. The decoder network outputs words in the caption sequentially.

When decoding word \( t+1 \), the encoder-decoder attention takes as input the visual encoding \( I \) and the current state of the decoder \( H \in \mathbb{R}^{1 \times S} \). We apply the attention operation with \( H \) as the query and \( I \) as both the key and the value. The encoder-decoder attention is then

\[
\text{EncDecAttn}(H, I) = \text{Attn}(H, I, I). \tag{2}
\]

After that, we apply the AddNorm operator, which contains a residual connection and layer normalization [3] and can be written as \( \text{LayerNorm}(\text{EncDecAttn}(H, I) + H) \).

Researchers have proposed other variants of the encoder-decoder attention. In Figure 3, we contrast these decoder architectures with the proposed VisualGPT model. The Attention-on-Attention (AoA) module [24] provides an alternative method for combining the visual encoding \( I \) and the linguistic information \( H \) from the decoder. For another method for combining visual and linguistic information, \( \mathcal{M}^2 \) Transformer [13] connects all decoder layers to all encoder layers. In Figure 3, it is represented by the box labeled as Meshed Connection Sum.

4. VisualGPT

Pretrained language models (PLMs) such as GPT-2 [59] are trained on data from a single modality. We use a PLM as the caption decoder and feed visual information to the PLM via the encoder-decoder attention, which plays a crucial role in quickly adapting the PLMs.

With the design of the encoder-decoder attention, we aim to carefully balance visual information from the encoder and linguistic knowledge stored in the PLM. During the generation of visual words, such as “person”, “truck”, or “dog”, the model should attend to visual information. In contrast, the generation of determiners or connectives requires only linguistic knowledge. Ideally, we would like to exploit the massive amount of linguistic knowledge stored in the PLM weights (e.g., [44]), while referring to the visual input only when required. To achieve this goal, we introduce a pair of specialized gating units.

4.1. Self-Resurrecting Activation Unit

The encoder-decoder attention \( \text{EncDecAttn}(H, I) \) may be seen as encoding the linguistic information \( H \) with visual
information $I$. In VisualGPT, we control the balance between these two modalities using two complementary gates $B^{vis}$ and $B^{lan}$. The output of this module is

$$B^{vis} \otimes \text{EncDecAttn}(H, I) + B^{lan} \otimes H,$$  \tag{3}

where $\otimes$ denotes element-wise multiplication. Letting $B^{vis}[i, j]$ and $B^{lan}[i, j]$ denote the elements in the matrices, they are computed in pairs as

$$B^{vis}[i, j] = \sigma(H[i, j])\mathbb{I}(\sigma(H[i, j]) > \tau),$$

$$B^{lan}[i, j] = (1 - \sigma(H[i, j]))\mathbb{I}(1 - \sigma(H[i, j]) > \tau),$$  \tag{4}

where $\tau$ is a predefined threshold hyperparameter and $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if the inner statement is true and 0 otherwise.

An alternative to SRAU is ordinary complementary gates (OCG), computed as $\sigma(H[i, j])$ and $1 - \sigma(H[i, j])$ (see Figure 4, top left). OCG can output values that are very close to zero. In contrast, with the indicator functions SRAU directly sets values less than the threshold $\tau$ to zero, thereby introducing sparsity. When $\tau$ is set to 0, SRAU becomes OCG. As the gradient cannot backpropagate through zero gates, SRAU prevents optimization from disrupting pretrained weights that capture linguistic knowledge. This property is crucial in effective utilizing of pretrained models. In contrast, when the OCG gates output near-zero values, some small but non-zero gradients may still overwrite existing linguistic knowledge.

Another advantage of SRAU is its ability to escape from zero outputs. It is possible for one gate to output zero and the other to output non-zero values, this removes the gradient. For this reason, we name these gates self-resurrecting activation units.

The asymmetry of SRAU may appear counter-intuitive. We contrast SRAU with a “normalized” version where the two gates $\tilde{B}^{vis}[i, j]$ and $\tilde{B}^{lan}[i, j]$ become symmetric.

$$\tilde{B}^{vis}[i, j] = \frac{B^{vis}[i, j]}{B^{vis}[i, j] + B^{lan}[i, j]},$$

$$\tilde{B}^{lan}[i, j] = \frac{B^{lan}[i, j]}{B^{vis}[i, j] + B^{lan}[i, j]}.$$  \tag{5}

These gates lose the asymmetry that enables the self-resurrecting property.

In Figure 4, we visualize OCG, SRAU, and normalized SRAU. In ablation experiments, we show that SRAU outperforms than both OCG and normalized SRAU.

4.2. The Architecture and Training of VisualGPT

For completeness, we introduce the overall architecture for VisualGPT. The image encoder comprises $K$ Transformer layers. Given an image, we extract objects in the image using an off-the-shelf object detection network. After that, we feed the spatial location into the image encoder. As such, the image encoder outputs $I$ of dimension $S \times O \times K$.

The caption decoder contains $M$ layers and its parameters are initialized from a PLM. We insert the encoder-decoder module, which is randomly initialized. We apply meshed connections between the encoder and the decoder like in $M^2$ Transformer. The network is trained to maximize the probability of the next token $w_t$ conditioned on tokens $w_1, \ldots, w_{t-1}$ and the encoder output $I$. After a predefined number of epochs on supervised learning, we switch to self-critical reinforcement learning [61] with CIDEr as the reward.

5. Experiments

5.1. Datasets and Evaluation Metrics

We evaluate our model on three datasets, MS COCO [43], Conceptual Captions [63], and IU X-ray [15]. MS COCO contains 123,287 images and each of them is annotated with 5 different captions. We follow the Karpathy split [28] for the validation and test set. The Conceptual Captions dataset [63] contains around 3.3M images for training and 28K for validation, with much higher diversity than COCO. As the test data is not publicly available, we instead use the public validation data as our test set, and randomly sample 5000 different image-caption pairs from the training set as the validation set. To create the small training data setup for MS COCO and Conceptual Captions, we randomly sample 0.1%, 0.5% and 1% image-caption pairs as training data, which matches to (567, 2,835 and 5,670 pairs) for COCO and (3,300, 16,500, and 33,000 pairs) for Conceptual Captions. We repeat the experiments 4 times with different random seeds, and report the average performance. We report metrics for BLEU [54], METEOR [4], and CIDEr [61].
<table>
<thead>
<tr>
<th>Method</th>
<th>PLM</th>
<th>COCO</th>
<th>Conceptual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.1% training data</td>
<td>0.5% training data</td>
</tr>
<tr>
<td>Transformer [68]</td>
<td>None</td>
<td>57.4 13.1 16.7 40.7 40.8</td>
<td>12.4 2.4 4.9 15.2 21.2</td>
</tr>
<tr>
<td>(M^2) Transformer [13]</td>
<td>None</td>
<td>56.9 13.1 16.9 40.6 40.9</td>
<td>13.1 2.8 4.8 15.5 23.5</td>
</tr>
<tr>
<td>AoA Transformer [24]</td>
<td>None</td>
<td>56.6 13.5 15.9 40.7 38.4</td>
<td>11.4 2.4 4.6 14.7 20.9</td>
</tr>
<tr>
<td>X-Transformer [53]</td>
<td>None</td>
<td>56.7 12.9 16.5 40.6 40.4</td>
<td>12.8 2.7 4.7 15.3 23.1</td>
</tr>
<tr>
<td>OSCAR [39] BERT</td>
<td></td>
<td>53.8 11.9 17.1 39.5 41.0</td>
<td>12.2 2.4 4.3 14.8 21.9</td>
</tr>
<tr>
<td>Transformer</td>
<td>GPT</td>
<td>56.8 15.3 17.0 41.2 42.9</td>
<td>13.2 2.5 5.0 15.1 21.9</td>
</tr>
<tr>
<td>(M^2) Transformer</td>
<td>GPT</td>
<td>54.9 14.7 16.6 41.1 40.9</td>
<td>13.1 2.8 4.9 15.4 24.0</td>
</tr>
<tr>
<td>AoA Transformer</td>
<td>GPT</td>
<td>55.5 14.4 16.2 40.7 40.1</td>
<td>11.8 2.4 4.6 13.9 20.5</td>
</tr>
<tr>
<td>VisualGPT (Normalized SRAU)</td>
<td>GPT</td>
<td>55.7 15.0 16.8 41.2 42.4</td>
<td>13.3 2.9 5.1 15.8 25.8</td>
</tr>
<tr>
<td>VisualGPT (Our SRAU) GPT</td>
<td></td>
<td>58.2 16.4 18.5 41.9 45.1</td>
<td>13.9 3.2 5.6 16.7 27.7</td>
</tr>
<tr>
<td>Transformer</td>
<td>GPT</td>
<td>62.8 18.8 19.4 25.2 59.2</td>
<td>13.2 3.3 5.5 16.3 29.6</td>
</tr>
<tr>
<td>(M^2) Transformer</td>
<td>GPT</td>
<td>63.3 19.4 19.8 45.6 61.3</td>
<td>14.5 3.6 6.0 17.1 32.0</td>
</tr>
<tr>
<td>AoA Transformer</td>
<td>GPT</td>
<td>63.5 20.2 19.4 45.8 63.9</td>
<td>13.8 3.3 5.6 17.9 31.8</td>
</tr>
<tr>
<td>X-Transformer</td>
<td>GPT</td>
<td>62.9 19.0 19.6 45.7 62.0</td>
<td>14.2 3.5 5.8 17.3 32.1</td>
</tr>
<tr>
<td>OSCAR BERT</td>
<td></td>
<td>59.2 18.0 21.0 45.3 60.2</td>
<td>14.4 3.7 6.1 17.2 33.5</td>
</tr>
<tr>
<td>Transformer</td>
<td>GPT</td>
<td>65.1 21.8 20.6 46.6 69.5</td>
<td>16.2 3.8 6.5 18.3 35.6</td>
</tr>
<tr>
<td>(M^2) Transformer</td>
<td>GPT</td>
<td>64.7 21.8 20.7 47.1 68.5</td>
<td>13.9 3.6 6.0 17.2 34.1</td>
</tr>
<tr>
<td>AoA Transformer</td>
<td>GPT</td>
<td>64.2 21.2 20.5 46.5 67.2</td>
<td>14.8 3.6 6.2 17.6 34.1</td>
</tr>
<tr>
<td>VisualGPT (Normalized SRAU)</td>
<td>GPT</td>
<td>65.3 21.8 20.9 47.0 69.3</td>
<td>14.9 3.9 6.1 18.0 35.9</td>
</tr>
<tr>
<td>VisualGPT (Our SRAU)</td>
<td>GPT</td>
<td>66.2 22.1 21.1 47.3 70.3</td>
<td>15.9 4.2 6.7 18.5 37.2</td>
</tr>
</tbody>
</table>

Table 1. Performance of the compared methods training on 0.1%, 0.5% and 1% of MS COCO and Conceptual Caption image-caption pairs. The best performance in each configuration is in bold. Ablated models are marked in gray.

ROUGE [42], and CIDEr [69].

IU X-ray [15] is a radiography dataset containing 7,470 chest X-ray images and 3,955 human-written reports. As the dataset is already small, we follow the original split, which has a training set of 5,226 images and 2,770 reports. Most reports have two images corresponding to the frontal and lateral viewpoints.

5.2. Experimental Settings

Baselines. We compare our model with several state-of-the-art transformer-based models, including:

- Plain Transformer [68].
- AoA Transformer, which inserts an attention-on-attention (AoA) module [24] into every transformer layer, as depicted by Figure 3 (b). Following [13], we slightly update the original AoA network in [24] by replacing the LSTM with Transformers in order to create a fair Transformer-to-Transformer comparison.
- \(M^2\) Transformer [13], which proposes a meshed connection between encoder and decoder and is one of the best-performing models on MS COCO.
- X-Transformer [53], which employs bilinear pooling to selectively capitalize on visual information and is
against state-of-the-art medical report generation models.

We compared VisualGPT and confirm our selection choice of using GPT as a decoder. VisualGPT is better than OSCAR in both datasets, which speaks to the value of leveraging GPT for medical report generation. We also initialize their decoder with 12-layer GPT and keep the same encoder as VisualGPT. We contrast between plain cross-attention, meshed cross-attention, and attention-on-

<table>
<thead>
<tr>
<th>Models</th>
<th>B-1</th>
<th>B-2</th>
<th>B-3</th>
<th>B-4</th>
<th>R</th>
<th>M</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Att2in</td>
<td>22.4</td>
<td>12.9</td>
<td>8.9</td>
<td>6.8</td>
<td>30.8</td>
<td>-</td>
<td>29.7</td>
</tr>
<tr>
<td>CoAtt</td>
<td>45.5</td>
<td>28.8</td>
<td>20.5</td>
<td>15.4</td>
<td>36.9</td>
<td>-</td>
<td>27.7</td>
</tr>
<tr>
<td>HRGR</td>
<td>43.8</td>
<td>29.8</td>
<td>20.8</td>
<td>15.1</td>
<td>32.2</td>
<td>-</td>
<td>34.3</td>
</tr>
<tr>
<td>CMAS-RL</td>
<td>46.4</td>
<td>30.1</td>
<td>21.0</td>
<td>15.4</td>
<td>37.1</td>
<td>-</td>
<td>27.5</td>
</tr>
<tr>
<td>Chen et al.</td>
<td>47.0</td>
<td>30.4</td>
<td>21.9</td>
<td>16.5</td>
<td>37.1</td>
<td>18.7</td>
<td>-</td>
</tr>
<tr>
<td>VisualGPT (ours)</td>
<td><strong>48.0</strong></td>
<td><strong>31.3</strong></td>
<td><strong>22.2</strong></td>
<td><strong>15.9</strong></td>
<td><strong>37.4</strong></td>
<td><strong>20.5</strong></td>
<td><strong>49.7</strong></td>
</tr>
</tbody>
</table>

Table 2. Performance on the IU X-ray dataset.

one of best-performing models on MS COCO.

- OSCAR [39], which finetunes BERT initialization on image-language dataset.

Since VisualGPT has GPT as the pretrained decoder, for fair comparisons, we also create variants of Transformer, AoA Transformer and $M^2$ Transformer with GPT as the decoder. For VisualGPT, we set $\tau$ to 0.2 in all experiments. We also explored the effect of different $\tau$ and find $\tau$ in the range of $[0,0.2]$ to offer the right level of sparsity. For all other baselines, we tune the hyperparameters on the validation set of MS COCO. We train our model and all the baselines in reinforcement learning setting following the work in [13]. Please see the supplemental material for more details on hyperparameters and experimental results.

5.3. Quantitative Results

Small In-domain Training Data. Results on MS COCO and Conceptual Captions are presented in Tables 1. VisualGPT outperforms the best-performing baseline model by 4.1 CIDEr when trained on 0.1% of MS COCO data, 6.4 CIDEr when trained on 0.5% data and 2.5 CIDEr with 1% training data. On Conceptual Caption dataset, VisualGPT also outperforms all the baselines. It outperforms the best baseline model by 4.2 CIDEr under 0.1% training data, 3.5 CIDEr under 0.5% data and 0.3 CIDEr under 1% data.

Comparison with BERT-based model. We compared with OSCAR [39] which is a BERT-based [16] model with good performing results in many benchmarks. We run their model without pretraining on a large-scale image-language corpus for the fair comparison with our model. The main difference between BERT and GPT is their different pretraining objectives, where BERT uses masked language modeling and GPT is the autoregressive prediction of the next word. GPT has more similar learning behaviors to the image captioning model compared to BERT since they are both optimized by autoregressively generating the next language word. The experimental result in Table 1 shows that VisualGPT is better than OSCAR in both datasets, which confirms our selection choice of using GPT as a decoder.

Medical Report Generation. We compared VisualGPT against state-of-the-art medical report generation models including Att2in [61], CoAtt [26], HRGR [35], CMAS-RL [25] and the model from Chen et al. [10]. This dataset only contains around 2,770 medical reports in the training set, which is less than 1% COCO data and poses a data-efficiency challenge. We follow the same experimental setting as in [10]. The results show that VisualGPT outperforms the baselines for most evaluation metrics and creates a new state-of-the-art. It shows the value of leveraging GPT knowledge into the highly specific domain which has very “expensive” and insufficient paired data. We hope our finding could inspire future work in other domains.

Comparison Against Semi-supervised and Unsupervised Methods. Kim et al. [30] proposed a semi-supervised learning method to improve the data efficiency of image captioning. They used 1% of images and all their captions as training data, rather than 1% of all the image-caption pairs in Table 1, hence they cover less images since each image is associated to more than 1 caption. For Kim et al. + unpaired, they also employ the other 99% of MS COCO as unpaired images and captions for training. We replicate their setup by only training with 1% of images. As shown in Table 3, without using additional unpaired images and captions, the proposed VisualGPT method outperforms Kim et al. [30] by 20.6 CIDEr score.

We also compare VisualGPT against unsupervised methods of Gu et al. [22] and Feng et al. [20], which use tens of millions of unpaired images and captions. Even though these are not fair comparisons, it is encouraging to see VisualGPT surpassing these baselines by utilizing the supervision of only 1133 training images.

5.4. Ablation Studies

Ablation on cross-attention: To fairly compare our SRAU with other cross-attention mechanisms in the baselines, we also initialize their decoder with 12-layer GPT and keep the same encoder as VisualGPT. We contrast between plain cross-attention, meshed cross-attention, and attention-on-

<table>
<thead>
<tr>
<th>Models</th>
<th>B-1</th>
<th>B-2</th>
<th>B-3</th>
<th>B-4</th>
<th>R</th>
<th>M</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim et al. [30]</td>
<td>58.1</td>
<td>13.4</td>
<td>15.9</td>
<td>-</td>
<td>36.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim et al. + unpaired</td>
<td>63.0</td>
<td>18.7</td>
<td>20.7</td>
<td>-</td>
<td>55.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gu et al. [22]</td>
<td>46.2</td>
<td>5.4</td>
<td>13.2</td>
<td>-</td>
<td>17.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feng et al. [20]</td>
<td>58.9</td>
<td>18.6</td>
<td>17.9</td>
<td>-</td>
<td>54.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VisualGPT (ours)</td>
<td><strong>67.1</strong></td>
<td><strong>24.3</strong></td>
<td><strong>21.9</strong></td>
<td><strong>48.6</strong></td>
<td><strong>75.8</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Comparison with unsupervised and semi-supervised learning methods using Kim et al.’s split of MS COCO. Kim et al. employ only 1% images for training in contrast to 1% image-caption pairs from Table 1. Note that Kim et al. + unpaired also use the rest of training data as unpaired images and texts. The gray shading denotes baselines that use a large amount of unpaired images and texts during training.
attention (AoA) modules. For AoA Transformer, we add the AoA module on top of cross-attention. Table 1 shows the results, which demonstrate that SRAU is better than other cross-attention modules in exploiting the GPT knowledge within the image-caption task.

Ablation on SRAU: We create an ablation called Normalized SRAU, where we replace the SRAU with the normalized SRAU (see Figure 4) and use GPT2 initialization. We provided the results in table 1. The normalized SRAU results in substantially lowered performance, decreasing CIDEr from full VisualGPT by 2.7, 1.0, and 0.3 respectively on the three setups on MS COCO, and it also decreases from Full VisualGPT by 2.2, 1.3 and 0.6 respectively on Conceptual Caption. This demonstrates that the self-resurrecting property is beneficial for learning from small data. We experimented with Leaky ReLU and GELU, which ameliorate zero gradients, but the training crashed due to the lack of upper limits for function values.

We explored different $\tau$ among (0, 0.1 0.2) and show their CIDEr performance on different percentage of COCO training data in the Figure 5. $\tau=0$ is equivalent to ordinary complementary sigmoid gates. We can observe that $\tau=0.2$ can give us the best performance in most cases, indicating the usefulness of incorporating sparsity in our SRAU complementary gates.

5.5. Human Study

In addition to automatic evaluation metrics, we conduct two human studies to further evaluate the quality of generated captions. In the first study, we asked participants directly for preference over generated captions. We randomly selected 250 test images from the three setups of 0.1%, 0.5%, and 1% training data. For every image, we generated one caption from VisualGPT and each of three high-performing baselines from Table 1, Transformer [68], $M^2$ Transformer [13], and AoA Transformer [24], all with three decoder layers. Every image was evaluated by 5 different Turkers, who chose the caption that most accurately described the image content. We received 3750 (250 images

![Figure 5. CIDEr performance v.s. different thresholds $\tau$ with 0.1% 0.5%, 1% and 5% training data.](image)

![Table 4. The percentage of votes received by VisualGPT and baseline models under different quantity of training data.](table)

**Q1. Does the caption miss things shown in the image?**

<table>
<thead>
<tr>
<th>Answer</th>
<th>Ours</th>
<th>$M^2$ Transformer</th>
<th>Transformer</th>
<th>AoA</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>719</td>
<td>624</td>
<td>633</td>
<td>621</td>
<td>973</td>
</tr>
<tr>
<td>Yes</td>
<td>367</td>
<td>438</td>
<td>456</td>
<td>447</td>
<td>73</td>
</tr>
<tr>
<td>No Rate</td>
<td>0.66</td>
<td>0.59</td>
<td>0.58</td>
<td>0.58</td>
<td>0.93</td>
</tr>
</tbody>
</table>

**Q2. Does the caption describe things not in the image?**

<table>
<thead>
<tr>
<th>Answer</th>
<th>Ours</th>
<th>$M^2$ Transformer</th>
<th>Transformer</th>
<th>AoA</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>720</td>
<td>692</td>
<td>633</td>
<td>655</td>
<td>448</td>
</tr>
<tr>
<td>Yes</td>
<td>360</td>
<td>418</td>
<td>423</td>
<td>412</td>
<td>43</td>
</tr>
<tr>
<td>No Rate</td>
<td>0.67</td>
<td>0.62</td>
<td>0.60</td>
<td>0.61</td>
<td>0.96</td>
</tr>
</tbody>
</table>

![Figure 6. Visual scores of words in generated captions.](image)
We summarize the results in Table 4. Overall, the captions generated by VisualGPT received the largest share of votes, 39.2% for the 0.1% training data split, 39.1% for the 0.5% split, and 37.4% for the 1% split. For each training setup, we conducted Pearson’s Chi-square test [55], which shows the differences are statistical significant with \( p < 0.05 \) in all cases.

In the second study, we evaluate if using pretrained language models introduces excessive linguistic prior that could cause the known object hallucination problem [62]. From the models trained using 1% COCO data, we randomly sampled 250 images with the generated caption from each model. For each image, we asked 5 different participants if the caption (1) described non-existent objects or (2) missed objects existing in the image. To catch random clickers, we created 5 images with verified captions, so that we knew the right answers of these questions. Participants who answered these questions wrongly were considered unreliable and removed from the results.

The results are in Table 5. Compared to the baselines, VisualGPT has less hallucination and higher coverage of objects. The study also finds that the ground-truth captions has the least amount of hallucination and highest coverage of objects in the image. This finding lends positive support to the validity of the experimental protocol.

5.6. Analysis

In this section, we visually examine examples from the VisualGPT model trained on 1% of MS COCO. First, we show example captions generated by VisualGPT in Figure 6 and the associated \( B^{vis} \) at the last decoder layer. Note that for every word generated, we have a 768-dimensional visual gate vector, which is a slice of \( B^{vis} \) at different decoding time steps. We take the mean of the gate vector as the visual score for that word. After that, we normalize the visual scores across the dataset to the \([0, 1]\) interval and highlight the words accordingly. Blue indicates high visual scores and red indicates low visual scores. We observe that, in agreement with our intuition, VisualGPT assigns high visual scores to words like “desk” and “snowy surface” and low visual scores to determiners and prepositions.

In Figure 7, we plot the distribution of \( B^{vis} \) and \( B^{lan} \) at every decoding layer as a box-and-whisker diagram. We also show the words with the highest and lowest visual scores, which are again in line with our expectations. Additionally, we observe that, going from layer 0 to layer 9, the decoder makes increasing use of visual information, but the uppermost layers, 10 and 11, make more balanced use of information. We hypothesize that the low layers focus on low-level linguistics like syntax, whereas the middle layers learn to fuse linguistic information with visual information. Finally, the two information sources become balanced in the uppermost layers.

5.7. Limitation

One limitation of our proposal is that, as experiments in the supplementary material show, the gap between baseline models and VisualGPT gradually vanishes as in-domain training data increase. The phenomenon is more pronounced in COCO than Conceptual Captions, which has a more diverse vocabulary. We hypothesize that linguistic knowledge from pretrained models is the most useful when the training data are small and do not provide sufficient coverage of the vocabulary.

6. Conclusions

We present VisualGPT, a data efficient image captioning model which leverages the linguistic knowledge from the pretrained language model. To bridge the semantic gap between different modalities, we design a novel encoder-decoder attention mechanism with an unsaturated rectified gating function. We evaluate our model on 0.1%, 0.5% and 1.0% of MS COCO and Conceptual Captions, and IU X-ray, a small medical imaging report dataset. VisualGPT achieves the state-of-the-art result on IU X-ray and outperforms strong baseline models.

VisualGPT may solve the realistic need when training captioning models on low-resource languages or highly specialized domains, where it could be challenging to find annotators to collect a large amount of data.

Acknowledgments. This work is funded by KAUST BAS/1/1685-01-0, KAUST-FCC/1/2533-17-01, and National Research Foundation Fellowship (NRF-NRFF13-2021-0006), Singapore.
References


[29] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, EMNLP-IJCNLP, Association for Computational Linguistics, 2019. 2


[38] Xiangyang Li and Shuqiang Jiang. Know more say less: Image captioning based on scene graphs. IEEE Transactions on Multimedia, 21(8), 2019. 2


[41] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Pointing novel objects in image captioning. In CVPR, 2019. 2


[55] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50(302), 1900. 8


[64] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In ICCV, 2017. 2
[67] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, EMNLP-IJCNLP. ACL, 2019. 2
[71] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. TPAMI, 39(4), 2016. 2
[75] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015. 1, 2
[77] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In NeurIPS, 2019. 1