BERT Can See Out of the Box: On the Cross-modal Transferability of Text Representations

Thomas Scialom [*1 2] Patrick Bordes [*2] Paul-Alexis Dray [1] Jacopo Staiano [1] Patrick Gallinari [2 3]

Abstract

Pre-trained language models such as BERT have recently contributed to significant advances in Natural Language Processing tasks. Interestingly, while multilingual BERT models have demonstrated impressive results, recent works have shown how monolingual BERT can also be competitive in zero-shot cross-lingual settings. This suggests that the abstractions learned by these models can transfer across languages, even when trained on monolingual data. In this paper, we investigate whether such generalization potential applies to other modalities, such as vision: *does BERT contain abstractions that generalize beyond text?* We introduce *BERT-gen*, an architecture for text generation based on BERT, able to leverage on either mono- or multi- modal representations. The results reported under different configurations indicate a positive answer to our research question, and the proposed model obtains substantial improvements over the state-of-the-art on two established Visual Question Generation datasets.

1. Introduction

The BERT language model (Devlin et al., 2019) is a Deep Bidirectional Transformer (Vaswani et al., 2017) pre-trained on textual corpora (BookCorpus and Wikipedia) using a Masked Language Model (MLM) objective – predicting some words that are randomly masked in the sentence, along with a sentence entailment loss. Recent research efforts (Artetxe et al., 2019) have shown how BERT encodes abstractions that generalize across languages, even when trained on monolingual data only. This contradicts the common belief (Pires et al., 2019; Wu & Dredze, 2019) that a shared vocabulary and joint training on multiple languages are essential to achieve cross-lingual generalization capabilities. In this work, we further investigate the generalization potentials of large pre-trained LMs, this time moving to a cross-modal setup: *does BERT contain abstractions that generalize beyond text?*

In the Artificial Intelligence community, several works have investigated the longstanding research question of whether textual representations encode visual information. On the one hand, a large body of research called *language grounding* considers that textual representations lack visual commonsense (Baroni, 2016), and intend to *ground* the meaning of words (Lazaridou et al., 2015; Collell et al., 2017) and sentences (Kiela et al., 2018; Bordes et al., 2019) in the perceptual world. In another body of work, textual representations have successfully been used to tackle multi-modal tasks (Baltrusaitis et al., 2019) such as Zero-Shot Learning (Zablocki et al., 2019), Visual Question Answering (Malinowski & Fritz, 2014) or Image Captioning (Socher et al., 2014). Following the latter line of research, in this paper we evaluate the potential of pre-trained language models to generalize in the context of Visual Question Generation (VQG) (Mostafazadeh et al., 2016).

The Visual Question Generation task allows us to investigate the cross-modal capabilities of BERT: unlike Image Captioning (where the input is only visual) or VQA (where the input is visual *and* textual), VQG is a multi-modal task where input can be textual *and/or* visual. VQG data usually includes images and the associated captions, along with corresponding questions about the image; thus, different experimental setups can be designed to analyze the impact of each modality. Indeed, the questions can be generated using *i)* textual (the caption), *ii)* visual (the image), or *iii)* multi-modal (both the caption and the image) input.

From a practical standpoint, the VQG task has several applications: robots or AI assistants could ask questions rooted in multi-modal data (e.g. fusing conversational data with visual information from captors and cameras), in order to refine their interpretation of the situation they are presented with. It could also allow systems relying on knowledge-bases to gain visual common sense and deal with the Human Reporting Bias (Misra et al., 2016), which states that the content of images and text are intrinsically different, since visual common sense is rarely explicitly stated in text.

[*]Equal contribution [1]reciTAL, Paris, France [2]Sorbonne Université, CNRS, LIP6, F-75005 Paris, France [3]Criteo AI Lab, France. Correspondence to: Thomas Scialom <thomas@recital.ai>.

Preliminary work. Under review.

Recently, BERT-based Multi-Modal Language Models have been proposed (Lu et al., 2019; Tan & Bansal, 2019; Li et al., 2019b; Su et al., 2019) to tackle multi-modal tasks, using different approaches to incorporate visual data within BERT. From these works, it is left to explore whether the cross-modal alignment is fully learned, or it is to some extent already encoded in the BERT abstractions. Therefore, in contrast with those approaches, we explicitly avoid using the following complex mechanisms:

- *Multi-modal supervision*: all previous works exploit an explicit multi-modal supervision through a pre-training step; the models have access to text/image pairs as input, to align their representations. In contrast, our model can switch from text-only to image-only mode without any explicit alignment.
- *Image-specific losses*: specific losses such as Masked RoI (Region of Interest) Classification with Linguistic Clues (Su et al., 2019) or sentence-image prediction (Li et al., 2019b) have been reported helpful to align visual and text modalities. Instead, we only use the original MLM loss from BERT (and not its entailment loss).
- *Non-linearities*: we explore a scenario in which the only learnable parameters, for aligning image representations to BERT, are those of simple linear projection layer. This allows us to assess whether the representations encoded in BERT can transfer *out-of-the-box* to another modality.

Furthermore, to the best of our knowledge, this paper is the first attempt to investigate multi-modal text *generation* using pre-trained language models. We introduce *BERT-gen*, a text generator based on BERT, that can be applied both in mono and multi-modal settings. We treat images similarly to text: while a sentence is seen as a sequence of (sub)word tokens, an image is seen as a sequence of objects associated to their corresponding positions (bounding boxes). We show how a simple linear mapping, projecting visual embeddings into the first layer, is enough to ground BERT in the visual realm: text and image object representations are found to be effectively aligned, and the attention over words transfers to attention over the relevant objects in the image.

Our contributions can be summarized as follows:

1. we introduce *BERT-gen*, a novel method for generating text using BERT, that can be applied in both mono and multi-modal settings;
2. we show that the semantic abstractions encoded in pre-trained BERT can generalize to another modality;
3. we report state-of-the art results on the VQG task;
4. we provide extensive ablation analyses to interpret the behavior of *BERT-gen* under different configurations (mono- or multi- modal).

2. Related Work

Unsupervised Pre-trained Language Models Learning unsupervised textual representations that can be applied to downstream tasks is a widely investigated topic in the literature. Text representations have been learned at different granularities: words with Word2vec (Mikolov et al., 2013), sentences with SkipThought (Kiros et al., 2015), paragraphs with ParagraphVector (Le & Mikolov, 2014) and contextualized word vectors with ELMo (Peters et al., 2018). Other methods leverage a transfer-learning approach by *fine-tuning* all parameters of a pre-trained model on a target task, a paradigm which has become mainstream since the introduction of BERT (Devlin et al., 2019). BERT alleviates the problem of the uni-directionality of most language models (i.e. where the training objective aims at predicting the next word) by proposing a new objective called Masked Language Model (MLM). Under MLM, some words, that are randomly selected, are masked; the training objective aims at predicting them.

Multi-modal Language Models Following the successful application of BERT (Devlin et al., 2019), and its derivatives, across a great majority of NLP tasks, several research efforts have focused on the design of multi-modal versions of BERT. VideoBERT (Sun et al., 2019a), a joint *video* and text model, is pre-trained on a huge corpus of YouTube videos, and applied to action classification and video captioning tasks on the YouCook II dataset (Zhou et al., 2018b). The video is treated as a "visual sentence" (each frame being a "visual word") that is processed by the BERT Transformer.

Concerning models jointly treating information from images and text, visual features extracted from the image are used as "visual words", and a `[SEP]` special token is employed to separate textual and visual tokens. In the literature, visual features are object features extracted with a Faster R-CNN (Ren et al., 2017) – with the notable exception of Kiela et al. (2019) who used pooling layers from a CNN. A first body of work exploit *single-stream* Transformers in which visual features are incorporated in a BERT-like Transformer: this is the case for VisualBERT (Li et al., 2019b), VL-BERT (Su et al., 2019), Unicoder-VL (Li et al., 2019a) and B2T2 (Alberti et al., 2019). Other works, such as ViLBERT (Lu et al., 2019) and LXMERT (Tan & Bansal, 2019) have investigated *two-stream* approaches: these models employ modality-specific encoders built on standard Transformer blocks, which are then fused into a cross-modal encoder. Interestingly, none of the aforementioned models have been used for generation tasks such as VQG, tackled in this work.

Visual Question Generation The text-based Question Generation task has been largely studied by the NLP community (Rus et al., 2010; Rajpurkar et al., 2016; Zhou et al., 2017; Du et al., 2017; Song et al., 2017; Zhao et al., 2018; Scialom et al., 2019). However, its visual counterpart, Visual Question Generation (VQG), has been comparatively less explored than standard well-known multi-modal tasks such as Visual Question Answering (VQA) (Gao et al., 2015; Xu & Saenko, 2016; Ren et al., 2015; Ma et al., 2016), Visual Dialog (Das et al., 2017; 2019), or Image Captioning (Vinyals et al., 2015; Yan et al., 2016; Karpathy & Li, 2015).

The VQG task was first introduced by Yang et al. (2015) in their Neural Self Talk model: the goal is to gain knowledge about an image by iteratively generating questions (VQG) and answering them (VQA). The authors tackle the task with a simple RNN conditioned on the image, following Image Captioning works such as Karpathy & Li (2015).

Suitable data for the VQG task can come from standard image datasets on which questions have been manually annotated, such as VQG_{COCO}, VQG_{Flickr}, VQG_{Bing} (Mostafazadeh et al., 2016) , each consisting of 5000 images with 5 questions per image. Alternatively, VQG samples can be derived from Visual Question Answering datasets, such as $VQA1.0$ (Antol et al., 2015), by "reversing" them (taking images as inputs and questions as outputs).

A variety of approaches have been proposed. Mostafazadeh et al. (2016) use a standard Gated Recurrent Neural Network, *i.e.* a CNN encoder followed by a GRU decoder to generate questions. Zhang et al. (2017) aim at generating, for a given image, multiple visually grounded questions of varying types (*what*, *when*, *where*, etc.); similarly, Jain et al. (2017) generate diverse questions using Variational Autoencoders. In Li et al. (2018), VQG is jointly tackled along its dual task (VQA), just as Yang et al. (2015). In (Patro et al., 2018; Patro & Namboodiri, 2019), the image (processed by a CNN) and the caption (processed by a LSTM) are combined in a mixture module, followed by a LSTM decoder to generate the question, leading to state-of-the-art results on the VQG task on $VQA1.0$ data. More recently, Patro et al. (2020) incorporate multiple cues – place information obtained from PlaceCNN (Zhou et al., 2018a), caption, tags – and combine them within a deep Bayesian framework where the contribution of each cue is weighted to predict a question, obtaining the current state-of-the-art results on VQG_{COCO}.

3. Model

In VQG, the objective is to generate a relevant question from an image and/or its caption. The caption X_{txt} is composed of M tokens $txt_1, ..., txt_M$; these tokens can be words or subwords (smaller than word) units depending on the tokenization strategy used. As BERT uses subword tokenization, throughout this paper we will refer to subwords as our tokenization units.

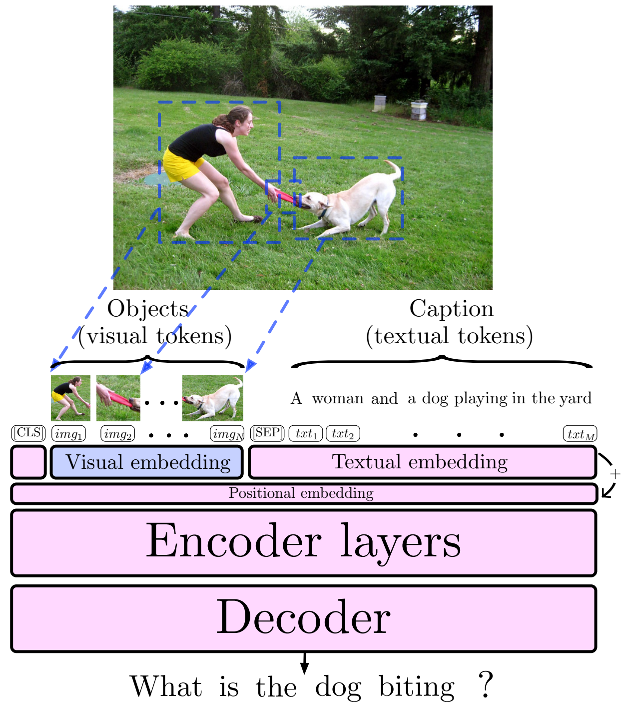

Figure 1. Model overview. Captions are encoded via BERT embeddings, while visual embeddings (blue) are obtained via a linear layer, used to project image representations to the embedding layer dimensions.

The proposed model is illustrated in Figure 1. In 3.1, we detail how images are incorporated in the Transformer framework. In 3.2, we present *BERT-gen*, a novel approach to use BERT for text generation.

3.1. Representing an Image as Text

In this work, we treat textual and visual inputs similarly, by considering both as sequences. Since an image is not a priori sequential, we consider the image X_{img} as a sequence of object regions $img_1, ..., img_N$, as described below.

The images are first processed as in Tan & Bansal (2019): a Faster-RCNN (Ren et al., 2017), pre-trained on Visual Genome (Krishna et al., 2017), detects the $N = 36$ most salient regions (those likely to include an object) per image. The weights of the Faster-RCNN are fixed during training, as we use the precomputed representations made publicly available[1] by Anderson et al. (2018). Each image is thus represented by a sequence of $N = 36$ semantic embeddings $f_1, ...f_N$ (one for each object region) of dimension 2048,

[1] https://github.com/peteanderson80/bottom-up-attention

along with the corresponding bounding box coordinates $b_1, ...b_N$ of dimension 4. With this approach, the BERT attention can be computed at the level of objects or salient image regions; had we represented images with traditional CNN features, the attention would instead correspond to a uniform grid of image regions without particular semantics, as noted in Anderson et al. (2018). To build an object embedding o_j encoding both the object region semantics and its location in the image, we concatenate f_j and b_j ($j \in [1, N]$). Hence, an image is seen as a sequence of $N = 36$ visual representations (each corresponding to an object region) $o_1, ..., o_N$. Object region representations o_i are ordered by the relevance of the object detected, and the model has access to their relative location in the image through the vectors b_i.

To investigate whether our BERT-based model can transfer knowledge beyond language, we consider image features as simple visual tokens that can be presented to the model *analogously* to textual token embeddings. In order to make the o_j vectors (of dimension $2048 + 4 = 2052$) comparable to BERT embeddings (of dimension 768), we use a simple linear *cross-modal projection* layer W of dimensions 2052×768. The N object regions detected in an image, are thus represented as $X_{img} = (W.o_1, ..., W.o_N)$. Once mapped into the BERT embedding space with W, the image is seen by the rest of the model as a sequence of units with no explicit indication if it is a text or an image embedding.

3.2. ***BERT-gen***: Text Generation with BERT

We cast the VQG task as a classic sequence-to-sequence (Sutskever et al., 2014) modeling framework:

$$P_{\Theta,W}(Y|X) = \prod_{t=1}^{T} P_{\Theta,W}(y_t|X, y_{<t}) \qquad (1)$$

where the input $X = X_{txt}$ in caption-only mode, $X = X_{img}$ in image-only mode, and $X = X_{img} \oplus X_{txt}$ in a multi-modal setup; $Y = y_1, ..., y_T$ is the question composed of T tokens. Θ are the parameters of the BERT model;[2] W represents the weights of the linear layer used for projecting visual input to the BERT embedding layer.

As mentioned earlier, BERT is a Transformer (Vaswani et al., 2017) encoder pre-trained using the Masked Language Model (MLM) objective: tokens within the text are replaced with a `[MASK]` special token, and the model is trained to predict them. Since BERT was not trained with an unidirectional objective, its usage for text generation is not straightforward.

To generate text, Liu & Lapata (2019) propose to stack a Transformer decoder, symmetric to BERT. However, the authors report training difficulties since the stacked decoder is not pre-trained, and propose a specific training regime, with the side-effect of doubling the number of parameters. Dong et al. (2019) opt for an intermediate step of self-supervised training, introducing a unidirectional loss. As detailed below, we propose a relatively simpler, yet effective, method to use BERT *out-of-the-box* for text generation.

Decoder We simply use the original BERT decoder as is, initially trained to generate the tokens masked during its pre-training phase. It consists in a feed-forward layer, followed by normalization, transposition of the embedding layer, and a softmax over the vocabulary.

Next Token Prediction At inference time, to generate the first token of the question y_1, we concatenate `[MASK]` to the input tokens X, then encode $X \oplus$ `[MASK]` with the BERT encoder, and feed the output of the encoder to the decoder; y_1 is the output of the decoder for the `[MASK]` token. Subsequently, given y_1, we concatenate it to the input tokens and encode $X \oplus y_1 \oplus$ `[MASK]` to predict the next token y_2. This procedure is repeated until the generation of a special token `[EOS]` signaling the end of the sentence.

Attention Trick As we iteratively concatenate the generated tokens, the BERT *bi-directional* self-attention mechanism would impact, at every new token, the representations of the previous tokens. To counter that, we use a *left-to-right* attention mask, similar to the one employed in the original Transformer decoder (Vaswani et al., 2017). For the input tokens in X, we apply such mask to all the target tokens Y that were concatenated to X, so that input tokens can only attend to the other input tokens. Conversely, for target tokens y_t, we put an attention mask on all tokens $y_{>t}$, allowing target tokens y_t to attend only to the input tokens and the already generated target tokens.

This novel method allows to use pre-trained encoders for text generation. In this work, we initialize our model with the parameters from BERT-base. Nonetheless, the methodology can be applied to any pre-trained Transformer encoders such as RoBERTa (Liu et al., 2019), or Ernie (Sun et al., 2019b).

Modality-specific setups The proposed model can be used in either mono- or multi- modal setups. This is accomplished by activating or deactivating specific modules.

4. Experimental Protocol

Our main objective is to measure whether the textual knowledge encoded in pre-trained BERT can be beneficial in a cross-modal task. Thus, we define the three following experimental setups, which we refer to as Step 1, 2, and 3:

[2] We use the smaller architecture released, `BERT-base` (12 layers), pre-trained on English cased text.

1. Caption only Deactivating the *Visual embedding* module (see Figure 1), the model has only access to textual input, *i.e.* the caption. The model is initialized with the BERT weights and trained according to Equation 1.

2. Image only Conversely, deactivating the *Textual embedding* module (see Figure 1), the model has only access to the input image, not the caption. To indicate the position t of img_t in the sequence, we sum the BERT positional embedding of t to the visual representation of img_t, just as we would do for a text token txt_t. The model is initialized with the weights learned during step 1. All *BERT-gen* Θ weights are frozen, and only the linear layer W is learnable. Hence, *if the model is able to learn to generate contextualized questions w.r.t. the image, it shows that a simple linear layer is enough to bridge the two modalities*.

3. Image + Caption The full model is given access to both image and caption inputs. In this setup, we separate the two different inputs by a special BERT token `[SEP]`. Thus, the input sequence for the model takes the form of `[CLS]`$, img_1, ..., img_N,$ `[SEP]`$, txt_1, ..., txt_M$. In step 1, only *BERT-gen* Θ parameters are learned, as no image input was given. In step 2, W is trained while keeping Θ frozen. Finally then, in step 3, we fine-tune the model using both image and text inputs: the model is initialized with the parameters Θ learned during step 1 and the W learned during step 2, and we unfreeze all parameters.

Ablations Additionally, we report results obtained with: *Image only (unfreeze)*, where the *BERT-gen* parameters Θ are not frozen; and *Image+Caption (from scratch)* where the model is learned without the intermediate steps 1 and 2: the *BERT-gen* parameters Θ are initialized with the weights from pre-trained BERT while W is randomly initialized.

4.1. Datasets

We conduct our experiments using two established datasets for Visual Question Generation:

VQG_{COCO} Introduced by Mostafazadeh et al. (2016), it contains 2500 training images, 1250 validation images and 1250 test images from MS COCO (Lin et al., 2014); each image has 5 corresponding questions and 5 ground-truth captions.[3]

VQA The Visual Question Answering (Antol et al., 2015) dataset can be used to derive VQG data (Li et al., 2018). The task is reversed: instead of answering the question based on the image (VQA), models are called to generate a relevant question given the image (VQG). Also based on MS COCO, it contains 82783 training images, 40504 validation images and 81434 testing images. In $VQA1.0$,[4] each image has 3 associated questions. Since the test set of MS COCO does not contain ground-truth captions, we generated artificial captions for it using NeuralTalk2 (Karpathy & Li, 2015): for fair comparison, we used exactly the same model[5] as Patro & Namboodiri (2019) (MDN-Joint).

4.2. Baselines

We compare the proposed model to the following:

Sample (Yang et al., 2015) Questions are generated by a RNN conditioned on the image: at each generation step, the distribution over the vocabulary is computed and used to sample the next generated word. This baseline enables to generate diverse questions over the same image, as the word selection process is non-deterministic.

Max (Yang et al., 2015) Using the above model, selecting words with maximum probability from the computed distribution.

MDN-Joint (Patro & Namboodiri, 2019) State-of-the-art model on $VQA1.0$, based on joint usage of caption and image information.

MC-SBN (Patro et al., 2020) State-of-the-art on VQG_{COCO}. The model jointly leverages on multiple cues (the image, place information, caption, tags) to generate questions.

4.3. Metrics

We report the following metrics for all experiments, consistently with previous works:

BLEU (Papineni et al., 2002) A precision-oriented metric, originally proposed to evaluate machine translation. It is based on the counts of overlapping n-grams between the generated sequences and the human references.

ROUGE (Lin, 2004) The recall-oriented counterpart to BLEU metrics, again based on n-gram overlaps.

METEOR (Banerjee & Lavie, 2005) The harmonic mean between precision and recall w.r.t. unigrams. As opposed to the other metrics, it also accounts for stemming and synonymy matching.

[3]Publicly available at https://www.microsoft.com/en-us/download/details.aspx?id=53670

[4]Publicly available at https://visualqa.org/vqa_v1_download.html

[5]Publicly available at https://github.com/karpathy/neuraltalk2

		BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L	METEOR	CIDEr
Sample		38.8	-	-	-	34.2	12.7	13.3
Max		59.4	-	-	-	49.3	17.8	33.1
MDN-Joint		65.1	-	-	-	52.0	22.7	33.1
Caption only	Step 1	75.41	56.49	43.26	32.28	66.18	26.51	43.56
Image only	Step 2 (freeze)	73.62	53.54	39.37	27.44	64.34	24.36	29.58
Image only	Step 2 (unfreeze)	73.97	55.07	42.20	31.76	65.70	26.36	41.43
Image + Caption	Step 3	75.59	**56.88**	**43.96**	**33.35**	**66.71**	**26.76**	**44.99**
Image + Caption	Step 3 (from scratch)	**75.84**	56.42	43.53	32.85	66.30	25.92	38.81

Table 1. Quantitative VQG results on $VQA1.0$. We report results from previous works in the upper block, and those obtained by our proposed models in the bottom block.

CIDEr (Vedantam et al., 2015) Originally designed for Image Captioning, it uses human consensus among the multiple references, favoring rare words and penalizing frequent words. This feature is particularly relevant for our task, as the automatically generated questions often follow similar patterns such as "What is the [...] ?". Indeed, we verify experimentally (cf Table 1 and Table 2) that the CIDEr metric is the most discriminant in our quantitative results.

4.4. Implementation details

All models are implemented in PyText (Aly et al., 2018). For all our experiments we used a single NVIDIA RTX 2080 Ti GPU, a batch size of 128 and 5 epochs. We used the Adam optimizer with the recommended parameters for BERT: learning rate is set at $2e^{-5}$ with a warmup of 0.1. The most computationally expensive experiment is the step 3 described above: for this model, completion of one epoch demands 30 seconds and 2 minutes for VQG_{COCO} and VQA datasets, respectively. Metrics were computed using the Python package released by Du et al. (2017).[6]

5. Results

In Table 1, we report quantitative results for the VQG task on $VQA1.0$. The *Caption only* model already shows strong improvements for all metrics over state-of-the-art models. For this text-only model, the impressive performance can mostly be attributed to BERT, demonstrating once again the benefits obtained using pre-trained language models. In our second step (*Image only*), the BERT Θ parameters are frozen and only those of the cross-modal projection matrix W are learned. Despite using a simple linear layer, the model is found to perform well, generating relevant questions given only visual inputs.

This suggests that the conceptual representations encoded in pre-trained language models such as BERT can effectively be used beyond text. Further, we report an additional *Image only* experiment, this time unfreezing the BERT parameters Θ – see *Step 2 (unfreeze)* in Table 1. As could be expected, since the model is allowed more flexibility, the performance is found to further improve.

Finally, in our third step (*Image + Caption*), we obtain the highest scores, for all metrics. This indicates that the model is able to effectively leverage the combination of textual and visual inputs. Indeed, complementary information from both modalities can be exploited by the self-attention mechanism, making visual and textual tokens interact to generate the output sequences. Again, we additionally report the results obtained bypassing the intermediate steps 1 and 2: for the model denoted as *Step 3 (from scratch)* (last row of Table 1), Θ parameters are initialized with the original weights from pre-trained BERT, while the W matrix is randomly initialized. Under this experimental condition, we observe lower performances, a finding that consolidates the importance of the multi-step training procedure we adopted.

In Table 2, we report quantitative VQG results on VQG_{COCO}. These are globally consistent with the ones above for $VQA1.0$. However, we observe two main differences. First, a bigger relative improvement over the state-of-the-art. As the efficacy of pre-trained models is boosted in small-data scenarios (Radford et al., 2018), this difference can be explained by the smaller size of VQG_{COCO}. Second, we note that the *Caption only* model globally outperforms all other models, especially on the discriminant CIDEr metric. This can be explained by the fact that, in VQG_{COCO}, the captions are human-written (whereas they are automatically generated for $VQA1.0$) and, thus, of higher quality; moreover, the smaller size of the dataset could play a role hindering the ability to adapt to the visual modality. Nonetheless, the strong performances obtained for *Step 2* compared to the baselines highlight the effectiveness of our method to learn a cross-modal projection even with a relatively small number of training images.

[6] https://github.com/xinyadu/nqg/tree/master/qgevalcap

		BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L	METEOR	CIDEr
MDN-Joint		36.0	24.9	16.8	10.4	41.8	23.4	50.7
MC-SBN		40.7	-	-	-	-	22.6	-
Caption only	Step 1	**74.58**	54.94	43.33	34.36	64.09	**29.76**	**77.70**
Image only	Step 2 (freeze)	69.57	49.93	38.23	29.54	61.01	27.03	57.38
Image only	Step 2 (unfreeze)	74.34	**55.26**	**43.47**	**34.41**	**64.63**	29.17	72.18
Image + Caption	Step 3	70.96	50.83	39.20	30.29	61.87	27.65	62.77
Image + Caption	Step 3 (from scratch)	64.18	42.88	30.19	20.14	56.99	23.32	30.99
Human Performance		*86*	-	-	-	-	*60*	-

Table 2. Quantitative VQG results on VQG_{COCO}. We report results from previous works in the upper block, and those obtained by the our proposed models in the middle block. Human Performance is taken from Mostafazadeh et al. (2016).

	Read.	Caption Rel.	Image Rel.
Caption only	**4.9**	**4.72***	4.25*
Image only	4.77	3.87	4.32*
Image + Caption	4.89	4.06*	**4.69***
Human	*4.83*	*3.64*	*4.9*

Table 3. Human evaluation results for three criterions: *readability*, *caption relevance* and *image relevance*. Two-tailed t-test results are reported in comparison to "Human" (*: $p < 0.05$).

Human Evaluation To get more in-depth understanding of our models, we report human assessment results in Table 3. We randomly sampled 50 images from the test set of $VQA1.0$. Each image is paired with its caption, the human-written question used as ground-truth, and the output for our three models: *Caption only*, *Image only* and *Image+Caption*. We asked 3 human annotators to assess the quality of each question using a Likert scale ranging from 1 to 5, for the following criteria: *readability*, measuring how well-written the question is; *caption relevance*, how relevant the question is w.r.t. to the caption; and, *image relevance*, how relevant the question is toward the image. For caption and image relevance, the annotators were presented with only the caption and only the image, respectively.

We observe that all evaluated models produce well-written sentences, as *readability* does not significantly differ compared to human's questions. Unsurprisingly, the *Caption only* model shows a higher score for *caption relevance*, while the relatively lower *image relevance* score can be explained by the automatically generated and thus imperfect captions in the $VQA1.0$ dataset. Comparatively, the *Image only* model obtains lower *caption relevance* and higher *image relevance* scores; this indicates that the cross modal projection is sufficient to bridge modalities, allowing BERT to generate relevant questions toward the image. Finally, the *Image + Caption* model obtains the best *image relevance* among our models, consistently the quantitative results reported in Tables 1 and 2.

6. Model Discussion

What does the model look at? To interpret the behavior of attention-based models, it is useful to look at which tokens are given higher attention (Clark et al., 2019). In Figure 2, we present two images A and B, along with their captions and the three generated questions corresponding to our three experimental setups (*Caption only*, *Image only* and *Image + Caption*). For this analysis, we average the attention vectors of all the heads in the last layer, and highlight the textual and visual tokens most attended by the models.

For both images, the *Caption only* model attends to salient words in the caption. The *Image only* model remains at least as much relevant: on image A, it generates a question about a table (with an unclear attention). Interestingly, for image B, the *Image only* model corrects a mistake from step 1: it is a *woman* holding an umbrella rather than a *man*, and the attention is indeed focused on the woman in the image. Finally, the *Image + Caption* model is able to generate fitting questions about the image, with relatively little relevance to the caption: for image A, *Image + Caption* the model generates "What time is it?" while paying attention to the clock; for image B, *Image + Caption* generates "What is the color of the umbrella ?", focusing the attention on the umbrella. The captions of either samples include no mentions of clocks or umbrellas, further indicating effective alignment between visual and textual representations.

Cross-modal alignment We carry out an additional experiment to analyze the text/vision alignment for each model. Figure 3 shows the *cross-modal* similarity X_{sim} for different model scenarios, computed at each BERT-base layer from 1 to 12. We define the cross-modal similarity X_{sim} as the cosine similarity between the vector representations of both modalities. These vectors are the two continuous space representations from a model when given as input either *i)* an image, or *ii)* its corresponding caption. We represent these captions and images vectors with the special BERT token `[CLS]`, following previous works (Reif et al., 2019) where `[CLS]` is used to represent the entire sequence.

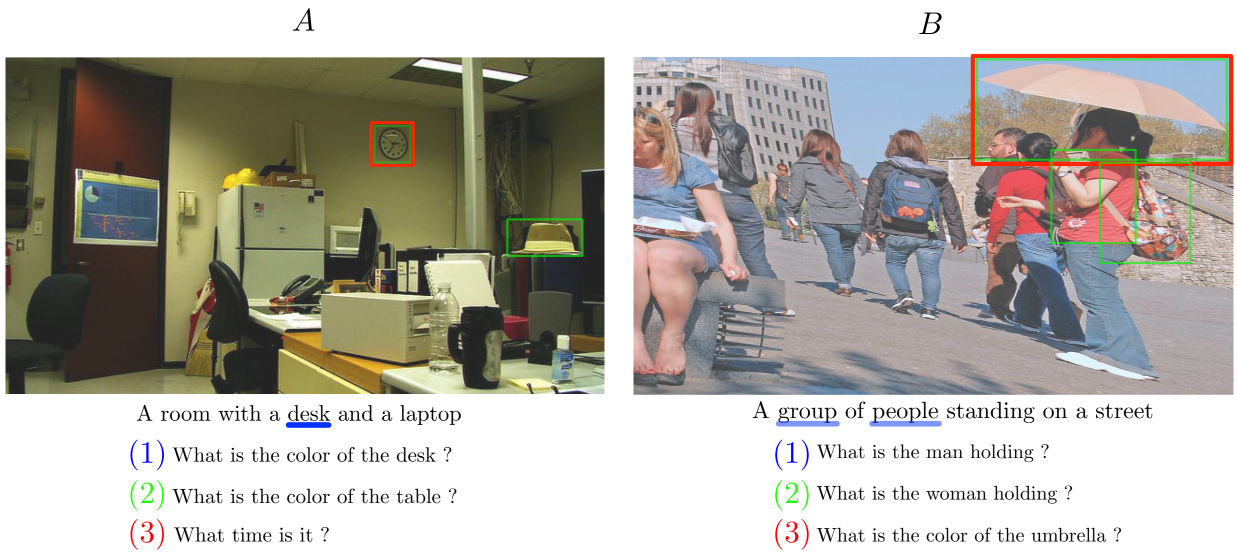

Figure 2. Qualitative Analysis. We show the outputs of the three steps of our model, using two samples from the $VQA1.0$ test set. *1)* Caption only; *2)* Image only; *3)* Image + Caption. Words and object regions with maximum attention are underlined and marked, respectively. Color intensity is proportional to attention.

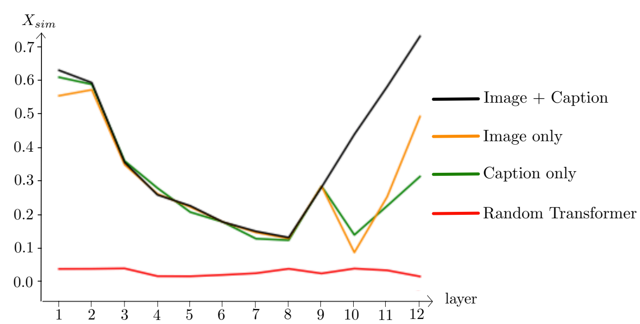

Figure 3. Cross-modal similarity X_{sim} between images in VQG_{COCO} and corresponding captions at each BERT encoding layer. Captions and images are embedded here using the `[CLS]` special token.

The reported values correspond to the average cross-modal similarity calculated for all the examples of VQG_{COCO} test set. In addition to the setups described in Section 4 (*Caption-only*, *Image-only* and *Image + Caption*), we also report X_{sim} for *Random Transformer*, a BERT architecture with random weights. As expected, its X_{sim} is close to zero.

All the other models are based on BERT. As suggested by Tenney et al. (2019), the first layers in BERT tend to encode lower-level language information. This might explain why the models show similar X_{sim} scores up to the 9th layer, and diverge afterwards: the weights for those layers remain very similar between our fine-tuned models.

For the last layer ($l = 12$), we observe that *Caption only* $<$ *Image only* $<$ *Image + Caption*. The *Caption only* model has never seen images during training, and therefore is not able to encode semantic information given only images as input. Still, its reported $X_{sim} > 0$ can be attributed to the fact that, when fine-tuned on VQG during Step 1, *BERT-gen* encodes task-specific information in the `[CLS]` token embedding (*e.g.* a question ends with a "?" and often begins with "What/Where/Who"). *Image only* $>$ *Caption only* can be explained by the learning of the cross-modal projection W. However, since BERT is not fine-tuned, the model learns a "contortion" allowing it to align text and vision. Finally, *Image + Caption* $>$ *Image only* can be attributed to BERT fine-tuning, contributing to an increase in the observed gap, and its emergence in earlier layers.

7. Conclusion and Perspectives

We investigated whether the abstractions encoded in a pre-trained BERT model can generalize beyond text. We proposed *BERT-gen*, a novel methodology that allows to directly generate text from *out-of-the-box* pre-trained encoders, either in mono- or multi- modal setups. Moreover, we applied *BERT-gen* to Visual Question Generation, obtaining state-of-the-art results on two established datasets. We showed how a simple linear projection is sufficient to effectively align visual and textual representations.

In future works, we plan to extend *BERT-gen* to other modalities, such as audio or video, exploring the potential interactions that can emerge in scenarios where more than two modalities are present.

References

Alberti, C., Ling, J., Collins, M., and Reitter, D. Fusion of detected objects in text for visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 2131–2140, 2019. doi: 10.18653/v1/D19-1219.

Aly, A., Lakhotia, K., Zhao, S., Mohit, M., Oguz, B., Arora, A., Gupta, S., Dewan, C., Nelson-Lindall, S., and Shah, R. Pytext: A seamless path from nlp research to production. *arXiv preprint arXiv:1812.08729*, 2018.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6077–6086, 2018. doi: 10.1109/CVPR.2018.00636.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Artetxe, M., Ruder, S., and Yogatama, D. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856, 2019.

Baltrusaitis, T., Ahuja, C., and Morency, L. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Banerjee, S. and Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Baroni, M. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13, 2016. doi: 10.1111/lnc3.12170.

Bordes, P., Zablocki, E., Soulier, L., Piwowarski, B., and Gallinari, p. Incorporating visual semantics into sentence representations within a grounded space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 696707, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1064.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Collell, G., Zhang, T., and Moens, M. Imagined visual representations as multimodal embeddings. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pp. 4378–4384, 2017.

Das, A., Agrawal, H., Zitnick, L., Parikh, D., and Batra, D. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. doi: 10.1016/j.cviu.2017.10.001.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Lee, S., Moura, J. M. F., Parikh, D., and Batra, D. Visual dialog. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(5): 1242–1256, 2019. doi: 10.1109/TPAMI.2018.2828437.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H.-W. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pp. 13042–13054, 2019.

Du, X., Shao, J., and Cardie, C. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1342–1352, 2017. doi: 10.18653/v1/P17-1123.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., and Xu, W. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2296–2304, 2015.

Jain, U., Zhang, Z., and Schwing, A. G. Creativity: Generating diverse questions using variational autoencoders. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5415–5424, 2017. doi: 10.1109/CVPR.2017.575.

Karpathy, A. and Li, F. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3128–3137, 2015. doi: 10.1109/CVPR.2015.7298932.

Kiela, D., Conneau, A., Jabri, A., and Nickel, M. Learning visually grounded sentence representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 408–418, 2018.

Kiela, D., Bhooshan, S., Firooz, H., and Testuggine, D. Supervised multimodal bitransformers for classifying images and text. *CoRR*, abs/1909.02950, 2019.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Urtasun, R., Torralba, A., and Fidler, S. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 3294–3302, 2015.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. doi: 10.1007/s11263-016-0981-7.

Lazaridou, A., Pham, N. T., and Baroni, M. Combining language and vision with a multimodal skip-gram model. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pp. 153–163, 2015.

Le, Q. V. and Mikolov, T. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pp. 1188–1196, 2014.

Li, G., Duan, N., Fang, Y., Jiang, D., and Zhou, M. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *CoRR*, abs/1908.06066, 2019a.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C., and Chang, K. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019b.

Li, Y., Duan, N., Zhou, B., Chu, X., Ouyang, W., Wang, X., and Zhou, M. Visual question generation as dual task of visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6116–6124, 2018. doi: 10.1109/CVPR.2018.00640.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Liu, Y. and Lapata, M. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3721–3731, 2019.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, abs/1908.02265, 2019.

Ma, L., Lu, Z., and Li, H. Learning to answer questions from image using convolutional neural network. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pp. 3567–3573, 2016.

Malinowski, M. and Fritz, M. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 1682–1690, 2014.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 3111–3119, 2013.

Misra, I., Zitnick, C. L., Mitchell, M., and Girshick, R. B. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *2016 IEEE Conference on Computer Vision and Pattern Recognition,* *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2930–2939, 2016. doi: 10.1109/CVPR.2016.320.

Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., and Vanderwende, L. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

Patro, B. N. and Namboodiri, V. P. Deep exemplar networks for VQA and VQG. *CoRR*, abs/1912.09551, 2019.

Patro, B. N., Kumar, S., Kurmi, V. K., and Namboodiri, V. P. Multimodal differential network for visual question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 4002–4012, 2018.

Patro, B. N., Kurmi, V. K., Kumar, S., and Namboodiri, V. P. Deep bayesian network for visual question generation, 2020.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 2227–2237, 2018.

Pires, T., Schlinger, E., and Garrette, D. How multilingual is multilingual bert? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4996–5001, 2019.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. 2018.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 2383–2392, 2016.

Reif, E., Yuan, A., Wattenberg, M., Viégas, F. B., Coenen, A., Pearce, A., and Kim, B. Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 8592–8600, 2019.

Ren, M., Kiros, R., and Zemel, R. S. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2953–2961, 2015.

Ren, S., He, K., Girshick, R. B., and Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6): 1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031.

Rus, V., Wyse, B., Piwek, P., Lintean, M. C., Stoyanchev, S., and Moldovan, C. The first question generation shared task evaluation challenge. In *INLG 2010 - Proceedings of the Sixth International Natural Language Generation Conference, July 7-9, 2010, Trim, Co. Meath, Ireland*, 2010.

Scialom, T., Piwowarski, B., and Staiano, J. Self-attention architectures for answer-agnostic neural question generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 6027–6032, 2019.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2:207–218, 2014.

Song, L., Wang, Z., and Hamza, W. A unified query-based generative model for question generation and question answering. *CoRR*, abs/1709.01058, 2017.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. VL-BERT: pre-training of generic visual-linguistic representations. *CoRR*, abs/1908.08530, 2019.

Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C. Videobert: A joint model for video and language representation learning. *CoRR*, abs/1904.01766, 2019a.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., and Wang, H. Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*, 2019b.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Tan, H. and Bansal, M. LXMERT: learning cross-modality encoder representations from transformers. *CoRR*, abs/1908.07490, 2019.

Tenney, I., Das, D., and Pavlick, E. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3156–3164, 2015. doi: 10.1109/CVPR.2015.7298935.

Wu, S. and Dredze, M. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *CoRR*, abs/1904.09077, 2019.

Xu, H. and Saenko, K. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, pp. 451–466, 2016. doi: 10.1007/978-3-319-46478-7_28.

Yan, X., Yang, J., Sohn, K., and Lee, H. Attribute2image: Conditional image generation from visual attributes. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pp. 776–791, 2016. doi: 10.1007/978-3-319-46493-0_47.

Yang, Y., Li, Y., Fermüller, C., and Aloimonos, Y. Neural self talk: Image understanding via continuous questioning and answering. *CoRR*, abs/1512.03460, 2015.

Zablocki, E., Bordes, P., Soulier, L., Piwowarski, B., and Gallinari, P. Context-aware zero-shot learning for object recognition. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 7292–7303, 2019.

Zhang, S., Qu, L., You, S., Yang, Z., and Zhang, J. Automatic generation of grounded visual questions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 4235–4243, 2017. doi: 10.24963/ijcai.2017/592.

Zhao, Y., Ni, X., Ding, Y., and Ke, Q. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 3901–3910, 2018.

Zhou, B., Lapedriza, À., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6): 1452–1464, 2018a. doi: 10.1109/TPAMI.2017.2723009.

Zhou, L., Louis, N., and Corso, J. J. Weakly-supervised video object grounding from text by loss weighting and object interaction. *arXiv preprint arXiv:1805.02834*, 2018b.

Zhou, Q., Yang, N., Wei, F., Tan, C., Bao, H., and Zhou, M. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*, pp. 662–671, 2017. doi: 10.1007/978-3-319-73618-1_56.