Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers

Lisa Anne Hendricks John Mellor Rosalia Schneider
Jean-Baptiste Alayrac Aida Nematzadeh
DeepMind, United Kingdom
{lmh, johnme, rgschneider, jalayrac, nematzadeh}@google.com

Abstract

Recently, multimodal transformer models have gained popularity because their performance on downstream tasks suggests they learn rich visual-linguistic representations. Focusing on zero-shot image retrieval tasks, we study three important factors that can impact the quality of learned representations: pretraining data, the attention mechanism, and loss functions. By pretraining models on six datasets, we observe that dataset noise and language similarity to our downstream task are important indicators of model performance. Through architectural analysis, we learn that models with a multimodal attention mechanism can outperform deeper models with modality-specific attention mechanisms. Finally, we show that successful contrastive losses used in the self-supervised learning literature do not yield similar performance gains when used in multimodal transformers.

1 Multimodal Pretraining

Significant progress in pretraining of natural language processing (NLP) models has been made through both architectural innovations (e.g., transformers; Vaswani et al., 2017) as well as a huge increase in the size of pretraining data and the model (e.g., Devlin et al., 2019; Brown et al., 2020). This success in language pretraining has inspired parallel multimodal vision–language efforts; in particular, multimodal image–language transformers, pretrained on large noisy image–text datasets, have achieved state-of-the-art results on a range of downstream tasks such as image retrieval, visual question answering, and visual reasoning (e.g., Lu et al., 2019; Chen et al., 2020; Tan and Bansal, 2019; Li et al., 2020a,b).

However, even though many variants of multimodal image–language transformer models have been proposed recently, it is unclear how learned representations are impacted by the large amounts of pretraining data, the transformer architecture and self-attention, or their specific losses. We address this gap, by first establishing a baseline that is trained on the same pretraining data as multimodal transformers but with a different architecture. We then perform an investigative analysis to better understand the extent to which these aspects contribute to models' performance.

Our evaluation mainly focuses on zero-shot tasks where evaluation data is taken from a dataset unseen during pretraining. Measuring zero-shot performance enables us to evaluate whether a pretrained model learns general representations. Previous work in NLP has considered probing classifiers to evaluate representations; however, this approach can be misleading as the performance of probing classifiers does not solely depend on the quality of representations (e.g., Hewitt and Liang, 2019; Voita and Titov, 2020). Similarly, evaluation after fine-tuning is a less direct measure of strength of representations since performance on these tasks is highly dependent on the fine-tuning experimental set-up and the size of fine-tuning data (Yogatama et al., 2019).

We first study the importance of different properties of multimodal datasets such as their size and their noise level (i.e., how closely the language describes a given image's content). Recent work has introduced image–text datasets with different qualities—for example, noisy but very large ones (Sharma et al., 2018) as well as carefully annotated but smaller ones (Pont-Tuset et al., 2019). Better understanding of what aspect of a dataset is more important can result in better task performance and also guide us in future dataset curation efforts. We find that a dataset's size does not always predict multimodal transformers' performance; its noise level and language similarity to the evaluation task are both important contributing factors. We also show that multimodal transformers can achieve competitive results without relying on

Transactions of the Association for Computational Linguistics, vol. 9, pp. 570–585, 2021. https://doi.org/10.1162/tacl_a_00385
Action Editor: Marco Baroni. Submission batch: 10/2020; Revision batch: 12/2020; Published 7/2021.

language-only or image-only pretraining for weight initialization or feature extraction.

We also dissect multimodal transformers' architecture, analyzing the effectiveness of different attention mechanisms, depth, and number of parameters. We show that *multimodal* attention, where both language and image transformers attend to each other, are crucial for these models' success. Multimodal attention achieves the best results when combined with multi-level (deep) interactions. Moreover, models with other types of attention (even with more depth or parameters) fail to achieve comparable results to shallower and smaller models with multimodal attention.

Additionally, inspired by the success of self-supervised representation learning (e.g., van den Oord et al., 2018), we examine whether using a contrastive image–text matching loss instead of a classification one improves the quality of representations in our models. Surprisingly, we find that the choice of image–text matching loss does not matter much in multimodal transformers. On the other hand, models *without* multimodal attention (a multi-level ''cross-talk'' between modalities) benefit significantly from a contrastive loss.

Finally, we believe that advances in multimodal pretraining can have significant impacts on a wide range of downstream applications; however, it is important to form a clear understanding of how and why multimodal transformer models perform well to avoid overfitting to a set of downstream evaluation tasks. Our analysis of pretraining data, attention, and loss functions is an important step towards gaining a deeper understanding of these powerful models.

2 Multimodal Transformers

The success of transformer-based language models on a variety of language tasks (e.g., Devlin et al., 2019) has inspired similar multimodal efforts (e.g., Lu et al., 2019; Chen et al., 2020; Tan and Bansal, 2019; Li et al., 2020a,b).[1] The main distinction is that image-text multimodal transformers take image-text pairs as input, attend over both modalities, and are trained with additional losses. Similar to the language models, multimodal transformers are often fine-tuned on down-stream tasks but multimodal ones; e.g., image retrieval (Young et al., 2014) or visual question answering (Goyal et al., 2017).

We give a brief overview of the BERT model (Devlin et al., 2019), which forms the backbone of multimodal transformers. The BERT architecture consists of a stack of transformer blocks (Vaswani et al., 2017) and has three main components. First, the input text is tokenized and three embedding functions are used to embed the token, its position in the sentence (i.e., positional encoding), and the sentence it belongs to. The final language embedding is a summation of these three vectors. The BERT model also includes a <SEP> token to separate different sentences and a <CLS> token, which can be thought of as an aggregate representation of the input text. Second, the sequence of token embeddings are input into a series of transformer layers where tokens are combined through self-attention. Third, two different losses are applied to the model output: a *masked language modeling* loss, in which the model predicts a masked word (denoted by a <MASK> token), and a *next sentence prediction* loss which, given two sentences, predicts if the second sentence follows the first.

Multimodal transformer models facilitate learning from multimodal data via three changes to the BERT architecture: *multimodal data preprocessing* (more specifically images), adding *multimodal attention* by changing self-attention such that it combines image and text modalities, and introducing image and multimodal *loss functions*.

2.1 Multimodal Data Processing

Training multimodal transformers requires image–text pairs such that the text for a given image, at least to some degree, describes the image.Recent work attempts to remove the annotation cost by automatically collecting datasets (e.g., Web images and their alt-text as in Sharma et al., 2018). In Section 4.2, we examine whether the quality of text descriptions impacts these models' performance.

The text input processing is the same as language models; in fact, many of the existing models (such as Lu et al., 2019) are initialized with BERT pretrained weights. We show that this initialization is not important in our experiments (see Section 4.2). Processing images into a sequence involves defining ''visual tokens''

[1] We use the term multimodal transformers to refer to image–text transformer–based models. Note that similar architectures are applied to other modalities such as videos (Sun et al., 2019) but are outside of the scope of this work.

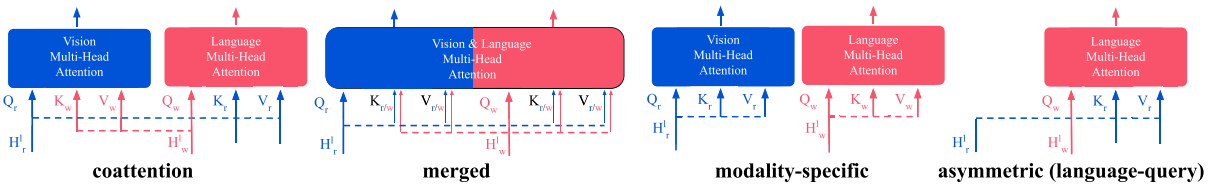


Figure 1: Different attention types (see Section 2.2). Queries, keys, and values are shown by Q, K, and V; w and r index language words and image regions, respectively. H^l is the activation at layer l.

analogously to language tokens. Almost all image-text multimodal transformer models consider a bounding box from a pretrained object detection model to be a ‘‘visual token’’. Similar to the positional encodings in language models, for each visual token, the spatial position of each bounding box is also encoded.

Although most multimodal transformers require training a supervised model (a detector) to extract bounding-box features, there are other possible ways to represent visual tokens—for example, Huang et al. (2020) bypass training a detector by using regions from a high-level feature map in an image classification network as visual tokens. We focus our studies on models that use bounding-box features as this reflects the majority of recent work, though we achieve comparable results when learning directly from images without a detector (or even a pretrained classifier) in Section 4.2.

2.2 Multimodal Attention

Each transformer block consists of a multi-head attention module (Vaswani et al., 2017) that for a given token embedding produces a weighted representation of all other tokens in a sentence. This weighted representation is then combined with the input representation of the given token and is passed to the next layer. More specifically, for the token i at layer l, each attention head takes as input a key k_l^i, value v_l^i, and query q_l^i, which are computed by passing the representation from the previous layer h_{l-1}^i through a linear layer. The output of the attention module for token i is:

$$A(q_l^i, K_l, V_l) = \text{softmax}\left(\frac{q_l^i K_l}{\sqrt{d_k}}\right) V_l, \quad (1)$$

where d_k is the dimension of the key and K_l and V_l matrices contain all tokens' keys and values.

Given this definition, there are a few possible ways to implement multi-head attention over image and language modalities as shown in Figure 1. For a given query (from one modality), we can simply consider keys and values from all input tokens regardless of the modality type (e.g., Chen et al., 2020). We refer to this multimodal attention as *merged attention* because it simply merges inputs from the two modalities.

Alternatively, given queries from one modality (e.g., image), keys and values can be taken *only* from the other modality (e.g., language). Following Lu et al. (2019), we refer to this multimodal attention as *coattention*. We also consider cases where this attention is asymmetric, that is, queries are *either* from language or image, while keys and values are from image or language, respectively. We call these two attention types *language-query attention* or *image-query attention*.

Another possibility is to consider single-modality transformers where queries, keys, and values all come from either the image or text modality; we refer to this attention as *modality-specific attention*, where each modality has its own multi-head attention. Single-modality transformers with modality-specific attention allow us to study the role of ‘‘cross-talk’’ between modalities in multimodal transformer models.

We note that we use the term *multimodal* attention to refer to both *merged attention* and *coattention* and discuss the importance of different attention types in Section 4.3.

2.3 Multimodal Loss Functions

Broadly, multimodal transformers have three loss types, language and image losses that are applied to the language and image outputs, respectively, as well as an image-text matching loss applied to image–language pairs. Let $\boldsymbol{r} = \{r_1, \cdots, r_N\}$ be the N input image regions and $\boldsymbol{w} = \{w_1, \cdots, w_T\}$ be the T word tokens representing an image–text pair. A subset of input image regions and word tokens are masked (e.g., set to zero) before being passed through the

transformer layers. After applying the mask, we refer to the *unmasked* image regions as $\boldsymbol{r}^{\boldsymbol{m}}$ and to the *unmasked* word tokens as $\boldsymbol{w}^{\boldsymbol{m}}$. We use N_m and T_m to denote the set of image region and word token indices that are masked, respectively. Similar to the BERT model, the language loss is a masked-language modeling (MLM) loss:

$$-\sum_{t \in T_m} \log P_\theta^{\boldsymbol{w}}(w_t | \boldsymbol{w}^{\boldsymbol{m}}, \boldsymbol{r}^{\boldsymbol{m}}), \tag{2}$$

where $P_\theta^{\boldsymbol{w}}$ corresponds to the output probability distribution over words in the vocabulary from the transformer model parameterized by θ.

Most models also include an analogous *masked region modeling loss* (MRM) for images. One popular region modeling loss, for each bounding box, minimizes the KL-divergence between the predicted distribution over object classes and the distribution over classes obtained from a pretrained detector $D(l|r_n)$ (e.g., Chen et al., 2020; Lu et al., 2019).

$$\sum_{n \in N_m} \mathrm{KL}(D(l|r_n) || P_\theta^{\boldsymbol{r}}(r_n | \boldsymbol{r}^{\boldsymbol{m}}, \boldsymbol{w}^{\boldsymbol{m}})), \tag{3}$$

where $P_\theta^{\boldsymbol{r}}$ corresponds to the predicted probability distribution over object classes from the transformer model parameterized by θ.

Finally, multimodal transformer models include an image–text matching (ITM) loss, which predicts whether an image and text pair match; this is generally posed as a binary classification problem:

$$\begin{aligned} &-y \log(\sigma(s_\theta(\boldsymbol{r}^{\boldsymbol{m}}, \boldsymbol{w}^{\boldsymbol{m}}))) \\ &-(1-y) \log(1 - \sigma(s_\theta(\boldsymbol{r}^{\boldsymbol{m}}, \boldsymbol{w}^{\boldsymbol{m}}))), \end{aligned} \tag{4}$$

where y is equal to 1 for positive pairs and 0 otherwise and s_θ corresponds to the confidence score of the model that a pair $(\boldsymbol{r}, \boldsymbol{w})$ are matched and σ is the sigmoid function. Recently, contrastive image–text matching losses have been successful in self-supervised representation learning (e.g., van den Oord et al., 2018); thus, we also explore whether a contrastive formulation of ITM can improve the performance of multimodal transformers and discuss the challenges of using these losses for multimodal transformer models. Our contrastive loss is formulated as:

$$-\log\left(\frac{e^{s_\theta(\boldsymbol{r}^{\boldsymbol{m}}, \boldsymbol{w}^{\boldsymbol{m}}))}}{e^{s_\theta(\boldsymbol{r}^{\boldsymbol{m}}, \boldsymbol{w}^{\boldsymbol{m}})} + \sum_{(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{w}}) \sim \mathcal{N}} e^{s_\theta(\tilde{\boldsymbol{r}}^{\boldsymbol{m}}, \tilde{\boldsymbol{w}}^{\boldsymbol{m}})}}\right), \tag{5}$$

where $\mathcal{N}$ is a set of negative image-text pairs. Section 4.4 outlines our findings on loss ablations.

3 Experimental Setup

Here we outline the details of our experimental setup: the base multimodal transformer model used in most of our experiments, our baseline model, and the pretraining datasets.

3.1 Base Multimodal Transformer

Our base multimodal transformer model (MMT) most closely resembles the ViLBERT model (Lu et al., 2019). For text inputs, we first tokenize sentences using SentencePiece (Kudo and Richardson, 2018) and truncate sentences into a fixed length of 22 for pretraining datasets and 25 for datasets used to fine-tune and evaluate retrieval models. We then include a separator (<SEP>) and an aggregator (<CLS>) token. Unless otherwise stated, we do not transfer weights from a pretrained BERT model.

For image inputs, we represent ''visual tokens'' as region of interest pooled features corresponding to bounding boxes from an object detector (Ren et al., 2015) trained on Visual Genome (Krishna et al., 2017) images with labels parsed as was done in Anderson et al. (2018). The detection model is trained using a multi-label sigmoid cross-entropy loss to simultaneously predict objects and attributes. The highest 36 or 100 scoring bounding boxes are input when pretraining or evaluating, respectively. Like ViLBERT, we include an ''average'' feature, which is computed by averaging features across bounding boxes and serves a similar role to the <CLS> token in the text input.

In addition to the positional encoding added to text embeddings before the first transformer layer, we also add the positional encoding to the text embedding *at each layer* of the language-only transformer blocks as in XLNet (Yang et al., 2019) because this led to improvements on a language-only BERT model. For image inputs, we embed bounding box coordinates and add this to our image embedding.

In our model, following ViLBERT, a multimodal coattention layer consists of an image-only and a language-only transformer, each followed by a transformer with *coattention* (see Section 2.2). We use the term ''layer'' to refer to this multimodal layer. Like VilBERT, our model consists of

6 language-only layers, followed by 6 multimodal ones. We train the model by minimizing masked language modeling (Equation (2)), masked region modeling (Equation (3)), and binary classification image–text matching (Equation (4)) losses. To calculate the image-text loss, we apply an element-wise multiplication to the <CLS> language features and output corresponding to the averaged image feature input. The resulting ''multimodal feature'' is input into a classification model. We create negative image-text examples by sampling text from another image in our batch. Unless otherwise noted, we have an equal number of negative and positive image-text pairs.

We train our models with a global batch size of 1024 distributed over 64 Google Cloud TPU v3 cores.[2] We use the LAMB optimizer (You et al., 2019) with an initial learning rate of 0.00176 and 20,000 warm-up steps. Learning rate is decayed with polynomial decay with a minimum learning rate ratio of 0.004. We use gradient clipping (1) and dropout (0.1) as well as weight decay (0.1). We find weight decay particularly important in ensuring that our loss did not diverge. We train our models for a maximum of 1,000,000 iterations.

3.2 The Baseline Model

Multimodal transformers are different from most prior image–text models because they are pretrained on a large dataset (millions of image-text pairs). To better understand if data alone can lead to better image–text representations, we train a strong baseline model, which does not include a multimodal attention mechanism, with the same data as our multimodal transformer.

Our baseline model learns a joint space between language and vision (Weston et al., 2011; Frome et al., 2013; Kiros et al., 2014) by minimizing the distance between image and text features taken from a positive pair (where text describes the image) and at the same time increasing that distance for a negative pair. Despite lacking a multimodal attention mechanism, this approach has been popular in image and video domains because of its simplicity and effectiveness for retrieval applications (e.g., Gong et al., 2014; Wang et al., 2016; Chowdhury et al., 2018; Miech et al., 2018).

To implement our baseline, we encode word tokens $\boldsymbol{w}$ into a fixed-size sentence representation $S \in \mathbb{R}^{768}$ and image regions $\boldsymbol{r}$ into a fixed-size image representation $I \in \mathbb{R}^{768}$. To encode sentence representations, we input words into a randomly initialized BERT model and extract sentence representations S from the <CLS> output. To extract image representations I, we mean-pool features across detected bounding boxes then pass the features into a one-layer MLP with an output of size 768. Finally, we element-wise multiply I and S and input the resulting vector into a two-layer MLP parameterized by θ which outputs a score, s_θ, indicating whether I and S match. The baseline model is trained with the contrastive loss defined in Equation (5) with 1024 negative examples. The detector weights are fixed during training.

3.3 Pretraining Datasets

Conceptual Captions (CC) consists of over 3 million image-text pairs harvested from the Web where the caption corresponding to an image is its alt-text description (Sharma et al., 2018). Image–text pairs are filtered and preprocessed such that text is more image relevant than raw alt-text; however, the dataset is still ''noisy'' and includes pairs where the text is not relevant to the image's content. We were able to download 81% of the training set of CC; unless otherwise stated, we train our models on this subset of CC.

The *SBU* dataset (Ordonez et al., 2011) consists of 1 million image-text pairs sourced from Flickr with text taken from users' captions. As a result, similar to CC, not all text is image-relevant. We also use datasets that were collected by asking annotators to describe images, resulting in more image relevant language including the *MSCOCO* dataset (Chen et al., 2015) and *Visual Genome (VG)* (Krishna et al., 2017), which includes descriptions for bounding boxes in images.

When using VG, we consider each bounding box description to be a caption for the entire image. We also experiment with the *Localized Narratives* dataset (Pont-Tuset et al., 2019). This dataset includes rich annotations collected by asking users to describe an image while pointing to each part of the image being described (using their mouse). The resulting ''narratives'' often consist of multiple sentences. We break the narratives into individual sentences and treat each sentence as a caption paired with the image. We use the localized narratives collected for the Open Images (Kuznetsova et al., 2018) and MSCOCO

[2] https://cloud.google.com/tpu/.

Dataset	# images	Caption Type	Caption #
MSCOCO	83K	Annot.	592K
Visual Genome (VG)	110K	Annot.	5.4M
MSCOCO-narratives	83K	Narration	230K
OI-narratives	500K	Narration	1.3M
SBU	1M	Web	1M
Conceptual Captions	2.7M	Alt-text	2.7M

Table 1: The pretraining datasets: the type and number of images and captions.

datasets, and refer to them as OI-narratives and MSCOCO-narratives. This allows us to compare models that are trained with the same images (MSCOCO) with different language (MSCOCO captions vs. localized narratives). Table 1 provides an overview of our pretraining datasets.

We combine datasets using two sampling approaches: *instance sampling*, where we mix all datasets together and sample from this mix for each batch, and *dataset sampling*, where we sample evenly from datasets so that each batch contains the same number of examples from each dataset. For datasets with multiple captions, we first sample an image, then sample a caption for the given image. We combine all six datasets described here as well as the four datasets combined in Chen et al. (2020) (MSCOCO, VG, SBU, and Conceptual Captions) which we refer to as UNITER data.

3.4 Evaluation Tasks

We focus on *zero-shot* evaluation as it enables us to examine the representations without confounding our findings with the side-effects of fine-tuning (Yogatama et al., 2019) or probing classifiers (e.g., Zhang and Bowman, 2018; Hewitt and Liang, 2019). Following Lu et al. (2019) and Chen et al. (2020), we use the term *zero-shot* to refer to experiments where we test our models on a dataset different from our pretraining data *without fine-tuning*. For example, we use the MSCOCO dataset to test the models that are pretrained on Conceptual Captions. This is considered as a zero-shot task since the properties of the dataset used for testing (for example, its language) differ from those in the pretraining dataset. We use zero-shot image retrieval tasks since image retrieval directly measures what our pretraining data and objectives encourage our models to learn: whether an image and a sentence are aligned.

We evaluate on the Flickr30k dataset (Young et al., 2014) (referred to as *zero-shot Flickr*) and use the splits defined in Karpathy and Fei-Fei (2015). We evaluate checkpoints after 1 million steps as well as when the loss on the CC validation set is lowest. When varying the pretraining data, our models sometimes overfit quickly on smaller datasets; as a result, we evaluate checkpoints every 100K steps. We select the best checkpoint according to zero-shot performance on Flickr30k validation split and use it for all other downstream tasks. We also report retrieval numbers on MSCOCO (Chen et al., 2015) (which we call *zero-shot MSCOCO*) using the splits of Karpathy and Fei-Fei (2015). We report retrieval numbers on the test split of datasets. Flickr30k and MSCOCO images are annotated with 5 captions.

In addition to the zero-shot image retrieval tasks, we use the fine-tuned Flickr30k image-retrieval task to examine whether our observations transfer when fine-tuning the MMT model. We fine-tune our models for 10,000 steps and use MLM, MRM, and ITM losses. All results for image retrieval are reported using Recall@K (R@K), which measures whether the ground-truth image is among the top K images retrieved by our model.

When comparing pretraining datasets, we hypothesize that which pretraining dataset is best depends on the downstream task, so we additionally consider VQA (Antol et al., 2015; Goyal et al., 2017). To fine-tune for VQA, we replace the image–text matching loss with a 2-layer MLP and train with a binary cross-entropy loss against soft answer scores (Teney et al., 2018). We use similar hyper-parameters as when pretraining and report results on the validation set. We report the average score across 3 random initializations of the MLP.

We use Flickr IDs to filter out images appearing in the Flickr30k and MSCOCO validation/test sets from our pretraining sets. Conceptual Captions is not collected from Flickr, so we could not filter out images using this method. Table 2 provides an overview of our evaluation datasets.

4 Experimental Results

We first compare MMT to a baseline and then investigate how pretraining data, attention, and loss functions impact model performance.

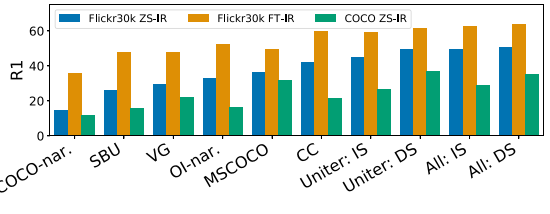
Dataset	# images train	# images test	ZS	FT
Flickr30k	29K	1K	✓	✓
MSCOCO	n/a	5K	✓	
VQA	440K	210K		✓

Table 2: Number of images in evaluation tasks and whether datasets were used in a zero-shot (ZS) or fine-tuned (FT) setting.

	Flickr30k ZS R1	Flickr30k ZS R10	Flickr30k FT R1	Flickr30k FT R10	MSCOCO ZS R1	MSCOCO ZS R10
Baseline	25.4	64.9	40.9	81.8	13.0	44.5
− contrastive	21.7	61.0	39.0	80.6	10.2	40.9
+ BERT PT	24.8	65.1	39.9	79.9	12.7	43.1
MMT	**41.9**	**79.0**	**59.1**	**91.5**	**21.3**	**57.9**
ViLBERT	31.9	72.8	58.2	**91.5**	–	–

Table 3: Comparison of our proposed baseline to our multimodal transformer model (MMT).

(a) Zero-shot (ZS) & fine-tuned (FT) image retrieval (IR)



(b) Visual question answering (VQA v2)

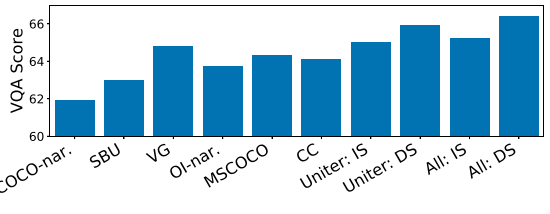


Figure 2: Effect of pretraining data. The datasets on the x axis are ordered based on their *zero-shot Flickr* scores. IS:Instance Sampling, DS: Dataset Sampling.

4.1 Comparison to a Baseline

We compare our multimodal transformer (MMT) against a strong baseline inspired by recent success in visual retrieval (e.g., Miech et al., 2018). To disentangle the effect of pretraining data and architecture, we investigate whether our baseline (described in Section 3.2), without *multimodal* attention or MLM and MRM losses but pretrained on the same data (i.e., Conceptual Captions) as multimodal transformers, produces competitive results.

In Table 3, we compare MMT to our proposed baseline, verifying that MMT learns better representations not only because it is pretrained on a large dataset, but because of architectural choices. Our MMT results are on par with existing models trained with the same data: comparing to ViLBERT, the most similar model to ours, on the *zero-shot Flickr*, we achieve an R@1 of 41.9 in comparison to 31.9. As expected, retrieval numbers on *zero-shot MSCOCO* are lower than *zero-shot Flickr* because MSCOCO has more images in its evaluation set (see Table 2) and is therefore harder. On the fine-tuned image retrieval task, we achieve comparable performance to ViLBERT (our R@1 is 59.1 vs. 58.2), even though we do not sample hard negatives when training. We emphasize that our goal is not to outperform existing work, but to build a strong multimodal transformer model to analyze the role of data, attention, and losses.

On our baseline, we verify that a contrastive loss (Equation (5)) leads to stronger results than a classification one. As shown in Table 3, replacing the contrastive loss with a classification loss consistently decreases performance. Initializing our baseline with BERT weights marginally decreases performance, for example, R@1 on *zero-shot Flickr* decreases by 0.6.

4.2 Multimodal Data Preprocessing

We investigate how pretraining datasets, supervised image features, and weights from a pretrained language model impact our results.

Pretraining Datasets. Figure 2 reports our results when we pretrain the MMT on the individual and combined datasets introduced in Section 3.3. We observe that in all our tasks, *larger datasets usually lead to better performance, but not always.* For example, SBU consistently performs worse than MSCOCO, despite being substantially larger.

Additionally, *when combining datasets, how datasets are sampled matters*. In our experiments, dataset sampling (DS) is more effective than instance sampling (IS). In dataset sampling, smaller datasets (like MSCOCO) will be sampled more frequently than in instance sampling. Because MSCOCO pretraining leads to good performance, more exposure to MSCOCO samples is beneficial.

We consider combining all datasets as well as datasets combined in UNITER (Chen et al., 2020). Figure 2a shows that combining all datasets performs better than UNITER data on the *zero-shot Flickr* task, but not on the *zero-shot MSCOCO*, showing that more data is not always better. On *zero-shot MSCOCO* the impact of the sampling mechanism is even more evident: Given UNITER data, dataset sampling performs better than instance sampling by over 10 points (37.1 vs 26.4).

Next, we compare datasets that have a similar number of images to investigate the role of the type of language used in each dataset. As an extreme example, MSCOCO and MSCOCO-narratives contain the same images, but the former does substantially better on our downstream tasks. To better understand this observation, we quantify the difference between the language of pretraining and evaluation datasets: we trained a language model (a 6-layer Transformer) on a given pretraining dataset, and use that model to compute the perplexity of the evaluation dataset. For our three datasets with the same number of images (MSCOCO, MSCOCO-narratives, and VG), the perplexity of the evaluation dataset (Flickr or MSCOCO) explains their performance—the perplexities are the lowest on MSCOCO, then VG, and lastly on MSCOCO-narratives. *This shows that the similarity between the language of pretraining and evaluation datasets is important.*

However, not all performance differences are explained by the number of images or perplexity: Pretraining on SBU results in poorer performance than OI-narratives on our downstream tasks, despite SBU having twice the number of images and lower perplexity on both evaluation datasets. We conjecture that SBU's poor performance is due to noise: SBU text is scraped from captions and may not match the images as well as the manually annotated text in OI-narratives. To investigate this, we calculate an *overlap metric* for an image–text pair as the ratio of text words overlapping with predicted bounding box labels. For each dataset, we calculate the average overlap for 3000 images, providing an approximation of how much the language describes the images in the dataset. The overlap is much lower for SBU compared to OI-narratives (0.14 vs. 0.25), showing that SBU is indeed noisier, which can decrease its utility for pretraining multimodal representations.[3]

[3]The *overlap metric* for other datasets: VG: 0.82, MSCOCO: 0.42, MSCOCO-narratives: 0.27, and CC: 0.11.

Figure 3: Comparing models trained with the MSCOCO and CC datasets. We provide the top-1 ranked retrieved image given an input query sentence on the Flickr val dataset. Correctly retrieved images are framed in green and the incorrect ones in red.

Moreover, we observe that the *goodness of a pretraining dataset for one task does not always transfer to a different task.* For example, CC is a better pretraining dataset than VG when fine-tuning for image retrieval, but they perform similarly when fine-tuning for VQA, a substantially different task.In fact, we note that VQA performance varies less across pretraining datasets (e.g., CC, VG, and MSCOCO), likely because the VQA training split is large. We also observe differences between zero-shot and fine-tuned image retrieval. Though MSCOCO performs 3.8 points better on *zero-shot Flickr* than OI-narratives, OI-narratives performs 2.9 points better after fine-tuning.

Finally, to visually illustrate the difference between the learned representations, we compare qualitative examples of models trained with our best two pre-training datasets: MSCOCO and CC (see Figure 3). Though the model trained with MSCOCO retrieves examples with some semantic relevance, our model trained with CC is able to retrieve images with more correct details like ''enjoying a view'' and ''black fleece jacket''.

Language-only Pretraining. Many multimodal transformers initialize language weights from a pretrained BERT model. Similar to LXMERT, we find this hurts performance on our retrieval task; R@1 on *zero-shot Flickr* decreases to 39.7 and R@1 on *zero-shot MSCOCO* decreases to 20.4.

Image-only Pretraining. The object detector used to extract image features is another source

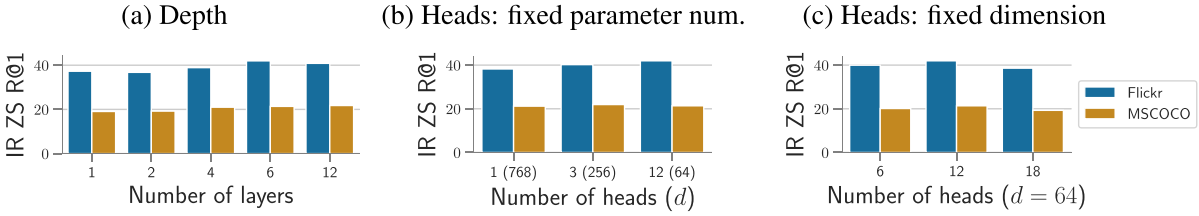


Figure 4: Ablation studies on number of layers and heads.

of modality-specific pretraining. We replace detection features with grid features taken from the last residual block of a ResNet-50 trained from scratch.[4] Similarly to Huang et al. (2020), this model is trained *without* the MRM loss since features aggregate information in the whole image, and as a result, masking specific regions is not straightforward. This model performs slightly better than our base MMT on *zero-shot Flickr* (43.4 vs. 41.9) and comparably on *zero-shot MSCOCO* (21.3 vs. 20.6). Though Huang et al. (2020) showed a detector can be replaced with an image classifier, we show that comparable results can be achieved without any image-only pretraining.

We conclude that careful consideration of pretraining datasets and their sampling methods is important in a model's performance—the level of noise and the type of language in a dataset can be more significant than its size. Finally, the image-only and language-only pretraining are not crucial in training strong multimodal representations.

4.3 Multimodal Attention

We explore the impact of the number of attention heads and *coattention* layers in our base multimodal transformer model before investigating the effect of different attention mechanisms.

Number of Heads and Layers. We test the importance of the number of heads in multi-head attention when fixing the total number of parameters by comparing models trained with one head, 3 heads, and 12 heads with query/key size of 768, 256, and 64, respectively. Increasing the number of heads to 12 leads to an improvement (Figure 4b). Next, we vary the number of heads (6, 12, and 18) but fix the query/key size to 64. We observe that increasing the number of heads up to 12 still leads to an improvement, but further increase results in poorer performance (see Figure 4c).

Consistent with Lu et al. (2019), increasing the number of layers (Figure 4a) helps up to a point, and then adding more layers degrades performance.

[4]We fit images into a 384×384 square by resizing and padding to preserve the aspect ratio. As the total stride of ResNet-50 is 32, a feature grid is of size 12×12, which we flatten to 144 features and give as input along with the averaged features (for the <CLS> token) to our MMT.

R@1	Co	Merge	Asym. Attn.				Mod.
			L-12	I-12	L-24	I-24	Spec.
F. ZS	**41.9**	40.0	24.4	31.3	33.6	31.6	16.9
F. FT	**59.1**	57.0	45.1	48.4	52.5	46.3	15.4
M. ZS	**21.3**	19.6	13.8	16.1	17.0	16.0	8.0

Table 4: MMT trained with *coattention* (Co), *merged attention* (Merge), *language-query attention* (L-12 and L-24), *image-query attention* (I-12 and I-24) (the number indicates the number of attention heads) and *modality-specific attention*.

Type of Attention Mechanism. We perform an in-depth analysis on different types of attention explained in Section 2.2 (see Table 4). We compare *coattention* with *merged attention*—these mechanisms both ''combine'' the image and language modalities; however, *coattention* does so by taking keys/values and queries from opposite modalities, while *merged attention* shares keys and values across the modalities. When controlled for the number of parameters, *coattention* performs marginally better than *merged attention*. Both perform considerably better than asymmetric attention where attention queries are over one modality.

The number of heads in an asymmetric attentions is half of the equivalent *coattention*, so we experiment with asymmetric attention mechanisms with 12 heads (L-12, I-12) as well as 24 heads (L-24, I-24). Increasing the number of attention heads for the asymmetric attention improves

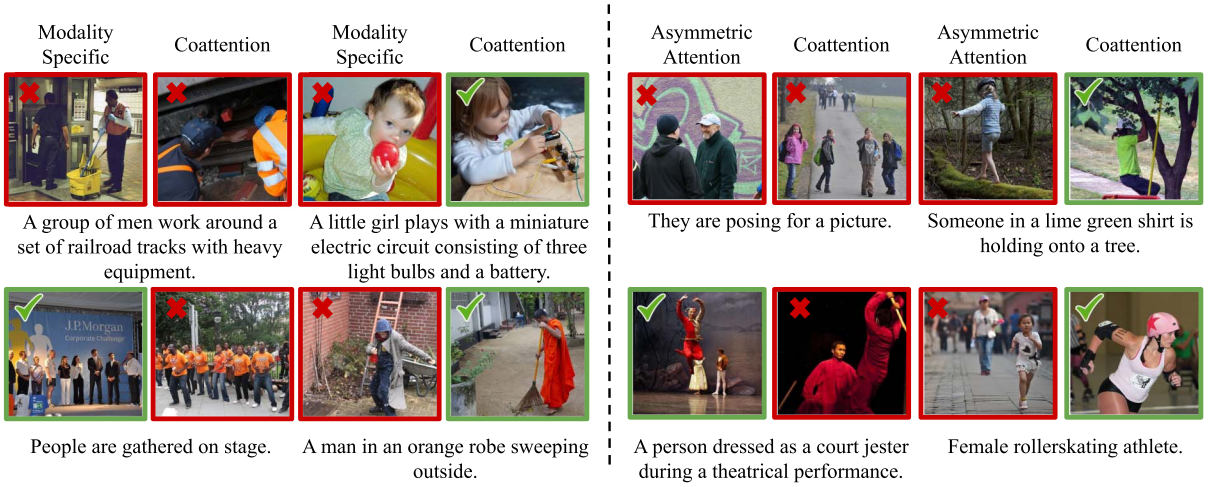


Figure 5: Comparing top-1 ranked images retrieved with models trained with the different attention mechanisms on the Flickr dataset. Correctly retrieved images are framed in green and the incorrect ones in red.

results, but the gap between our best-performing model with asymmetric attention (L-24) and *coattention* is still quite large.

We also consider transformers with *modality-specific attention* where there is no cross-talk between the modalities through attention, but the model has the same number of parameters as our MMT with *coattention* and is trained with the same losses (Table 4, Mod. Spec. column). This model performs substantially worse than MMT.

To better demonstrate the strength of multimodal attention compared to asymmetric and modality-specific attention, we compare our models in Table 4 to shallower and smaller models with *coattention* on the *zero-shot Flickr* task. Strikingly, our best-performing model *without* multimodal attention with 24 attention heads and 12 layers (R@1 of 33.6; L-24 in Table 4) performs worse than the *coattention* model with only one head (R@1 of 38.2; Figure 4b) or one multimodal layer (R@1 of 37.2; Figure 4a).

Figure 5 shows example retrieval results comparing the asymmetric and modality specific attention to our coattention mechanism. When the coattention mechanism retrieves the incorrect image, the image frequently includes important content from the sentence (e.g., in Figure 5 lower left, the image shows ‘‘people gathered’’, but they are not on stage). Though other attention mechanisms retrieve images with some similarities to the text, the coattention mechanism retrieves fine details like ‘‘lime green shirt’’ and ‘‘miniature electric circuit’’.

A modality specific transformer model is computationally more efficient than models with multimodal attention because image and language features can be computed once and reused across image–text pairs; this means that single-modality transformers are faster for retrieval and thus would be more appealing in large-scale applications if their accuracy were higher. We therefore investigate whether we can improve the single-modality transformer’s poor performance by combining five *modality-specific attention* layers followed by one *coattention* layer to introduce multimodal interaction.This model is as deep as our MMT, but performs worse than our MMT with one *coattention* layer: R@1 of 33.1 vs 37.2 on *zero-shot Flickr* and 16.7 vs 19.0 on *zero-shot MSCOCO*.

We conclude that *multimodal attention* mechanisms, either *coattention* or *merged attention*, are a key component to multimodal transformers’ success. Moreover, a shallow or small model with multimodal attention outperforms deeper models with an inferior attention mechanism yet more parameters. Finally, we show that a model’s depth alone is not important; both multimodal attention and depth are needed for best performance.

4.4 Losses

We explore the degree to which MLM, MRM, and ITM losses contribute to our MMT results. We then explore whether a contrastive formulation of the ITM loss—used commonly in self-supervised

	Flickr-ZS	COCO-ZS
MRM + ITM	20.2	9.7
MLM + ITM	41.1	22.4
MRM + MLM + ITM	41.9	21.3

Table 5: Zero-shot retrieval results (R@1) on models trained with different losses.

representation learning and important for our baseline—improves MMT's performance.

Comparing MLM, MRM, and ITM. Table 5 shows performance of our models with different combinations of the masked modeling losses and the image-text loss. With careful hyper-parameter tuning (in particular, decreasing the learning rate from 0.00176 to 0.001 and using cosine decay instead of polynomial decay) we can remove the MRM loss during pretraining and achieve comparable performance on our image retrieval tasks. We found negligible difference when training our base MMT with the different hyper-parameters. We note that our multimodal transformer trained on pixels (Section 4.2) is also trained without a region modeling loss, yet performs similarly to our base MMT. Additionally, our finding is in line with the results of Li et al. (2020b), who achieve strong results without a region modeling loss.

Contrastive ITM Loss. Contrastive losses (e.g., Equation (5)) require sampling many negative examples to achieve good performance and thus can be computationally expensive (e.g., Tian et al., 2019; Miech et al., 2020). In models without multimodal attention (e.g., our baseline model), the computational cost is reduced by caching and reusing negative examples; in such models, since image and text input are processed independently, once image and text features are calculated, they can be considered as negatives for all other training examples in the batch. Due to their multimodal attention, multimodal transformers process image and text examples as pairs and thus cannot share image or text features across training examples. This limits the number of negatives available for these models to the maximum batch size that fits in memory. As a result, to study the role of a contrastive loss with a reasonable number of negatives, we consider our MMT with one multimodal layer. We also examine whether a model with only *modality-specific attention* (here, we use 6 image and 12 language layers) benefits from a contrastive loss since it is easier to increase the negatives in a model without multimodal attention. In both models, we replace the image–text matching classification loss, Equation (4), with a contrastive one, Equation (5).

Model	Loss	Negatives	Flickr-ZS	COCO-ZS
MSA	Cls.	1	15.0	6.9
MSA	Con.	32	17.9	8.3
MSA	Con.	1024	19.7	9.5
MMT-1	Cls.	1	37.3	19.1
MMT-1	Con.	32	35.7	19.1

Table 6: R@1 with a classification ITM loss (cls) and contrastive ITM loss (con) for a MMT with one multimodal layer (MMT-1) and a model which only has modality specific attention (MSA).

Table 6 compares the performance of a single-modality transformer trained with a classification loss to a model trained with a contrastive loss and 32 or 1024 negatives. We observe a notable improvement with the contrastive loss and adding more negatives. We next compare the performance of our one-layer MMT trained with a classification loss and a contrastive loss with 32 negatives (the max we could fit into memory). When training with the contrastive loss, we see no performance difference on *zero-shot MSCOCO* and a small performance degradation on *zero-shot Flickr*. This is surprising given the large body of research demonstrating the benefit of contrastive losses. We conclude that the multimodal attention and MLM loss can help the model learn better representations without relying on stronger image–text losses.

5 Related Work

Multimodal transformers are the first family of multimodal models to be pretrained on large data and applied to a range of different language and vision tasks (Lu et al., 2019; Chen et al., 2020; Tan and Bansal, 2019; Li et al., 2020b,a). The recent image-text transformers share the same backbone but have slight differences in data preprocessing and other architectural choices. Notably, the UNITER model (Chen et al., 2020) achieves state-of-the-art results on most existing image–language benchmarks by using a larger dataset and a number of different loss functions. Huang et al. (2020) removes the need for using image features (taken

from a pretrained object detector) by training models on raw images (pixels). To combine image and text modalities, LXMERT (Tan and Bansal, 2019) and ViLBERT (Lu et al., 2019) propose coattention mechanisms, similar to the coattention originally proposed for VQA (Lu et al., 2016). In ViLBERT, feed-forward layers are applied after the coattention and self-attention layers, whereas in LXMERT, a feed-forward layer is only applied after the self-attention layer.

A few of our findings are similar to observations in prior work: **(i)** LXMERT and ViLBERT show that more layers improve results, **(ii)** ViLBERT and UNITER show that more data boosts performance, and **(iii)** LXMERT shows that transferring BERT weights is not beneficial. In contrast to UNITER, we show that with the right hyper-parameters, the MRM loss is not needed.

Finally, while joint-space approaches to multimodal training are applied to multilingual data (Gella et al., 2017; Sigurdsson et al., 2020), all existing multimodal transformers are applied to English; an interesting future direction is to extend these models to other languages.

Analyzing Multimodal Transformers. Recent analysis work (Singh et al., 2020; Cao et al., 2020) has shed light on different aspects of multimodal transformer models. Singh et al. (2020) study which pretraining data is best when fine-tuning two different multimodal transformer variants—ViLBERT (Lu et al., 2019) and VisualBERT (Li et al., 2019)—on four fine-tuned tasks, whereas we mainly focus on a zero-shot retrieval task across a variety of pretraining datasets, architectural choices, and loss functions. Our results are complementary to this work: Singh et al. (2020) observe that dataset size is not the only factor for good performance and pretraining datasets are better when they match the domain of a downstream task. We take a first step towards quantifying what it means for a pretraining dataset to be similar to a downstream task by analyzing the language used in the pretraining datasets and tasks (Section 4.2).

Cao et al. (2020) consider various probing methods on two models (UNITER and LXMert, Chen et al., 2020; Tan and Bansal, 2019) to study what information is captured in pretraining. Cao et al. (2020) show that while representations become more similar in the last layers of models with merged attention, in coattention models, they are most similar at the first multimodal layer. They also observe that attention heads in merged attention models mostly focus on the language modality, only a few heads are specialized for cross-modality processing, and that attention heads are able to capture some image-text alignment. Our comparisons of merged and coattention is performed in a more controlled setting than that of Cao et al. (2020) and Singh et al. (2020): They compare two models with many small differences other than the attention mechanism; in contrast, we compare the attention mechanisms in the same modeling framework.

6 Discussion

We rigorously examined different aspects of training multimodal transformers (datasets, attention, and losses) that contribute to the quality of their learned representations. We focused on zero-shot image retrieval tasks to evaluate learned representations. Zero-shot tasks are advantageous because they directly measure what a model has learned and do not introduce confounds such as the size of a fine-tuning dataset and its experimental setup. At the same time, datasets do not always capture what they are designed to measure; e.g., Akula et al. (2020) show that models can do well on a referring expression task while ignoring the linguistic structure.Thus, we argue that designing and curating specialized zero-shot evaluation tasks and datasets is an important future direction that will allow us to better understand our models' limitations.

We find the quality of language and the degree to which the language describes its corresponding image (noisiness) plays an important role in our results. Moreover, language-only and image-only pretraining do not notably contribute to the performance of multimodal transformers. These suggest curating less noisy image–text datasets to be more important than relying on single-modality datasets. Previous work has successfully removed some of the noise in automatically harvested datasets through preprocessing (e.g., Sharma et al., 2018) but such approaches are still limited in their robustness to noise, and the far from negligible degree of noise in large-scale real-world datasets (e.g., Ordonez et al., 2011; Miech et al., 2019) still poses a challenge. An alternative approach is to aim to remove this noise by designing models that better tap into statistical regularities of image–text pairs (e.g., Duygulu et al., 2002) and thus are more robust to noise.

We show that multimodal attention—where each modality is informed by both modalities—is crucial in these models' performance. Smaller models with multimodal attention outperform deeper models with no or other multi-head attention mechanisms. This suggests that we can potentially train smaller models (than the existing multimodal transformers) for a given task, especially when the pretraining data is chosen carefully. Moreover, with multimodal attention, we can achieve the best zero-shot retrieval results using a classification loss which uses only one negative example per image–text pair (compare to a contrastive loss with 16384 negatives used in Tian et al., 2019) and also removes the need for mining more hard negatives (Faghri et al., 2017).

Additionally, we observe that comparable results can be achieved without the image (masked region modeling) loss in multimodal transformers. This suggests that our current models are not tapping into the useful signal in the image modality, presumably because of the image loss formulation. An interesting future direction is designing better generative pretraining losses for images; previous work shows that the choice of loss significantly impacts the quality of language representations (Voita and Titov, 2020).

Finally, we believe that examining why and how multimodal transformers perform so well can guide future work in more effectively measuring progress in learning rich visual-linguistic features.

Acknowledgments

Special thanks to Aishwarya Agrawal for detailed comments on our paper. Also, thanks to Angeliki Lazaridou, Andrew Zisserman, Phil Blunsom, Laura Rimell, Stephen Clark, and the anonymous reviewers for their helpful feedback and Sebastian Borgeaud and Cyprien de Masson d'Autume for providing a BERT codebase.

References

Arjun R. Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. *arXiv preprint arXiv:2005.01655*. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.586

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. **DOI:** https://doi.org/10.1109/CVPR.2018.00636

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433. **DOI:** https://doi.org/10.1109/ICCV.2015.279

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. *arXiv preprint arXiv:2005.07310*. **DOI:** https://doi.org/10.1007/978-3-030-58539-6_34

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*. **DOI:** https://doi.org/10.1007/978-3-030-58577-8_7

Mithun Chowdhury, Panda Rameswar, Evangelos Papalexakis, and Amit Roy-Chowdhury. 2018. Webly supervised joint embedding for cross-modal image-text retrieval. In *ACM International Conference on Multimedia*. **DOI:**

https://doi.org/10.1145/3240508.3240712

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics (NAACL)*.

P. Duygulu, K. Barnard, J.F.G. Freitas, and D.A. Forsyth. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, editors, In *European Conference on Computer Vision (ECCV)*, pages 97–112. **DOI:** https://doi.org/10.1007/3-540-47979-1_7

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. *arXiv preprint arXiv:1707.07601*. **DOI:** https://doi.org/10.18653/v1/D17-1303

Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*. **DOI:** https://doi.org/10.1007/s11263-013-0658-4

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913. **DOI:** https://doi.org/10.1109/CVPR.2017.670

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743. **DOI:** https://doi.org/10.18653/v1/D19-1275

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

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. **DOI:** https://doi.org/10.1109/CVPR.2015.7298932

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

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. **DOI:** https://doi.org/10.1007/s11263-016-0981-7

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*. **DOI:** https://doi.org/10.18653/v1/D18-2012

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*.

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*. **DOI:** https://doi.org/10.1609/aaai.v34i07.6795

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. *arXiv preprint arXiv:1908.03557*.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Cho, and Jianfeng Gao. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*.

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, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*, pages 289–297.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual Representations from uncurated instructional videos. In *Computer Vision and Pattern Recognition*. **DOI:** https://doi.org/10.1109/CVPR42600.2020.00990

Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2630–2640. **DOI:** https://doi.org/10.1109/ICCV.2019.00272

Aaron van den Oord, Yazhe Li, Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv: 1807.03748*.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2019. Connecting vision and language with localized narratives. *arXiv preprint arXiv: 1912.03098*. **DOI:** https://doi.org/10.1007/978-3-030-58558-7_38

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

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. **DOI:** https://doi.org/10.18653/v1/P18-1238

Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, João Carreira, Phil Blunsom, and Andrew Zisserman. 2020. Visual grounding in video for unsupervised word translation. In *Computer Vision and Pattern Recognition*. **DOI:** https://doi.org/10.1109/CVPR42600.2020.01086

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. Are we pretraining it right? digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv:2004.08744*.

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 International Conference on Computer Vision*, pages 7464–7473. **DOI:** https://doi.org/10.1109/ICCV.2019.00756

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Empirical Methods in Natural Language Processing*. **DOI:** https://doi.org/10.18653/v1/D19-1514

Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, pages 4223–4232. **DOI:** https://doi.org/10.1109/CVPR.2018.00444

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.

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.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. *arXiv preprint arXiv:2003.12298*. **DOI:** https://doi.org/10.18653/v1/2020.emnlp-main.14

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *CVPR*. **DOI:** https://doi.org/10.1109/CVPR.2016.541

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*.

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*, pages 5753–5763.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2019. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*.

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. **DOI:** https://doi.org/10.1162/tacl_a_00166

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361. **DOI:** https://doi.org/10.18653/v1/W18-5448