What BERT Sees: Cross-Modal Transfer for
Visual Question Generation

Thomas Scialom[*‡*]   **Patrick Bordes**[‡*]   **Paul-Alexis Dray**[*]
Jacopo Staiano[*]   **Patrick Gallinari**[‡]

‡ Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
* reciTAL, Paris, France
{thomas, paul-alexis, jacopo}@recital.ai
{patrick.bordes, patrick.Gallinari}@lip6.fr

Abstract

Pre-trained language models have recently contributed to significant advances in NLP tasks. Recently, multi-modal versions of BERT have been developed, using heavy pre-training relying on vast corpora of aligned textual and image data, primarily applied to classification tasks such as VQA. In this paper, we are interested in evaluating the visual capabilities of BERT *out-of-the-box*, by avoiding pre-training made on supplementary data. We choose to study Visual Question Generation, a task of great interest for grounded dialog, that enables to study the impact of each modality (as input can be visual and/or textual). Moreover, the generation aspect of the task requires an adaptation since BERT is primarily designed as an encoder. We introduce *BERT-gen*, a BERT-based architecture for text generation, able to leverage on either mono- or multi- modal representations. The results reported under different configurations indicate an innate capacity for *BERT-gen* to adapt to multi-modal data and text generation, even with few data available, avoiding expensive pre-training. The proposed model obtains substantial improvements over the state-of-the-art on two established VQG datasets.

1 Introduction

In Artificial Intelligence, several works have investigated the longstanding research question of whether textual representations encode some sort of visual information. This has been done primarily for word embeddings, e.g, by applying them to Zero-Shot Learning (Zablocki et al., 2019), or sentence embeddings, e.g, by applying them to Image Captioning (Socher et al., 2014). In this paper, we are interested in evaluating the visual capacities of *pre-trained language models*; in our case, BERT (Devlin et al., 2019).

To do so, we choose the Visual Question Generation (VQG) (Mostafazadeh et al., 2016) task, for the following reasons. First, 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. Second, 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. Moreover, 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: this is of particular interest to analyze the impact of each modality. Finally, VQG relies on textual generation, which is challenging since BERT is not primarily designed for generation.

BERT-based Multi-Modal Language Models have been proposed (Li et al., 2019; Su et al., 2019) to tackle multi-modal tasks, relying on heavy pre-training and large corpora of aligned textual and visual data. From these works, it is left to explore whether the cross-modal capacities come from the pre-traning, or are to some extent already encoded in BERT's representations.

It has recently been shown that BERT can generalize to another *language*, with great results, in a zero-shot manner (Artetxe et al., 2019), i.e. without supervision between languages. In preliminary experiments, we extended this work to another *modality*: we found out that, in VQG, without any supervision between the images and the questions, the cross-modal alignment was not successfully learnt. This discrepancy between multi-*lingual* and multi-*modal* results might find its root cause in the

---

intrinsic semantic difference between textual and visual modalities (Bruni et al., 2014). Nonetheless, we hypothesize that BERT contains some abstractions that generalize across modalities. If so, it may transfer knowledge to the visual modality even with few training data rather than expensive pre-training and complex architectures.

Thus, in contrast with latter Multi-Modal BERT approaches, *we explicitly avoid* using the following complex mechanisms: (1) *Multi-modal supervision*: we do not exploit explicit supervision between images and captions through a pre-training step; (2) *Image-specific losses*: specific pre-training losses such as Masked RoI Classification with Linguistic Clues (Su et al., 2019) or sentence-image prediction (Li et al., 2019); (3) *Non-linearities*: we explore a scenario in which the only learnable parameters, for aligning image representations to BERT, are those of a simple linear projection layer.

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 report state-of-the art results on the VQG task; (3) we show that the semantic abstractions encoded in pre-trained BERT can generalize to another modality without pre-training of any sort; (4) we provide extensive ablations and qualitative analyses to interpret the behavior of *BERT-gen* under different configurations (mono- or multi- modal).

2 Related Work

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. The first attempt was VideoBERT (Sun et al., 2019a), a joint *video* and text model pre-trained on a huge corpus of YouTube videos, where the video is treated as a "visual sentence" (each frame being a "visual word") 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., 2019) and VL-BERT (Su et al., 2019). Other works, such as ViLBERT (Lu et al., 2019) and LXMERT (Tan and 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 (Zhao et al., 2018; Scialom et al., 2019). However, its visual counterpart, Visual Question Generation, has been comparatively less explored than standard well-known multi-modal tasks such as Visual Question Answering (Gao et al., 2015), Visual Dialog (Das et al., 2017), or Image Captioning (Vinyals et al., 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 and 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 VQA datasets, such as $VQA1.0$ (Antol et al., 2015), by "reversing" them (taking images
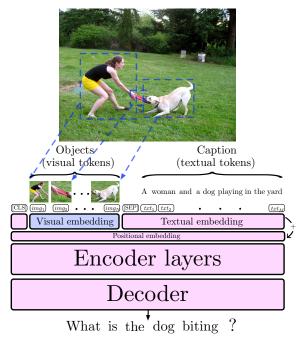
Figure 1: Model overview. Captions are encoded via BERT embeddings, while visual embeddings (blue) are obtained via a linear layer, used to project image representations to the embedding layer dimensions.

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 and 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., 2018), 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.

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 and 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, 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.

---

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

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 tokens. 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 BERT's 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 of a textual or visual nature.

3.2 *BERT-gen*: Text Generation with BERT

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

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

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.

Wang and Cho (2019) demonstrate BERT capacity to infer text, however there method does not allow to train BERT for a text generation task. To that purpose, Liu and 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 an unidirectional loss.

---

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

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 BERT-base parameters. 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:

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} Mostafazadeh et al. (2016) 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] The Visual Question Answering (VQA) (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 and Li, 2015): for fair comparison, we used exactly the same model[5] as Patro and 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 and 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 over-

---

[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
Cap. only	Step 1	75.41	56.49	43.26	32.28	66.18	26.51	43.56
Im. only	Step 2 (freeze)	73.62	53.54	39.37	27.44	64.34	24.36	29.58
Im. only	Step 2 (unfreeze)	73.97	55.07	42.20	31.76	65.70	26.36	41.43
Im. + Cap.	Step 3	75.59	**56.88**	**43.96**	**33.35**	**66.71**	**26.76**	**44.99**
Im. + Cap.	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.

lapping n-grams between the generated sequences and the human references.

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

METEOR (Banerjee and 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.

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 SOTA models. For this text-only model, the impressive performance can mostly be attributed to BERT, demonstrating once again the benefits of pre-trained language models. In our Step 2 (*Image only*), BERT's 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 experi-

---

[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	-
Cap. only	Step 1	**74.58**	54.94	43.33	34.36	64.09	**29.76**	**77.70**
Im. only	Step 2 (freeze)	69.57	49.93	38.23	29.54	61.01	27.03	57.38
Im. only	Step 2 (unfreeze)	74.34	**55.26**	**43.47**	**34.41**	**64.63**	29.17	72.18
Im. + Cap.	Step 3	70.96	50.83	39.20	30.29	61.87	27.65	62.77
Im. + Cap.	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.	Cap. Rel.	Im. 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" (*: $p < 0.05$).

mental 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.

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

A B

A room with a <u>desk</u> and a laptop

(1) What is the color of the desk ?

(2) What is the color of the table ?

(3) What time is it ?

A <u>group</u> of <u>people</u> standing on a street

(1) What is the man holding ?

(2) What is the woman holding ?

(3) What is the color of the umbrella ?

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.

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 cross-modal alignment for each version of our model.
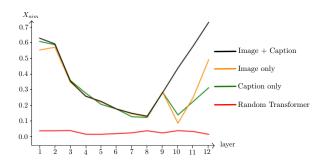

Figure 3: Cross-modal similarity X_{sim} between images in VQG_{COCO} and corresponding captions at each BERT encoding layer. Captions and images are embedded here using the [CLS] special token.

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.

Reported values are averaged over all 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. W is set at random for models where visual data has not been used (*Random Transformer*, *Caption only*).

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* comes from the learning of the cross-modal matrix 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* comes from BERT's 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 multimodal 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.

References

Ahmed Aly, Kushal Lakhotia, Shicong Zhao, Mrinal Mohit, Barlas Oguz, Abhinav Arora, Sonal Gupta, Christopher Dewan, Stef Nelson-Lindall, and Rushin Shah. 2018. Pytext: A seamless path from nlp research to production. *arXiv preprint arXiv:1812.08729*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.*, 49:1–47.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13042–13054.

Xinya Du, Junru Shao, and Claire Cardie. 2017. 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*, pages 1342–1352.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. 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*, pages 2296–2304.

Unnat Jain, Ziyu Zhang, and Alexander G. Schwing. 2017. 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*, pages 5415–5424.

Andrej Karpathy and Fei-Fei Li. 2015. 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*, pages 3128–3137.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *CoRR*, abs/1909.02950.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. 2018. 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*, pages 6116–6124.

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

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755.

Yang Liu and Mirella Lapata. 2019. 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)*, pages 3721–3731.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, abs/1908.02265.

Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross B. Girshick. 2016. 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*, pages 2930–2939.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. 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*.

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

Badri N. Patro, Vinod K. Kurmi, Sandeep Kumar, and Vinay P. Namboodiri. 2020. Deep bayesian network for visual question generation.

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

Badri Narayana Patro, Sandeep Kumar, Vinod Kumar Kurmi, and Vinay P. Namboodiri. 2018. 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*, pages 4002–4012.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viégas, Andy Coenen, Adam Pearce, and Been Kim. 2019. 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*, pages 8592–8600.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.

Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2019. 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, pages 6027–6032.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2:207–218.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: pre-training of generic visual-linguistic representations. *CoRR*, abs/1908.08530.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019a. Videobert: A joint model for video and language representation learning. *CoRR*, abs/1904.01766.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019b. Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. 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*, pages 5998–6008.

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

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. 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*, pages 3156–3164.

Alex Wang and Kyunghyun Cho. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*.

Yezhou Yang, Yi Li, Cornelia Fermüller, and Yiannis Aloimonos. 2015. Neural self talk: Image understanding via continuous questioning and answering. *CoRR*, abs/1512.03460.

Eloi Zablocki, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. 2019. 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*, pages 7292–7303.

Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. 2017. 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*, pages 4235–4243.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. 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*, pages 3901–3910.

Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464.