

# BERTGEN: Multi-task Generation through BERT

Faidon Mitzalis<sup>1</sup>, Ozan Caglayan<sup>1</sup>, Pranava Madhyastha<sup>1</sup>, Lucia Specia<sup>1,2</sup>

<sup>1</sup>Department of Computing, Imperial College London, UK

<sup>2</sup>Department of Computer Science, University of Sheffield, UK

phaedonmit@gmail.com, {o.caglayan, pranava, lspecia}@ic.ac.uk

## Abstract

We present BERTGEN, a novel generative, decoder-only model which extends BERT by fusing multimodal and multilingual pre-trained models VL-BERT and M-BERT, respectively. BERTGEN is auto-regressively trained for language generation tasks, namely image captioning, machine translation and multimodal machine translation, under a multi-task setting. With a comprehensive set of evaluations, we show that BERTGEN outperforms many strong baselines across the tasks explored. We also show BERTGEN’s ability for zero-shot language generation, where it exhibits competitive performance to supervised counterparts. Finally, we conduct ablation studies which demonstrate that BERTGEN substantially benefits from multi-tasking and effectively transfers relevant inductive biases from the pre-trained models.

## 1 Introduction

Recent work in unsupervised and self-supervised pre-training has revolutionised the field of natural language understanding (NLU), resulting in high performance ceilings across multiple tasks (Devlin et al., 2019; Yang et al., 2019; Dong et al., 2019). The recent success of language model pre-training with masked language modelling (MLM) such as BERT (Devlin et al., 2019) further paved the way for more complex approaches that combine language pre-training with images (Tan and Bansal, 2019; Su et al., 2020; Lu et al., 2020), video (Sun et al., 2019), and speech (Chuang et al., 2020). Most of these approaches follow a task-specific fine-tuning step after the model is pre-trained.

However, there has been little work on exploiting pre-trained MLMs for natural language generation (NLG) tasks. Previous work argues that the MLM objective is ill-suited for generation tasks such as machine translation (Yang et al., 2019; Rothe et al.,

2020). Recent work in this direction has predominantly investigated the use of pre-trained models to either initialise Transformer-based encoder-decoder models (Imamura and Sumita, 2019; Clinchant et al., 2019; Yang et al., 2020; Rothe et al., 2020) or to distill knowledge for sequence generation tasks (Chen et al., 2020).

In this work, we present BERTGEN, which extends BERT in a generative setting (§ 2.1). This results in a single generator – without a separation between the encoder and the decoder – capable of consuming multiple input modalities and generating in multiple languages. The latter features are achieved by transferring knowledge from state-of-the-art pre-trained models, namely VL-BERT (Su et al., 2020) and multilingual BERT (M-BERT) (Devlin et al., 2019). We train BERTGEN on various tasks, including image captioning, machine translation and multimodal machine translation, and datasets in four different languages (§ 2.2).

Based on a number of experiments, our findings (§ 3) show that BERTGEN (i) is surprisingly versatile as it is capable of describing images and performing translation in unimodal and multimodal settings, across all languages, (ii) generalises well across zero-shot image captioning, multimodal machine translation, and out-of-domain news translation tasks, and finally (iii) is parameter efficient when compared to state-of-the-art models for each of the tasks combined together.

## 2 Method

In this section, we describe BERTGEN and the tasks we explore. We then detail the baselines and SoTA systems that we compare against.

### 2.1 Model

This section details the main aspects of BERTGEN that distinguish it from the existing work on vision & language pre-training.

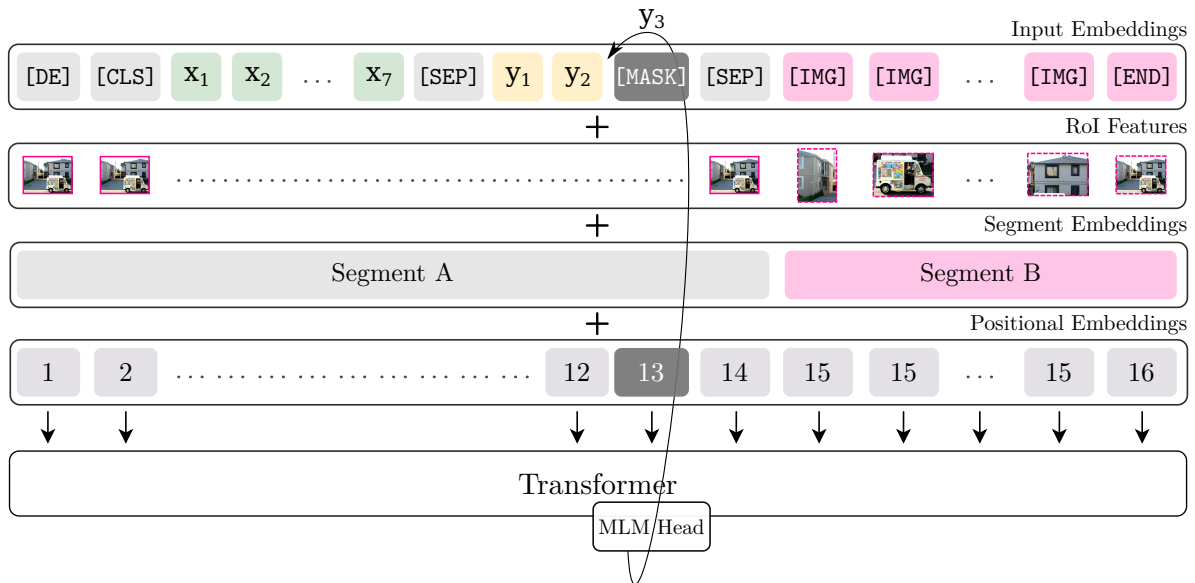


Figure 1: A view of the BERTGEN model during the training of an MMT sample: solid and dashed borders around the images represent full-image features and regional features, respectively. At **test time**, the most likely token  $y_3 = \text{argmax}(P(y_t|\mathbf{x}, \mathbf{v}, \mathbf{y}_{<t}))$  is placed back into the sequence and the [MASK] token is shifted right by one.

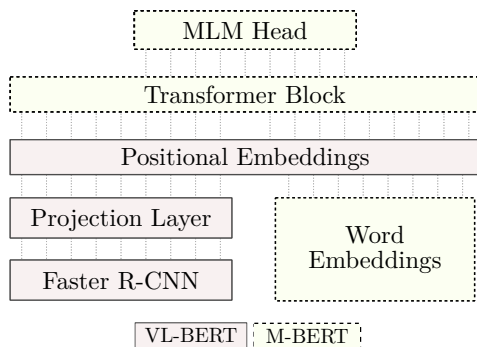


Figure 2: Hybrid initialisation: the **solid** and **dashed** blocks are transferred from pre-trained VL-BERT and M-BERT checkpoints, respectively.

**Initialisation.** We take advantage of the previous successes in large-scale pre-training and propose a hybrid initialisation for BERTGEN (Figure 2). This involves using the VL-BERT (Su et al., 2020) checkpoint and initialising the word embeddings, the Transformer weights and the MLM head with M-BERT (Devlin et al., 2019). We conjecture that this primes BERTGEN to be aware of the visual modality and of multiple languages. This is simply due to VL-BERT being pre-trained on English monolingual and image captioning corpora, as well as M-BERT offering a 119K WordPiece vocabulary, trained on the entire Wikipedia in 104 languages<sup>1</sup>.

<sup>1</sup>We adopt the ‘BERT-Base, Multilingual Cased’ version from the Transformers toolkit (Wolf et al., 2020).

**Input configuration.** While BERTGEN is potentially capable of modeling a variety of generative tasks, we focus on three particular tasks, namely machine translation (MT), multimodal MT (MMT) and image captioning (IC). Therefore, depending on the task, the input configuration of the model may change during both training and testing. To clarify further, let us first denote a sequence of embeddings representing a source sentence by  $\mathbf{x}^{(i)} = [x_1^{(i)}, \dots, x_m^{(i)}]$ , its target translation by  $\mathbf{y}^{(i)} = [y_1^{(i)}, \dots, y_n^{(i)}]$ , and a collection of  $k$  regional visual features extracted from an associated image by  $\mathbf{v}^{(i)} = [v_1^{(i)}, \dots, v_k^{(i)}]$ . Figure 1 depicts BERTGEN when processing a sample from the MMT task. This task’s input configuration is a triplet that involves all the three sequences i.e.  $\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}, \mathbf{v}^{(i)}\}$ . Using this notation, the MT and IC tasks’ configurations would correspond to  $\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}$  and  $\{\mathbf{v}^{(i)}, \mathbf{y}^{(i)}\}$ , respectively.

**Visual embeddings.** We follow VL-BERT and represent images as a collection of  $k$  features  $\mathbf{v}^{(i)}$  defined for regions of interest (RoI). After pre-extracting the 2048-dimensional RoI features using the *bottom-up-top-down* object detector (Anderson et al., 2018), we keep between 10 and 100 (i.e.  $k \in [10, 100]$ ) of them depending on the confidence score. The final visual embedding for an RoI is obtained by summing its feature vector and its geometric embedding (i.e. the projection of the

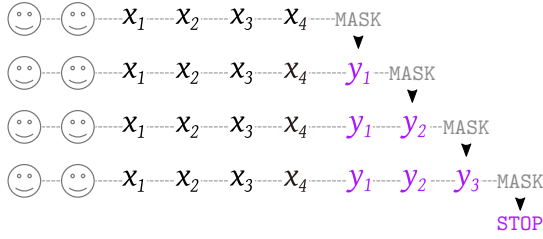


Figure 3: A look at BERTGEN’s self-attention: the connections denote that self-attentive representations are re-computed in every step. The generation ends when STOP is predicted. The smileys refer to RoI features.

bounding box coordinates). When encoding the non-visual positions, the same RoI feature vector for the full image is repeated (see Figure 1). We note that we do **not fine-tune** the object detector during training.

**Sequence unrolling.** An important aspect of BERTGEN is that it does not explicitly distinguish between the encoder and the decoder blocks usually seen in sequence-to-sequence models. This is accomplished by formalising both encoding and generation using the MLM framework. Formally, let us consider the MMT task and define the maximum log-likelihood objective for a given triplet  $\{\mathbf{x}^{(i)}, \mathbf{v}^{(i)}, \mathbf{y}^{(i)}\}$  where the target  $\mathbf{y}^{(i)}$  has  $n$  tokens:

$$\mathcal{L}^{(i)} = \sum_{t=1}^n \log \left( P(y_t^{(i)} | \mathbf{x}^{(i)}; \mathbf{v}^{(i)}; \mathbf{y}_{<t}^{(i)}) \right) \quad (1)$$

In a typical sequence-to-sequence model, each log-probability term would be computed by a *decoder* within the forward-pass of the *same* training example. In contrast, BERTGEN explicitly unrolls the example  $n$  times, forming  $n$  new training examples. In other words, each conditional term in Equation 1 is observed independently within an epoch of training. Therefore, sequence unrolling has a data **augmentation** effect since a training corpus with  $D$  examples is approximately augmented by a factor of the average length of the target sequences. Moreover, the unified encoder-decoder formalism halves the number of parameters, making BERTGEN parameter efficient.

**Self attention.** Given that a single Transformer (Vaswani et al., 2017) performs both encoding and decoding, sequence unrolling affects **self-attention** as well (Figure 3). First, all positions attend to each other for a given unrolled example i.e. the attention is bi-directional. Second, since each unrolled case is an independent example, the

self-attentive representations of early positions are naturally **re-computed**, in contrast to typical Transformer decoders. Finally, due to how inputs/outputs are represented in a single stream and encoded through shared self-attention, BERTGEN enforces an inductive bias towards a truly multi-modal and multi-lingual representation space.

**Target language specifiers.** Finally, to select the language during generation, input sequences begin with special target language specifiers (Ha et al., 2016; Johnson et al., 2017) (Figure 1). The specifier is task-agnostic, i.e. the same specifier [DE] is used both when captioning into German and when translating into German.

**Training & hyper-parameters.** We extend<sup>2</sup> the base configuration of VL-BERT which is a Transformer with 12 self-attention layers and 12 heads. The model and feed-forward dimensions are 768 and 3072, respectively. On a single 32GB V100 GPU, one epoch (§ 3) takes approximately two days to complete as we could only fit one example per task (i.e. batch size equal to 13) into the memory<sup>3</sup>. We use AdamW optimiser (Loshchilov and Hutter, 2019) with base learning rate set to  $1.3 \times 10^{-5}$ . The learning rate is warmed up in the first 16K steps and then decays linearly. We set the weight decay to  $10^{-4}$ . During training, we let the model update the positional embeddings as BERTGEN needs to learn new positions not covered by VL-BERT pre-training. The final model has  $\sim 89.3$ M parameters excluding the word embeddings.

**Decoding.** At test time, we incrementally add the most likely prediction (i.e. greedy search) into the previously masked position (Figure 1) and shift the [MASK] token right by one. The reason we chose greedy over beam search is because the latter would make decoding much slower due to self-attentive representations being re-computed. The decoding ends when [STOP] is predicted.

## 2.2 Tasks & Systems

To evaluate BERTGEN’s generative abilities, we explore a diverse set of tasks: image captioning, text-only MT and multimodal MT. Table 1 summarises the training statistics for the various datasets we use.

<sup>2</sup><https://github.com/ImperialNLP/BertGen>

<sup>3</sup>With careful optimisation of the training code and mixed precision multi-GPU training, the training time can be substantially reduced.

Name	Type	Task	Sents	Augm.
FLICKR8K	IC	IM→TR	13,828	263K
MULTI30K	MMT	DE→EN	29,000	464K
		FR→EN		464K
		EN→FR		560K
MULTI30K	MMT	EN→DE	29,000	582K
FLICKR30K	IC	IM→EN	145,000	2.39M
		IM→DE		2.48M
IWSLT	MT	DE→EN	158,388	3.85M
		EN→DE		4.43M
IWSLT	MT	FR→EN	163,328	4.01M
		EN→FR		4.78M
SETIMES	MT	TR→EN	185,318	6.01M
		EN→TR		8.40M

Table 1: Training statistics of BERTGEN: the last column is the number of samples after sequence unrolling.

### 2.2.1 Image Captioning

Image captioning (IC) involves describing images in a specified natural language. We train BERTGEN for English, German and Turkish captioning tasks. Specifically, we use the FLICKR30K dataset (Young et al., 2014) that provides 29K training images, each with five **English** captions collected through crowd-sourcing. The validation and test sets contain approximately 1K images each. We use the MULTI30K dataset (Elliott et al., 2016), which annotates FLICKR30K images with five **German** captions. Finally, we use the TASVIRET dataset (Unal et al., 2016) which provides two **Turkish** captions for each of the 8,092 images in the FLICKR8K dataset (Rashtchian et al., 2010). Since FLICKR8K is a subset of FLICKR30K, we create a new split of TASVIRET to avoid data leakage between training and test splits. The resulting training, validation and test splits contain 6914, 543, and 543 images, respectively.

To evaluate BERTGEN’s performance on IC, we compare it against previous work with strong performance on COCO (Chen et al., 2015) and FLICKR30K. More precisely, ADAPTIVE ATTENTION (SENTINEL) (Lu et al., 2017), which uses a *sentinel* token to distinguish between visual and non-visual representations, and NEURAL BABY TALK (NBT), which follows a slot-filling approach through explicit object region information (Lu et al., 2018).

### 2.2.2 Multimodal Machine Translation

Multimodal Machine Translation (MMT) attempts to improve MT quality by incorporating information from modalities other than language (Sulubacak et al., 2020). In our case, we train BERTGEN for EN↔DE and EN↔FR MMT tasks and use the MULTI30K dataset, the main dataset for image-informed translation, which provides caption translations for FLICKR30K images in German and French. To evaluate BERTGEN on MMT tasks, we use the original 2016 test set which contains 1,000 examples.

For a comprehensive comparison with previous work, we train a SoTA recurrent MMT (Caglayan et al., 2020) *solely* on the MULTI30K dataset, which applies a secondary (visual) attention in the decoder over the RoI features i.e. the same features that are also used by BERTGEN (§ 2.1). There are two GRU (Cho et al., 2014) layers in both the encoder and the decoder and the embedding & hidden dimensions in the model are set to 200 and 320, respectively. Each model has ~5.6M parameters excluding the word embeddings.

Besides the state-of-the-art *constrained* recurrent MMT model described above, we further compare BERTGEN – which is trained on various other MT and IC corpora – to an *unconstrained* Transformer-based MMT trained on ~9M additional EN→DE sentences (Libovický, 2019)<sup>4</sup> in addition to MULTI30K.

### 2.2.3 Text-only Machine Translation

We incorporate six text-only MT tasks into our training protocol. We use EN↔DE and EN↔FR MT datasets from IWSLT’14 (Cettolo et al., 2012) which consists of TED Talks’ subtitles and their translations. We take the `prepare-iwslt14` recipe from FAIRSEQ (Ott et al., 2019) to prepare the *dev* and *test* sets. This yields an EN↔DE test set of 6,750 sentences which consists of *dev2010*, *dev2012.TEDX*, *tst2010*, *tst2011* and *tst2012*. Similarly, the EN↔FR test set consists of *dev2010*, *tst2010*, *tst2011* and *tst2012*, which amounts to 4,493 sentences.

For EN↔TR directions, we use the SETIMES2 (Tiedemann, 2012) news dataset for training. For development and test sets, we take the official WMT test sets (Bojar et al., 2018), namely, *newstest2016* and *newstest2017* as the development

<sup>4</sup>We obtained test set outputs from the author and pre-processed with M-BERT tokeniser to ensure comparability.

set (6,007 sentences), and *newstest2018* (6,000 sentences) as the test set. Both IWSLT and SETIMES2 corpora are medium-scale resources often used in MT research community, and have much harder test sets than the MMT and IC tasks, due to a significant domain shift.

Finally, for each translation direction, we train a Transformer NMT model (Vaswani et al., 2017) using the IWSLT-DE-EN recipe of the FAIRSEQ toolkit (Ott et al., 2019). This recipe has six encoders and six decoders, each equipped with 4-head self-attention layers. The model and feed-forward dimensions are set to 512 and 1024, respectively. Each model has  $\sim 31.5$ M parameters excluding the word embeddings. Since BERTGEN is a general purpose multilingual and multimodal generator, we expect it to perform in the same ballpark as these strong NMT baselines, but not necessarily be SoTA compared to novel & sophisticated NMT models, which also make use of a lot more training data.

### 3 Results and Findings

We train BERTGEN on lowercased sentences for 45 epochs, after which the overall performance on the tasks reached a plateau. We define one BERTGEN epoch as a single pass over all of the training data for the MULTI30K EN $\rightarrow$ DE MMT task and denote this task as the *reference task*. We use greedy search for all systems that we trained and merge back the *word pieces* before evaluation. We compute tokenised<sup>5</sup> BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and CIDEr (Vedantam et al., 2015) using *coco-caption*<sup>6</sup>. In what follows, we provide detailed quantitative and qualitative findings.

#### 3.1 Image Captioning

Table 2 provides an overview of BERTGEN’s image captioning performance on different test sets and languages. First of all, on **English** FLICKR30K, BERTGEN is clearly able to outperform strong captioning models (§ 2.2.1) SENTINEL (Lu et al., 2017) and NBT (Lu et al., 2018), even though they use beam search for decoding. On COCO (Chen et al., 2015), an image captioning corpus much larger and diverse than FLICKR30K, we evaluate BERTGEN on Karpathy’s test split (Karpathy and Fei-Fei, 2015) and notice that the scores are reasonable

<sup>5</sup>Since M-BERT is aggressive on splitting apostrophes and hyphens, our results may slightly differ from other work.

<sup>6</sup><https://github.com/tylin/coco-caption>

TASK	APPROACH	BL	MT	CR
F30K EN	BERTGEN	27.0	<b>23.2</b>	<b>0.587</b>
	SENTINEL <sup>‡</sup>	25.1	20.4	0.531
	NBT <sup>‡</sup>	<b>27.1</b>	21.7	0.575
COCO EN	BERTGEN	15.9	20.4	0.487
	NBT <sup>‡</sup>	<b>34.7</b>	<b>27.1</b>	<b>1.072</b>
F30K FR	BERTGEN	5.2	18.1	0.397
F8K TR	BERTGEN	8.5	14.5	0.363
F30K DE	BERTGEN	17.8	34.2	0.500


Table 2: BLEU (BL), METEOR (MT) and CIDEr (CR) scores for image captioning: gray background indicates *zero-shot* generation whereas <sup>‡</sup> denotes the systems decoded with beam search.


given that BERTGEN **is not trained** on COCO: our model lags behind NBT (w/ beam search) by 6.7 METEOR.

For *zero-shot* French captioning (F30K FR), we resort to the reference MMT translations from the MULTI30K EN $\rightarrow$ FR task, as there are no human references for French. Although this is problematic as the metrics will penalise captions that are not translations of English captions, we provide the scores to show that the **zero-shot** outputs are valid descriptions. We note that the low range of scores reported here is also due to having one reference caption instead of five references<sup>7</sup> as in FLICKR30K. Finally we report results for our custom **Turkish** split (§ 2.2.1) (F30K TR) and **German** (F30K DE). Even though there are no comparable results in the literature for these three tasks, we demonstrate through some qualitative examples that BERTGEN produces sensible outputs.

**Qualitative examples.** We now focus on a few examples to examine the multilingual image captioning ability of BERTGEN in action (Table 3). For the first image, all captions are almost the same as the image has few salient points. For the second image however, we observe much more variation across captions, in line with the complexity of the scene. We are particularly surprised by the **zero-shot French** captioning performance, a task that BERTGEN is not trained for at all. Upon manual inspection, we noticed that the captions are often short, objective gists of the images. These observations also hold for the captions generated for the

<sup>7</sup>As a reference, evaluating English captions using one reference at a time, yields 7.9 BLEU on average, compared to 27.0 BLEU in Table 2.

	
EN	a man wearing a hat and glasses.
DE	ein mann mit hut und brille. <i>a man with hat and glasses</i>
TR	şapkalı ve gözlüklü bir adam. <i>a man with a hat and glasses.</i>
FR	un homme avec un chapeau et des lunettes. <i>a man with a hat and glasses.</i>

	
EN	two men are on a rooftop working on something.
DE	zwei männer arbeiten auf einem dach. <i>two men working on a roof</i>
TR	iki binanın inşasında oturmuş, yanyana yerde duran iki kişi. <i>two people seated in the construction of two buildings, standing next to each other on the ground.</i>
FR	trois ouvriers du bâtiment construisent un toit. <i>three construction workers build a roof.</i>


	
EN	a man in a red shirt and helmet is riding a motorbike on a dirt road.
DE	ein mann fährt mit einem motorrad auf einem weg an einem fluß entlang. <i>a man rides a motorcycle on a path along a river.</i>
TR	çamurlu bir yolda motoruyla ilerlemekte olan kırmızı üstlü bir adam ve arkasındaki dağ manzarası. <i>A man in a red top riding his bike down a muddy road with a mountain landscape behind him.</i>
FR	un homme avec un casque fait du motocross. <i>a man with a helmet rides motocross.</i>

Table 3: Multilingual image captioning examples: The *italicised* sentences are Google Translate translations of DE, TR, FR sentences into English. Gray background indicates *zero-shot* outputs. The last example is from COCO while the others are from FLICKR30K.

COCO test set, as we can see in the third example. A set of additional examples in the Appendix shows that BERTGEN does not simply retrieve caption translations learned from the EN→FR task. Overall, both quantitative and qualitative results provide evidence of the utility of multimodal and multilingual initialisation as well as the efficacy of knowledge transfer across different tasks for image captioning.

MMT	APPROACH	BL	MT
EN→DE	BERTGEN <sup>★</sup>	<b>42.2</b>	<b>61.6</b>
	Libovický (2019) <sup>★‡</sup>	40.8	59.2
	Caglayan et al. (2020)	37.8	56.9
	FAIRSEQ NMT	37.5	56.1
EN→FR	BERTGEN <sup>★</sup>	<b>68.0</b>	<b>81.2</b>
	Libovický (2019) <sup>★‡</sup>	63.4	77.3
	FAIRSEQ NMT	61.5	75.5
	Caglayan et al. (2020)	61.0	75.3
DE→FR	BERTGEN <sup>★</sup>	<b>44.8</b>	<b>64.1</b>
	Caglayan et al. (2020)	43.8	62.1
	FAIRSEQ NMT	41.7	60.7
FR→DE	BERTGEN <sup>★</sup>	<b>35.1</b>	<b>56.9</b>
	FAIRSEQ NMT	<b>35.1</b>	53.6
	Caglayan et al. (2020)	33.5	53.1

Table 4: BLEU (BL) and METEOR (MT) scores for MMT: *zero-shot* systems are highlighted with gray. Systems marked with <sup>★</sup> and <sup>‡</sup> denote the use of auxiliary resources (i.e. unconstrained) and beam-search decoding, respectively.

### 3.2 Multimodal Machine Translation

Table 4 summarises BERTGEN’s performance on MMT. First of all, BERTGEN consistently outperforms the Transformer-based FAIRSEQ NMT models and the recurrent MMT (Caglayan et al., 2020) models on both the EN→DE and the EN→FR language pairs. Furthermore, BERTGEN is also substantially better than a state-of-the-art *unconstrained* MMT (Libovický, 2019) model trained on a ~6x larger parallel corpus.

**Adversarial evaluation.** Following Elliott (2018), we probe BERTGEN’s ability for integrating multiple modalities effectively. Specifically, we decode translations by shuffling {image, source caption} mappings so that the images do not correspond to the sentences to be translated. The EN→DE results showed that the incongruence leads to 1.1 and 0.9 point **drops** in BLEU and METEOR, respectively. For EN→FR, the drops are much more prominent with 3.1 and 2.3 points again for BLEU and METEOR. This indicates that the features are not ignored at all, unlike in (Caglayan et al., 2019), where they showed that sequence-to-sequence MMT models can learn to ignore the images when the linguistic signal is sufficient to perform the task.

**Zero-shot performance.** The results in Table 4 show the surprising ability of BERTGEN to perform MMT on directions unseen during training.

TASK	FAIRSEQ		BERTGEN	
	BL	MT	BL	MT
IWSLT EN→DE	27.4	47.1	<b>27.8</b>	<b>48.4</b>
IWSLT DE→EN	33.6	33.8	<b>35.6</b>	<b>34.7</b>
IWSLT EN→FR	<b>41.0</b>	59.8	40.2	<b>60.5</b>
IWSLT FR→EN	39.1	36.4	<b>40.0</b>	<b>36.8</b>
SETIMES EN→TR	<b>14.1</b>	18.9	13.5	<b>19.1</b>
SETIMES TR→EN	17.3	25.8	<b>19.0</b>	<b>26.9</b>

Table 5: Comparison of text-only MT performance of BERTGEN to each dedicated FAIRSEQ NMT system: BERTGEN outperforms single models in most cases.

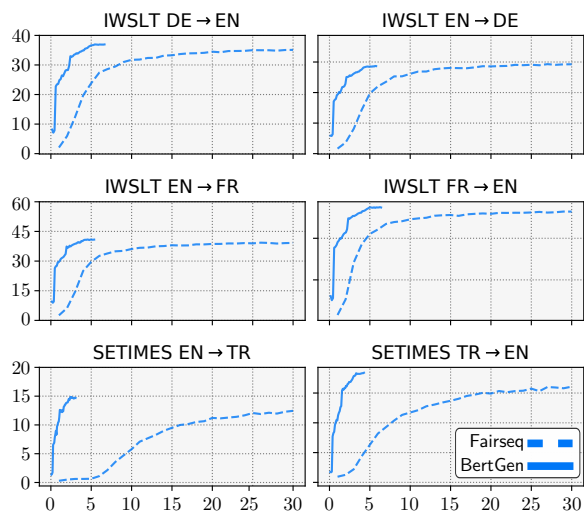


Figure 4: BERTGEN’s learning efficiency on MT: validation scores are plotted against the number of full passes completed by BERTGEN and **each** FAIRSEQ model, over the *corresponding* task’s training set. Best checkpoints’ test set performances are given in Table 5.

Moreover, the zero-shot performance surpasses strong MMT and NMT systems by up to 2 and 3.3 METEOR for DE→FR and FR→DE, respectively. Similar to the image captioning results, this demonstrates the potential of BERTGEN to generalise over a variety of language pairs and tasks.

### 3.3 Machine Translation

First, we compare BERTGEN’s performance to each task-specific FAIRSEQ system. According to Table 5, we observe that the translation quality of BERTGEN is generally superior compared to the strong FAIRSEQ systems, especially in METEOR, where BERTGEN leads in all pairs.

Second, we look at the learning efficiency by comparing the training curves between BERTGEN and **each** task-specific FAIRSEQ system (Figure 4). Here, the x axis represents how many times the spe-

	DE→FR		FR→DE	
	BL	MT	BL	MT
BERTGEN	19.6	40.5	13.1	36.7
TARTU <sup>‡</sup>	39.5	59.0	26.3	47.3
MSRA <sup>‡</sup>	46.5	64.2	38.2	56.4

Table 6: *Zero-shot* BERTGEN performance on WMT’19 test set: TARTU and MSRA systems are not *zero-shot* as they are trained on DE↔FR corpora. Systems marked with <sup>‡</sup> are beam search outputs.

cific task’s training set has been seen by the models. BERTGEN is trained for 45 reference epochs (§ 3), and this corresponds to only a few complete passes over the training sets of NMT tasks<sup>8</sup>. This is in contrast to the single-task systems that usually require a large number of epochs for convergence. We notice a general trend and observe that BERTGEN tends to outperform single-task systems usually after only a few passes over the corresponding training set. Many factors could be contributing to this observation such as sequence unrolling, multi-tasking, shared input space or relevant inductive biases transferred from M-BERT. We partly address these in the ablation studies (§ 3.4) and leave further investigation to future work.

**Zero-shot performance.** We use the DE↔FR test set from the WMT’19 *shared task on news translation* (Barrault et al., 2019) to assess the **zero-shot translation** capability of BERTGEN. This test set includes 1,701 sentences from news data regarding *European Elections*. We compare our results to two shared task systems, namely TARTU (baseline) and MSRA (state-of-the-art) (Barrault et al., 2019), after re-tokenising them accordingly with M-BERT<sup>9</sup>. Although BERTGEN is expected to obtain lower scores than the dedicated WMT systems due to the domain mismatch of the test set, we consider both the quantitative (Table 6) and the qualitative results (Table 7) extremely encouraging.

### 3.4 Ablation Studies

#### 3.4.1 Impact of initialisation

We train *single-task* MMT systems on the MULTI30K EN→DE language pair. Specifically, we begin with a baseline system which is initialised with **random** weights. We then train a second baseline where only the **visual** processing layers are

<sup>8</sup>For example, only  $\sim 3$  passes over SETIMES EN→TR.

<sup>9</sup>TARTU is the baseline and MSRA is the best performing system for the shared task

BERTGEN:	la décision est tombée au 70ème anniversaire de ma femme. <i>the decision fell on my wife's 70th birthday.</i>
WMT REF	la décision est tombée le jour du 70ème anniversaire de ma femme. <i>the decision fell on my wife's 70th birthday.</i>
BERTGEN:	en espagne, on s'est malheureusement habitué à une rôle double et passive. <i>in spain, we unfortunately got used to a <b>double</b> and passive role.</i>
WMT REF	en espagne, on s'est malheureusement habitué à un rôle secondaire, passif. <i>in spain, we unfortunately got used to a <b>secondary</b>, passive role.</i>
BERTGEN:	pas parce que le président du fdp a dit quelque chose qu' ils ont défaillant leur vote. <i>not because the fdp president said something that they <b>messed their vote.</b></i>
WMT REF	ce n' est pas parce que le président fédéral du fdp a dit quelque chose qu' ils ont refusé d' approuver. <i>it is not because the federal president of the fdp said something that they <b>refused to approve.</b></i>

Table 7: Zero-shot DE→FR translations on WMT'19 test set. The *italicised* sentences are Google Translate translations of French sentences into English. **Bold** indicates important differences between BERTGEN and references.

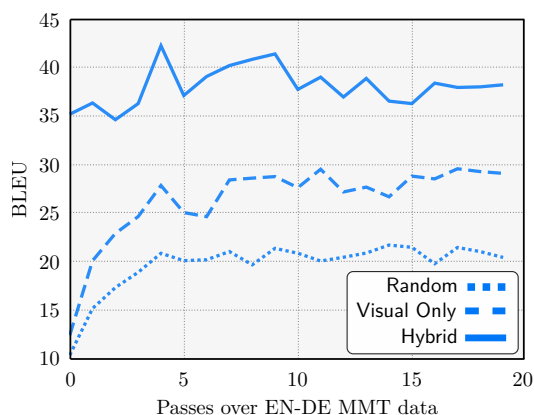


Figure 5: Validation scores on MULTI30K EN→DE MMT for the **initialisation** ablation: Hybrid initialisation is the most beneficial strategy for BERTGEN.

transferred from VL-BERT. Finally, we train a third baseline that is initialised similar to BERTGEN, i.e. using the **hybrid** initialisation (§ 2.1). Figure 5 compares the validation BLEU scores of these three systems. We observe that the benefits of knowledge transfer from pre-trained models are incrementally positive, however, BERTGEN’s hybrid initialisation outperforms the other two ablations.

### 3.4.2 Impact of multi-task training

We now remove the multi-tasking aspect from BERTGEN to investigate the extent to which the performance improvements are related to other tasks. Similar to § 3.4.1, we focus on the MULTI30K EN→DE MMT task and train a single-task, *hybrid-initialised* BERTGEN. Figure 6 compares the validation BLEU scores obtained by the default BERTGEN and the single-task variant. We observe that BERTGEN benefits from multi-task training and, more importantly, does not seem to exhibit patterns

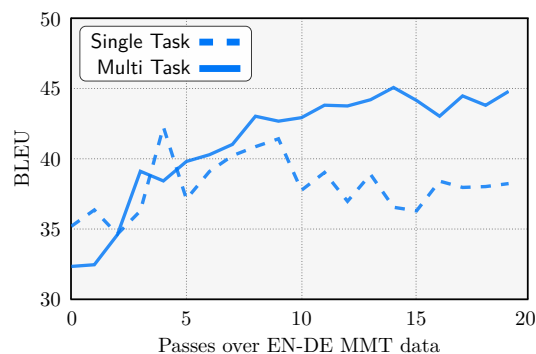


Figure 6: Validation scores on MULTI30K EN→DE MMT for the **multi-tasking** ablation: The default multi-task BERTGEN outperforms the single-task one.

of catastrophic forgetting (French, 1999). Based on these observations, we expect similar model behavior to hold for other tasks.

## 4 Related Work

### 4.1 Multimodal multilingual pre-training

Research in NLP and related fields has been increasingly focusing on transfer learning approaches where a model is first pre-trained on a data-rich task, and then transferred to downstream tasks (McCann et al., 2017; Peters et al., 2018; Devlin et al., 2019). This framework presumably allows the model to capture useful inductive biases that generalise to a variety of NLP tasks, often after performing a task-specific fine-tuning (Raffel et al., 2020). Of these, the most relevant studies to our work are BERT (Devlin et al., 2019) and its multilingual version M-BERT, which pre-train a Transformer (Vaswani et al., 2017) on large monolingual corpora using the masked language modelling (MLM) objective.



Recent research has also attempted to combine linguistic inputs with other modalities such as vision and speech, to achieve a grounded understanding of meaning. Successful approaches including LXMERT (Tan and Bansal, 2019), VL-BERT (Su et al., 2020) and others (Lu et al., 2019; Li et al., 2020a,b) achieve this by combining BERT’s MLM objective with auxiliary tasks such as masked region classification and image sentence matching, and pre-train their model on large-scale image captioning corpora (Chen et al., 2015; Sharma et al., 2018). Similarly, SpeechBERT extends BERT by jointly training on speech and text data (Chuang et al., 2020). Although SoTA results are reported by these approaches, they focus on unimodal and multimodal natural language understanding (NLU) tasks, with a strong emphasis in English. The backbone of BERTGEN combines VL-BERT (Su et al., 2020) with M-BERT (Devlin et al., 2019) to realise *a multilingual and multimodal generator* that can be used for a diverse set of generative tasks and languages rather than NLU tasks.

#### 4.2 Pre-training for generative tasks

Previous work has studied how to benefit from pre-trained BERT models in generative tasks such as NMT (Imamura and Sumita, 2019; Clinchant et al., 2019; Zhu et al., 2020). BERTGEN differs from these as it is not fine-tuned for a particular MT corpus and it exhibits multi-lingual and multi-modal properties for general purpose generation.

Another related branch of work explores pre-training strategies specific to sequence-to-sequence tasks. This includes MASS (Song et al., 2019), which exploits an encoder-decoder framework with the MLM objective for task-specific generative pre-training and UniLM (Dong et al., 2019), which introduces uni-directional, bi-directional and sequence-to-sequence LM objectives by carefully adjusting the self-attention masks during training. Zhou et al. (2020) extend UniLM to vision & language pre-training using Conceptual Captions (Sharma et al., 2018) as the pre-training dataset. However, these models require a further fine-tuning step for generative tasks, unlike BERTGEN that is trained only *once*.

#### 4.3 Multi-task learning for generation

Several approaches exist for multi-task learning & generation (Dong et al., 2015; Luong et al., 2016) in NLP, especially in multilingual NMT, where tasks denote different language pairs (Zoph and

Knight, 2016; Firat et al., 2016). The multi-task (and zero-shot) generation ability of BERTGEN is mostly inspired by Ha et al. (2016) and Johnson et al. (2017). Both of these introduced target language specifiers to select the output language when decoding translations from their model.

Our multilingual & multimodal take on multi-task generation is most similar to Kaiser et al. (2017), where a single Transformer model is trained on different tasks including image captioning, object classification, machine translation, speech recognition and parsing. However, their architecture depends on particular structures such as encoders, decoders, modality-specific networks and I/O mixers, unlike BERTGEN which does not require task-specific modules.

## 5 Conclusions

In this paper, we presented BERTGEN, a novel generative, decoder-only model which extends BERT by combining multimodal and multilingual pre-trained models. Our findings show that BERTGEN obtains strong performance on a variety of generative tasks and further generalises over unseen tasks. Importantly, our model demonstrates the potential for general-purpose (instead of task-specific) generation that is above and beyond the traditional pre-training and fine-tuning practices. BERTGEN is also parameter efficient as it has 89.3M total parameters and is trained on thirteen tasks encompassing MT, multimodal MT and image captioning. On the other hand, each of the single-task FAIRSEQ NMT baselines has 31.5M parameters.

Our ablation studies show that BERTGEN is able to efficiently transfer relevant inductive biases from the pre-trained models and benefits from multi-task learning without suffering from catastrophic forgetting. We hope that these findings will motivate future research in exploiting more sophisticated pre-trained models in place of M-BERT and VL-BERT and others.

## Acknowledgments

This paper is a follow-up work to the MSc. Thesis of Faidon Mitzalis, co-supervised by Prof. Lucia Specia and Dr. Ozan Caglayan. Lucia Specia, Pranava Madhyastha and Ozan Caglayan received support from MultiMT project (H2020 ERC Starting Grant No. 678017). Lucia Specia also received support from the Air Force Office of Scientific Research (under award number FA8655-20-1-7006).

## References

- 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.
- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.
- Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.
- Ozan Caglayan, Julia Ive, Veneta Haralampieva, Pranava Madhyastha, Loïc Barrault, and Lucia Specia. 2020. Simultaneous machine translation with visual context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2350–2361, Online. Association for Computational Linguistics.
- Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. 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 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.
- Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server.
- Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020. Distilling knowledge learned in BERT for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7893–7905, Online. Association for Computational Linguistics.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.
- Yung-Sung Chuang, Chi-Liang Liu, Hung yi Lee, and Lin shan Lee. 2020. SpeechBERT: An Audio-and-Text Jointly Learned Language Model for End-to-End Spoken Question Answering. In *Proc. InterSpeech 2020*, pages 4168–4172.
- Stephane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. 2019. On the use of BERT for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 108–117, Hong Kong. Association for Computational Linguistics.
- Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.
- 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, Minneapolis, Minnesota. Association for Computational Linguistics.
- Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.
- 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*, volume 32, pages 13063–13075. Curran Associates, Inc.
- Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

- Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. [Multi30K: Multilingual English-German image descriptions](#). In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.
- Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. [Multi-way, multilingual neural machine translation with a shared attention mechanism](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.
- Robert M. French. 1999. [Catastrophic forgetting in connectionist networks](#). *Trends in Cognitive Sciences*, 3(4):128–135.
- Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*.
- Kenji Imamura and Eiichiro Sumita. 2019. [Recycling a pre-trained BERT encoder for neural machine translation](#). In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31, Hong Kong. Association for Computational Linguistics.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. [One Model To Learn Them All](#).
- Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3128–3137.
- Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11336–11344. AAAI Press.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Computer Vision – ECCV 2020*, pages 121–137, Cham. Springer International Publishing.
- Jindřich Libovický. 2019. *Multimodality in Machine Translation*. Ph.D. thesis, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czechia.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- J. Lu, C. Xiong, D. Parikh, and R. Socher. 2017. [Knowing when to look: Adaptive attention via a visual sentinel for image captioning](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3250.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.
- Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7219–7228.
- Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task Sequence to Sequence Learning. In *International Conference on Learning Representations*.
- Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6297–6308, Red Hook, NY, USA. Curran Associates Inc.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. [fairseq: A fast, extensible toolkit for sequence modeling](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.
- 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. [Collecting image annotations using Amazon’s Mechanical Turk](#). In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 139–147, Los Angeles. Association for Computational Linguistics.
- Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. [Leveraging pre-trained checkpoints for sequence generation tasks](#). *Transactions of the Association for Computational Linguistics*, 8:264–280.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. [Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. [MASS: Masked sequence to sequence pre-training for language generation](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.
- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. [VL-BERT: Pre-training of Generic Visual-Linguistic Representations](#). In *International Conference on Learning Representations*.
- Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. [Multimodal machine translation through visuals and speech](#). *Machine Translation*, pages 1–51.
- Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. [Videobert: A joint model for video and language representation learning](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.
- Hao Tan and Mohit Bansal. 2019. [LXMERT: Learning cross-modality encoder representations from transformers](#). 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 5100–5111, Hong Kong, China. Association for Computational Linguistics.
- Jörg Tiedemann. 2012. [Parallel data, tools and interfaces in OPUS](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).
- Mesut Erhan Unal, Begum Citamak, Semih Yagcioglu, Aykut Erdem, Erkut Erdem, Nazli Ikizler Cinbis, and Ruket Cakici. 2016. [Tasviret: A benchmark dataset for automatic turkish description generation from images](#). In *2016 24th Signal Processing and Communication Application Conference (SIU)*, pages 1977–1980. IEEE.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in neural information processing systems*, 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.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. [Towards making the most of BERT in neural machine translation](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9378–9385.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. [XLNet: Generalized Autoregressive Pretraining for Language Understanding](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. [From image descriptions to visual](#)

denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.

## A Qualitative Examples



BERTGEN: un groupe de jeunes sont réunis, et ils sont debout à l 'extérieur d'un bâtiment.  
(a group of young people are gathered, and they are standing outside a building.)

MT REF: des gens debout devant un bâtiment.  
(people standing outside of a building.)

---



BERTGEN: une petite fille lit un livre.  
(a little girl reads a book.)

MT REF: un jeune enfant dormant dans son lit avec un livre ouvert sur sa poitrine.  
(a young child sleeping in her bed with an open book on her chest.)

---



BERTGEN: des manifestants avec des pancartes.  
(demonstrators with placards.)

MT REF: des familles de militaires défilent dans new york un jour de pluie.  
(military families are marching through new york on a rainy day.)

---

Table 8: Zero-shot French image captioning examples for the FLICKR30K test set.



- EN: a group of people are riding on **elephants** through a river.  
FR: un groupe de personnes sur des chevaux sur un bateau dans un ruisseau.  
*a group of people on **horses** on a boat in a stream.*  
DE: eine gruppe reiter fährt auf einem fluss.  
*a group of riders is riding on a river.*  
TR: bir grup insan bir derede duran dört tane at ile ilerliyorlar.  
*a group of people are moving with **four horses** standing in a stream.*
- 



- EN: a **black** dog is playing with a **yellow toy** in the grass.  
FR: un chien avec une frisbee dans la pelouse.  
*a dog with a **frisbee** in the lawn.*  
DE: ein schwarzer hund mit rotem halsband spielt mit einem gelben ball auf einer wiese.  
*a **black** dog with a **red collar** is playing with a **yellow ball** in a meadow.*  
TR: yeşil bir topu ısırmaaya çalışan siyah bir köpek.  
*a **black** dog trying to bite a **green ball**.*
- 



- EN: a boy in a red shirt and white shorts is playing tennis.  
FR: un **tennisteur** frappe une balle.  
*a **tennisteur** hits a ball.*  
DE: ein junge spielt tennis.  
*a boy is playing tennis.*  
TR: tenis raketi ile topa vuran çocuk.  
*boy hitting the ball with a tennis racket.*
- 

Table 9: COCO captioning examples: *Italicised captions are Google’s translations into English for non-English examples.* **Bold** phrases highlight lexical variations or errors related to the salient visual concepts in the images. The last example shows a morphological error that BERTGEN does when trying to generate a **tennis player** in French.

BERTGEN	mehrere ngos, darunter die mozilla - und greenpeace - stiftung, schätzen, dass diese neuen werkzeuge unfähig sind und zu spät kommen. <i>several ngos, including the mozilla and greenpeace foundations, estimate that these new tools are <b>incapable</b> and come too late.</i>
WMT REF	mehrere ngos, unter denen die mozilla - stiftung und greenpeace, schätzen, dass diese neuen tools unzureichend sind und zu spät kommen. <i>several ngos, including the mozilla foundation and greenpeace, estimate that these new tools are <b>inadequate</b> and come too late.</i>
BERTGEN	immigration wird als ein großes problem für die ue betrachtet, für 45 prozent der deutschen und 40 prozent aller europäischen. <i>immigration is seen as a big problem for the <b>ue</b>, for 45 percent of germans and 40 percent of all european ones.</i>
WMT REF	die einwanderung halten 45 prozent der deutschen und 40 prozent aller europäer für das größte problem der eu. <i>45 percent of germans and 40 percent of all europeans consider immigration to be the biggest problem in the <b>eu</b>.</i>
BERTGEN	das ist der grund, warum er in seinem buch die frage erforscht, ob es alternativen zu wahlen gibt. <i>that is the reason why he explores the question of whether there are alternatives to choose from in his book.</i>
WMT REF	deshalb geht er in seinem buch der frage nach, ob es alternativen zu wahlen gibt. <i>therefore, in his book, he investigates the question of whether there are alternatives to choose from.</i>

Table 10: Zero-shot FR→DE translations of WMT’19 test set. The *italicised* sentences are Google Translate translations of the German outputs into English.