Towards Unsupervised Image Captioning with Shared Multimodal Embeddings

Iro Laina  
Technische Universität München  
iro.laina@tum.de

Christian Rupprecht  
University of Oxford  
chrisr@robots.ox.ac.uk

Nassir Navab  
Technische Universität München  
nassir.navab@tum.de

Abstract

Understanding images without explicit supervision has become an important problem in computer vision. In this paper, we address image captioning by generating language descriptions of scenes without learning from annotated pairs of images and their captions. The core component of our approach is a shared latent space that is structured by visual concepts. In this space, the two modalities should be indistinguishable. A language model is first trained to encode sentences into semantically structured embeddings. Image features that are translated into this embedding space can be decoded into descriptions through the same language model, similarly to sentence embeddings. This translation is learned from weakly paired images and text using a loss robust to noisy assignments and a conditional adversarial component. Our approach allows to exploit large text corpora outside the annotated distributions of image/caption data. Our experiments show that the proposed domain alignment learns a semantically meaningful representation which outperforms previous work.

1. Introduction

Generating natural language descriptions for images has gained attention as it aims to teach machines how humans see, understand and talk about the world. Assisting visually impaired people [23, 62] and human-robot interaction [12, 39] are some examples of the importance of image captioning. Even though it is straightforward for humans to describe the contents of a scene, machine generation of image descriptions is a challenging problem that requires compositional perception of images translated into semantically and grammatically correct sentences.

Traditionally, image captioning has been carried out using full supervision in the form of image-caption pairs, given by human annotators. Crowd-sourcing captions is a cumbersome task that requires extensive quality control and further manual cleaning. Since annotators are often paid per image, the captions tend to be short and repetitive. In addition, current captioning benchmarks [38, 49] consist of a limited number of object categories and are focused on performance under imperfect evaluation metrics. Thus, methods developed on such datasets might not be easily adopted in the wild. Nevertheless, great efforts have been made to extend captioning to out-of-domain data [3, 9, 69] or different styles beyond mere factual descriptions [22, 55].

In this work we explore unsupervised captioning, where image and language sources are independent. The unsuper-
vised setting can benefit from an almost unlimited amount of unlabeled or weakly labeled images as well as readily available large text corpora, without the need of bias-prone and costly human annotations. Although significant progress has been achieved in other unsupervised tasks [15, 34, 54, 70], unsupervised generation of image descriptions remains mostly unexplored.

The building blocks of our method are a language model and the translation of the image to the language domain. On the language side, we first learn a semantically structured embedding space, i.e., sentences describing similar visual concepts (e.g., woman and person) and similar context are encoded with similar embeddings. We then perform a weakly supervised domain alignment between image features and the learned text embeddings leveraging visual concepts in the image. This alignment allows to exploit co-occurrence statistics of visual concepts between sentences and images. For example, the words boat and water might often appear together in the language domain, similar to the fact that most images that contain a boat also contain water.

When language and images come from different sources, some weak supervisory signal is needed to align the manifold of visual concepts to the textual domain. Similar to previous work [18], we use a pre-trained object detector to generate an initial noisy alignment between the text source and visual entities that can be detected in the image.

We show that we can indeed learn to predict meaningful captions for images that extend beyond the limited capabilities of the object detector. Due to visual concept co-occurrence, the model learns to produce text descriptions including concepts that are not necessarily contained in the object detector’s fixed set of labels (e.g., beach). This shows that the alignment is meaningful and the statistics of both domains help to discover more visual concepts. Quantitatively, our unsupervised approach nearly matches the performance of some early supervised methods and outperforms previous unsupervised methods. Finally, our approach makes it possible to leverage various language sources, for instance from a different language or with a particular style — poetic (Shakespeare), funny, story-telling — that cannot be easily obtained by crowdsourcing.

2. Related Work

Fully supervised. Pioneering work in neural-based image captioning [27, 60] established the commonly used framework of a Convolutional Neural Network (CNN) image encoder, followed by a Recurrent Neural Network (RNN) language decoder. There has been significant progress improving over the standard CNN-RNN approach. Xu et al. [63] introduced the concept of attention to image captioning and, subsequently, several methods focused on attention mechanisms to visualize the grounding of words on image context and effectively guide the generation process [4, 41, 64, 68]. Noteworthy efforts also include generating video descriptions [14] or dense captions on image regions [26], exploiting additional information such as attributes [67] or visual relationships [66] and optimizing evaluation metrics [40, 50]. Other methods focus on generating diverse and natural captions with adversarial models [11, 36, 53, 61], moving beyond just factual descriptions [19, 55] or addressing gender bias [5].

Novel object captioning. Recent approaches have also explored the task of novel object captioning to exploit large-scale visual object recognition from readily available datasets, such as ImageNet [51]. Their goal is to address the limitations of conventional models in integrating new entities into image descriptions without explicit training pairs. In [45] the problem is addressed by learning from few labeled pairs for novel categories. Copying mechanisms are employed in [6, 65] to transfer knowledge from the paired data to out-of-domain objects, while [59] jointly exploits semantic information from independent images and text sources. Another approach is to produce sentence templates and fill in the slots with detected concepts [42]. Instead of training the model to handle new concepts, [2] proposes to constrain the model to handle new concepts.

Partial supervision. Recent work has further advanced the field towards generating image descriptions under more challenging settings, for example unpaired or unsupervised.

Chen et al. [9] address cross-domain captioning, where the source domain consists of image-caption pairs and the goal is to leverage unpaired data from a target domain through a critic. In [69], the cross-domain problem is addressed with a cycle objective. Similarly, unpaired data can be used to generate stylized descriptions [22, 46]. Anderson et al. [3] propose a method to complete partial sequence data, e.g., a sequence of detected visual concepts, without the need for paired image-caption datasets. Gu et al. [20] address unpaired image captioning from a different perspective, using an intermediary language where paired data is available, and then translating the captioner to the target language using parallel corpora. However, the goal of these methods is different to ours, as they typically align a target domain that contains limited paired or unpaired data with a source domain. A generic image captioner is first built from full supervision in the source domain and then adapted to a different language domain or novel object categories.

Most closely related to our work is [18] which does not require any image-sentence pairs. In this case, it is optimal to use a language domain which is rich in visual concepts. Therefore, their (and our) goal is to exploit image and language sources that are disjoint yet compatible, instead of aligning different language sources as in cross-domain approaches. Supervision comes in only through image recognition models, which are used to detect objects in the image.
Multimodal embeddings. A key component of our approach is the alignment of latent representations from two independent modalities. In unsupervised machine translation, although unimodal, [34, 35] create a shared latent space (interlingua) for both source and target languages. Kiros et al. [29] pose captioning as a translation problem and learn a multimodal embedding space that also allows them to perform vector arithmetics. Similarly, joint embedding spaces have been used in [16] for cross-modality retrieval and in [47] for video captioning. Finally, Fang et al. [17] predict visual words from images to produce caption candidates and use the similarity between images and sentences in a joint space to rank the captions.

3. Methods

An overview of our method is shown in Figure 2. The proposed approach consists of two components, a language model and a domain alignment model between images and text. The language model independently encodes samples from the language domain into a semantic-aware representation. The goal of the domain alignment is to translate image representations into the embedding space learned by the language model and decode these embeddings into meaningful image descriptions. In absence of paired image-caption data this is a challenging task.

We consider a visual domain \( I \) and an image \( I_i \in I \), represented by the set of visual entities that it encloses:

\[
V_i = \{v_k \mid k \in \mathbb{N}, 1 \leq k \leq N_i\},
\]

where \( i \) iterates over the total number of image samples and \( N_i \) is the total number of visual concepts in image \( i \).

Similarly, in the language domain \( \mathcal{L} \), a text sequence \( s_j \in \mathcal{L} \) can be described by a bag of words

\[
W_j = \{w_k \mid k \in \mathbb{N}, 1 \leq k \leq M_j\},
\]

where \( j \) enumerates sequences of length \( M_j \).

For the purpose of this work, we assume that the image and language domains are not entirely disjoint. For example, it would seem unreasonable to attempt describing natural images based on text corpora of economics. Thus, we assume a universal set of concepts \( \Omega = V \cap W \) that language and images have in common. We refer to joint concepts, such as person, as visual concepts.

3.1. Language Model

To create a basis for domain alignment, our first step is to create a meaningful textual domain. We learn an unsupervised sentence embedding by training a language model on the text corpus, following a standard sequence-to-sequence approach with maximum likelihood estimation [57]. The encoder \( f \) embeds an input sentence \( s \) into a \( d \)-dimensional latent representation which is reconstructed back into the same sentence by a decoder \( g \):

\[
f(s) = \phi, \quad g(\phi) = \hat{s}, \quad \phi \in \Phi \subseteq \mathbb{R}^d.
\]

RNNs are the most common choice for \( f \) and \( g \). Typically, language models of this structure are trained by minimizing the negative log-likelihood between \( s \) and \( \hat{s} \) per word.

A model without any constraints on the latent space would learn a grammatical and syntactic embedding. Instead, we are primarily interested in creating a representation that encodes visual semantics. This means that we have to encourage the model to learn a manifold structured by visual concepts. As we show later, our representation encodes strong semantic properties in the sense that sentences with similar contents have a low distance in the embedding space. Since our goal is image captioning, our notion of similar sentence contents stems from visual concepts—words in a sentence that have visual grounding—and their co-occurrence statistics. We impose a visual concept-based structure on the manifold of \( \phi \) with a triplet loss, defined as

\[
L_t(\phi, \phi^+, \phi^-) = \max(0, \|\phi - \phi^+\|_2^2 - \|\phi - \phi^-\|_2^2 + \lambda n)
\]

that operates on triplets of embeddings \( \phi \). The loss is minimized when the distance from an anchor embedding \( \phi \) to a positive pair \( \phi^+ \) is smaller than the distance to a negative pair \( \phi^- \) by at least a margin \( m \in \mathbb{R}^+ \).

The positive and negative pairs can be defined based on the visual concepts that exist in the sentences. For a given sentence \( s_j \) we define the set of negative pairs \( S_j^- \) as the set of sentences that do not have any concepts in common

\[
S_j^- = \{s_k \mid k \in \mathbb{N}, W_k \cap W_j = \emptyset\}.
\]

Analogously, we define the set of positive pairs \( S_j^+ \) as the set of sentences that have at least two concepts in common

\[
S_j^+ = \{s_k \mid k \in \mathbb{N}, k \neq j, |W_k \cap W_j| \geq 2\}.
\]

We ignore sentence pairs that only have one overlapping concept to reduce bad alignments. For example, since many language datasets are human-centered, every sentence involving a person would be a positive pair to each other regardless of the context. The language model’s total loss is

\[
L_{LM}(s_j) = L_{CE}(g(\phi), s_j) + \lambda_i L_t(\phi_j, \phi_j^+, \phi_j^-).
\]

During training, a positive sentence \( s_j^+ \in S_j^+ \) is sampled from a multinomial distribution with probability proportional to the number of overlapping concepts. This favors positive pairs of sentences with many similar concepts. We sample a negative sentence \( s_j^- \) uniformly from \( S_j^- \).

The triplet loss imposes a visually aware structure on the embedding space. Sentences with similar visual contents are encouraged to be close to each other, while sentences with different context will be pushed apart. This
external emphasis on structure is important, since unconstrained language models are more likely to group sentences with similar words and grammar. Intuitively, generating image descriptions relies on visual content and thus the structured embedding space is presumably a more meaningful basis for the task at hand. A comparison between the visually constrained and unconstrained embedding space can be found in the supplementary material.

3.2. Joint Image and Language Domain

We have learned an encoder that projects text into a structured embedding. The next step is to project image features into the same embedding space so that they can be similarly decoded into sentences by the decoder. To do this, we need an initial alignment between the independent image and text sources for which we rely on the visual concepts they have in common. We build a bipartite graph \( G(L, I, P) \) with images \( I_i \) and sentences \( s_j \) as nodes. The edges \( P_{i,j} \) represent weak assignments between \( I_i \) and \( s_j \), weighted by the number of overlapping concepts

\[
P_{i,j} = |V_i \cap W_j|.
\]

During training, for \( I_i \) we sample \( s_j \) with probability

\[
p(s_j \mid I_i) = P_{i,j} \left( \sum_k P_{i,k} \right)^{-1}.
\]

For sentence-image pairs without overlap \( p(s_j \mid I_i) = 0 \) and they are excluded from training. Highly visually correlated pairs will be sampled with higher probability. At this point, we have created a stochastic training set, which we could use to train a standard captioning model by sampling an image-caption pair at each iteration. Training this model with teacher forcing alone, collapses to certain caption-modes describing sets of images.

Visual concepts can be extracted from the images using any pretrained image recognition method. However, this would often result in only a limited number of categories. To lexically enrich the search space for matching sentences, we also query hyponyms of the predicted visual concepts \( V_i \), i.e. words among the text source concepts \( W_j \) that have a kind-of relationship with the predicted concepts (for example, man to person, puppy to dog).

3.3. Learning the Semantic Alignment

The initial alignment allows us to learn a mapping from images to text. We extract image features \( \psi_i \) from \( I_i \) using a standard pretrained CNN. The task is now to translate between the image feature domain \( \psi_i \in \Psi \) to the visually structured text domain \( \phi_j \in \Phi \). The stochastic alignment graph \( G \) is expected to be very noisy and full of imprecise correspondences. We thus propose a robust training scheme to exploit the underlying co-occurrence information while ignoring problematic matches. We learn the translation function \( h : \Psi \rightarrow \Phi \), where \( h \) can be a simple multilayer perceptron (MLP), using the correspondences \((s_j, I_i)\) and the following objectives.

**Robust Alignment.** If we train the alignment using a simple loss \( L_2 = \sum_j \| h(\psi_i) - \phi_j \|_2^2 \) loss the optimal mapping \( h^* \) would be the conditional average \( h^*(\psi_i) = \sum_j p(\phi_j \mid I_i) \phi_j \) which might not be an optimal or verbally rich sentence embedding as it could land between modes of the distribution.
Thus, we propose to learn the feature alignment using a robust formulation that encourages the mapping to be close to a real sentence embedding:

\[
L_R(\psi_i) = \min_{\phi_j \sim \rho(s_j | I_i)} \| h(\psi_i) - \phi_j \|^2_2.
\] (10)

Since the set of matches is very large, we approximate the loss by sampling a fixed amount \( K \) of \( \phi_j \) for each image and by computing the minimum in this subset.

**Adversarial Training.** So far, the robust alignment encourages to learn a translation \( h \) that adheres to the structure of the conceptual text embedding. However, we need to ensure that the mapping does not discard important concept information from the image feature vector. This is necessary so that the decoder can decode a caption that directly corresponds to the visual concepts in the image. To this end, we employ adversarial training using a conditional discriminator. Since adversarial training on discrete sequences is problematic [8, 56], we perform it in feature space \( \Phi \) similar to [56]. The discriminator \( D : \Phi \times \Omega \rightarrow \mathbb{R} \) is trained with a set of positive/real and a set of negative/fake examples. In our case a positive example is the concatenation of a translated feature \( h(\psi_i) \) with the one-hot encoding of the image concepts \( V_i \). A negative example analogously is the concatenation of the sampled pair’s text embedding \( \phi_j \) and the image concepts \( V_i \). Thus, the discriminator learns the correlation of image concepts and text embeddings, which in turn encourages the mapping \( h \) to encode image concepts correctly. Otherwise the discriminator can easily identify a real sentence feature from a translated image feature.

In practice, we use a WGAN-GP formulation [21] to train the discriminator \( D \) to maximize its output for fake examples and minimize it for real. When training \( h \) we thus maximize the discriminator for the translation.

\[
L_{adv} = -D(h(\psi_i), V_i)
\] (11)

**Total loss.** Our final model is trained with all three aforementioned objectives:

\[
L_{total} = \lambda_{CE} L_{CE} + \lambda_R L_R + \lambda_{adv} L_{adv},
\] (12)

where the weight factors \( \lambda_{CE}, \lambda_R, \lambda_{adv} \in \mathbb{R} \) balance the contributions of the three losses.

4. Experiments and Results

The evaluation is structured as follows. First, we present ablation experiments in an *unpaired* setting on Microsoft COCO [38] to evaluate the effect of each component of our method. Second, we report the results in the *unsupervised* setting with independent image and language sources. We experiment with Flickr30k Images [49] paired with COCO captions and COCO images paired with Google’s Conceptual Captions dataset (GCC) [52]. Finally, we show qualitative results for image descriptions with varying text sources.

**Implementation details.** We tokenize and process all natural language datasets, replacing the least frequently used words with *unk* tokens. The next step is to extract visual word synsets. We use the Visual Genome [31] object synsets as reference and look up nouns (or noun phrases) extracted by parsing each sentence with the Stanford CoreNLP toolkit [44]. This results in 1415 synsets for COCO and 3030 synsets for GCC which describe visual entities. During the semantic-aware training of the language model with Equation 4, positive and negative pairs of captions are defined using this synset vocabulary.

The encoder and decoder of the language model are implemented using Gated Recurrent Units (GRUs) [10] with 200 hidden units. The last hidden state of the encoder is projected through a linear layer into 256-d text features \( \phi \). The decoder is followed by a linear layer that maps its output into a fixed-size vocabulary vector. We use 200-d GloVe embeddings [48] as inputs to the language model.

Similar to sentence pairs, we build weak image-sentence assignments based on (visual) synsets to train the image captioneer. For richness in visual concepts, we use the OpenImages-v4 dataset [30, 33], which consists of 1.74 million images and 600 annotated object categories. Visual concepts are extracted using a Faster R-CNN detector [25] trained on OpenImages, which has been made publicly available\(^1\). Please note that we only make use of class labels and do not rely on image regions (bounding boxes) in order to keep the amount of supervision minimal. Thus, any multi-label classifier could be used instead.

The baseline for our image captioneer is based on [60] and uses image features extracted by ResNet-101 [24] pretrained on ImageNet, without finetuning. The translator \( h \) is implemented with a single-layer MLP of size 512 to map \( \psi \in \mathbb{R}^{2048} \) into \( \phi \in \mathbb{R}^{256} \).

**Training details.** We train the language model until convergence with a batch size of 64. The initial learning rates of the encoder and decoder are set to \( 10^{-4} \) and \( 10^{-3} \) respectively and \( \lambda_t = 0.1 \). When training the alignment model, we further finetune the decoder so that it adapts to the joint embedding space. We optimize using Adam [28] with a learning rate of \( 10^{-3} \) and \( \lambda_{CE} = \lambda_R = 1, \lambda_{adv} = 0.1 \).

**Evaluation metrics.** We evaluate our method with the official COCO evaluation code and report performance under the commonly used metrics, BLEU 1-4 [10], ROUGE [37], METEOR [13], CIDEr [58], SPICE [1] and WMD [32].

4.1. Unpaired Captioning

The unpaired setting on COCO allows us to evaluate the effectiveness of the proposed method and to compare to previous work [18] using the same controlled setup. This is a

\(^1\)https://github.com/tensorflow/models/tree/master/research/object_detection
Table 1. Ablation Experiments on COCO test set [27]. Image and language data are unpaired; COCO ground truth object categories are used for the initial alignment. Every component of our domain alignment model improves the performance on the captioning task.

<table>
<thead>
<tr>
<th>Component Evaluation</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abbreviation</strong></td>
<td><strong>L_{CE}</strong></td>
</tr>
<tr>
<td>Supervised baseline</td>
<td>67.4</td>
</tr>
<tr>
<td>Oracle</td>
<td>49.1</td>
</tr>
<tr>
<td>Alignment only</td>
<td>✓</td>
</tr>
<tr>
<td>MLE only</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>Joint, baseline</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>Joint, robust</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>Joint, robust (λ₁ = 0)</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>Joint, adversarial</td>
<td>✓ ✓ ✓</td>
</tr>
</tbody>
</table>

Ablation study. We evaluate the proposed components through ablation experiments (Table 1). In these experiments, we use the 80 available COCO object categories as visual concepts. We compare the following models.

Oracle: We first evaluate the weak assignments using an oracle that selects the highest probability candidate among the ground truth captions assigned to an image. This candidate has the highest overlap of visual concepts with the image. Since there can be multiple captions with equally high probability, we randomly sample and report the best out of 100 runs. This baseline scores generally low as the initial assignments are very noisy.

Alignment only: The alignment is performed by training only the mapping h of image features into the sentence manifold. We keep the decoder frozen, using the weights from the pretrained language model. The model shows understanding of the major visual concepts in the scene, meaning that relevant classes appear in the output sentence. However, the sentences are grammatically incoherent because the decoder cannot adapt to the latent space difference between the projected image features and the real sentence embeddings it was trained with. Thus, for subsequent experiments we also jointly finetune the decoder.

MLE only: The full model is trained using the weak pairs of image-captions and teacher forcing, in the standard supervised manner, but without any constraints to encourage a shared domain. The model is prone to the bias often seen in MLE models such as repeating sub-phrases.

Joint, baseline: In addition to MLE training, domain alignment is performed by minimizing the L_{2}-distance between h(ψ) and φ. This naive alignment of the two domains does not improve over the MLE-only baseline.

Joint, robust: Instead of L_{2}, the model is trained with the proposed robust alignment loss (10) which gives a significant boost in performance. We randomly sample K = 10 sentences as candidate pairs for each training image.

Joint, robust (λ₁ = 0): To evaluate the importance of the embedding space, we also train the above model against sentence embeddings that come from a language model trained only with L_{LM}=L_{CE}, i.e. without the triplet loss. It performs worse, suggesting that the semantic structure of the language model is indeed beneficial for captioning.

Joint, adversarial: The full model additionally includes adversarial training conditioned on visual concepts as categorical inputs. We observe that our unpaired model reaches performance close to its fully supervised counterpart [60] and is comparable to early work on image captioning.

The consistent improvement shows that our model is able to learn concepts beyond the initial weak assignments.

Comparison to the State of the Art. The field of image captioning without image-caption pairs has only been explored very recently. In Table 2, we compare our approach to previous methods. We follow the same unpaired setup on COCO as in [18]. We use the object detector trained on OpenImages (OID) to predict visual concepts for both creating the image-caption assignments and conditioning the discriminator during adversarial training. The reported results correspond to the predictions from our full model trained with K = 10 samples and evaluated using a beam size of 3. Our method sets a new state of the art on this problem.

Qualitative Evaluation. We show qualitative results of our full model in Figure 3, comparing captions predicted in the unpaired setting with two variants trained with different visual concept extractors (COCO and OID). We find
that both the COCO model and the OID model capture the image contents well, whereas the OID model clearly benefits from the richer object detections. For example, in the last image the COCO model produces a description about a man—potentially due to bias. This is because only person is a category in COCO, but not man or woman, and therefore there can be no gender distinction in the captions that are weakly assigned to each image. The model trained with OID concepts has the capacity to resolve such ambiguities and correctly identifies woman in the last image. We note that the object detector is only used during training (for the weak assignments and the discriminator), but not during inference. The captioner learns to extrapolate from the labeled categories of the image domain; e.g., the generated words {tracks, airport, tower, passenger, grass} are unlabeled concepts that the model inferred due to co-occurrence with labeled concepts such as train, airplane, clock, etc.

4.2. Unsupervised Captioning

When training the image captioner in an unsupervised manner, the language model is pre-trained using an external text source and all other settings remain identical. We perform two cross-domain experiments: COCO images with GCC sentences and Flickr30k images with COCO captions. Quantitative results can be seen in Table 3 for the model variants with and without adversarial training. Adversarial training consistently improves our model. Naturally, we do not expect to match the performance of the unpaired setting since a different language domain implies vocabulary, content and style that differs from the ground truth captions in COCO.

Qualitatively, we show the predicted captions of the model trained on COCO images and GCC captions in Figure 3 (denoted as GCC). When using GCC as the language domain, we find that the initial image-caption assignments are even more noisy, which leads the model to produce short and simple descriptions. However, we also see that this model has learned some interesting concepts, not present in the unpaired setting, such as the difference between a plane being on the ground or in the air.

To produce descriptions with different styles that extend beyond captioning datasets, the choice of the language domain is not trivial, as it should be rich in visual descriptions. We thus experiment with VQA-v2 [7] as the language domain, using the questions provided by the dataset as the sentence source. Instead of captioning, the model learns to ask questions about the image content (Figure 3, VQA).

4.3. Joint Embedding Visualization

Finally, to verify that our training creates a meaningful joint latent space, we visualize the t-SNE embedding [43] of both the sentences (marked with [L]) and image-projected features ([I]) in Figure 4. The overall embedding is structured by visual categories due to the constraints we impose on the model during training. Within clusters, image and text features are well mixed. This means that the model has learned a joint embedding where it is not possible to separate text form images.
5. Limitations and Discussion

Although our approach sets the state of the art in unsupervised image captioning, there are still several limitations. As mentioned before, to generate the initial assignments, the language source needs to contain sufficient visual concepts overlapping with the image domain. We believe it is possible to alleviate this problem by learning from a combination of text sources with varying contents and styles.

Another limitation is the capability of the model to extend to novel compositions and atypical scene descriptions. We observe two factors that decide the model’s behavior in this respect. First, the capabilities of the base captioner itself, i.e., unsupervised training will not solve limitations that are present even for the supervised model [60]. In our experiments, the output often collapses into caption modes that are generic enough to describe a set of images; this results in approximately 20% of the generated captions actually being unique and 16% novel captions, not found in the training set. This is on par with the findings of [60].

The second factor is the amount of discoverable visual concepts. For example, it is not possible to discover the difference between a whole pizza and a slice of pizza, when only the concept pizza is known, unless slice also appears in other context. Naturally, learning from more concepts holds the potential for more diversity. One could enrich the search space of weak assignments by including predicates in the set of known visual concepts, thus relying on relationship detection. This could greatly help in resolving ambiguities such as a person riding a bike or carrying a bike, however it goes against the idea of weak or no supervision.

6. Conclusion

We have presented a novel method to align images and text in a shared latent representation that is structured through visual concepts. Our method is minimally supervised in the sense that it requires a standard, pre-trained image recognition model to obtain initial noisy correspondences between the image and the text domain. Our robust training scheme and the adversarial learning of the translation from image features to text allows the model to successfully learn the captioning task. In our experiments we show different combinations of image and text sources and improve the state of the art in the unpaired COCO setting.

For the future we are interested in investigating several directions. One could improve the decoder architecture with typical components, such as attention, or follow a template approach to encourage novel compositions of objects. Overall, unsupervised image captioning is an upcoming research direction that is gaining traction in the community.

Acknowledgements. Christian Ruprecht is supported by ERC Stg Grant IDIU-638009.
References


[22] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. MSCap: Multi-style image captioning with unpaired stylized text. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 1, 2


