Supervised Multimodal Bitransformers for Classifying Images and Text

Douwe Kiela[†], Suvrat Bhooshan[†], Ethan Perez[‡], Hamed Firooz[†], Davide Testuggine[†]
[†]Facebook AI; [‡]New York University
{dkiela,sbh,ethanperez,mhfirooz,davidet}@fb.com

Abstract

Self-supervised bidirectional transformer models such as BERT have led to dramatic improvements in a wide variety of textual classification tasks. The modern digital world is increasingly multimodal, however, and textual information is often accompanied by other modalities such as images. We introduce a supervised multimodal bitransformer model that fuses information from text and image encoders, and obtain at or near state-of-the-art performance on various multimodal classification benchmark tasks, outperforming strong baselines, including on hard test sets specifically designed to measure multimodal performance. Surprisingly, we find that our straightforward method is competitive on these tasks with self-supervised ViLBERT, a multimodal "BERT for vision-and-language" approach.

1. Introduction

Many of the classification problems that we face in the modern digital world are multimodal in nature: textual information on the web rarely occurs alone, and is often accompanied by images, sounds, videos, or other modalities. Recent advances in representation learning for natural language processing, such as BERT [12], have led to dramatic improvements in text-only classification problems. In this work, we propose and examine a straightforward yet highly effective method for making bidirectional transformers capable of going beyond text-only data, allowing them to handle the type of multimodal classification settings commonly found in real-world internet data.

Various self-supervised multimodal architectures have recently been proposed, such as ViLBERT [27], VisualBERT [26], LXMERT [41] and VL-BERT [39]. Contrary to those architectures, our model relies on nothing but individually pre-trained unimodal encoders, which are subsequently fused in a supervised fashion. We include a comparison against ViLBERT as a representative of the alternative.

We evaluate on the following three multimodal classification tasks, which have been used in the past specifically for evaluating multimodal classification architectures [21]: MM-IMDB [2], Food101 [47] and V-SNLI [45]. The reason for choosing these tasks is that 1) we argue that real-world multimodal classification on internet data is somewhat different from popular tasks like VQA and image-caption retrieval, taking these tasks as representatives of that goal; and 2) we are interested in exploring how bitransformer models perform beyond the "standard" text-only or vision-and-language tasks, in tasks like these where the data might be less clean, the text longer or the modalities less balanced.

A desired characteristic of multimodal models is improved performance on cases where high-quality multimodal information is available—i.e., the whole should strictly outperform the sum of its parts. To measure if this is indeed the case, we construct novel hard test sets consisting of examples that unimodal systems failed to classify correctly, specifically designed to measure the multimodal performance of a system.

Our findings indicate that the proposed multimodal bitransformer model outperforms the competitive approach of a deep network on top of concatenated image and text-only bitransformer features, even if we give that model strictly more parameters. We argue that this is due to the multimodal bitransformer's ability to employ self-attention over both modalities simultaneously, providing earlier and more fine-grained multimodal fusion. Furthermore, we find that our straightforward method approaches or matches the more sophisticated ViLBERT model on these tasks. These results show that the proposed method constitutes a not-to-be-ignored baseline for future work in multimodal classification, as it is not only competitive but straightforward to implement using existing self-supervised methods.

2. Multimodal Bitransformers

There is a long history, both in natural language processing and computer vision, of transfer learning from pre-trained representations. Self-supervised word and sentence embeddings [8, 28, 24] have become ubiquitous in natural language processing. In computer vision, transferring from supervised ImageNet features is the de facto standard in computer vision [29, 37].

While supervised data in NLP has also proven useful for

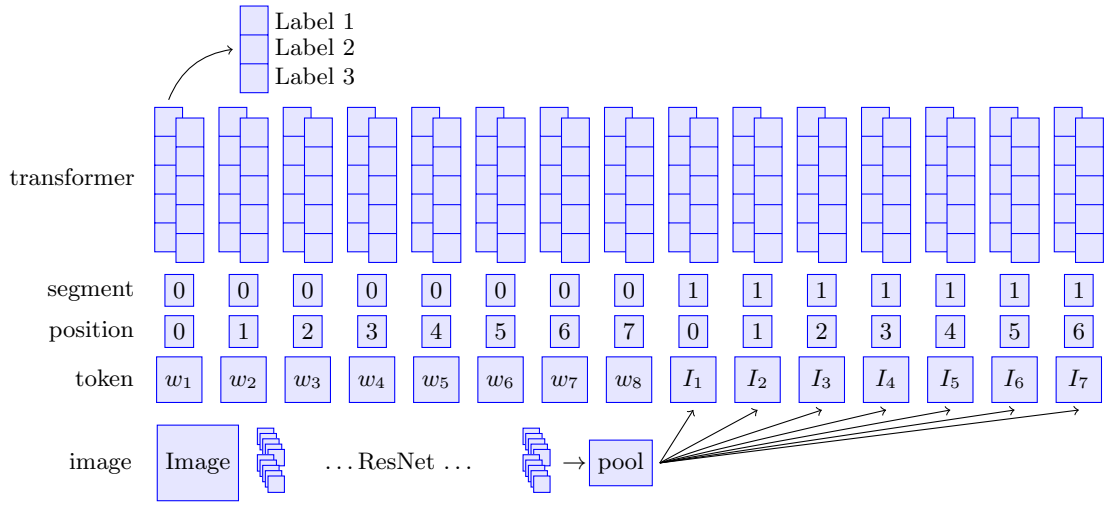

Figure 1: Illustration of the multimodal bitransformer architecture.

universal sentence representations [9], the field was recently revolutionized by the idea of fine-tuning self-supervised language modeling systems [10]. Language modeling enables systems to learn embeddings in a contextualized fashion, leading to improved performance on a variety of tasks [34, 18]. Training transformers [43] on large quantities of data yielded even better results [36]. BERT [12] improved on this further by training transformers bidirectionally (which we refer to as bitransformers) and changing the objective to masking, leading to state-of-the-art performance on a wide variety of important tasks.

We introduce a straightforward yet highly effective multimodal bitransformer model that combines the text-only self-supervised representations from natural language processing with the power of state-of-the-art convolutional neural network architectures from computer vision. See Figure 1 for an illustration of the architecture. In what follows, we describe the different components in more detail.

2.1. Image Encoder

In computer vision it is common to transfer the fully connected penultimate layer of a pre-trained convolutional neural network [37], where the output is often the result of a pooling operation over feature maps. Within the multimodal bitransformer architecture, however, we can handle arbitrary lengths and are not committed to a particular number of inputs. Thus, we generalize the final pooling layer to yield not one single output vector, but N separate image embeddings, unlike in a regular convolutional neural network. In this case we use a ResNet-152 [17] with average pooling over $K \times M$ grids in the image, yielding $N = KM$ output vectors of 2048 dimensions each, for every image. Input images are resized, center-cropped at 224x224 and normalized.

2.2. Multimodal Transformer Input Layer

We use a bidirectional transformer architecture initialized with pre-trained BERT weights. The architecture takes contextual embeddings as input, where each contextual embedding is computed as the sum of separate D-dimensional segment, position and token embeddings. We learn weights $W_n \in \mathbb{R}^{P \times D}$ to project each of the N image embeddings to D-dimensional token input embeddings:

$$I_n = W_n f(img, n), \tag{1}$$

where $f(\cdot, n)$ is the n-th output of the image encoder's final pooling operation.

For tasks that consist of a single text and single image input, we assign token inputs to one segment ID and image embeddings to another. We use 0-indexed positional coding for each segment, i.e., we start counting from 0 for each segment. The architecture can be straightforwardly generalized to an arbitrary number of modalities, as we show for the V-SNLI task, which consists of three inputs. Since pretrained BERT itself has only two segment embeddings, in those cases we initialize additional segment embeddings as $s_i = \frac{1}{2}(s_0 + s_1) + \epsilon$ where s_i is a segment embedding for $i \geq 2$ and $\epsilon \sim \mathcal{N}(0, 1e^{-2})$. Note that a strong advantage of our method is that it works even if not every modality is present in each example (i.e., if we only have text, or only an image, the bidirectional transformer still learns an appropriate representation for classification).

2.3. Classification

We use the first output of the final layer of the bitransformer as input to a classification layer $\texttt{clf}(x) = Wx + b$ where $W \in \mathbb{R}^{D \times C}$, with D as the transformer dimensionality and C as the number of classes. For multilabel tasks,

Dataset	Source	Type	Train	Dev	Test	# Inputs	# Classes
MM-IMDB	[2]	Multilabel	15552	2608	7799	2	23
FOOD101	[47]	Multiclass	60101	5000	21695	2	101
V-SNLI	[45]	Multiclass	545620	9842	9842	3	3

Table 1: Evaluation tasks used for evaluating performance.

which can have more than one right answer, we apply a sigmoid on the logits and train with a binary cross-entropy loss for each output class (during inference time, we set the threshold at .5); for multiclass tasks we apply a softmax on the logits and train with a regular cross-entropy loss.

2.4. Pre-training

The image encoder was pre-trained on ImageNet [11]. We use the ResNet-152 [17] implementation and weights available in PyTorch [30] through torchvision. We use the pre-trained 12-layer 768-dimensional base-uncased model for BERT [12], trained on the English version of Wikipedia.

2.5. Fine-tuning and Multimodal Optimization

Our architecture consists of a mixture of pre-trained and randomly initialized components. In NLP, BERT is commonly fine-tuned in its entirety, and not transfered as an encoder with fixed parameters, as used to be the case in e.g. SkipThought [24] and InferSent [9]. In computer vision, the convolutional network is often kept fixed [37], although it has been found that unfreezing the convolutional network during later stages of training leads to significant improvements, e.g. in image-caption retrieval [14].

Training multimodal models is not at all trivial, especially when it comes to the optimization strategy [46]. In the multimodal bitransformer model we propose here, ResNet outputs are mapped to BERT's token space using a set of randomly initialized mappings W_n. An additional contribution of this work is to explore a solution for optimization across multiple modalities, namely: we freeze and unfreeze the image and text encoding components at different stages, which we treat as a hyperparameter. If we first learn to map image embeddings to an appropriate subspace of the text encoder's input space, we may expect the network to make more use of visual information than otherwise. In other words, since the text modality is likely to dominate, we want to give the visual modality a chance. We experiment with different settings.

2.6. Availability

The code used to train the models in this paper is available at https://github.com/facebookresearch/mmbt. It is built using PyTorch [30] and on top of HuggingFace Transformers [49].

3. Approach

In this section, we describe how we evaluate performance, discuss the baselines and provide other experimental details.

3.1. Evaluation

We evaluate on a diverse set of multimodal classification tasks. We compare against two tasks also used in [21]: MM-IMDB [2] and FOOD101 [47]. To illustrate that the architecture generalizes beyond two input types, we additionally evaluate on V-SNLI [45], which consists of (premise, hypothesis, image) triplets. Visual grounding has shown to improve NLI performance [23]. In what follows, we describe the tasks in more detail. See Table 1 for dataset statistics and Table 2 for examples.

- **MM-IMDB** The MM-IMDB dataset [2] consists of movie plot outlines and movie posters. The objective is to classify each movie by genre. This is a multilabel prediction problem, i.e., one movie can have multiple genres. The dataset was specifically introduced by [2] to address the relative scarcity of high-quality multimodal classification datasets.
- **FOOD101** The UPMC FOOD101 dataset [47] contains textual recipe descriptions for 101 food labels. The recipes were scraped from web pages and subsequently cleaned to extract text data. Each page was matched with a single image, where the images were obtained by querying Google Image Search for the given category (which might be noisy). The objective is to find the corresponding food label for each recipe-image combination.
- **V-SNLI** The V-SNLI dataset is based on the SNLI dataset [6]. The objective is to classify a premise and hypothesis, with associated image, into one of three categories: entailment, neutral or contradition. The SNLI dataset was created by having Turkers provide hypotheses for premises that were derived from captions in the Flickr30k dataset [50]. [45] put the original images and the premise-hypothesis pairs back together in order to create what they refer to as a grounded entailment task, called V-SNLI. V-SNLI also comes with a hard subset of the test set, originally created for SNLI, where a hypothesis-only classifier fails [16].

Dataset	Label	Image	Text
MM-IMDB	Comedy		Brian is born in a stable on Christmas, right next to You Know Who. The wise men appear and begin to distribute gifts. The star moves further, so they take it all back and move on. This is how Brian's life goes. [...] He joins the Peoples' Front of Judea, one of several dozen separatist groups who actually do nothing, but really hate the Romans. While not about Jesus, it is about those who hadn't time, or interest to listen to his message. Many Political and Social comments.
FOOD101	Cup cakes		[...] simple and oh so delicious these basic cupcakes make a lovely birthday treat makes 24 ingredients 200g unsalted butter softened 1 teaspoon vanilla extract 1 cup caster sugar 3 eggs 2 1 2 cups self raising flour [...] bake for 15 to 17 minutes alternatively for 1 tablespoon capacity mini muffin pans use 1 tablespoon mixture bake for 10 to 12 minutes 4 stand cakes in pans for 2 minutes transfer to a wire rack to cool 5 decorate to suit your party theme [...]
V-SNLI	Entailment		**Premise**: Children smiling and waving at camera. **Hypothesis**: There are children present.

Table 2: Example data for each of the datasets.

3.2. Baselines

It is important to establish strong baselines for our methods. For example, [21] found that in many cases, text-only systems like FastText [19] perform surprisingly well. Here, we compare against strong unimodal baselines, as well as the highly competitive baseline of concatenating multimodal features as direct features for the classifier. In all cases we use a single layer classifier, fine-tuning the entire model end-to-end. We describe each of the baselines in more detail below.

- **Bag of words (Bow)** We sum 300-dimensional GloVe embeddings [31] (trained on Common Crawl) for all words in the text, ignoring the visual features, and feed it to the classifier.
- **Text-only BERT (Bert)** We take the first output of the final layer of a pre-trained base-uncased BERT model, and feed it to the classifier.
- **Image-only (Img)** We take a standard pre-trained ResNet-152 with average pooling as output, yielding a 2048-dimensional vector for each image, and classify it in the same way as the other systems.
- **Concat Bow + Img (ConcatBow)** We concatenate the outputs of the Bow and the Img baselines. Concatenation is often used as a strong baseline in multimodal methods. In this case, the input to the classifier is 2048+300-dimensions.
- **Late Fusion** We take our two best Bert and Img models, and average their scores to get the final prediction.
- **FiLMBert** We combine FiLM [32] with BERT, where the BERT model predicts feature-wise gains and biases for a ConvNet classifier. We use fixed ResNet-152 features as input to the ConvNet, similar to [32].
- **Concat BERT + Img (ConcatBert)** We concatenate the outputs of the Bert and the Img baselines. In this case, the input to the classifier is 2048+768-dimensions. This is a competitive baseline, since it combines the best encoder for each modality such that the classifier has direct access to the encoder outputs.

3.3. Making the Problem Harder

While we evaluate on a diverse set of multimodal classification tasks, there are actually surprisingly few high-quality tasks of this nature. In many cases, the textual modality is overly dominant, sometimes making it difficult to tease apart differences between different multimodal methods, or to identify if it is actually worthwhile to incorporate multimodal information in the first place. As we observed earlier, [16] created hard subsets of the SNLI dataset where a hypothesis-only baseline was unable to correctly classify the example, rectifying artifacts in the original SNLI test set. Here, we follow a similar approach, and create hard multimodal test sets for our other two tasks.

	MM-IMDB	FOOD-101	V-SNLI
GMU [2]	51.4 / 63.0	-	-
CentralNet [44]	56.1 / 63.9	-	-
W2V + VGG Fusion [47]	-	85.1	-
Bilinear-gated [21]	- / 62.3	90.8	-
V-BiMPM [45]	-	-	86.99
Bow	38.1±.2 / 45.6±.2	72.4±.3	48.6±.3
Img	32.5±.7 / 44.4±.3	63.2±.6	33.8±.3
Bert	59.9±.3 / 65.4±.1	87.2±.1	90.1±.3
Late Fusion	59.4±.1 / 66.2±.0	91.1±.1	90.1±.0
ConcatBow	43.8±.4 / 53.6±.4	79.0±.9	49.5±.1
FiLMBert	59.7±.4 / 65.1±.2	90.2±.3	89.1±.2
ConcatBert	60.5±.3 / 65.9±.2	90.0±.6	90.2±.4
MMBT	**61.6±.2 / 66.8±.1**	**92.1±.1**	**90.4±.1**

Table 3: Main Results. MM-IMDB is Macro F1 / Micro F1; others are Accuracy.

	MM-IMDB Hard	FOOD-101 Hard	V-SNLI Hard
Bow	50.6±.4 / 54.7±.4	72.7±.5	27.2±.2
Img	39.1±.9 / 48.2±.9	63.4±.6	32.3±.3
Bert	64.7±.5 / 67.0±.3	87.3±.2	79.7±.4
Late fusion	61.7±.9 / 66.4±.5	91.3±.5	79.6±.4
ConcatBert	64.9±.4 / 67.2±.2	90.4±.3	79.9±.9
MMBT	**65.3±.4 / 68.6±.4**	**92.4±.3**	**80.3±.1**

Table 4: Hard Subsets. MM-IMDB is Macro F1 / Micro F1; others are Accuracy.

We construct hard test sets by take the examples where the Bert and Img classifier predictions are most different from the ground truth classes in the test set, i.e. examples that maximize $p(a \neq t|I)p(a \neq t|T)$, where I and T are the image and textual information respectively, a is the predicted answer and t is the correct answer. We take the top 10% of the most-different examples as the hard cases in the new test sets. The idea is that these are the examples that require more sophisticated multimodal reasoning, allowing us to better examine multimodal-specific performance.

3.4. Other Implementation Details

For all models, we sweep by over the learning rate (in $\{1e^{-4}, 5e^{-5}\}$) and early stop on validation accuracy for the multiclass datasets, and Micro-F1 for the multilabel dataset. We additionally sweep over the number of epochs to keep the text and visual encoders fixed, as well as the number of image embeddings to use as input (see also Section 5 for a detailed analysis of these hyperparameters). For the Bert models, we use BertAdam [12] with a warmup rate of 0.1; for the other models we use regular Adam [22]. Since not all datasets are balanced, we weigh the class labels by their inverse frequency. Code, models and the benchmark suite will be made available at [GITHUB-URL-ANONYMIZED].

4. Results

The main results can be found in Table 3. In each case, we show mean performance over 5 runs with random seeds together with the standard deviation. We compare against the results of [21] on MM-IMDB and FOOD101. They found that a bilinear-gated model worked best, meaning that one of the two input modalities is sigmoided and then gates over the other input bilinearly, i.e. by taking an outer product. Note that in our case, with 2048-dimensional ResNet outputs and 768-dimensional Bert outputs, bilinear gated would need a $2048 \times 768 \times 101$-dimensional output layer (approximately 158M parameters just for the classifier on top), which is not practical. Still, it is a useful comparison to see if we can beat it with a deeper model.

On MM-IMDB, we also compare against Gated Multimodal Units, as introduced by [2], which are a special recurrent unit specifically designed for multimodal fusion (which similarly has one modality gate over the other). In addition, we compare to CentralNet [44], a multilayer approach for multimodal fusion that currently holds the state of the art

Image	Text
	Mulan is a girl, the only child of her honored family. When the Huns invade China, one man from every family is called to arms. Mulan's father, who has an old wound and cannot walk properly, decides to fight for his country and the honor of his family though it is clear that he will not survive an enemy encounter. [..] After being spotted and pursued by the enemies, an impasse situation in the mountains forces Mulan to come up with an idea. But her real gender will no longer be a secret. She decides to risk everything in order to save China.
Gold labels: Animation, Adventure, Family, Fantasy, Musical, War **Bow**: Adventure, Drama — **Img**: Action, Drama, Romance — **MMBT**: Animation, Adventure, Family, War	
	Izo (Kazuya Nakayama) is an assassin in the service of Hanpeida (Ryosuke Miki), a Tosa lord and Imperial supporter. After killing dozens of the Shogun's men, Izo is captured and crucified. Instead of being extinguished, his rage propels him through the space-time continuum to present-day Tokyo, where he finds himself one with the city's homeless. Here Izo transforms himself into a new, improved killing machine, his entire soul still enraged by his treatment in his past life. His response to the powers-that-be, is the sword.
Gold labels: Action, Drama, Fantasy, Horror, Sci-Fi, Thriller, War **Bow**: Action — **Img**: Drama, Horror — **MMBT**: Action, Drama, Fantasy, Sci-Fi	

Table 5: Example data for the MM-IMDB Hard (ground truth) test set.

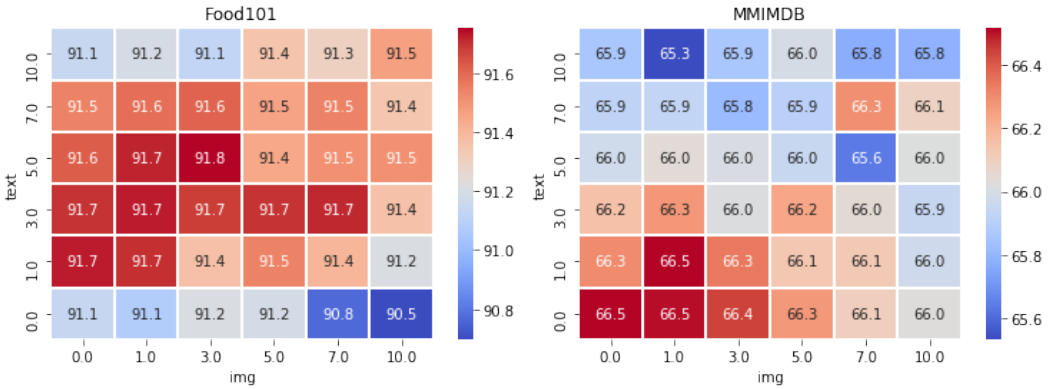

Figure 2: Analysis of freezing pre-trained text and image components for N epochs of training.

on this dataset. For FOOD101, we include the original results from the paper [47], which were obtained by concatenating word2vec and VGGNet features and classifying. For V-SNLI, we compare to the state-of-the-art Visual Bilateral Multi-Perspective Matching (V-BiMPM) model of [45].

We find that the multimodal bitransformer (MMBT) outperforms all other models by a significant margin. Late fusion, FiLMBert and ConcatBert perform similarly, with the latter probably being the strongest baseline. We speculate that the cause of MMBT's improvement over ConcatBert is its ability to let information from different modalities interact at different levels, via self-attention, rather than only at the final layer. Part of the improvement comes from Bert's superior performance (which makes sense, given text's dominance), but even then MMBT improves over Bert by e.g. $\sim$3% on MM-IMDB Macro-F1 and an impressive $\sim$6% on Food101. In all cases, multimodal models outperform their direct unimodal counterparts.

4.1. Hard Testsets

Table 4 reports the results on the hard test sets. Recall that these were created by selecting examples where unimodal (Bert and Img) classifiers differed the most from the ground truth, meaning that these results provide insight into true multimodal performance. We also report results on VSNLI$_{hard}$ [16].

We observe a similar pattern to the main results, with MMBT outperforming the alternatives. Note that on V-SNLI$_{hard}$, [45] report a score of 73.75 for their best-performing architecture, compared to our 80.4. It is also interesting to observe that on that hard test set, the image-only classifier already outperforms the text-only one, which

is definitely not the case for the normal (non-hard) V-SNLI test set. We include example predictions on MM-IMDB together with the ground truth in Table 5.

5. Analysis

In this section, we further explore the appropriate multimodal optimization strategy for (un)freezing unimodal encoders during training. We also compare ConcatBert and MMBT in terms of parameters, and show that MMBT still outperforms ConcatBert when that model has a deeper, multi-layer feedforward neural network classifier.

5.1. Freezing Strategy

We conduct an analysis of whether it helps to initially freeze the different pre-trained components (we keep the number of image embeddings fixed). This would help for instance in learning to map from visual space to the expected token input space of the transformer. The idea is to see if it helps to first learn something about the task outputs and, importantly, how to map to the bitransformer token space from the image embeddings. We can then unfreeze the image encoder, to make the image information maximally useful, before we unfreeze the bitransformer to tune the entire system on the task. Figure 2 shows the results, and indeed corroborates the intuition that it is useful to first learn to put the components together, then unfreeze the image encoder, and only after that unfreeze the pre-trained bitransformer. How many epochs to freeze the text encoder for appears to be task-dependent, while unfreezing the image encoder early works best.

5.2. Number of Parameters

A possible explanation for the superior performance of the multimodal bitransformer over ConcatBert could be that it has slightly more parameters (i.e., an additional $2048 \times D$ versus $2048 \times N$, where D is the embedding dimensionality and N is the number of classes), although the difference is small: 168M vs 170M parameters. To investigate this, we also compare against a ConcatBert with a 2-layer and 3-layer multi-layer perceptron (MLP) classifier on top, of 174M and 175M parameters respectively, rather than the single-layer logistic regression in MMBT. For MM-IMDB, ConcatBert-2 and ConcatBert-3 get a Macro-F1 of $60.21 \pm .5$ and $59.71 \pm .4$ and a Micro-F1 of $65.08 \pm .3$ and $64.82 \pm .2$ respectively; while for Food101 they get $91.13 \pm .2$ and $90.27 \pm .2$. This clearly demonstrates (cf. Table 3) that MMBT is superior to ConcatBert, even when we give an already highly competitive baseline even more parameters and a deeper classifier. The results suggest that ConcatBert is more prone to overfitting (we also tried giving it more image embeddings, and the result was the same).

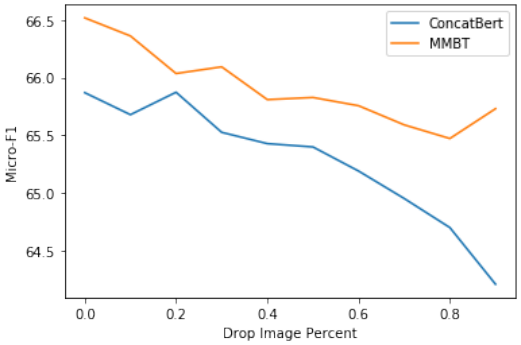

Figure 3: Performance (MicroF1) on MM-IMDB when we drop the image for a percentage of the training set, measuring robustness to missing images.

5.3. Robustness to Missing Modalities

We compare ConcatBert and MMBT in a setting where only a subset of the dataset has images. To our knowledge, this setting has not been explored thoroughly in the literature. It is unclear a priori which of the two models would be more robust to this data regime, and this experiment provides a useful extra dimension for comparing mid-level fusion with the more sophisticated type of fusion provided by MMBT. Figure 3 shows that performance drops with fewer images. It is interesting to observe that MMBT is much more robust to missing images than ConcatBert.

5.4. Comparison to ViLBERT

We examine the effectiveness of fusing unimodally pretrained components by comparing to self-supervised multimodally pretrained models. We take ViLBERT [27] as the canonical example of that class of models. ViLBERT was trained multimodally on images and captions, and is meant to be the "BERT of vision and language". It uses Faster RCNN-extracted bounding boxes , kept fixed during training. Our focus on these somewhat out-of-the-ordinary tasks now proves fruitful, since it allows us to compare these models on a level playing field.

Table 6 shows the results. We compare against a variety of ViLBert models, both the standard pre-trained version as well as the versions fine-tuned for particular tasks like VQA. The latter approach is not proposed in the original ViLBert paper, but similar "two-stage pre-training" approaches have proven effective for fine-tuning BERT on unimodal tasks [35]. We tune using the hyperparameter sets used in that paper: (batch size, learning rate) $\in \{(64, 2e^{-5}), (256, 4e^{-5})\}$. We observe that our straightforward MMBT model is surprisingly competitive. On MM-IMDB, it matches the task-specific ViLBERT models on Macro-F1. On the Hard subset of that dataset, which more accurately measures multimodal performance,

	MM-IMDB	-Hard	FOOD-101	-Hard
MMBT	61.6±.2 / 66.8±.1	**65.3±.4 / 68.6±.4**	92.1±.1	**92.4±.5**
ViLBert-VQA	60.0±.3 / 66.4±.2	62.7±.6 / 66.2±.4	92.1±.1	92.4±.3
ViLBert-VCR	61.6±.3 / 67.6±.2	63.4±.9 / 66.9±.4	92.1±.1	92.1±.3
ViLBert-Refcoco	61.4±.3 / 67.7±.1	63.4±.5 / 67.1±.4	92.2±.1	92.1±.3
ViLBert-Flickr30k	61.4±.3 / 67.8±.1	63.4±.9 / 67.0±.5	92.2±.1	92.2±.3
ViLBert	**63.0±.2 / 68.6±.1**	**65.4±1. / 68.6±.4**	**92.9±.1**	**92.9±.3**

Table 6: Comparison of MMBT to ViLBert on MM-IMDB and FOOD-101.

MMBT matches ViLBert's performance. For FOOD-101, we observe a similar story, with performance being remarkably close, occasionally outperforming task-specific models, in particular on the Hard subset. Our results suggest that self-supervised multimodal pre-training has more room for improvement, and that the supervised fusion of unimodally-pretrained components is remarkably competitive.

Since the proposed method is unimodally pre-trained, it may be more preferable depending on the constraints: if a new breakthrough happens in NLP or CV, it is easy to incorporate that model to get even stronger multimodal classification. This is trivial to do in our setting, but for ViLBERT would require retraining from scratch.

6. Related Work

Neural methods are the standard for almost every modern text and vision classification task. Transformers [43] have been used to encode sequential data for classification with great success when pre-trained for language modeling or language masking and subsequently fine-tuned [36, 12].

The question of how to effectively combine multimodal information, also known as multimodal fusion, has a long history [3]. While concatenation can be considered the default, other fusion methods have been explored e.g. for lexical representation learning [7, 25]. In classification, [21] examine various fusion methods for pre-trained fixed representations, and find that a bilinear combination of data with gating worked best. Our supervised multimodal bitransformer can be seen as incorporating a particular type of fusion mechanism, with interaction between the modalities via self-attention over many different layers.

Applications of multimodal research in NLP range from classification to cross-modal retrieval [48, 15, 38] to image captioning [5] to visual question answering [1] and multimodal machine translation [13]. Multimodal information is also useful in learning human-like meaning representations [4, 20]. Multimodal bitransformers provide what is effectively a deep fusion method. Related deep fusion methods include multimodal transformers [42], CentralNet [44], MFAS [33] and Tensor Fusion Networks [51].

Concurrently with the work presented in this paper, various self-supervised multimodal architectures have been published, e.g. ViLBERT [27], VisualBERT [26], LXMERT [41], VL-BERT [39], VideoBERT [40], and others. Our model differs from these self-supervised architectures in that the individual components are trained unimodally. This has pros and cons: our method is straightforward and intuitive, easy to implement even for existing self-supervised encoders, and already obtains impressive improvements. If a new and better text or vision model comes out, it is trivial to replace components. On the other hand, it is not able to fully leverage multimodal information during self-supervised pre-training. That said, it does potentially have access to orders of magnitude more unimodal data. In other words, if anything, these supervised multimodal bitransformers should provide a strong baseline for gauging if self-supervised multimodal bitransformers actually outperform their unimodal peers.

7. Conclusion

In this work, we introduced a supervised multimodal bitransformer model. We compared against several baselines on a variety of tasks, including on hard test sets created specifically for examining multimodal performance (i.e., where unimodal performance fails). We find that the proposed architecture significantly outperforms the existing state of the art, as well as strong baselines. We then conducted an analysis of multimodal optimization, exploring a freezing/unfreezing strategy, and looked at the number of parameters, showing that the strong baseline with more parameters and a deeper classifier was still outperformed.

Our architecture consists of components that were pretrained individually as unimodal tasks, which already showed great improvements over alternatives. It is as of yet unclear if multimodal self-supervised models are going to be generally useful. We compared to ViLBERT and showed that the proposed model performs competitively. The methods outlined here should serve as a useful and powerful baseline to gauge the performance of self-supervised multimodal models. Supervised multimodal bitransformers are straightforward and intuitive, and importantly, are easy to implement even for existing self-supervised encoders.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 8

[2] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017. 1, 3, 5

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. 8

[4] Marco Baroni. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13, 2016. 8

[5] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016. 8

[6] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015. 3

[7] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2014. 8

[8] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008. 1

[9] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017. 2, 3

[10] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087, 2015. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL*, 2019. 1, 2, 3, 5, 8

[13] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark, September 2017. 8

[14] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 3

[15] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 8

[16] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018. 3, 4, 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3

[18] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *Proceedings of ACL*, 2018. 2

[19] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016. 4

[20] Douwe Kiela. *Deep Embodiment: Grounding Semantics in Perceptual Modalities*. PhD thesis, University of Cambridge, Computer Laboratory, 2017. 8

[21] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1, 3, 4, 5, 8

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] Jamie Kiros, William Chan, and Geoffrey Hinton. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 922–933, 2018. 3

[24] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015. 1, 3

[25] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*, 2015. 8

[26] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1, 8

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv preprint arXiv:1908.02265*, 2019. 1, 7, 8

[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 1

[29] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations

using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014. 1

[30] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. PyTorch: Tensors and dynamic neural networks in python with strong GPU acceleration. Technical report, PyTorch, 2017. 3

[31] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 4

[32] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of AAAI*, 2018. 4

[33] Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. MFAS: multimodal fusion architecture search. arxiv preprint 1903.06496, 2019. 8

[34] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *Proceedings of NAACL*, 2018. 2

[35] Jason Phang, Thibault Févry, and Samuel R. Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *CoRR*, abs/1811.01088, 2018. 7

[36] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. 2, 8

[37] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014. 1, 2, 3

[38] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013. 8

[39] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 1, 8

[40] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *arXiv preprint arXiv:1904.01766*, 2019. 8

[41] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 1, 8

[42] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. *arxiv preprint 1906.00295*, 2019. 8

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 8

[44] Valentin Vielzeuf, Alexis Lechervy, Stephane Pateux, and Frederic Jurie. Centralnet: a multilayer approach for multimodal fusion. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018. 5, 8

[45] Hoa Trong Vu, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt. Grounded textual entailment. In *Proceedings of COLING*, page 2354–2368, 2018. 1, 3, 5, 6

[46] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal networks hard? *arXiv preprint arXiv:1905.12681*, 2019. 3

[47] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2015. 1, 3, 5, 6

[48] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011. 8

[49] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019. 3

[50] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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, 2014. 3

[51] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. 8