

Supervised Multimodal Bitransformers for Classifying Images and Text

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Davide Testuggine

Facebook AI

{dkiela,sbh,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 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.

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 (Devlin et al. 2019), have led to dramatic improvements in text-only classification problems. In this work, we propose 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.

Vision and language research is dominated by visual question answering (Antol et al. 2015), which is seen as requiring a deep understanding of vision, language and commonsense knowledge. Here, we take a somewhat different stance, and argue that data on the internet often is not that balanced across modalities: usually the textual modality tends to dominate, and text may or may not be augmented or accompanied by images. That is to say, we focus on the rather different problem of straight-up text-dominated multimodal classification, and evaluate on the following three tasks, which have been used in the past specifically for evaluating multimodal classification architectures (Kiela et al. 2018): MM-IMDB (Arevalo et al. 2017), Food101 (Wang et al. 2015) and V-SNLI (Vu et al. 2018).

A desired characteristic of multimodal models is improved performance on cases where high-quality multi-

modal information is indeed available. To that end, 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 naive but highly competitive approach of concatenating pre-trained image features with text-only bitransformer features, even if we put a deeper architecture on top of that model and give it 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.

Concurrently with this work, various self-supervised multimodal architectures have been proposed, such as ViL-BERT (Lu et al. 2019), VisualBERT (Li et al. 2019), LXMERT (Tan and Bansal 2019) and VL-BERT (Su et al. 2019). Contrary to those architectures, our model relies on nothing but individually pre-trained unimodal encoders, which are subsequently fused in a supervised fashion. We show that this straightforward and intuitive approach, which is easy to implement even for existing self-supervised encoders, already obtains impressive improvements.

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 (Collobert and Weston 2008; Mikolov et al. 2013; Kiros et al. 2015) have become ubiquitous in natural language processing. In computer vision, transferring from supervised ImageNet features is the de facto standard in computer vision (Oquab et al. 2014; Razavian et al. 2014).

While supervised data in NLP has also proven useful for universal sentence representations (Conneau et al. 2017), the field was recently revolutionized by the idea of fine-tuning self-supervised language modeling systems (Dai and Le 2015). Language modeling enables systems to learn embeddings in a contextualized fashion, leading to improved performance on a variety of tasks (Peters et al. 2018; Howard and Ruder 2018). Training transformers (Vaswani et al. 2017) on large quantities of data yielded even better

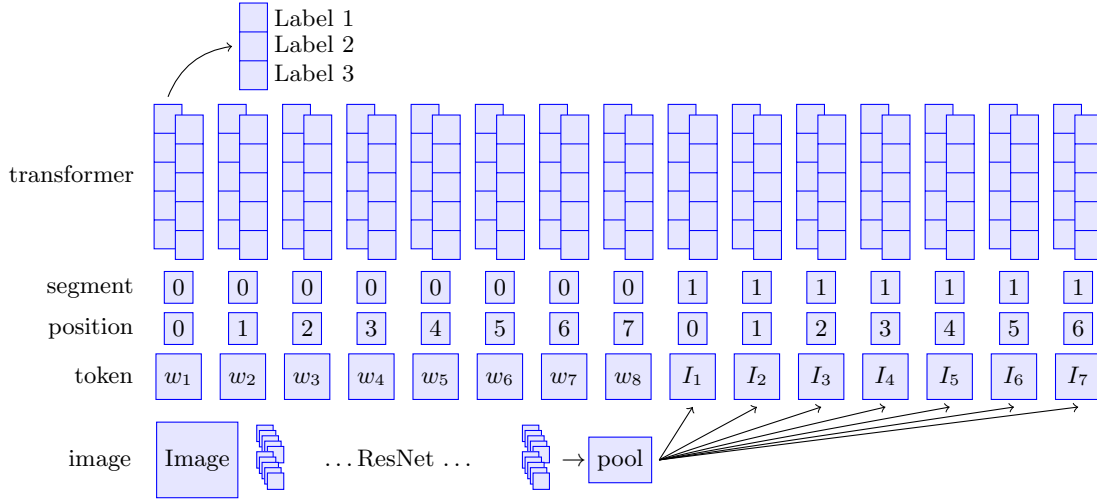


Figure 1: Illustration of the multimodal bitransformer architecture.

results (Radford et al. 2018). BERT (Devlin et al. 2019) 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 multimodal bitransformer model that enhances the strength of 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 this straightforward and intuitive model architecture. In what follows, we describe the different components in more detail.

Image Encoder

In computer vision it is common to transfer the fully connected **penultimate** layer of a pre-trained convolutional neural network (Razavian et al. 2014), 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 (He et al. 2016) 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, as is the standard for these networks.

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 $W_n \in \mathbb{R}^{2048 \times D}$ to map each of the N image embeddings to

D -dimensional token input embeddings via an affine transformation:

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

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 pre-trained 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).

Classification

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

Pre-training

The image encoder was pre-trained on ImageNet (Deng et al. 2009). We use the ResNet-152 (He et al. 2016) implementation and weights available in PyTorch (Paszke et al.

Dataset	Source	Type	Train	Dev	Test	# Inputs	# Classes
MM-IMDB	(Arevalo et al. 2017)	Multilabel	15552	2608	7799	2	23
FOOD101	(Wang et al. 2015)	Multiclass	60101	5000	21695	2	101
V-SNLI	(Vu et al. 2018)	Multiclass	545620	9842	9842	3	3

Table 1: Evaluation tasks used for evaluating performance.

2017) through torchvision. We use the pre-trained 12-layer 768-dimensional base-uncased model for BERT (Devlin et al. 2019), trained on the English version of Wikipedia.

Fine-tuning and Multimodal Optimization

Our architecture consists of a mixture of pre-trained and randomly initialized components. In NLP it is common to fine-tune BERT in its entirety, and not to transfer the encoder while keeping its parameters fixed, as used to be the case in e.g. SkipThought (Kiros et al. 2015) and InferSent (Conneau et al. 2017). In computer vision, the convolutional network is often kept fixed (Razavian et al. 2014), 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 (Faghri et al. 2017).

Training multimodal models is not at all trivial, especially when it comes to the optimization strategy (Wang, Tran, and Feiszli 2019). 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 simple 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. For example, 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 it otherwise would. 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.

Approach

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

Evaluation

We evaluate on a diverse set of multimodal classification tasks. We compare against two tasks also used in (Kiela et al. 2018): MM-IMDB (Arevalo et al. 2017) and FOOD101 (Wang et al. 2015). To illustrate that the architecture generalizes beyond two input types, we additionally evaluate on V-SNLI (Vu et al. 2018), which consists of (premise, hypothesis, image) triplets. 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 (Arevalo et al. 2017) 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

(Arevalo et al. 2017) to address the relative scarcity of high-quality multimodal classification datasets.

- **FOOD101** The UPMC FOOD101 dataset (Wang et al. 2015) 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 (Bowman et al. 2015). The objective is to classify a premise and hypothesis, with associated image, into one of three categories: entailment, neutral or contradiction. The text-only SNLI dataset was created by having Turkers provide hypotheses for premises that were derived from captions in the Flickr30k dataset (Young et al. 2014). (Vu et al. 2018) 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 (Gururangan et al. 2018).

Baselines

It is important to establish strong baselines for our methods. For example, (Kiela et al. 2018) found that in many cases, text-only systems like FastText (Joulin et al. 2016) 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. We describe each of the baselines in more detail below.

- **Bag of words (Bow)** We sum 300-dimensional GloVe embeddings (Pennington, Socher, and Manning 2014) (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 a single average pooling operation 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

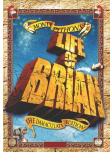
Dataset	Label	Image	Text
MM-IMDB	Comedy		Brian is born in a stable on Christmas, right next to You Know Who. The wise men appear and begin to distribute gifts. The star moves further, so they take it all back and move on. This is how Brian's life goes. [...] He joins the Peoples' Front of Judea, one 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.

is often used as a strong baseline in multimodal methods. In this case, the input to the classifier is 2048+300-dimensions.

- **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 highly competitive baseline, since it combines the best encoder for each modality such that the classifier has direct access to the encoder outputs.

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, (Gururangan et al. 2018) 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.

We examine two ways of constructing hard test sets. In the **hard ground-truth test set**, we take the examples where the Bow 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. In the **hard disagreement test set** alternative, we do not compare to the ground truth, but instead look at cases where the Bow and Img classifiers disagree with *each other* the most. 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 examples that require more sophisticated multimodal reasoning, allowing us to examine more closely whether a new architecture works better.

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 for a detailed analysis of these hyperparameters). For the Bert models, we use **BertAdam** (Devlin et al. 2019) with a warmup rate of 0.1; for the other models we use regular Adam (Kingma and Ba 2014). Since not all datasets are balanced, we **weight the class labels by their inverse frequency**. Code, models and the benchmark suite will be made available soon.

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 (Kiela et al. 2018) 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

	MM-IMDB	FOOD101	V-SNLI
GMU (Arevalo et al. 2017)	54.1 / 63.0	-	-
Word2vec+VGG Fusion (Wang et al. 2015)	-	85.1	-
Bilinear-Gated (Kiela et al. 2018)	- / 62.3	90.8	-
V-BiMPM (Vu et al. 2018)	-	-	86.99
Bow	38.5 \pm .6 / 46.3 \pm .7	72.4 \pm .8	49.4 \pm .2
Img	32.3 \pm .5 / 44.7 \pm .5	63.1 \pm .3	34.0 \pm .4
Bert	59.5 \pm .3 / 65.2 \pm .2	87.3 \pm .1	90.2 \pm .3
ConcatBow	43.4 \pm .4 / 53.3 \pm .3	79.2 \pm .7	49.1 \pm .3
ConcatBert	60.6 \pm .2 / 66.1 \pm .1	90.3 \pm .4	89.9 \pm .3
MMBT	61.1\pm.4 / 66.4\pm.1	92.2\pm.1	90.5\pm.1

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

	MM-IMDB - Hard		FOOD101 - Hard		V-SNLI
	ground truth	disagreement	gr. truth	disagr.	hard
Bow	50.1 \pm .7 / 53.8 \pm .6	41.2 \pm .6 / 48.0 \pm .8	74.6 \pm .8	71.0 \pm .5	27.6 \pm .3
Img	38.9 \pm .1 / 46.4 \pm .1	33.2 \pm .1 / 44.2 \pm .1	63.3 \pm .2	62.1 \pm .6	32.8 \pm .8
Bert	64.2 \pm .7 / 67.2 \pm .3	59.2 \pm .3 / 65.8 \pm .5	88.6 \pm .2	86.4 \pm .3	80.2 \pm .6
ConcatBert	64.9 \pm .5 / 67.5 \pm .2	60.9\pm.9 / 66.5 \pm .5	90.8 \pm .3	89.2 \pm .8	79.5 \pm .3
MMBT	65.7\pm.5 / 68.5\pm.4	60.6 \pm .8 / 67.0\pm.5	92.3\pm.3	91.1\pm.4	80.4\pm.3

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

to see if we can beat it with a deeper model. On MM-IMDB, we also compare against **Gated Multimodal Units**, as introduced by (Arevalo et al. 2017), which are a special recurrent unit specifically designed for multimodal fusion (which similarly has one modality gate over the other). For FOOD101, we include the original results from the paper (Wang et al. 2015), 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 (Vu et al. 2018).

We find that the multimodal bitransformer (MMBT) outperforms all other models by a significant margin. ConcatBert, the strongest baseline, is closest in performance. 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.

Hard Testsets

Table 4 reports the results on the hard test sets. Recall that for MM-IMDB and FOOD101, we created two different versions: one where unimodal (Bow and Img) classifiers disagree the most from the ground truth; and one where they disagree the most from each other. We also report results on $V-SNLI_{hard}$ (Gururangan et al. 2018).

We observe that MMBT outperforms the other methods

here too, this time by an even larger margin. This intuitively should make sense, as the datasets are constructed specifically to make it so that unimodal classifiers individually have a hard time getting the answer right. The difference between ConcatBert and MMBT appears to be particularly prevalent in the ground truth hard test sets. Note that on $V-SNLI_{hard}$, (Vu et al. 2018) 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 V-SNLI test set. We include example outputs from the different classifiers on MM-IMDB, as well as the ground truth, in Table 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 if we give it a deeper feedforward neural network classifier, consisting of multiple layers.

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



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.

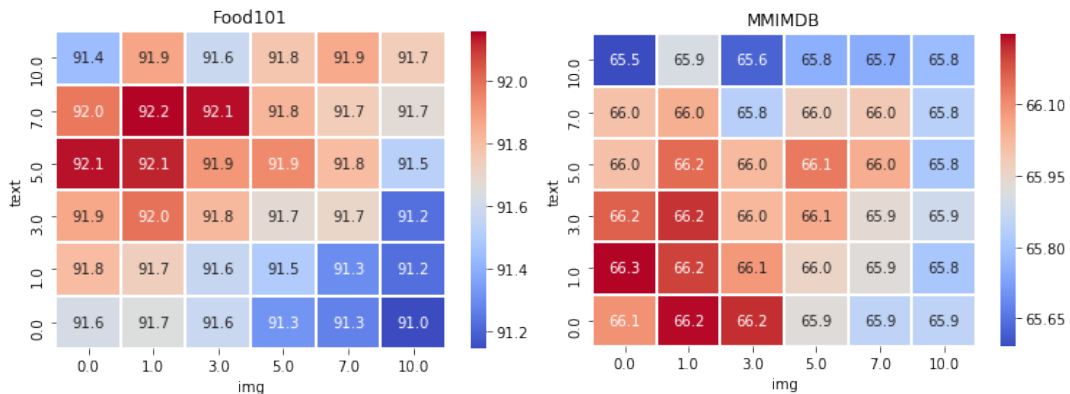


Figure 2: Analysis of freezing pre-trained text and image components for N epochs of training.

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.

Number of Parameters

While multimodal bitransformers definitely are able to perform more fine-grained earlier fusion, a possible alternative explanation for the fact that the architecture performs better than 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 correct for 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.

Related Work

Neural methods are the standard for almost every modern text and vision classification task. Transformers (Vaswani et al. 2017) 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 (Radford et al. 2018; Devlin et al. 2019).

The question of how to effectively combine multimodal information, also known as **multimodal fusion**, has a long history (Baltrušaitis, Ahuja, and Morency 2019). While concatenation can be considered the default, other fusion methods have been explored e.g. for lexical representation learning (Bruni, Tran, and Baroni 2014; Lazaridou, Pham, and Baroni 2015). In classification, (Kiela et al. 2018) 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** (Weston, Bengio, and Usunier 2011; Frome et al. 2013; Socher et al. 2013) to image captioning (Bernardi et al. 2016) to visual question answering (Antol et al. 2015) and multimodal machine translation (Elliott et al. 2017). Multimodal information is also useful in learning human-like meaning representations (Baroni 2016; Kiela 2017). Since text rarely occurs in isolation in the real world, it makes sense to use all available information in classification settings.

Concurrently with the work presented in this paper, various self-supervised multimodal architectures have been published, e.g. ViLBERT (Lu et al. 2019), VisualBERT (Li et al. 2019), **LXMERT** (Tan and Bansal 2019), VL-BERT (Su et al. 2019), **VideoBERT** (Sun et al. 2019). 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. 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.

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 pre-trained individually as unimodal tasks, which already showed great improvements over alternatives. It is as of yet unclear if self-supervised multimodal models are going to be generally useful. The methods outlined here should serve as a useful and powerful baseline to gauge their performance.

Supervised multimodal bitransformers are straightforward and intuitive, and importantly, are easy to implement even for existing self-supervised encoders.

References

- [Antol et al. 2015] Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. **Vqa**: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.
- [Arevalo et al. 2017] Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; and González, F. A. 2017. **Gated** multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.
- [Baltrušaitis, Ahuja, and Morency 2019] Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2019. **Multimodal** machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(2):423–443.
- [Baroni 2016] Baroni, M. 2016. Grounding distributional semantics in the visual world. *Language and Linguistics Compass* 10(1):3–13.
- [Bernardi et al. 2016] Bernardi, R.; Cakici, R.; Elliott, D.; Erdem, A.; Erdem, E.; Ikizler-Cinbis, N.; Keller, F.; Muscat, A.; and Plank, B. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research* 55:409–442.
- [Bowman et al. 2015] Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- [Bruni, Tran, and Baroni 2014] Bruni, E.; Tran, N.-K.; and Baroni, M. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49:1–47.
- [Collobert and Weston 2008] Collobert, R., and Weston, J. 2008. A unified architecture for natural language processing: Deep neural networks with **multitask learning**. In *Proceedings of the 25th international conference on Machine learning*, 160–167. ACM.
- [Conneau et al. 2017] Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.
- [Dai and Le 2015] Dai, A. M., and Le, Q. V. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*, 3079–3087.
- [Deng et al. 2009] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- [Devlin et al. 2019] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL*.
- [Elliott et al. 2017] Elliott, D.; Frank, S.; Barrault, L.; Bougares, F.; and Specia, L. 2017. 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*, 215–233.
- [Faghri et al. 2017] Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. **Vse++**: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.
- [Frome et al. 2013] Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2121–2129.
- [Gururangan et al. 2018] Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. **Annotation artifacts** in natural language inference data. *arXiv preprint arXiv:1803.02324*.
- [He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- [Howard and Ruder 2018] Howard, J., and Ruder, S. 2018. Universal language model fine-tuning for text classification. *Proceedings of ACL*.
- [Joulin et al. 2016] Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.
- [Kiela et al. 2018] Kiela, D.; Grave, E.; Joulin, A.; and Mikolov, T. 2018. Efficient large-scale multi-modal classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [Kiela 2017] Kiela, D. 2017. *Deep Embodiment: Grounding Semantics in Perceptual Modalities*. Ph.D. Dissertation, University of Cambridge, Computer Laboratory.
- [Kingma and Ba 2014] Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [Kiros et al. 2015] Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, 3294–3302.
- [Lazaridou, Pham, and Baroni 2015] Lazaridou, A.; Pham, N. T.; and Baroni, M. 2015. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.
- [Li et al. 2019] Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. **Visualbert**: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
- [Lu et al. 2019] Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. **ViLBERT**: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv preprint arXiv:1908.02265*.
- [Mikolov et al. 2013] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.
- [Oquab et al. 2014] Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1717–1724.
- [Paszke et al. 2017] Paszke, A.; Gross, S.; Chintala, S.; and Chanan, G. 2017. PyTorch: Tensors and dynamic neural networks in python with strong GPU acceleration. Technical report, PyTorch.
- [Pennington, Socher, and Manning 2014] Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.
- [Peters et al. 2018] Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *Proceedings of NAACL*.
- [Radford et al. 2018] Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.
- [Razavian et al. 2014] Razavian, A. S.; Azizpour, H.; Sullivan, J.; and Carlsson, S. 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 806–813.
- [Socher et al. 2013] Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, 935–943.
- [Su et al. 2019] Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. **Vi-bert**: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.
- [Sun et al. 2019] Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; and Schmid, C. 2019. **Videobert**: A joint model for video and language representation learning. *arXiv preprint arXiv:1904.01766*.
- [Tan and Bansal 2019] Tan, H., and Bansal, M. 2019. **Lxmert**: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.
- [Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.
- [Vu et al. 2018] Vu, H. T.; Greco, C.; Erofeeva, A.; Jafarizadehjan, S.; Linders, G.; Tanti, M.; Testoni, A.; Bernardi, R.; and Gatt, A. 2018. Grounded textual entailment. In *Proceedings of COLING*, 23542368.
- [Wang et al. 2015] Wang, X.; Kumar, D.; Thome, N.; Cord, M.; and Precioso, F. 2015. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1–6. IEEE.
- [Wang, Tran, and Feiszli 2019] Wang, W.; Tran, D.; and

Feiszli, M. 2019. What makes training multi-modal networks hard? *arXiv preprint arXiv:1905.12681*.

[Weston, Bengio, and Usunier 2011] Weston, J.; Bengio, S.; and Usunier, N. 2011. **Wsabie**: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

[Young et al. 2014] Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 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.