ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic

Yoad Tewel Yoav Shalev Idan Schwartz Lior Wolf
School of Computer Science, Tel Aviv University

Abstract

Recent text-to-image matching models apply contrastive learning to large corpora of uncurated pairs of images and sentences. While such models can provide a powerful score for matching and subsequent zero-shot tasks, they are not capable of generating caption given an image. In this work, we repurpose such models to generate a descriptive text given an image at inference time, without any further training or tuning step. This is done by combining the visual-semantic model with a large language model, benefiting from the knowledge in both web-scale models. The resulting captions are much less restrictive than those obtained by supervised captioning methods. Moreover, as a zero-shot learning method, it is extremely flexible and we demonstrate its ability to perform image arithmetic in which the inputs can be either images or text and the output is a sentence. This enables novel high-level vision capabilities such as comparing two images or solving visual analogy tests. Our code is available at: https://github.com/YoadTew/zero-shot-image-to-text.

1. Introduction

Deep learning has led to at least three major revolutions in computer vision: (i) machines that achieve, in multiple domains, what is considered a human level of performance earlier than anticipated [39, 64], (ii) effective transfer learning, which supports rapid modeling of new domains [74], and (iii) a leap in unsupervised learning through the use of adversarial and self-supervised learning [9, 23].

A fourth revolution that is currently taking place is that of zero-shot learning. A seminal work by OpenAI presented the transformer-based [65] GPT-3 model [6]. This model is trained on extremely large text corpora and can then generate text given a prompt. If the prompt contains an instruction, GTP-3 can often carry it out. For example, given the prompt "Translate English to French: typical $\rightarrow$ typique, house $\rightarrow$ ... " would generate the word "maison."

Impressive zero-shot capability was later on demonstrated, also by OpenAI, in computer vision. While state-of-the-art computer vision models are often trained as task-specific models that infer a fixed number of labels, Radford *et al*. [55] have presented the CLIP image-text transformer model, which can perform tens of downstream tasks, without further training, with an accuracy comparable to the state of the art. This is done by selecting, given an image, the best match out of sentences of the form "This is an image of X." Subsequently, Ramesh *et al*. [57] presented a bimodal Transformer termed DALL-E, which generates images that match a given description in unseen domains with unprecedented performance.

In this work, we employ CLIP to perform the inverse task of DALL-E, namely zero-shot image captioning. Given an image, we employ CLIP together with the GPT-2 language model [56] (we do not have access to GPT-3) to generate a textual description of the input image. This adds a new image-analysis capability to CLIP, beyond the fixed-prompt zero-shot learning demonstrated by Radford *et al*.

As a zero-shot method, our approach does not involve any training. One can argue that the underlying CLIP model is trained with exactly the same type of supervision that image captioning methods [62, 76] are trained on, i.e., pairs of matching images and captions. However, image captioning methods are trained from curated sources, such as MS-COCO [43] or Visual Genome [38], while CLIP is trained on WebImageText (WIT), which is an automatically collected web-scale dataset. Previous attempts to train a captioning model on WIT have led to poor performance in recognizing the objects in the image, see Sec. 2.

As a result of the difference in both methodology and underlying data, the captions produced by our method are very different from those obtained by the supervised captioning methods. While supervised methods can mimic human annotators and provide similar sentences, in terms of conventional NLP metrics (such as BLEU [53]) to the ground truth sentences, our results exhibit much more freedom and match the image better in the visual-semantic CLIP embedding space (ours is optimized for this). Moreover, the semantic knowledge incorporated into CLIP and GPT-2 is manifested in the resulting caption, see Fig. 1.

In addition to the different nature of the obtained captions, our method is also more flexible, since all the computing occurs at inference time. Specifically, we show the

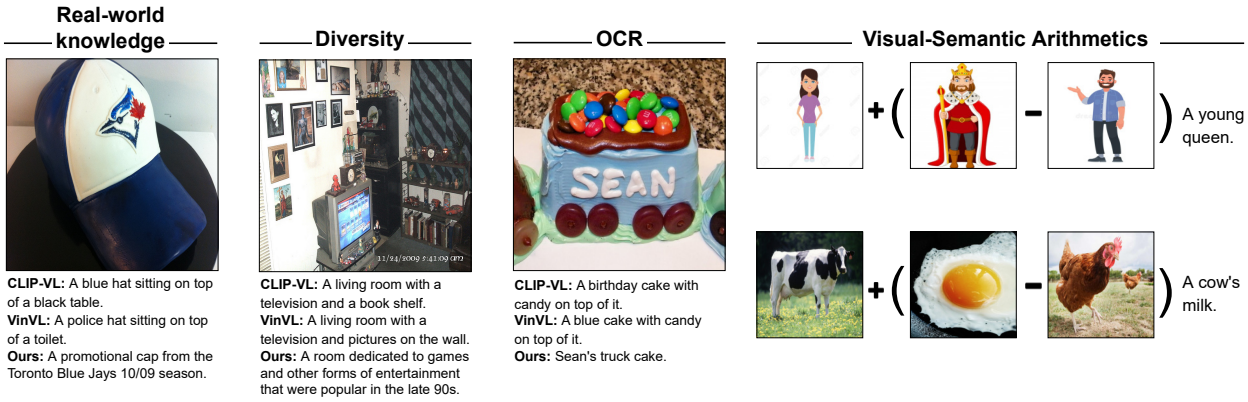

Figure 1. Our novel captioning method ZeroCap exhibits real-world knowledge, generates text that is more diverse and less scripted than existing methods, can address the written content of an image, and can perform visual-semantic arithmetic.

ability to perform semantic analysis in image space by using a new form of arithmetic. A well-known example for concept arithmetic in NLP is that of retrieving the word 'queen' as the closest word, in the embedding space, to the equation involving the embedding vectors associated with 'king,' 'man,' and 'woman,' after subtracting the 2nd from the 1st and adding the 3rd. We present the novel ability to do the same, only with images instead of words, such that the result is generated as a short sentences, and not just a word, see Fig. 1.

As a corollary, we can, for example, ask what the difference is between two scenes. This ability to compare two images semantically is a novel computer vision capability, which further demonstrates the power of zero-shot learning.

2. Related work

The first deep captioning methods applied RNNs to generate sequences of words [37, 48]. Attention was added to identify relevant salient objects [59, 61, 71]. Graph neural networks and scene graphs incorporated spatial as well as semantic relationships between objects [36, 72, 73]. Subsequently, Transformers modeled interactions among all image elements with self-attention [17, 51, 60, 65]. On the text modeling side of the problem, language models (LMs) have also advanced with the development of LSTMs [16, 69], CNNs [4] and Transformers [27, 29, 47]. Language improvements include devising better image grounding [46], decoding non-visual words (e.g., 'the,' 'and') [45], generating fine, novel and diverse sentences [7,24,70], and incorporating information from different semantic taggers [12, 33].

In recent years, significant improvements have been achieved by utilizing large-scale vision-language data sets. The unsupervised data is used as a pre-training phase, to initialize models with image-text correspondence [15, 41, 76]. With this technique, millions of image and text pairs from the web can be adopted. Nevertheless, in previous work we are aware of, all captioning models employ human-annotated datasets, such as MS-COCO or the Visual Genome, in the last stage of training.

It is likely impossible to construct a database of curated captions that is large enough, to describe even a modestly large fraction of plausible images and objects. This results in biases [20, 21, 68]. Several approaches focused on describing novel objects by conditioning the model on external unsupervised data during training [1, 28, 67]. Alternatively, external object taggers can be used during different phases (e.g., pre-training, training, or inference) [3, 18, 31, 42, 46]. Semi-supervised methods are also available [35]. Unsupervised approaches can be achieved by training with a visual concept detector or by learning a joint image-language embedding space [19, 40]. In contrast, our method makes use of an existing image-text alignment score to direct an existing large-scale LM toward a given image without training.

CLIP is trained on 400M images/sentence pairs from the web [55], resulting with a powerful text-image matching score. Originally CLIP's authors explored training an image-to-caption language model with this training set, but found that it struggled with zero-shot transfer. In a 16 GPU-day experiment, a language model only achieved 16% accuracy on ImageNet [14]. CLIP achieves the same level of accuracy roughly 10x faster.

Using prompts, it is possible to imitate some capabilities of text generation. For example, CLIP-based applications exhibit zero-shot solving capabilities in various scenarios never seen before. With careful engineering of the prompt, one can, for example, improve detection of unseen objects [26]. Zero-shot prompt engineering has also been used for higher-level tasks (e.g., VQA), but it is nowhere near the level of supervised methods [62].

CLIP also provides powerful means for supporting text-driven image manipulation with Generative Adversarial Networks (GANs) or other generative models [8, 54, 58]. Our work explores the other direction: generating text us-

ing an image, by guiding a large-scale LM with CLIP.

Guided language modeling has become a primary challenge, as researchers strive to tune prior knowledge within large-scale LMs, such as GPT-2 [56]. Fine-tuning is often accomplished by employing Reinforcement Learning [77] or GANs [75] for each attribute separately. Disentangling the latent representations into style and content is also relevant in terms of text style transfer [32, 63]. A controllable LM can also be formed using fixed control codes [34]. Ideally, conditioning should be applied directly to the existing large-scale LM, without the need for fine-tuning. Several studies have explored the idea of steering an LM using small neural networks [10, 25]. Following that, PPLM [13] demonstrated that a simple attribute classifier could steer a model without any further training. With our work, we present a novel visual LM guidance from visual cues.

3. Method

Visual captioning is the process of generating a descriptive sentence for an image. It can be formalized as a sequence generation problem given an input image I, i.e., as a conditional probability inference for the i-th word x_i of the sentence, i.e., $p(x_i|[x_t]_{t<i}, I)$.

This is typically accomplished in a supervised manner, by optimizing weights to reproduce ground truth sentences. However, since carefully curated datasets are small, and cannot adequately describe all images, the sentences generated often describe the content at the basic level of the objects present in the scene and sound artificial. Such problems can be mitigated with the use of web-scale datasets. We present a zero-shot method for guiding large-scale language models with a large-scale text-image alignment model.

Overview Our approach uses a transformer-based LM (e.g., GPT-2) to infer the next word from an initial prompt, such as "Image of a," as illustrated in Fig. 2. To incorporate image-related knowledge to the auto-regression process, a calibrated CLIP loss $\mathcal{L}_{\text{CLIP}}$ stimulates the model to generate sentences that describe a given image. An additional loss term $\mathcal{L}_{\text{CE}}$ is used to maintain the next token distribution similar to the original language model. Optimization occurs during auto-regression, and repeated for each token.

Furthermore, the flexibility of our method enables the capturing of semantic relations through simple arithmetic of visual cues in CLIP's embedding space. Finally, combining multi-modal encoders with our method allows knowledge to be extracted in a new way that mixes between text and images.

.

Language models In recent years, LMs have improved significantly and are getting closer to AI-complete capabilities, including broad external knowledge and solving a wide variety of tasks with limited supervision. A Transformer-based LM typically models interactions between the generated token and past tokens at each time-step.

Recall that the transformer block has three embedding functions K, Q, V [65]. The first two, K, Q, learn the token interactions that determine the distribution over V. The attention mechanism pools values based on the similarity between queries and keys. Specifically, the pooled value for each token i depends on the query associated with this token Q_i, which is computed using the function Q over the current embedding of this token. The result is obtained as the weighted average of the value vectors, based on the cosine similarity between Q_i and the keys associated with all tokens K_j.

While K and V are functions, the obtained key and values K_j and V_j are used repeatedly when generating text, one word at a time. K_j and V_j can therefore be stored in what is called a context cache, in order to keep track of past embedding outputs of K and V. The sequence generation process can then be written as

$$x_{i+1} = \text{LM}\left(x_i, [(K_j^l, V_j^l)]_{j<i, 1\leq l\leq L}\right), \tag{1}$$

where x_i is the i-th word of the generated sentence, K_j^l, V_j^l are the context transformer's key and value of the j-th token, and l indicates the index of the transformer layers, out of a total of L layers. Our method employs GPT-2, which has $L = 24$ layers.

We next describe how we align our LM with the input image. We do so by modifying, during inference, the values of the context cache $C_i = [(K_j^l, V_j^l)]_{j<i, 1\leq l\leq L}$ leaving the LM unchanged.

CLIP-Guided language modelling Our goal is to guide the LM towards a desired visual direction with each generation step. The guidance we propose has two primary goals: (i) alignment with the given image; and (ii) maintaining language attributes. The first goal is obtained through CLIP, which is used to assess the relatedness of a token to an image and adjust the model (or, rather, the cache) accordingly. For the second goal, we regularize the objective to be similar to the original target output, i.e., before it was modified.

The solved optimization problem adjusts the context cache C_i at each time point and is formally defined as $\arg\min_{C_i} \mathcal{L}_{\text{CLIP}}(\text{LM}\left(x_i, C_i\right), I)$

$$+ \lambda \mathcal{L}_{\text{CE}}\left(\text{LM}\left(x_i, C_i\right), \hat{x}_{i+1}\right), \tag{2}$$

where $\hat{x}_{i+1}$ is the token distribution obtained using the original, unmodified, context cache. The second term employs CE loss to ensure that the probability distribution across words with the modified context is close to the one of the original LM. The hyperparameter λ balances the two loss terms. It was set to 0.2 early on in the development process and was unmodified since. Next, we explain how the CLIP loss term is calculated.

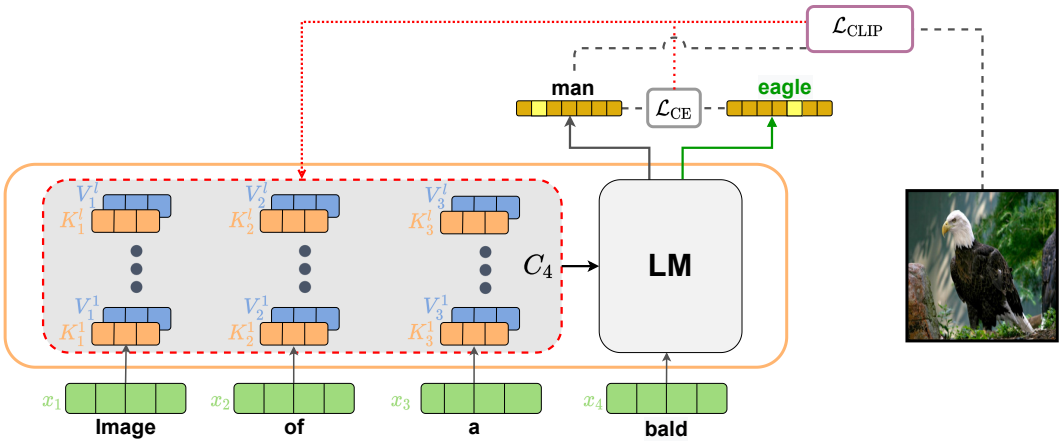

Figure 2. An overview of our approach. We guide the model towards the phrase 'eagle' instead of 'man'. We do this by adjusting the context (C_4), using the gradients of CLIP loss ($\mathcal{L}_{\text{CLIP}}$) illustrated with a red arrow. To maintain language attributes, we optimize the minimum distance to the original distribution of the LM ($\mathcal{L}_{\text{CE}}$).

CLIP loss We calculate image relevance for the possible tokens at time i. It is sufficient to compute potentials for the top 512 token candidates and set the rest to zero potential for efficiency. To this end, the corresponding candidate sentence $s_i^k = (x_1, ..., x_{i-1}, x_i^k)$ for the k-th candidate token is matched against the image I.

The clip potential of the k-th token is computed as

$$p_i^k \propto \exp(\text{D}_{\text{CLIP}}(E_{\text{Text}}(s_i^k), E_{\text{Image}}(I))/\tau_c)), \quad (3)$$

where D_{CLIP} is the cosine distance between CLIP's embeddings of the text (i.e., E_{Text}) and the image (i.e., E_{Image}), and $\tau_c > 0$ is a temperature hyperparameter that controls the sharpness of the target distribution. In all our experiments, we set τ_c to 0.01.

The CLIP loss is defined as the cross-entropy loss between the clip potential distribution and the target distribution of the next token x_{i+1} obtained by the language model:

$$\mathcal{L}_{\text{CLIP}} = \text{CE}(p_i, x_{i+1}). \quad (4)$$

This loss encourages words that lead to higher CLIP matching scores between the image and the generated sentence.

Inference As a zero-shot method, no training takes place. At inference time one optimizes the problem in Eq. (2), which we denote as $p(x_{i+1}|C_i)$, by conducting five steps of gradient descent, i.e.,

$$C_i \longleftarrow C_i + \alpha \frac{\nabla_{C_i} p(x_{i+1}|C_i)}{\|\nabla_{C_i} p(x_{i+1}|C_i)\|^2}. \quad (5)$$

This update rule is simplified for brevity. With each newly-generated token, the optimization is re-done. In our implementation, the gradients are normalized with Euclidean normalization before each step, separately for each transformer layer. We set the learning rate α to 0.3.

Beam search The byte-level tokenizer used employs 256 bytes of base tokens to represent every word in existence [56]. Any word can also be split into more than one subwords, e.g., the word 'zebra' is tokenized as 'zeb' and 'ra'. As a result, we found that images of zebras are described as striped animals, since the token 'zeb' is not picked. Beam search inference helps solve this problem by enabling the search to be conducted in a less myopic way.

4. Visual-Semantic Arithmetic

Recent studies suggested that CLIP multi-modal representation holds an elaborate concept taxonomy [22]. In accordance with this intuition, we find that our method can express CLIP's embedding in a textual way. For instance, subtracting between CLIP-encoded images and applying our method transcribes a relationship between the two images. Furthermore, by summing vectors we can steer the generated caption towards a conceptual direction.

To perform arithmetic in CLIP's embedding space, we first encode the image/text using CLIP's image/text encoder. For instance, let I_1, I_2 be two images. We encode the images with CLIP's encoder, i.e., $E_{\text{image}}(I_1), E_{\text{image}}(I_2)$. Next, we carry out the desired arithmetic, e.g., addition with $E_{\text{image}}(I_1) + E_{\text{image}}(I_2)$. Finally, we use the obtained result instead of the image encoding $E_{\text{image}}(I)$ within Eq. (3) to steer the generated sentence.

Consequently, we can generate detailed knowledge of the external world by moving in conceptual directions. This way, our method can answer questions expressed visually, for example, "who is the president of Germany?" To achieve this, we subtract "America's flag" from an image of "Obama" and obtain a presidential-direction, to which we can then add the image of a second country's flag.

Our approach extends beyond visual interactions alone. Using CLIP's textual encoder, interaction with a natural lan-

	Supervised Metrics					Diversity Metrics		Unsupervised Metric
Method	B@4	M	C	S	CLIP-S^{Ref}	Vocab	%Novel	CLIP-S
ClipCap [50]	32.15	27.1	108.35	20.12	0.81	1650	66.4%	0.77
CLIP-VL [62]	40.2	29.7	134.2	23.8	0.82	2464	85.1%	0.77
VinVL [76]	**41.0**	**31.1**	**140.9**	**25.2**	**0.83**	1125	77.9%	0.78
Ours	2.6	11.5	14.6	5.5	0.79	**8681**	**100%**	**0.87**

Table 1. For each method, we report supervised metrics (i.e., ones requiring human references): B@1 = BLEU-1, M = METEOR, C = CIDEr, S = SPICE. We also report diversity metrics, which measures the vocabulary size (Vocab), and the number of novel sentences w.r.t the training set (%Novel). Finally, we report semantic relatedness to the image (CLIP-S), and to the human references (CLIP-S^{Ref}) based on CLIP's embeddings.

guage is possible. In this case, one performs arithmetic operations in the embedding space such that the expression contains both image- and text-embeddings.

5. Experiments

For all the results reported in this section, we used a strategy for reducing repetitions, in which the probability for generating tokens that were generated at the last four time-steps was decreased by a factor of two. We also incorporated a mechanism that directly controls the length of the generated text by multiplying the probability of the end token by a factor of f_e, starting from time-step t_e. We use $f_e = 1.04$ and $t_e = 3$ for image captioning, and $f_e = 1.06$ and $t_e = 1$ for image arithmetic. On a single Titan X GPU, five beams and 512 candidate tokens can be generated in three seconds. Inference time is proportional to the number of candidates and beams.

5.1. Image Captioning

We begin by studying our zero-shot method for caption generation. Notably, we find our captions to exhibit human-like characteristics, such as generating diverse captions, reading, exploiting a wide range of external knowledge, and coping with abstract concepts. In Tab. 1, we present our results for COCO's test set [43]. Two recent baselines that use CLIP's embedding are compared to: ClipCap [50] and CLIP-VL [62]. In ClipCap, the image is encoded using CLIP and the representation is transferred and plugged as a token into a fine-tuned GPT-2. CLIP-VL incorporates spatial grid features from CLIP into a transformer network. Another method, VinVL [76] is a state-of-the-art technique.

We first consider supervised metrics, i.e., metrics requiring human references. These metrics include the BLEU [52], METEOR [5], CIDEr [66], SPICE [2], and CLIPScoreRef that we discuss below. As can be seen, our method lags in these metrics in comparison to the supervised captioning methods. Since the ground truth human annotation is obtained similarly to the training set, with the same group of annotators using similar terms, there is a clear advantage for methods trained on COCO annotations.

We next consider diversity metrics. Our vocabulary over COCO's test set is significantly larger than previous approaches (8681 vs. 2464). In addition, none of the generated sentences appear in the training set of COCO (100% on %Novel).

CLIPScore [30] is a reference-free method for evaluating relatedness between an image and its caption, using CLIP's alignment score. Evidently, our method is much better in this metric than the supervised method (87% vs. 77%). As an alternative to exact correspondence with human reference, we use CLIPScoreRef to measure the semantic distance from the references. Although supervised methods outperform our method in this score (similarity in the vocabulary and the sentence style still provide an advantage), the gap is narrower than in other supervised metrics.

Qualitative Analysis Fig. 3 compares our zero-shot approach with other baselines, demonstrating that our method can generate human-like captions, i.e., textually richer, better at image reasoning, and more effective at grounding objects. We discuss each image from left to right. First, as opposed to CLIP-VL, which assumes a toilet is in the bathroom, and VinVL, which disregards the background buildings and presumes it is on a sidewalk, our method determines it is on a rooftop. Next, our method attempts to generate the written text on a boat's side. The following image describes a flight meal as a regular tray of food with the baselines, whereas our method describes it as a flight meal. We accurately describe the next image as a bar restroom with portraits and not a bathroom. Our method and VinVL specify specific birds in the following photo (red falcons and hawks are hard to tell apart). Next, the baselines repeat the same sentence, while our method mentions an interesting mesh tile pattern. In the next photo, our method identifies a family rather than a general group. Last, our method accurately describes a room's interior, such as a bedroom with posters, and deduce that the posters depict bands. Note that the baselines' captions are generally of the same pattern, while our method generates novel sentences. Also, note that the images are taken from COCO dataset,

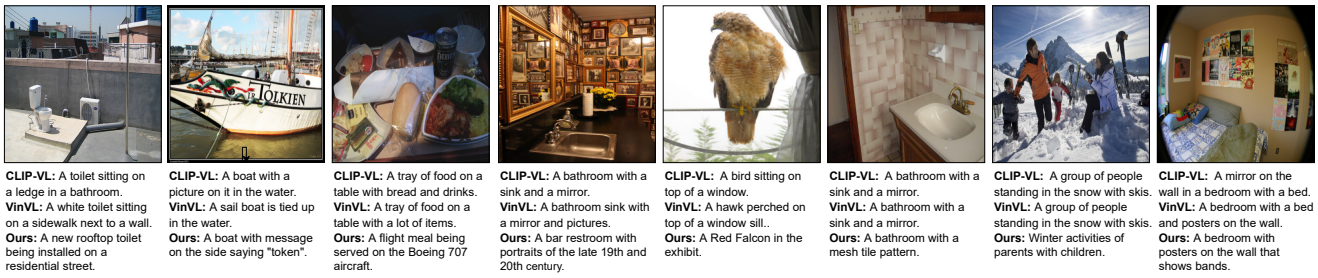

Figure 3. Examples of our zero-shot image captioning compared against supervised captioning methods.

Figure 4. Examples of OCR capabilities.

Figure 5. Examples of real-world knowledge.

which was used to fine-tune CLIP-VL and VinVL.

OCR The ability of CLIP to classify text within an image from a closed set of possible prompts is impressive [55]. We show in Fig. 4 that these capabilities can be exploited in a generative manner. To accomplish this, we change the prefix prompt we use in our method from "Image of a" to "Image of text that says." Results include impressive understanding. e.g., "The president Kennedy's death" from an image of a paper declaring it or generating "The University of Stanford" from a sign depicting its name.

External knowledge The generated captions comprise a wealth of real-world knowledge on a variety of topics. In Fig. 5 we show samples of famous people (e.g., Trump), animated shows (e.g., Simpsons), cities (e.g., Manhattan), movies (e.g., Avengers), games (e.g., Mario driving), and places (e.g., Stanford).

5.2. Visual-Semantic Arithmetic Study

We demonstrate how our method can generate text for subtraction to explain semantic directions. Next, we demonstrate that the summation operator allows guidance of the generated text through visual cues. One can then apply the above insights to solve visual analogy puzzles.

Subtraction Subtracting vectors intuitively represents a direction between the vectors. In Fig. 6 we demonstrate our method's ability to express relations through several examples. "A caricature illustration" is the result of subtracting a real photo of an airplane from a caricature. To put it another way, adding the concept "A caricature illustration" to the right image of a real plane will match the image on the left of a caricature plane. Concepts of quantity and color can also be seen, for example, a comparison of a green apple versus a red apple yields 'Red,' and vice-versa, subtracting one basketball from many basketballs results in "a bunch." Furthermore, we find directions related to a geographical area, e.g., 'Snow,' and 'Desert.' Further, a concept directly tied to day and night, and a concept of prison (i.e., 'Jailed'). It should be noted that the operator is not symmetric, and cannot always be derived textually. For instance, on the right images, the concept direction from a skateboard to a skateboard tournament can be generated as "The event." However, the direction from a skateboard tournament to a skateboard generated "schematic fossil view," which is irrelevant.

Summation Through the addition operation, the generated text can be guided through visual semantics. In Fig. 7 we show examples of guidance. On the left side, with the addition of a police officer's hat, the caption describes a man running as "A police officer...," if we add a hammer to a man, we get "The judge." On the right side, we show that a concept can be abstract. For example, the Apple company can be represented by an apple. Thus, adding an apple to a phone, results in the text "Apple's iPhone released." Additionally, a country's concept can be represented visually with flags. If Canada's flag is added to a tree, "Toronto Maple" results.

Guidance with Visual-Relations In the field of natural language processing, semantic relations have long been studied [49]. Previous efforts studied visual relations with expensive annotated language-priors [44]. With the introduction of CLIP, richer visual concepts from large-scale data became available [22]. Through visual arithmetic, we are able to exploit this richer embedding space.

In Fig. 8, we show our proposed strategy. Using subtraction, we first determine the direction. For example, the concept of leadership is represented by an image of Obama minus the American flag. With this direction in hand, we can now manipulate the case of other nations. By adding the direction to the German flag, we obtain "Angela Merkel." A different example is to examine the concept direction of CEO-to-company. With different images (e.g., Bill Gates

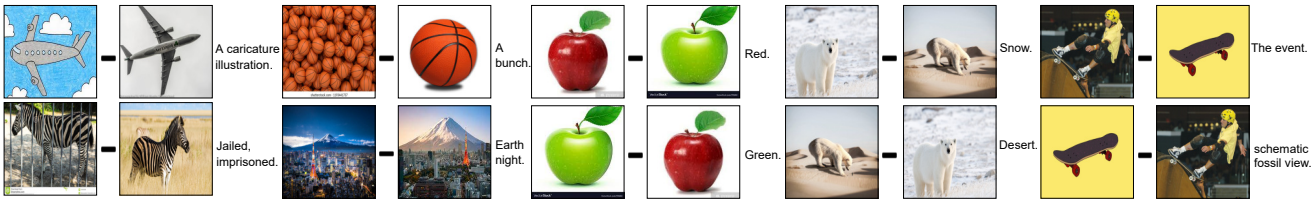

Figure 6. Examples of vector directions derived by subtracting representations from CLIP's embedding space. By generating text for the given direction, the concept is revealed.

Figure 7. Examples of caption guidance with an image through the addition operator.

and Microsoft, Jeff Bezos and Amazon), the direction can be summed to Mark Zuckerberg and Steve Jobs generating 'Facebook' and 'Apple,' respectively. On the right side, we study various interactions with country-related representations. We guide the image of a baguette to generate 'France' by taking photos of pizza and Italy and deriving the country-to-food direction.

The Visual Relations benchmark To further study the relation capabilities of our technique quantitatively, we introduce a new benchmark of visual relations, VR for short. This benchmark comprises 320 relations of the following templates: buildings→country, countries→capital, foods→country, leaders→country, and CEO→company. These were chosen because they are roughly many to one, i.e., a country has many buildings, but a building only relates to one country. The benchmark is designed to measure both the ability to model relations visually and to apply real-world knowledge to perform the task.

We constructed the benchmark through the following steps: (i) we created semantic directions by subtracting visual pairs and (ii) we then used each direction and added it to a visual element in another pair to create its corresponding text companion. As an example, we used images of ('japan,' 'sushi') to convey the direction of food→country, and then we added this direction to an image of a pizza and examined the appearance of Italy in the generated text.

We focused on single-word answers. The three evaluation metrics we find relevant to this setting are (1) BLEU-1, which measures unigram precision; (2) Recall@5, which indicates a word's appearance within the first five words generated; and (3) CLIP-score, which indicates semantic relatedness. To calculate the CLIP-score, we first add "Image of" as a prefix to the ground truth. Using CLIP's textual encoder, we then use a cosine distance. More details are provided in the supplementary material.

In Tab. 2, we show performance for each relation. While this task is challenging, our approach resulted in a significant success rate of 30% at R@5 in most relations. Note that, since the benchmark lacks multiple references, it is still limited, e.g., we mark a miss if the generated word is 'US,' while the ground truth is 'USA'. Observing the returned answers reveals that some mistakes are understandable, e.g., answering Sydney instead of Canberra or the Sinai province instead of the country Egypt. However, other cases return truncated sentences, e.g., returning 'flag' instead of a country name or returning general concepts such as "flickr image". See supplementary for a discussion. When employing the softer CLIP-Score metric, which is based on a semantic distance, a correlation of 70% is observed.

We compared our results with ClipCap [50] that encodes the image with CLIP's image encoder and uses it as an initial token for GPT-2. The method is fine-tuned based on COCO dataset. As can be seen, this method fails to retrieve the correct response, despite employing the same large-scale models as we do and performing arithmetic in the same CLIP embedding space. CLIP-VL [62] and the supervised captioning methods cannot be tested on this benchmark since it uses spatial grid features as embedding.

Multi-modal Arithmetic Our method enables multi-modal reasoning, which involves manipulating images and text simultaneously in the same embedding space. Using CLIP's textual encoder, E_{Text}. In Fig. 9, we show that a day-to-night direction can be obtained with text inputs, i.e., "image of a night," and "image of a day." The direction steers an image of breakfast to "Nighttime dinner."

6. Discussion and Limitations

The zero-shot capabilities presented by CLIP [55] pave a new path for computer vision. However, these are limited to multiclass classification. DALL-E [57] presents an impressive ability to generate images that are very different from its training images in what is termed zero-shot generation ability. However, this ability is exactly the generative task DALL-E was trained to do, only in new domains. No previous computer vision work, as far as we can ascertain, has presented a generative semantic zero-shot capability of the sort that is revolutionizing the NLP world with transformers, such as GPT-3 [6]. Our work is the first to present a generative visual-semantic work.

While the ability to rely on pre-trained models such as

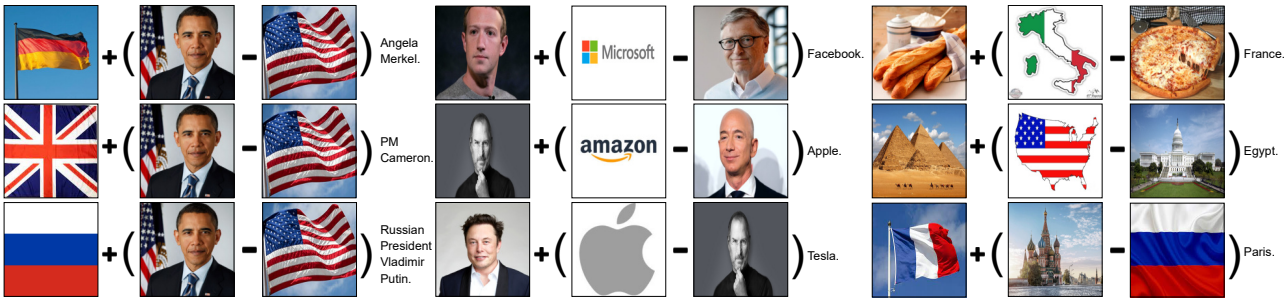

Figure 8. Image arithmetic with both summation and subtraction. For example, on the left side, by removing the American flag from Obama, a leadership direction results. The presidents of different countries are generated when the derived vector is added to their flags.

	Building → Country			**Country → Capital**			**CEO → Company**			**Food → Country**			**Leader → Country**		
Method	B@1	R@5	C-s	B@1	R@5	C-s	B@1	R@5	C-s	B@1	R@5	C-s	B@1	R@5	C-s
ClipCap [50]	0.003	0.035	0.24	0.0	0.0	0.22	0.004	0.05	0.18	0.0	0.0	0.24	0.008	0.24	0.26
Ours	**0.1**	**0.32**	**0.7**	**0.14**	**0.32**	**0.68**	**0.1**	**0.3**	**0.64**	**0.03**	**0.33**	**0.66**	**0.1**	**0.28**	**0.68**

Table 2. Comparison of our method and ClipCap baseline on our novel benchmark for visual relations. B@1 = BLEU-1, R@5 = Recall@5, C-s = CLIP-score.

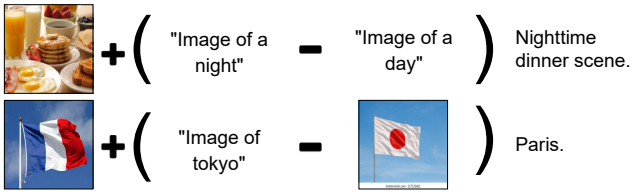

Figure 9. Our method is not only devoted to visual relations, but it also allows arithmetic between image and language.

GPT-2 [56] and CLIP allows us to achieve such new capabilities, they also highlight the uneven playing field AI has become. GPT-2 is far inferior to GPT-3 and other recent LMs in which resources far beyond the reach of most research labs are invested.

On a similar note, it is likely that combining zero-shot with supervised training would lead to a method that outperforms the baselines in all captioning metrics. However, the amount of resources currently used to train supervised captioning methods is becoming a deterring factor from pursuing this direction. For instance, UNITER uses 3645 hours of a V100 GPU [11].

The use of an LM and an image-language matching model trained on large corpora of collected data inevitably leads to biases. For example, the models we employ are clearly oriented towards Western knowledge and can recognize people, places, objects and concepts that are popular in Western media, while being much less knowledgeable about other cultures. For example, our model fails to form relations with the president of China, Xi Jinping.

7. Conclusions

The marriage between a language model and a visual-semantic matching model is a powerful union, with the potential to provide zero-shot captioning that brings together real-world variability in text, recognition abilities that are unrestricted by categories, and real-world knowledge that is embedded in the models through web-scale datasets.

We propose a zero-shot method for combining the two models, which does not involve optimizing over the weights of the models. Instead, we modify, for all layers and attention heads, the key-value pairs of the tokens generated by the language model up to each inference step.

As a captioning model, our method produces results that are less restrictive than those provided by the human annotators on the datasets used by supervised captioning methods. While this lowers the word-to-word metrics, the captions generated seem to be a good match to the image at the semantic level and exhibit real-world information. Moreover, the flexibility of using an embedding-space zero-shot method enables us to perform visual-semantic arithmetic.

We show how we can describe in words the difference between two images and how we can combine concepts from multiple images. Both are novel high-level recognition tasks. Combining these two capabilities, a powerful image analogy machine is obtained, which answers, by providing a text string, questions of the form "A is to B as C is to X" ($X \sim C + B - A$), in which A, B, and C can each be either textual or visual.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research, innovation programme (grant ERC CoG 725974). The contribution of the first author is part of a PhD thesis at Tel Aviv University.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019. 2

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 5

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. *EMNLP*, 2017. 2

[4] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Convolutional image captioning. In *CVPR*, 2018. 2

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL*, 2005. 5

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 1, 7

[7] Moitreya Chatterjee and Alexander G Schwing. Diverse and coherent paragraph generation from images. In *ECCV*, 2018. 2

[8] Hila Chefer, Sagie Benaim, Roni Paiss, and Lior Wolf. Image-based clip-guided essence transfer. *arXiv preprint arXiv:2110.12427*, 2021. 2

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[10] Yun Chen, Victor OK Li, Kyunghyun Cho, and Samuel R Bowman. A stable and effective learning strategy for trainable greedy decoding. *EMNLP*, 2018. 3

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *ECCV*, 2020. 8

[12] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020. 2

[13] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *ICLR*, 2020. 3

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 2

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019. 2

[16] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *CVPR*, 2015. 2

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[18] Qianyu Feng, Yu Wu, Hehe Fan, Chenggang Yan, Mingliang Xu, and Yi Yang. Cascaded revision network for novel object captioning. *TCSVT*, 2020. 2

[19] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *CVPR*, 2019. 2

[20] Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *NeurIPS*, 34, 2021. 2

[21] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *NeurIPS*, 2020. 2

[22] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. 4, 6

[23] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press. 1

[24] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. Stack-captioning: Coarse-to-fine learning for image captioning. In *AAAI*, 2018. 2

[25] Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK Li. Learning to translate in real-time with neural machine translation. *EACL*, 2017. 3

[26] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Zero-shot detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[27] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and geometry-aware self-attention network for image captioning. In *CVPR*, 2020. 2

[28] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016. 2

[29] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *NeurIPS*, 2019. 2

[30] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *EMNLP*, 2021. 5

[31] Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Vivo: Visual vocabulary pre-training for novel object captioning. *AAAI*, 2021. 2

[32] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *ICML*, 2017. 3

[33] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *AAAI*, 2021. 2

[34] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019. 3

[35] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. *EMNLP*, 2019. 2

[36] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017. 2

[37] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*, 2014. 2

[38] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *ICCV*, 2017. 1

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 1

[40] Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. In *ICCV*, 2019. 2

[41] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2

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

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 5

[44] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*. Springer, 2016. 6

[45] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning, corr abs/1612.01887 (2016). *CVPR*, 2017. 2

[46] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018. 2

[47] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. *AAAI*, 2021. 2

[48] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). *CoRR*, abs/1412.6632, 2014. 2

[49] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *NIPS*, 2013. 6

[50] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 5, 7, 8

[51] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*, 2020. 2

[52] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2001. 5

[53] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 1

[54] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 2

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *ICML*, 2021. 1, 2, 6, 7

[56] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1, 3, 4, 8

[57] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 1, 7

[58] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, and Marco Fumero. Clip-forge: Towards zero-shot text-to-shape generation. *arXiv preprint arXiv:2110.02624*, 2021. 2

[59] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. High-order attention models for visual question answering. *NIPS*, 2017. 2

[60] Idan Schwartz, Alexander G. Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *CVPR*, 2019. 2

[61] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. Factor graph attention. In *CVPR*, 2019. 2

[62] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 1, 2, 5, 7

[63] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. *NIPS*, 2017. 3

[64] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 2, 3

[66] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2014. 5

[67] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5753–5761, 2017. 2

[68] Sahil Verma, Michael Ernst, and Rene Just. Removing biased data to improve fairness and accuracy. *arXiv preprint arXiv:2102.03054*, 2021. 2

[69] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. corr abs/1411.4555 (2014). *ICML*, 2015. 2

[70] Liwei Wang, Alexander G Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. *NIPS*, 2017. 2

[71] 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 *International conference on machine learning*, pages 2048–2057, 2015. 2

[72] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019. 2

[73] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018. 2

[74] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792*, 2014. 1

[75] L Yu, W Zhang, J Wang, and Y Yu. Seqgan: Sequence generative adversarial nets with policy gradient. arxiv e-prints, page. *AAAI*, 2017. 3

[76] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021. 1, 2, 5

[77] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. 3