1

The Role of Image Representations in Vision to Language Tasks

Pranava Madhyastha, Josiah Wang, Lucia Specia

Dept. of Computer Science, University of Sheffield, Regent Court, 211 Portobello St., Sheffield, UK, S1 4DP
{*p.madhyastha, j.k.wang, l.specia*}*@sheffield.ac.uk*

Abstract

Tasks that require modeling of both language and visual information such as image captioning have become very popular in recent years. Most state-of-the-art approaches make use of image representations obtained from a deep neural network, which are used to generate language information in a variety of ways with end-to-end neural network-based models. However, it is not clear how different image representations contribute to language generation tasks. In this paper, we probe the representational contribution of the image features in an end-to-end neural modeling framework and study the properties of different types of image representations. We focus on two popular vision to language problems: the task of image captioning and the task of multimodal machine translation. Our analysis provides interesting insights into the representational properties and suggests that end-to-end approaches implicitly learn a visual-semantic subspace and exploit the subspace to generate captions.

1 Introduction

There has been a substantial interest in multimodal tasks that combine language and vision. One such a task is Image Captioning (IC) where given an image the goal is to generate a caption that describes it (Vinyals *et al.*, 2015; Karpathy & Fei-Fei, 2015; Kiros *et al.*, 2014). This interest has driven the community to create a series of datasets, including IAPR-TC12 (Grubinger *et al.*, 2006), UIUC PASCAL Sentences and Flickr8k (Rashtchian *et al.*, 2010), Flickr30k (Young *et al.*, 2014) and MSCOCO (Chen *et al.*, 2015), the largest of them all. This has also led to the very popular MSCOCO captioning challenges. The success in IC has inspired other, more advanced, vision to language problems, including Visual Question Answering (VQA) (Antol *et al.*, 2015) and Multimodal Machine Translation (MMT) (Specia *et al.*, 2016; Elliott *et al.*, 2017).

Recent advances in deep learning models in the area of sequence modeling using recurrent neural networks (RNN) have led to highly effective ways of learning sequential tasks (Elman, 1990). End-to-end deep neural models achieve impressive results for various tasks including language modeling (Mikolov *et al.*, 2010) and machine translation (Bahdanau *et al.*, 2015). For IC, most state-of-the-art models condition a deep recurrent sequence generator (i.e., an RNN) on some image information. The image information is usually the penultimate layer of a Convolutional Neural Network (CNN) that has been

pre-trained for object classification (Karpathy & Fei-Fei, 2015; Vinyals *et al.*, 2015). Alternatively, other layers in the network are used along with attention mechanisms on these representations to condition the RNN-based generator (Kiros *et al.*, 2014; Xu *et al.*, 2015; Wu *et al.*, 2016). The success obtained in these tasks comes to some surprise given the differences between the representational spaces of image embeddings and the language in RNN-based models. End-to-end deep neural IC methods are able to generate captions without resorting to higher-level semantic mappings of the image space into the language space. More recent work has also investigated representations of the image in the form of attributes, such as the objects potentially appearing in it, using class-based probabilistic distributions (Yao *et al.*, 2017). These methods achieve even better results on standard test sets for the tasks of IC and VQA (Wu *et al.*, 2016). In MMT, the results are less conclusive. This raises interesting questions about the informativeness of different types of representations, in particular, low versus high-level information in the context of vision to language tasks. A sparse, attribute-level representation is indicative of the presence of a pre-defined, limited number of attributes (often objects) given an image. On the other hand, dense, low- or mid-level or the CNN activation-based image representations are expected to capture more details in the images, such as abstract scene information.

Previous work utilizes several types of image representations coupled with different ways to use them in vision to language tasks. However, it is not clear what the representational contribution of these different types of image information is and why different representations lead to certain words being generated over others. In this work, we study the influence of different types of image information in a controlled setup and empirically probe the informativeness of the image representations. Our main **contributions** are:

- We study the effect of different image level representational features in the context of end-to-end IC and MMT systems.
- We show that end-to-end models conditioned on image representations mostly perform image matching in a common image-text space to generate sentences.
- We show that a low-dimensional, sparse and interpretable vector also performs competitively with higher-dimensional CNN image embeddings, suggesting that such low-dimensional features may be sufficient to generate sentences in the visual-semantic subspace.

2 Background and Related Work

In this section, we first describe various approaches used to tackle IC and MMT tasks (Sections 2.1 and 2.2 respectively). We then describe recent efforts in exploring different representations for vision to language tasks that provide some context for our study (Section 2.3).

2.1 Image Captioning Approaches

Approaches for IC can be categorised into three primary groups: (i) pipelined approaches; (ii) retrieval approaches; (iii) end-to-end approaches.
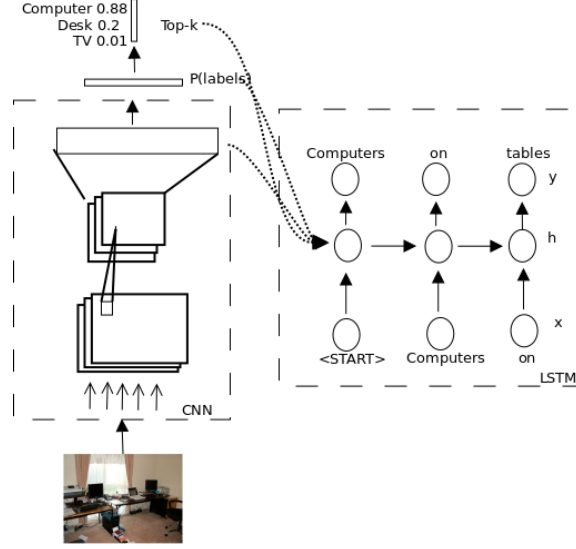
Fig. 1: RNN conditioned on different types of image representations: (a) penultimate layer; (b) posterior over object class labels; and (c) averaged word representations for the top-k object classes.

Pipelined approaches. We call early work on IC 'pipelined' as it follows a sequence of steps: first, object categories are explicitly detected with visual object detectors; then the output of such detectors is used as input to generate image descriptions through a generative model, such as template filling (Yao *et al.*, 2010; Kulkarni *et al.*, 2011; Yang *et al.*, 2011; Li *et al.*, 2011; Mitchell *et al.*, 2012; Elliott & de Vries, 2015), combining phrases from a corpus (Li *et al.*, 2011), generating trees (Mitchell *et al.*, 2012) or learning a statistical language model (Fang *et al.*, 2015). Such methods are capable of generating captions not seen at training time, although their performance depends on the quality of the visual detectors, whose outputs form the input 'representation' to the caption generator.

Retrieval approaches. Retrieval approaches to IC retrieve existing captions from the training set or an external dataset. These methods include projecting images and captions onto a common representation space (Farhadi *et al.*, 2010; Hodosh *et al.*, 2013; Socher *et al.*, 2014) and utilizing some image similarity measure (Ordonez *et al.*, 2011) among other methods. For example, Hodosh et al. (2013) use Kernel Canonical Correlation Analysis to project images and their captions into a joint representation space, in which images and captions can be related and ranked to perform illustration and annotation tasks. Such retrieval methods produce image captions that are fluent and expressive (since they are 'copied' from human-authored captions in the training set) but cannot produce novel captions. Work towards generating novel captions retrieves and combines existing text fragments (Kuznetsova *et al.*, 2012; Kuznetsova *et al.*, 2014) or prunes irrelevant fragments for better generalization (Kuznetsova *et al.*, 2013). The resulting captions, however, may still be irrelevant to the image content. On the image side, such methods mainly use a global im-

age representation (e.g., the penultimate layer of a CNN) or an intermediate representation such as a semantic tuple.

End-to-End approaches. Finally, end-to-end, deep neural network-based approaches are currently the most popular method for IC, yielding state-of-the-art results. These approaches were inspired by the success shown in transferring image representations to other tasks (Razavian *et al.*, 2014) using simple transfer learning approaches. End-to-end methods will be discussed in more detail in Section 3. In general, such approaches extract image-related features using a CNN, which are then fed to an RNN caption generator. A popular and simple approach to condition the RNN on the image representation is by initializing the *start* state of the RNN with the image encoding (Karpathy & Fei-Fei, 2015; Vinyals *et al.*, 2015) as shown in Figure 1. The CNN model used in most state-of-the-art approaches for IC (and MMT) is based on a classification model trained to optimally perform on an object classification task. The visual representation obtained as the activations of the penultimate layer have been shown previously to generalize to other tasks in the framework of transfer learning (Donahue *et al.*, 2014). Most previous approaches use pre-trained deep CNN networks, such as VGGNet (Karpathy & Fei-Fei, 2015), Inception CNN (Vinyals *et al.*, 2015) and ResNet (Yao *et al.*, 2017), to obtain an image representation that is fed into a continuous sequence generator. Attention mechanisms have also been used. For example, Xu et al. (2015) learns an IC model that attends to the output of a convolutional layer of a CNN.

Other ways of inducing representations in end-to-end approaches include attribute-level information. These correspond to the class-based predictions of the image network, i.e. the posterior probability distribution on a pre-defined set of classes that can correspond to objects in the image as shown in Figure 1. Wu et al. (2016) further fine-tune the pre-trained image network on a new label set. This fine-tuning helps the image network predict classes that correspond to the expected vocabulary.

Image captions generated by end-to-end systems can be novel to a certain extent depending on search configurations, e.g. the beam size used during decoding. In these approaches, the proportion of novel descriptions has been reported to be between 30% to 50% for optimally trained systems (Devlin *et al.*, 2015; Vinyals *et al.*, 2016; Karpathy, 2016). The number of unique captions generated by such systems has also been reported to be approximately 30%. Humans, in contrast, rarely repeat descriptions, having a rate of 95%–99% unique descriptions reported for the MSCOCO dataset (Devlin *et al.*, 2015; Karpathy, 2016). End-to-end systems also require a lot of parallel corpora (images with captions) for training, making it hard to adapt to different languages, styles or domains. Thus, end-to-end systems seem to predominantly 'memorize' parallel corpora, making it seemingly more like a 'retrieval machine' rather than genuinely generating image descriptions as in older pipelined approaches.

We refer readers to Bernardi et al. (2016) for an in-depth discussion on various image captioning approaches.
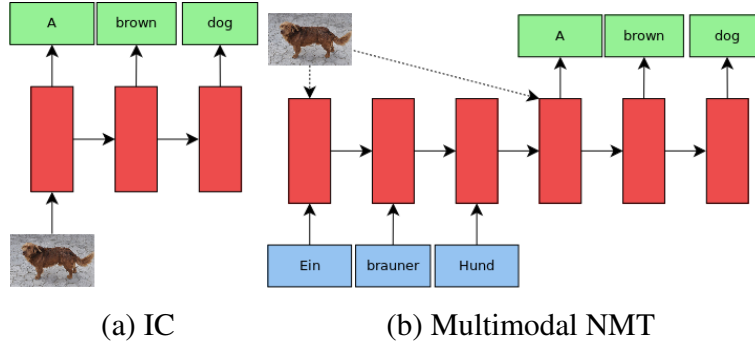
(a) IC   (b) Multimodal NMT

Fig. 2: Typical architecture of IC and MMT systems. In (a), the input image is encoded as a vector, and a description is decoded using an RNN. In (b), the source sentence encoding is used as decoder input, and the image embedding as input to either (or both) the source encoder or target decoder.

2.2 Multimodal Machine Translation Approaches

The task of MMT is closely related to that of IC. Most existing work focuses on end-to-end approaches, with an additional RNN used to encode the source sentence to produce a sequence of encoded vectors. Figure 2 illustrates the differences between typical IC and MMT architectures. In MMT, the visual information can be used to condition the source RNN, the target RNN, or both (Elliott *et al.*, 2015). Most existing work obtain the best results by combining the penultimate layer of the CNN (via concatenation, summation, etc.) with the final state of the source sentence representation and using it to initialize the target RNN (Caglayan *et al.*, 2016; Calixto *et al.*, 2016; Huang *et al.*, 2016).

Recent work also explores an attention mechanism where they use lower level CNN features of the images, such as a convolutional layer, and condition the source and the target sentences on the image features (Calixto *et al.*, 2016; Calixto *et al.*, 2017). The intuition here is that the lower-level CNN features capture information about different areas of the images and an attention mechanism could learn to attend to specific regions while both encoding the source and decoding the target sentence.

Alternative approaches rely on pre-generated candidate translations for each source sentence from a text-only MT model, which are then reranked based on visual information (Shah *et al.*, 2016), or use image information by pivoting on it to find relevant captions in external corpora (Hitschler *et al.*, 2016). Approaches that exploit multi-task learning to jointly model how to translate and learn visually grounded representations showed promising results (Elliott & Kádár, 2017).

2.3 Studying Visual Representations

Recent work in analyzing multimodal representations include (Devlin *et al.*, 2015; van Miltenburg & Elliott, 2017), which focus on linguistic regularities in the generated captions. They are interested in comparing different IC architectures and the properties of the

produced captions. In contrast, our work focuses on studying *visual* representations and their impact in vision to language tasks.

Focusing on MMT, Lala et al. (2017) show that, given reliable image information in the form of captions, an ideal MMT system would be able to significantly benefit and obtain better translations.

Vinyals et al. (2016) and Karpathy et al. (2016) present an analysis of lexical and syntactic properties of the generated captions. They conclude that almost 80% of the time the best caption for an image in the validation or test sets of MSCOCO can be retrieved from its training set, and that beam size often dictates the diversity in the output captions. Lebret et al. (2015) also analyzed the syntax of image captions in Flickr30k and MSCOCO and found that they comprise a simple and predictable structure.

The MSCOCO shared task (Chen *et al.*, 2015) showed that participating systems using variants of retrieval-based approaches (Devlin *et al.*, 2015; Kolář *et al.*, 2015) performed competitively with end-to-end approaches. Recent work seems to suggest that, in the end-to-end learning framework, using posterior distributions over a refined set of object classes (relevant to captions) performs better than using lower level dense image representations (Wu *et al.*, 2016; You *et al.*, 2016). Vinyals et al. (2016) note that using a better image network (a network that performs better on the image classification task) results in improvements in the generated captions.

In this paper, we concentrate on the *image* side of image captioning, and systematically investigate the contribution of different types of visual representations in these tasks and study plausible reasons that drive the language generation component. We focus on the currently dominant end-to-end approaches, which represent the state-of-the-art for both IC and MMT. We acknowledge that there might be other types of approaches, e.g. Fang et al. (2015) use different architectures and also achieve strong performance, but studying these is left for future work.

3 Model Setting

We base our IC implementation on a simple end-to-end approach by Karpathy & Fei-Fei (2015), and consider most state-of-the-art systems as predominantly variants of this architecture. We use the Long Short-Term Memory (LSTM) RNN (Hochreiter & Schmidhuber, 1997; Chung *et al.*, 2014) as our generative network, as described in Zaremba et al. (2014) for IC.

In order to use the image information, we first perform a linear projection of the image representation followed by a non-linearity as shown below:

$$Im_{feat} = \sigma(W \cdot I_m)$$

Here, $I_m \in \mathcal{R}^d$ is the d-dimensional initial image representation, $W \in \mathcal{R}^{d \times m}$ is the linear transformation matrix, σ is the non-linearity. We use exponential linear units as the non-linearity (Clevert *et al.*, 2015) since it is faster to compute. Following Vinyals et al. (2015), we initialize the LSTM generative sequence model with the projected image information.

For MMT, we first build an attention-based, encoder-decoder framework as described in Luong et al. (2015). We explore two approaches to use image information: (i) conditioning

the encoder on image information; (ii) conditioning the decoder on image information. Both (i) and (ii) are similar to the afore-described approach for IC.

The sentence generator is trained to generate sentences conditioned on the image representation (IC and MMT), and also on the source sentence representation for MMT. This is done by using the cross-entropy loss. That is, the sentence-level loss corresponds to the sum of the negative log likelihood of the correct word at each time step. For IC:

$$\Pr(S|Im_{feat};\theta) = \sum_t \log(\Pr(w_t|w_{t-1}..w_0; Im_{feat})) \tag{1}$$

where $\Pr(S|Im_{feat};\theta)$ is the sentence level loss conditioned on the image features Im_{feat} and $\Pr(w_t)$ is the probability of the word w_t at time step t.

For MMT, given a source sentence F and the image features Im_{feat}, we obtain the negative log-likelihood of the target sentence E as:

$$\Pr(E|F, Im_{feat};\theta) = \sum_i \log(\Pr(w_t|w_{t-1}..w_0; F, Im_{feat})) \tag{2}$$

where $\Pr(E|F, Im_{feat};\theta)$ is now conditioned on both the source sentence F and the image features Im_{feat} and w_t are words corresponding to the sentence in the target language.

The standard maximum likelihood objective is used to train the model, with teacher forcing as described in Sutskever et al. (2014) where the correct word information is fed to the next state in the LSTM. Inference is usually done using approximate techniques like beam search and sampling methods (Karpathy & Fei-Fei, 2015; Vinyals *et al.*, 2015). In this paper, as we are mainly interested in studying the effect of different image representations, we focus on the language output that the models can most confidently produce. Therefore, in order to isolate any other variables from the experiments, we generate captions using a greedy $\arg\max$ based approach, i.e. no beam search.

4 Image Representations

Various representations are explored in this paper to study the representational contribution of images for both IC and MMT. We first provide an overview of the various pre-trained image networks used to obtain image features (Section 4.1), which are then used to form image representations for IC (Section 4.2) and MMT (Section 4.3).

4.1 Pre-trained Image Networks

In computer vision, CNNs became the de facto choice for image representations after the successful performance of the AlexNet CNN (Krizhevsky *et al.*, 2012) in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC 2012) (Russakovsky *et al.*, 2015). Such networks are trained on the ILSVRC dataset for object classification, i.e. classifying images into a set of 1,000 pre-defined categories or synsets ("is this an image of a *cat*?"). Intermediate layers of the CNN are also often extracted and used as off-the-shelf features for various other vision tasks (Donahue *et al.*, 2014; Razavian *et al.*, 2014). For IC and MMT, it is worth noting that the object categories may not be directly relevant to the captions and vice versa (the captions may mention concepts that are not covered by the

1,000 categories). We explore the following two CNNs, both pre-trained on the ILSVRC dataset:

VGG19: VGGNet (Simonyan & Zisserman, 2015) achieved a top-5 accuracy of 92.7% in the ILSVRC 2014 challenge, making it among the top two best performing networks at the time. VGGNet is found to generalize well to different datasets and tasks, and is thus still widely used for different tasks. We use the pre-trained 19-layer version of VGGNet, which is reported to give slightly better performance in object classification over the 16-layer version, at the expense of being more complex.

ResNet152: ResNet (He *et al.*, 2016) reported a top-5 classification accuracy of 97.4% in the ILSVRC 2015 challenge, a significant improvement over VGGNet. The improvement resulted from drastically increasing the number of layers to 152, compared to VGGNet's 19. We also explore using the output of the pre-trained 152-layer version of ResNet for IC and MMT to investigate whether the improvement in classification accuracy on ILSVRC helps with downstream vision to language tasks.

We also explore two other variants of ResNet152:

Places365-ResNet152: Zhou et al. (2014) trained a CNN on the Places2 dataset (Zhou *et al.*, 2017) to classify 365 scene categories (*sky*, *baseball stadium*, etc.). We investigate whether these networks predicting scene-specific categories are useful for IC, despite not predicting object-specific categories. We experiment with ResNet152 pre-trained solely on the Places2 dataset. Similar to the 1,000 ILSVRC categories, the scene categories may not be relevant to the captions, and some scenes mentioned in the captions may not exist in the 365 scene categories.

Hybrid1365-ResNet152: Zhou et al. (2014) also proposed training a CNN on the concatenation of both ILSVRC and Places2 datasets, thus predicting both object and scene categories (1,365 classes). Therefore, we examine whether such a network combining both types of information can be helpful for vision to language tasks. This network is again based on the ResNet512 architecture.

4.2 Image Representations for IC

We now describe different representations explored for the task of IC. These include a lower-bound baseline (Section 4.2.1), representations derived from image classification (Section 4.2.2), and representations derived from object detectors (Section 4.2.3).

4.2.1 Lower-bound representation

Random: We condition the LSTM on a 300-dimensional vector containing random values sampled uniformly between $[0, 1)^1$. This represents a worst case image feature and provides an artificial lower bound.

4.2.2 Representations from image-level classification

We explore various representations derived from pre-trained CNNs (Section 4.1):

Penultimate layer (Penultimate): Most previous attempts for IC use the output of the penultimate layer of a CNN pre-trained on the ILSVRC data. Previous work motivates using 'off-the-shelf' feature extractors in the framework of transfer learning (Razavian *et al.*, 2014; Donahue *et al.*, 2014). Such features have been often applied to IC (Mao *et al.*, 2015; Karpathy & Fei-Fei, 2015; Gao *et al.*, 2015; Vinyals *et al.*, 2015; Donahue *et al.*, 2015) and have been shown to produce state-of-the-art results. Therefore, for each image, we extract the fc7 layer of *VGG19* ($4096D$) and the pool5 layer for the *ResNet152* variants ($2048D$) .

Class prediction vector (Softmax): We investigate higher-level image representations, where each element in the vector is an estimated posterior probability of object categories. As previously noted, the categories may not directly correspond to the captions in the dataset. While there are alternative methods that fine-tune the image network on a new set of object classes extracted in ways that are directly relevant to the captions (Fang *et al.*, 2015; Wu *et al.*, 2016; Yao *et al.*, 2017) we study the impact of off-the-shelf prediction vectors on the IC task. The intuition is that category predictions from pre-trained CNN classifiers may also be beneficial for IC, alongside the standard approach of using mid-level features from the penultimate layer. Therefore, for each image, we use the predicted category posterior distributions of VGG19 and ResNet152 (1000 *object* categories), Places365-ResNet152 (365 *scene* categories), and Hybrid-ResNet152 (1365 *object* and *scene* categories).

Object class word embeddings (Top-k): Here we experiment with a method that utilizes the averaged word representations of top-k predicted object classes. We first obtain *Softmax* predictions using *ResNet152* for 1000 object categories (synsets) per image. We then select the objects that have a posterior probability score $> 5\%$ and use the 300-dimensional pre-trained word2vec (Mikolov *et al.*, 2013) representations[2] to obtain the averaged vector over all top object categories. This is motivated by the central observation that averaged word embeddings can represent semantic-level properties and are useful for classification tasks (Arora *et al.*, 2017).

---

[1] We also tried using 1,000-dimensions, but it yielded slightly poorer results.
[2] `https://code.google.com/archive/p/word2vec/`

4.2.3 Representations from object-level detections

We also explore representing images using information from object *detectors* that identify *instances* of object categories present in an image, rather than a global, image-level classification as described earlier. The output of visual detectors can help form a more interpretable and informative image representation:

- *Ground truth (Gold)* region annotations for instances of 80 pre-defined categories provided with MSCOCO, the dataset we use for the IC experiments. It is worth noting that these were annotated independently of the image captions, i.e. people writing the captions had no knowledge of the 80 categories and the annotations (and vice versa). As such, there is no direct correspondence between the region annotations and image captions.

- The state-of-the-art object detector *YOLO* (Redmon & Farhadi, 2017) pre-trained on MSCOCO for 80 categories (YOLO-Coco), or pre-trained on MSCOCO and ILSVRC for 9000 categories (YOLO-9k) in a weakly supervised fashion (bounding boxes surrounding object instances are not provided).

We explore several representations derived from instance-level object class annotations/detectors above:

Bag of objects (BOO): We represent each image as a sparse *bag of objects* vector, where each element represents the frequency of occurrence for each object category in the image (Counts). We also explore an alternative representation where we only encode the presence or absence of the object category regardless of its frequency (Binary), to determine whether or not it is important to encode object counts in the image. These representations help us examine the importance of explicit object categories and in a sense interactions between object categories (*dog* and *ball*) in the image representation. We investigate whether such a sparse and high-level BOO representation is helpful for IC. It is also worth noting that BOO is different from the Softmax representation above as it encodes the *number* of object occurrences, not the *confidence* of class predictions at image level. We compare BOO representations derived from the *Gold* annotations (Gold-Binary and Gold-Counts) and both YOLO-Coco and YOLO-9k detectors (*Counts* only).

Pseudo-random vectors: To further probe the capacity of IC models to make use of image representations, we experiment with noisy vectors that contain object-level information. More specifically, we examine a type of representation where *similar objects are represented using similar random vectors*. We then form the representation of the image from BOO Gold-Counts and BOO Gold-Binary; formally, $Im_{feat} = \sum_{o \in \text{Objects}} f \times \phi_o$, where $\phi_o \in \mathcal{R}^d$ is an object-specific random vector and f is a scalar representing counts of the object category. We call these *pseudo-random* vectors. In the case of Pseudo-random-Counts, f is the frequency counts from Gold-Counts. In the case of Pseudo-random-Binary, f is either 0 or 1 based on Gold-Binary. We use $d = 120$. We investigate whether these seemingly random representations (but which have a latent structure) can generate reasonable captions.

4.3 Image Representations for MMT

Based on the observations from our experiments for IC, we explore the following image features for MMT:

Penultimate layer (Penultimate): As with previous successful approaches to MMT (Elliott *et al.*, 2015; Huang *et al.*, 2016; Libovický *et al.*, 2016), we use image information obtained from the penultimate layer of a pre-trained image network. Since we observed that *ResNet152*-based representations were slightly better for IC, we only use *ResNet152* pretrained on object classification for MMT, with representations from the penultimate layer (Pool5) of the network.

Class prediction vector (Softmax): As in IC, we also use the posterior distribution from *ResNet152* (1,000 object categories) as image information.

5 Experiments and Results

To study the efficacy of vision to language models and understand the contribution of image information, we perform a series of experiments on standard datasets. We explore end-to-end approaches to IC and MMT, and make our source code and models available for replicability.

5.1 Datasets

IC: We use the most widely used evaluation setup for IC, i.e. MSCOCO (Chen *et al.*, 2015). The dataset consists of 82,783 images for training, with five captions per image, thus totaling 413,915 captions in total. The validation set consists of 40,504 images and 202,520 captions. We perform model selection on a 5000-image development set and report the results on a 5000-image test set using standard, publicly available[3] splits of the MSCOCO validation dataset as in previous work (Karpathy & Fei-Fei, 2015).

Details about the collection of the images and captions can be found in (Chen *et al.*, 2015). While other image captioning datasets exist (Grubinger *et al.*, 2006; Rashtchian *et al.*, 2010; Young *et al.*, 2014), we focus on MSCOCO as it is more recent and has been extensively used and evaluated in an open platform[4]. More information on different image captioning or image description datasets can be found in (Ferraro *et al.*, 2015).

MMT: We use the Multi30k (Elliott *et al.*, 2016) English-German (en-de) MMT dataset which was released as part of the WMT 2016 shared task on MMT (Specia *et al.*, 2016). The dataset consists of English-German sentence pairs, where the English sentence is a caption belonging to the Flickr30k dataset (Young *et al.*, 2014) and the corresponding German sentence is a translation of this description professionally crafted. We also experiment with using the same data and flipping the translation direction, i.e. with a German-English

---

[3] http://cs.stanford.edu/people/karpathy/deepimagesent
[4] http://cocodataset.org/

(de-en) dataset. This dataset is reasonably small, containing 29K sentence pairs for training, 1K for development, and 1K for test. As in most datasets derived from IC tasks, sentences are very short: on average 11.9 tokens for English, and 11.1 tokens for German.

5.2 Evaluation Metrics

We evaluated system outputs using standard metrics for IC and MMT.

IC: The most common metrics for IC are BLEU (Papineni *et al.*, 2002), Meteor (Denkowski & Lavie, 2014) and CIDEr (Vedantam *et al.*, 2015). All of these metrics are based on some form of n-gram overlap between the system output and the reference captions (i.e. no image information is used). BLEU is computed from 1-gram to 4-gram precision scores (B-1 ... B-4); as n increases (longer phrases) there will be less chances of an n-gram match, resulting in a decrease in the overall score from B-1 to B-4. Meteor is an f-measure based metric that finds the optimal alignment between chunks of matched text and can incorporate semantic knowledge by allowing terms to be matched to stemmed words, synonyms and paraphrases, if such resources are available for the target language. CIDEr was developed specifically for image captioning, and measures the average cosine similarity between a generated caption and a reference, each represented as TF-IDF weighted bag of n-grams. We compare each system generated caption against five reference captions. We used the publicly available *cocoeval* script for evaluation[5]. Note that there are inherent weaknesses with these automatic metrics as they often do not correlate well with human judgements (Elliott & Keller, 2014; Kilickaya *et al.*, 2017; Anderson *et al.*, 2016). This is also reflected in the official MSCOCO metrics based on human judgements[6]. Other metrics have emerged in an attempt to address this issue (Anderson *et al.*, 2016), but they have not been widely adopted.

MMT: We use the official metrics of the WMT16 MMT task – 4-gram BLEU and Meteor – computed using the publicly available *multeval* script[7]. Each generated caption is computed against one reference (human) translation. These are the mostly widely used metrics by the machine translation community for translation evaluation.

5.3 Model Settings and Hyperparameters

IC: We use a 2-layer LSTM with 128-dimensional word embeddings and 256-dimensional hidden dimensions.

MMT: We use a single hidden layer encoder and decoder both with 128-dimensional word embeddings and 256-dimensional hidden dimensions. We train with dropout set to 0.3 for the RNNs.

For both IC and MMT, as training vocabulary we retain only words that appear at least twice.

---

[5] https://github.com/pdollar/coco
[6] http://cocodataset.org/#captions-leaderboard
[7] https://github.com/jhclark/multeval

5.4 Results

	Representation	B-1	B-2	B-3	B-4	M	C
	Random	0.48	0.24	0.11	0.07	0.11	0.07
Softmax	VGG19	0.62	0.43	0.29	0.19	0.20	0.61
	ResNet152	0.62	0.43	0.29	0.19	0.20	0.62
	Places365-ResNet152	0.60	0.41	0.28	0.19	0.19	0.56
	Hybrid1365-ResNet152	0.60	0.41	0.27	0.18	0.19	0.60
Penultimate	VGG19 (fc7)	0.65	0.46	0.32	0.22	0.21	0.69
	ResNet152 (Pool5)	0.66	0.48	0.33	0.23	0.22	0.74
	Places365-ResNet152	0.61	0.41	0.27	0.19	0.19	0.55
	Hybrid1365-ResNet152	0.65	0.46	0.32	0.23	0.22	0.72
Embeddings	Top-k	0.62	0.42	0.28	0.19	0.20	0.63
BOO	Gold-Binary	0.65	0.47	0.32	0.22	0.22	0.75
	Gold-Counts	0.66	0.48	0.33	0.23	0.22	0.80
	YOLO-Coco	0.65	0.46	0.32	0.22	0.22	0.75
	YOLO-9k	0.64	0.44	0.30	0.20	0.20	0.67
Pseudo-random	Pseudo-random-Binary	0.65	0.47	0.33	0.22	0.22	0.74
	Pseudo-random-Counts	0.65	0.46	0.31	0.20	0.21	0.78

Table 1: Results on the MSCOCO test split for IC, where we vary only the image representation and keep other parameters constant. The captions are generated with $beam = 1$
.

5.4.1 Image Captioning

We first report results of IC on MSCOCO in Table 1, where the IC model (Section 3) is conditioned on the various image representations described in Section 4. As expected, using random image embeddings clearly does not provide any useful information and performs poorly. The *Softmax* representations with similar sets of object classes (*VGG19*, *ResNet152*, and *Hybrid1365-ResNet152*) have very similar performance. However, the *Places365-ResNet* representations perform worse. We note that the posterior distribution may not directly correspond to captions as there are many words and concepts that are not contained in the set of object classes. Our results differ from those by (Wu *et al.*, 2016; Yao *et al.*, 2017; Fang *et al.*, 2015) where the object classes have been fine-tuned to correspond directly to the caption vocabulary. We posit that the degradation in performance is due to spurious probability distributions over object classes for similar looking images.

The performance of the *Pool5* image representations shows a similar trend for *VGG19*, *ResNet152*, and *Hybrid1365-ResNet152*, with *ResNet152* showing slightly better scores. Once again, the *Places365-ResNet* representation performs worse. The representations from the image network trained on object classes is probably able to capture more fine-grained image details from the images, whereas the image network trained with scene-based classes captures more coarse-grained information.

The performance of the averaged top-k word embeddings is similar to that of the *Softmax* representation. This is interesting, since the averaged word representational information is mostly noisy: we combine top-k synset-level information into one single vector. However, it still performs competitively.

The performance of the *Bag of Objects (BOO)* sparse 80-dimensional annotation vector is better than all other image representations, if we consider the CIDEr scores. This is despite the fact that the annotations may not directly correspond to the semantic information in the image or the captions. The sparse representational information is indicative of the presence of only a subset of potentially useful objects. We notice a marked difference with *Binary* and *Count*-based representations. This takes us back to the motivation that image captioning ideally requires information about objects, as well as interactions between objects, with attribute-level information such as number. Although our representation is really sparse on the object interactions, it captures the basic concept of the presence of more than one object of the same kind, and thus provides additional information. A similar trend is observed by Yin & Ordonez (2017), although in their models they further try to learn interactions using another RNN for encoding objects.

Using objects predicted with *YOLO-Coco* performs better than using objects predicted with *YOLO-9k*. This is expected as *YOLO-Coco* was trained on the same dataset hence obtaining better object proposals. With *YOLO-9k*, a significant number of objects were predicted for the test images that had not been seen in the training set (around 20%).

The most surprising result is the performance of the pseudo-random vectors. Both the *pseudo-random-Binary* and the *pseudo-random-Count* vectors perform almost as well as the *Gold* objects. This suggests that the RNN is able to isolate the noise and learn some form of a common 'visual-semantic' subspace.

5.4.2 Multimodal Machine Translation

Model	en-de		de-en	
	BLEU	Meteor	BLEU	Meteor
Pool5-enc	32.9	51.3	36.5	35.1
Pool5-dec	32.3	50.4	37.6	35.6
Softmax-enc	32.7	50.8	37.0	35.1
Softmax-dec	33.0	51.0	36.3	34.2
Caglayan et al. (2016)[†]	34.1	53.2	–	–

Table 2: Results for en-de and de-en MMT test sets. † are best WMT16 results taken from Caglayan et al. (2016) which are generated based on a combination of statistical machine translation and re-scoring.

For MMT, we summarize the results in Table 2. Our models do not reach the performance of the top system at WMT16, but such a system is actually a combination of multiple strategies. We compare with one of best performing systems — Caglayan et al., (2016). Their system uses a phrase-based statistical machine translation model, plus a re-scoring strategy using language model and visual information in the form of the penultimate layer of a pre-trained VGG network. The most interesting observation is that *Pool5* and *Softmax* perform similarly, as in the IC task, and that the efficacy of the use of the visual information in the encoding versus decoding seems to depend on the type of visual representation and also on the dataset. In fact, no clear trend could be observed and additional experiments are needed, ideally with more realistic translation (not image captioning) data.

6 Analysis and Discussion

In what follows we further analyze the results for the IC task, for which the representations and models studied in this paper seem to show a clearer trend than for MMT.
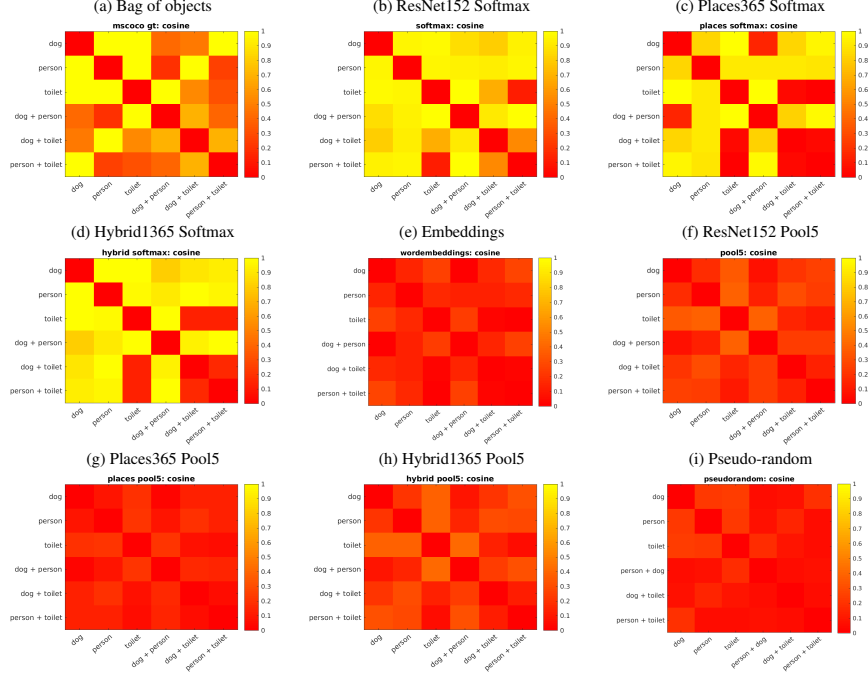

Fig. 3: The cosine distance matrix between six groups: three MSCOCO categories and pairwise combinations of the three categories) from the training dataset. Each group is represented by the average image feature of 25 randomly selected images from the category or combination of categories.

6.1 Image Representations

We first compare different image representations with respect to their ability to group and distinguish between semantically related images. For this, we selected three categories from MSCOCO ("dog", "person", "toilet") and also pairwise combinations of these ("dog+person", "dog+toilet", "person+toilet"). Up to 25 images were randomly selected for each of these six groups (single category or pair) such that the images are annotated with *only* the associated categories. Each group is represented by the average image feature of these images. Figure 3 shows the cosine distances between each group, for each of our image representations. The *Bag of Objects* representation forms the clearest clusters, as expected (e.g. the average image representation of "dog" correlates with images containing "dog" as a pair like "dog+person" and "dog+toilet"). The *Softmax* representations seem to also exhibit semantic clusters, although to a lesser extent. This can be observed

with "person", where the features are not semantically similar to any other groups. The most likely reason is that there is no "person" category in ILSVRC. Also, *Place365* and *Hybrid1365 Softmax* (Figure 3c) also showed very strong similarity for images containing "toilet", whether or not they contain "dog" or "person", possibly because they capture scene information. On the other hand, *Pool5* features seem to result in images that are more similar to each other than *Softmax* overall.
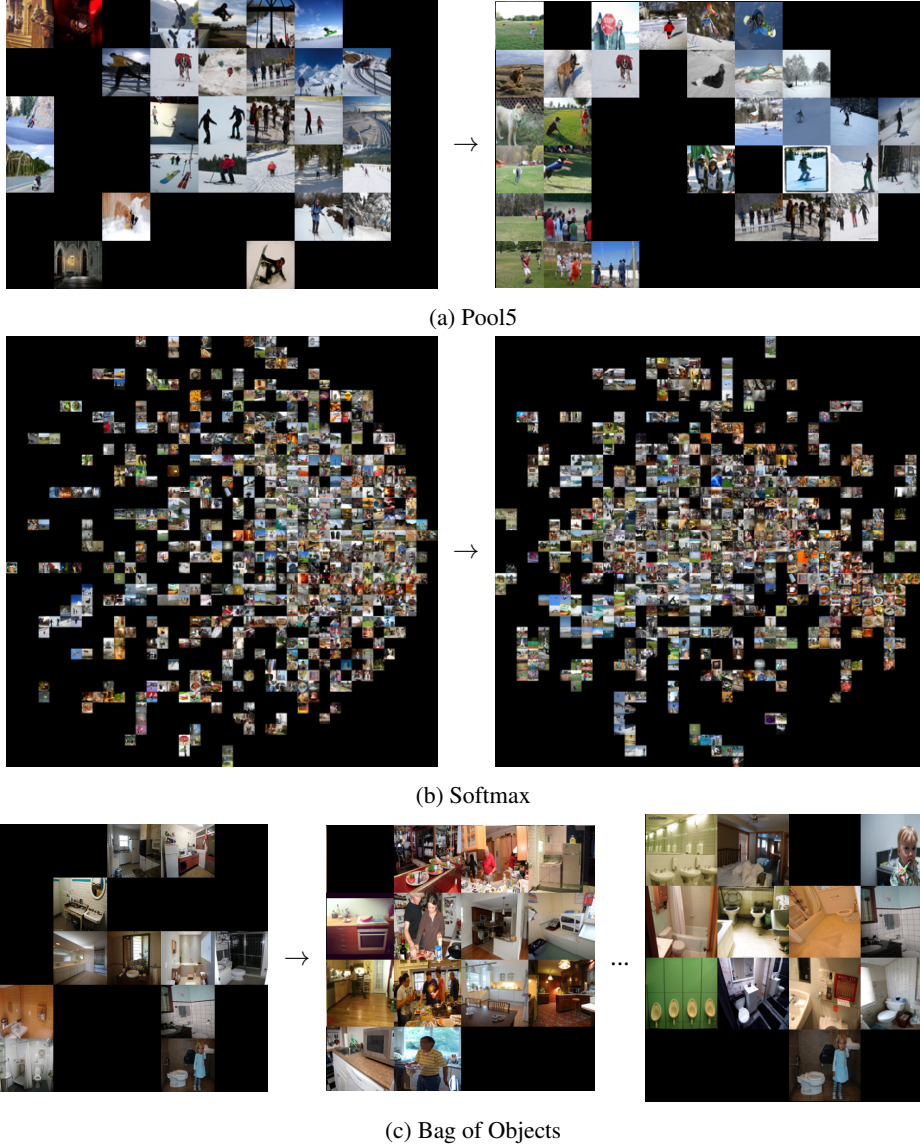

(a) Pool5

(b) Softmax

(c) Bag of Objects

Fig. 4: Visualization of the t-SNE projection of the initial representational space (left) vs. the transformed representational space (right).

6.2 Transformed Representations

To test the possibility that the RNN conditioned on visual information learns some sort of common 'visual-semantic' space, we explore the difference in representations between the initial representational space and the transformed representational space. The transformation is learned jointly as a subtask of image captioning. To visualize both representational spaces, we use *Barnes-Hut t-SNE* (van der Maaten & Hinton, 2008) to compute a 2-dimensional embedding over the test split. In general, we found that images are initially clustered by visual similarity (*Pool5*) and semantic similarity (*Softmax*, *Bag of Objects*). After transformation, linguistic information from the captions leads to different types of clusters.

Figure 4 highlights some interesting observations about the changes in clustering across three different representations. For *Pool5*, images seem to be clustered by their visual appearance, for example snow scenes in Figure 4a, regardless of the subjects in the images (people or dogs). After transformation, separate clusters seem to be formed for snow scenes involving a single person, groups of people, and dogs. Interestingly, images of dogs in fields and snow scenes are also drawn closer together.

Softmax (Figure 4b) shows many small, isolated clusters before transformation. After transformation, bigger clusters seem to be formed – suggesting that the captions have again drawn related images together despite being different in the *Softmax* space.

For *Bag of Objects* (Figure 4c), objects seem to be clustered by co-occurrence of object categories, for example toilets and kitchens are clustered since they share sinks. Toilets and kitchens seem to be further apart in the transformed space.

A similar observation was made by Vinyals et al. (2016) in which the authors observe that end-to-end IC models are capable of performing retrieval tasks with comparable performance to the task-specific models that are trained with ranking loss.

6.3 Generated Captions

In this section we provide a qualitative analysis of different image representations and gain insights into how they contribute to the IC task. *Bag of Objects* led to a strong performance in IC despite being extremely sparse and low-dimensional ($80D$). Analyzing the test split, we found that each vector consists of only 2.86 non-zero entries on average (standard deviation 1.8, median 2). Thus, with minimal information being provided to the RNN generator, we find it surprising that it is able to perform so well.

We compare the output of the remaining models against the *Bag of Objects* representation by investigating what each representation adds to or subtracts from this simple, yet strong model. We start by selecting images (from the test split) annotated with the exact same *Bag of Objects* representation – which should result in the same caption. For our qualitative analysis, several sets of one to three MSCOCO categories were manually chosen. For each set, images were selected such that there is exactly one instance of each category in the set and zero for others. We then shortlisted images where the captions generated by the *Bag of Objects* model produced the five highest and five lowest CIDEr scores (ten images per set). We compare the captions sampled for each of the other representations.

Figure 5 shows some example outputs from this analysis. In Figure 5a, *Bag of Objects*
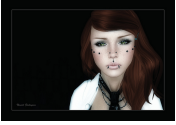
Representation	CIDEr (Δ)	Caption
Bag of objects	2.78 (+0.00)	a bird is perched on a branch in the sun .
VGG19 softmax	3.14 (+0.36)	a owl is perched on a branch of a tree .
ResNet softmax	3.67 (+0.89)	a owl is perched on a branch in a tree .
Places365 softmax	2.00 (-0.77)	a bear is sitting on a branch in the wilderness .
Hybrid1365 softmax	0.01 (-2.77)	a giraffe standing in a field of grass .
VGG19 fc7	0.18 (-2.59)	a black and white image of a bird sitting on a window sill .
ResNet pool5	0.38 (-2.40)	a large black bear standing in a forest .
Places365 pool5	0.34 (-2.43)	a giraffe standing in the middle of a forest .
Hybrid1365 pool5	3.03 (+0.26)	a bird is perched on a branch in a tree .
Embeddings	2.38 (-0.40)	a bird sitting on a branch in a window .

(a) Bag of objects: bird (1)

Representation	CIDEr (Δ)	Caption
Bag of objects	0.09 (+0.00)	a large airplane flying through a blue sky .
VGG19 softmax	0.00 (-0.09)	a man in a baseball cap and sunglasses is holding a baseball bat .
ResNet softmax	0.00 (-0.09)	a man is holding a baseball bat in a batting cage .
Places365 softmax	0.06 (-0.03)	a dog is standing in the grass with a ball in its mouth .
Hybrid1365 softmax	0.00 (-0.09)	a man holding a tennis racquet on a tennis court .
VGG19 fc7	0.73 (+0.63)	a plane is sitting on a runway with a few people .
ResNet pool5	0.01 (-0.08)	a train is on the tracks in a city .
Places365 pool5	0.00 (-0.09)	a giraffe standing in a fenced in enclosure .
Hybrid1365 pool5	0.01 (-0.08)	a man holding a baseball bat standing next to home plate .
Embeddings	0.01 (-0.09)	a baseball player holding a bat on a field .

(b) Bag of objects: airplane (1)

Representation	CIDEr (Δ)	Caption
Bag of objects	0.01 (+0.00)	a man wearing a suit and tie standing in front of a building .
VGG19 softmax	0.04 (+0.04)	a woman in a pink wig and a pink dress .
ResNet softmax	0.00 (-0.00)	a man in a suit and tie is smiling .
Places365 softmax	0.13 (+0.12)	a woman with a red polka dotted dress tie .
Hybrid1365 softmax	0.06 (+0.05)	a woman in a red dress is talking on a cell phone .
VGG19 fc7	0.24 (+0.24)	a woman with a cell phone in her hand .
ResNet pool5	0.08 (+0.08)	a woman in a red shirt and tie .
Places365 pool5	0.10 (+0.09)	a woman is holding a cell phone to her ear .
Hybrid1365 pool5	0.05 (+0.04)	a woman in a dress shirt and tie holding a parasol .
Embeddings	0.00 (-0.01)	a man wearing a tie and a shirt and a tie .

(c) Bag of objects: person (1), tie (1)

Fig. 5: Example outputs from our system with different representations, the sub-captions indicate the annotation along with the frequency in braces. We also show the CIDEr score and the difference in CIDEr score relative to the *Bag of Objects* representation.

achieved a high CIDEr score despite only being given "bird" as input, mainly by 'guessing' that the bird will be perching/sitting on a branch. The object-based *Softmax* (VGG and ResNet) models led to an even more accurate description as "owl" is the top-1 prediction of both representations (96% confidence for VGG, 77% for ResNet). *Places365* predicted "swamp" and "forest". The *Penultimate* features on the other hand struggled with representing the images correctly. In Figure 5b, *Bag of Objects* suffered from lack of information (only "airplane" is given), the *Softmax* features mainly predicted "chain-link fence", Places365 predicted "kennel" (hence the dog description), and it is most likely that *Penultimate* has captured the fence-like features in the image rather than the plane. In Figure 5c, *Softmax* features generally managed to generate a caption describing a woman despite not explicitly containing the "woman" category. This is because other correlated categories were predicted, such as "mask", "wig", "perfume" and "hairspray", and for *Places365* "beauty salon" and "dressing room". *ResNet* predicted categories like "stethoscope", "suit", "cloak", where we assume that doctor roles may be male-dominated in the dataset, thus generating 'man'.

6.4 Uniqueness of Captions

Model	Unique (%)
BOO Gold-Counts	29.5
Top-k Class(Embeddings)	29.0
Softmax(ResNet152)	28.7
Pool5 (ResNet152)	28.8
Human	99.4

Table 3: Unique captions with beam $= 1$.

Challenges with IC datasets have been well explored in previous work. Karpathy et al. (2016) perform both word level and syntactic level analysis on the MSCOCO and Flickr8k datasets and concludes they both lack diversity. This means that most of the captions are generic descriptions and can fit multiple images. This extends directly for our experiments on the IC and MMT datasets.

We now turn to the question on the ability of representations to produce unique captions for every distinct image. We use the validation portion of the MSCOCO dataset, which contains 40,504 images, and produce captions with four types of image representations. We report the results in Table 3. We observe that in almost all cases, the produced representations are far from unique. In most cases, there is a significant portion of the captions that are repeated. This has also been observed by Devlin et al. (2015) on different test splits, but using retrieval-based and pipeline methods for IC.

7 Conclusions

Our experiments probe the contribution of various types of image representations and shed some light on the utility of image representations for vision to language tasks. We observed

that a conditional RNN-based language model is capable of making sense of noisy information and correctly clustering the noisy representation in the projected space. However, the task datasets do not reflect the paucity of information content in the image representation and, in most cases, we obtain repeated captions for similar sets of images. Our empirical observations indicate that the direct use of lower-level image features may not be the only way to condition an RNN, and that higher-level, abstract, semantic features may also be beneficial in order to capture the semantic aspects of the images. As future work, we are interested in exploring more complex models that use attention-based architectures and those that exploit latent spaces.

Acknowledgments

References

Anderson, P., Fernando, B., Johnson, M., & Gould, S. 2016. SPICE: semantic propositional image caption evaluation. *In: Proc. of the European Conference on Computer Vision (ECCV)*.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. 2015. VQA: visual question answering. *In: Proc. of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*.

Arora, S., Liang, Y., & Ma, T. 2017. A simple but tough-to-beat baseline for sentence embeddings. *In: Proc. of the International Conference on Learning Representations, Workshop Contributions*.

Bahdanau, D., Cho, K., & Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. *In: Proc. of the International Conference on Learning Representation (ICLR)*.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., & Plank, B. 2016. Automatic description generation from images: a survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research (JAIR)*, **55**, 409–442.

Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F., Barrault, L., & van de Weijer, J. 2016. Does multimodality help human and machine for translation and image captioning? *In: Proc. of the Conference on Machine Translation (WMT)*.

Calixto, I., Elliott, D., & Frank, S. 2016. DCU-UvA multimodal MT system report. *In: Proc. of the Conference on Machine Translation (WMT)*.

Calixto, I., Liu, Q., & Campbell, N. 2017. Doubly-attentive decoder for multi-modal neural machine translation. *In: Proc. of the Association for Computational Linguistics (ACL)*.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. 2015. Microsoft COCO captions: data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *Deep Learning and Representation Learning Workshop*.

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. 2015. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv preprint arXiv:1511.07289*.

Denkowski, M., & Lavie, A. 2014. Meteor universal: language specific translation evaluation for any target language. *In: Proc. of the EACL Workshop on Statistical Machine Translation*.

Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., & Mitchell, M. 2015. Language models for image captioning: The quirks and what works. *In: Proc. of the Association for Computational Linguistics (ACL)*.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. 2014. Decaf: a deep convolutional activation feature for generic visual recognition. *In: Proc. of the International Conference on Machine Learning (ICML)*.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. *In: Proc. of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*.

Elliott, D., & de Vries, A. 2015. Describing images using inferred visual dependency representations. *In: Proc. of the Association for Computational Linguistics (ACL)*.

Elliott, D., & Kádár, A. 2017. Imagination improves multimodal translation. *In: Proc. of the International Joint Conference on Natural Language Processing (IJCNLP)*.

Elliott, D., & Keller, F. 2014. Comparing automatic evaluation measures for image description. *In: Proc. of the Association for Computational Linguistics (ACL)*.

Elliott, D., Frank, S., & Hasler, E. 2015. Multi-language image description with neural sequence models. *arxiv preprint arxiv:1510.04709*.

Elliott, D., Frank, S., Sima'an, K., & Specia, L. 2016. Multi30K: multilingual English-German image descriptions. *In: Proc. of the 5th Workshop on Vision and Language*.

Elliott, D., Frank, S., Barrault, L., Bougares, F., & Specia, L. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. *In: Proc. of the Conference on Machine Translation (WMT)*.

Elman, J. L. 1990. Finding structure in time. *Cognitive science*, **14**, 179–211.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., & Platt, J. C. 2015. From captions to visual concepts and back. *In: Proc. of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*.

Farhadi, A., Hejrati, M., Sadeghi, M., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. 2010. Every picture tells a story: generating sentences from images. *In: Proc. of the European Conference on Computer Vision (ECCV)*.

Ferraro, F., Mostafazadeh, N., Vanderwende, L., Devlin, J., Galley, M., & Mitchell, M. 2015. A survey of current datasets for vision and language research. *In: Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., & Xu, W. 2015. Are you talking to a machine? Dataset and methods for multilingual image question. *In: Proc. of the Advances in Neural Information Processing Systems (NIPS)*.

Grubinger, M., Clough, P., Müller, H., & Deselaers, T. 2006. The IAPR TC-12 benchmark: a new evaluation resource for visual information systems. *In: International Workshop on Language Resources for Content-Based Image Retrieval, OntoImage'2006*.

He, K., Zhang, X., Ren, S., & Sun, J. 2016. Deep residual learning for image recognition. *In: Proc. of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*.

Hitschler, J., Schamoni, S., & Riezler, S. 2016. Multimodal pivots for image caption translation. *In: Proc. of the Association for Computational Linguistics (ACL)*.

Hochreiter, S., & Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

Hodosh, M., Young, P., & Hockenmaier, J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research (JAIR)*, **47**, 853–899.

Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., & Dyer, C. 2016. Attention-based multimodal neural machine translation. *In: Proc. of the Conference on Machine Translation (WMT)*.

Karpathy, A. 2016. *Connecting images and natural language*. Ph.D. thesis, Department of Computer Science, Stanford University.

Karpathy, A., & Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. *In: Proc. of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*.

Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., & Erdem, E. 2017. Re-evaluating automatic metrics for image captioning. *In: Proc. of the European Chapter of the Association for Computational Linguistics (EACL)*.

Kiros, R., Salakhutdinov, R., & Zemel, R. S. 2014. Multimodal neural language models. *In: Proc. of the International Conference on Machine Learning (ICML)*.

Kolář, M., Hradiš, M., & Zemčík, P. 2015. Technical report: Image captioning with semantically similar images. *arXiv preprint arXiv:1506.03995*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. *In: Proc. of the Advances in Neural Information Processing Systems (NIPS).*

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. 2011. Baby talk: understanding and generating image descriptions. *In: Proc. of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR).*

Kuznetsova, P., Ordonez, V., Berg, A., Berg, T., & Choi, Y. 2012. Collective generation of natural image descriptions. *In: Proc. of the Association for Computational Linguistics (ACL).*

Kuznetsova, P., Ordonez, V., Berg, A., Berg, T., & Choi, Y. 2013. Generalizing image captions for image-text parallel corpus. *In: Proc. of the Association for Computational Linguistics (ACL).*

Kuznetsova, P., Ordonez, V., Berg, T. L., & Choi, Y. 2014. TREETALK: composition and compression of trees for image descriptions. *In: Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Lala, C., Madhyastha, P., Wang, J., & Specia, L. 2017. Unraveling the contribution of image captioning and neural machine translation for multimodal machine translation. *The Prague Bulletin of Mathematical Linguistics.*

Lebret, R., Pinheiro, P. O., & Collobert, R. 2015. Phrase-based image captioning. *In: Proc. of the International Conference on Machine Learning (ICML).*

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., & Choi, Y. 2011. Composing simple image descriptions using web-scale n-grams. *In: Proc. of the SIGNLL Conference on Computational Natural Language Learning (CoNLL).*

Libovický, J., Helcl, J., Tlustý, M., Bojar, O., & Pecina, P. 2016. CUNI system for WMT16 automatic post-editing and multimodal translation tasks. *In: Proc. of the Conference on Machine Translation (WMT).*

Luong, M.-T., Pham, H., & Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. 2015. Deep captioning with multimodal recurrent neural networks (m-RNN). *In: Proc. of the International Conference on Learning Representation (ICLR).*

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., & Khudanpur, S. 2010. Recurrent neural network based language model. *In: Proc. of the Annual Conference of the International Speech Communication Association (Interspeech).*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *In: Proc. of the Advances in Neural Information Processing Systems (NIPS).*

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., Mensch, A., Berg, A., Berg, T., & Daume III, H. 2012. Midge: generating image descriptions from computer vision detections. *In: Proc. of the European Chapter of the Association for Computational Linguistics (EACL).*

Ordonez, V., Kulkarni, G., & Berg, T. L. 2011. Im2Text: describing images using 1 million captioned photographs. *In: Proc. of the Advances in Neural Information Processing Systems (NIPS).*

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. *In: Proc. of the Association for Computational Linguistics (ACL).*

Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. 2010. Collecting image annotations using Amazon's Mechanical Turk. *In: Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.*

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. 2014. CNN features off-the-shelf: an astounding baseline for recognition. *In: Proc. of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR).*

Redmon, J., & Farhadi, A. 2017. YOLO9000: better, faster, stronger. *In: Proc. of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR).*

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV).*

Shah, K., Wang, J., & Specia, L. 2016. SHEF-Multimodal: grounding machine translation on images. *In: Proc. of the Conference on Machine Translation (WMT)*.

Simonyan, K., & Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *In: Proc. of the International Conference on Learning Representation (ICLR)*.

Socher, R., Karpathy, A., Le, Q., Manning, C., & Ng, A. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, **2**, 207–218.

Specia, L., Frank, S., Simaan, K., & Elliott, D. 2016. A shared task on multimodal machine translation and crosslingual image description. *In: Proc. of the Conference on Machine Translation (WMT)*.

Sutskever, I., Vinyals, O., & Le, Q. V. 2014. Sequence to sequence learning with neural networks. *In: Proc. of the Advances in Neural Information Processing Systems (NIPS)*.

van der Maaten, L., & Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, **9**, 2579–2605.

van Miltenburg, E., & Elliott, D. 2017. Room for improvement in automatic image description: an error analysis. *arXiv preprint arXiv:1704.04198*.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. 2015. Cider: consensus-based image description evaluation. *In: Proc. of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. 2015. Show and tell: A neural image caption generator. *In: Proc. of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. 2016. Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**(4), 652–663.

Wu, Q., Shen, C., Liu, L., Dick, A., & van den Hengel, A. 2016. What value do explicit high level concepts have in vision to language problems? *In: Proc. of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. 2015. Show, attend and tell: neural image caption generation with visual attention. *In: Proc. of the International Conference on Machine Learning (ICML)*.

Yang, Y., Teo, C., Daumé III, H., & Aloimonos, Y. 2011. Corpus-guided sentence generation of natural images. *In: Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yao, B. Z., Yang, X., Lin, L., Lee, M. W., & Zhu, S. C. 2010. I2T: image parsing to text description. *Proc. of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*.

Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T. 2017. Boosting image captioning with attributes. *In: Proc. of the IEEE International Conference on Computer Vision (ICCV)*.

Yin, X., & Ordonez, V. 2017. Obj2Text: generating visually descriptive language from object layouts. *In: Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. 2016. Image captioning with semantic attention. *In: Proc. of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. 2014. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, **2**, 67–78.

Zaremba, W., Sutskever, I., & Vinyals, O. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. 2014. Learning deep features for scene recognition using Places database. *In: Proc. of the Advances in Neural Information Processing Systems (NIPS)*.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. 2017. Places: a 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PP**(99).