Few-shot Multimodal Multitask Multilingual Learning

Aman Chadha
Department of Computer Science
Stanford University
amanc@stanford.edu

Vinija Jain
Department of Computer Science
Stanford University
vinija@stanford.edu

1 Abstract

While few-shot learning as a transfer learning paradigm has gained significant traction for scenarios with limited data, it has primarily been explored in the context of building unimodal and unilingual models. Furthermore, a significant part of the existing literature in the domain of few-shot multitask learning perform in-context learning which requires manually generated prompts as the input, yielding varying outcomes depending on the level of manual prompt-engineering. In addition, in-context learning suffers from substantial computational, memory, and storage costs which eventually leads to high inference latency because it involves running all of the prompt's examples through the model every time a prediction is made. In contrast, methods based on the transfer learning via the fine-tuning paradigm avoid the aforementioned issues at a one-time cost of fine-tuning weights on a per-task basis. However, such methods lack exposure to few-shot multimodal multitask learning. In this paper, we propose **f**ew-shot learning for a **m**ultimodal **m**ultitask **m**ultilingual (FM3) setting by adapting pre-trained vision and language models using task-specific hypernetworks and contrastively fine-tuning them to enable few-shot learning. FM3's architecture combines the best of both worlds of in-context and fine-tuning based learning and consists of three major components: (i) multimodal contrastive fine-tuning to enable few-shot learning, (ii) hypernetwork task adaptation to perform multitask learning, and (iii) task-specific output heads to cater to a plethora of diverse tasks. FM3 learns the most prominent tasks in the vision and language domains along with their intersections, namely visual entailment (VE) [1], visual question answering (VQA) [2], and natural language understanding (NLU) tasks such as neural entity recognition (NER) and the GLUE benchmark [3] including QNLI [4], MNLI [5], QQP [6], and SST-2 [7].

2 Introduction

Self-supervised pretraining has propelled the adoption of deep learning on tasks with limited labeled data. With their task-agnostic features and improved data efficiency, self-supervised pre-trained models have drastically reduced the opportunity cost to tackle tasks that earlier required a significant amount of data and thus proved intractable using supervised learning. As a result of the advancements in self-supervised pretraining, semi-supervised approaches that combine self-supervision with supervised learning on a task-specific dataset that tackles a related task, have emerged as a new paradigm that has enabled transfer learning.

One of the biggest open challenges for machine learning research is building models that can be rapidly adapted to novel tasks using only a handful of annotated examples. The domain of few-shot learning (FSL), which is a specific variant of transfer learning, has emerged as an attractive solution to label-scarce scenarios where data annotation can be time-consuming and costly. These methods are designed to work with a small number of labeled training examples, and typically involve adapting pre-trained models for specific downstream tasks. Several flavors of FSL methods exist, each with its pros and cons.

One such large-scale self-supervised approach, popularized by the arrival of the generative pre-trained transformer (GPT) series [8, 9] of NLP models, is transfer learning via in-context learning (ICL) which emerges from training at scale. ICL teaches a model to perform a downstream task by feeding in a prompt with a nominal set of supervised examples as input to the model along with a single unlabeled example for which a prediction is desired. In effect, few-shot prompting using a small collection of input-target

pairs offers a walk-through to the model on how to transform the input into the output. Notably, since ICL requires no parameter updates, i.e., no gradient-based training is required, a single model can effectively act as a swiss-army knife by being able to immediately perform a wide variety of tasks. ICL, therefore, solely relies on the capabilities that a model learned during pretraining. The ease of use and quick adaptability to target tasks are characteristic features that have caused widespread adoption of ICL [10, 11, 12, 13, 14].

While ICL offers a multitude of benefits, it also suffers from several major drawbacks. First, processing all the prompted input-target pairs every time the model makes a prediction incurs significant compute, memory, and latency costs. These costs stack up as the number of the inferences increases – in a situation where the goal is to perform inference over a batch of test examples rather than one-off predictions, ICL can prove to be impractical from a resource standpoint. Second, owing to a limited-length context window, the number of support examples k that the model can utilize are restricted to nominal numbers. This is because we must fit all k examples into the model's context window, which is limited to a specific number of tokens (1024 in case of GPT-2 and 2048 in case of GPT-3). Third, ICL typically produces inferior performance compared to fine-tuning [15, 8, 16]. Finally, while the model's performance is a function of semantic and structural aspects of the prompt which can cause a significant yet unpredictable impact on the model's performance [17], far beyond inter-run variation of fine-tuning. In particular, semantic changes such as the phrasing or choice of words in the prompt and syntactic changes such as the exact formatting of the prompt (including the wording [18] and ordering of examples [19]) can cause a significant, unintended, and difficult-to-estimate impact on the model's performance. Furthermore, recent work has also demonstrated that ICL can perform well even when provided with incorrect labels, raising concerns as to how much learning is taking place at all [13].

Another common semi-supervised learning paradigm is transfer learning via fine-tuning (FT) which follows a two-staged process: (i) utilize the parameters of a pre-trained large-scale self-supervised model learning for weight initialization, and (ii) perform gradient-based fine-tuning using data associated with the downstream task of interest. With the advent of representation-learning approaches such as BERT [20], the domain of NLP underwent a radical transformation from supervised to semi-supervised approaches for tasks such as sentiment analysis, neural entity recognition, question answering, summarization, conversational response generation, etc. Representation-learning approaches have now taken center-seat in NLP, with the learned contextualized representations from these pre-trained models serving as initial task-agnostic features that, in turn, offer a the starting point for learning task-specific features. While problems with limited labeled data have benefited significantly owing to the reduced data-appetite of semi-supervised approaches, tasks with abundant labeled data have also seen improved performance.

While FT has produced many state-of-the-art (SoTA) results [21] on a range of classification tasks, it results in a model that is specialized for a single task with an entirely new set of parameter values, which can become impractical when FT a model on many downstream tasks. In other words, such models typically perform one task at a time, and cannot learn new concepts or adapt to new tasks in a few shots. FM3 seeks to address this gap and enable multimodal FSL – much like how SETFIT contrastively fine-tunes pre-trained Sentence Transformer models [22] and dispenses with prompts altogether and does not require large-scale pre-trained LMs to achieve high accuracy. With only 8 labeled examples in the Customer Reviews (CR) sentiment dataset, SETFIT is competitive with RoBERTa finetuned on the full training set [23], despite the fine-tuned model being three times larger.

Both ICL and fine-tuning have been explored in a multimodal context. A slew of methods, notably *Flamingo* [15] and *Frozen* [24], perform ICL with the final objective to have the model rapidly adapt to a variety of multimodal tasks. While *Flamingo* achieves competitive performance with FSL, in some cases outperforming models fine-tuned on thousands of times more task-specific data, *Frozen* offers relatively lower performance in return for the flexibility of using an off-the-shelf pre-trained LM and keeping its weights frozen. On the other hand, Oscar [25] and Omninet [26] are multimodal multitask models that do not perform ICL. While Oscar is pre-trained using pre-trained with aligned data on task-agnostic cross-modal objectives (a masked token loss over words and visual tags, and a contrastive loss between visual tags and others) and then fine-tuned to specific tasks, Omninet is simultaneously trained on its target tasks and undergoes no finetuning. In the zero-/few-shot learning context, multimodal pretraining has recently shown to enable strong generalization in the discriminative setting using large-scale contrastive learning [27, 28].

An additional paradigm for enabling a model to perform a new task with minimal updates is parameter efficient fine-tuning (PEFT), where a pre-trained model is fine-tuned by only updating a small number of added or selected parameters. Recent methods have matched the performance of fine-tuning the full

model while only updating or adding a small fraction (e.g. 0.01%) of the full model's parameters [29, 30]. Furthermore, certain PEFT methods allow mixed-task batches where different examples in a batch are processed differently [30], making both PEFT and ICL viable for multitask models. While the benefits of PEFT address some shortcomings of fine-tuning (when compared to ICL), there has been relatively little focus on whether PEFT methods work well when very little labeled data is available. [16] closes this gap by proposing T-Few, a model that learns using PEFT and a fixed set of hyperparameters, attaining strong performance on novel, unseen tasks while only updating a tiny fraction of the model's parameters.

FM3 combines the best of both worlds of ICL- and FT-based transfer learning and offers an efficient and prompt-free framework that offers strong generalization to new multimodal vision-language tasks in a few-shot setting. Despite the flexibility offered by ICL, its limitations leave much to be desired, especially in situations where compute, latency, memory, batch inference, performance determinism, etc. are important. On the other hand, FT offers performance invariance since it does not require prompts, offers better performance than ICL-based methods [15], and is resource-efficient in terms of compute, latency, memory, etc. While zero-/few-shot generalization is a desirable by-product of ICL, the only significant downside to FT is that generalization to new tasks with limited data is challenging. FM3 is architected keeping the aforementioned drawbacks of ICL in mind and thus follows the FT approach but overcomes its limitations as follows: (i) multimodal contrastive fine-tuning to enable FSL, (ii) using hypernetworks with a limited parameter count to perform task adaptation for multitask learning, and (iii) task-specific output heads to cater to a plethora of diverse tasks.

FM3 achieves high accuracy with little labeled data - for instance, with only 16 labeled examples per class on the complex task of SNLI-VE [1], FM3 surpasses the current SoTA fine-tuned on the full training set of 430K examples! Compared to other FSL methods, FM3 has several unique features:

- **No prompts or verbalisers:** Current techniques for few-shot fine-tuning require handcrafted prompts or verbalisers to convert examples into a format that's suitable for the underlying language model. FM3 dispenses with prompts altogether by generating rich embeddings directly from text examples. This obliterates the need for manual prompt engineering, which in turn, results in performance determinism.
- **Resource efficiency:** Optimal use of compute, latency, memory, etc. compared to our baselines *Flamingo*, *Frozen*, and especially ICL-based methods.
- **Frozen pre-trained models:** FM3 uses pre-trained vision and language encoders without fine-tuning them. This implies that FM3 architecture enables drop-in plug-and-play replacement for modality encoders. Only small hypernetwork models need to be fine-tuned when experimenting with different encoders.
- **Fast to train:** FM3 doesn't require large models like *Flamingo* (80B) or *Frozen* (7B+) to achieve high accuracy. As a result, it is significantly faster to train and run inference with.
- **Multilingual support:** FM3 enables multilingual processing, and can be paired up with any multilingual text encoder such as multilingual Sentence Transformer [22] variants of MPNet [31], RoBERTa [32], ALBERT [33], LASER [34], etc., which enables multilingual learning in 50+ languages by simply fine-tuning a multilingual model checkpoint.

While proposals that address a subset of the areas of few-shot multimodal multitask multilingual learning exist, to our knowledge, FM3 is the first to explore the intersection of the domains of multimodal multitask multilingual learning in a FSL setting.

3 Related Work

3.1 Few-shot learning using pre-trained models

In the domain of NLP, SETFIT proposed by Tunstall et al. [23] is an efficient and prompt-free framework for few-shot fine-tuning of Sentence Transformers (ST). SETFIT works by fine-tuning a pre-trained ST on a small number of text pairs in a contrastive Siamese manner. The resulting model is then used to generate rich text embeddings, which are used to train a classification head. This simple framework requires no prompts, and achieves high accuracy with orders of magnitude less parameters than existing techniques. SETFIT obtains comparable results with parameter-efficient fine-tuning (PEFT) and parameter efficient tuning (PET) techniques, while being an order of magnitude faster to train. SETFIT achieves high accuracy with little labeled data - for instance, with only 8 labeled examples per class on the Customer Reviews

sentiment dataset [35], SETFIT is competitive with fine-tuning RoBERTa Large on the full training set of 3K examples. Owing to its practical utility in enabling FSL, we adopt the idea of contrastive fine-tuning from SETFIT and generalize it to a multimodal multitask multilingual setting as part of FM3.

3.2 Multitask fine-tuning using PEFT

In [36], Houlsby et al. propose a parameter-efficient fine-tuning method which introduces adapter modules between the layers of a pre-trained language model. Adapter modules yield a compact and extensible model; they add only a few trainable parameters per task, and new tasks can be added without revisiting previous ones. The parameters of the original network remain fixed, yielding a high degree of parameter sharing. They achieve SoTA performance on GLUE [3] whilst adding only a few parameters per task. However, the downside of this approach is that they are trained separately for each task and thus do not enable sharing information across tasks.

To circumvent the aforementioned issue of knowledge sharing across tasks, Mahabadi et al. [37] learn adapter parameters for all layers and tasks using shared hypernetworks, which condition on task, adapter position, and layer ID in a transformer model. This parameter-efficient multitask learning framework achieves the best of both worlds by sharing knowledge across tasks via hypernetworks while enabling the model to adapt to each individual task through task-specific adapters. Experiments on the GLUE benchmark show improved performance in multitask learning while adding only 0.29% parameters per task. Given the fact that hypernetworks enable easy multitask fine-tuning of pre-trained models without having to actually fine-tune the model's weights (i.e., they remain frozen in this process), we adopt multitask finetuning in our proposed architecture.

3.3 Multitask multimodal learning

Hu and Singh propose UniT [38], a Unified Transformer encoder-decoder model that learns 7 tasks jointly across 8 datasets spread over different vision and lanuage domains, ranging from object detection to natural language understanding and multimodal reasoning. UniT achieves strong performance with significantly few parameters in some cases outperforming separately trained single task models. While the architecture offers joint end-to-end training of each task, it requires a substantial amount of data across all tasks for the model to generalize. Our approach, on the other hand, utilizes FSL to efficiently learn a task with a small fraction of data.

In [26], Pramanik et al. propose OmniNet, a single model to support tasks with multiple input modalities as well as asynchronous multitask learning. OmniNet is powered by a spatio-temporal cache that enables learning spatial dimension of the input in addition to the hidden states corresponding to the temporal input sequence. Even though OmniNet is $3\times$ parameter-efficient, there is a significant performance gap on most tasks it was trained on compared to the individual model counterparts.

3.4 Multitask multilingual multimodal learning

M^3P, proposed in [39], is a multitask multilingual multimodal pre-trained model that combines multilingual pre-training and multimodal pre-training into a unified framework via multitask pre-training. M^3P learns universal representations that can map objects occurred in different modalities or texts expressed in different languages into a common semantic space. In addition, to alleviate the issue of lack of sufficient labeled data for non-English multimodal tasks, they propose multimodal code-switched training (MCT) [40] which replaces each word in the caption with a translated word with a probability of β. If a word has multiple translations, a random one is chosen. Experiments on the multilingual image retrieval task across MS COCO [41] and Multi30K [42] show competitive results for English and new establish SoTA results for non-English languages. While M^3P tackles a similar problem as FM3 , it does not assume any restrictions on the annotation budget – in other words, it does not consider the few-shot setting for learning its tasks.

3.5 Few-shot multimodal multitask learning

In [15], Alayrac et al. introduce Flamingo, a family of Visual Language Models (VLM) trained on large-scale multimodal web corpora with an ability to rapidly adapt to a variety of image and video tasks. *Flamingo* proposes the following key architectural innovations: (i) bridge powerful pre-trained vision-only and language-only models, (ii) handle sequences of arbitrarily interleaved visual and textual data, and

(iii) seamlessly ingest images or videos as inputs. The end result is a single *Flamingo* model that can achieve a new SoTA with FSL, simply by prompting the model with task-specific examples. On numerous benchmarks, *Flamingo* outperforms models fine-tuned on thousands of times more task-specific data. These include open-ended tasks such as visual question-answering, where the model is prompted with a question which it has to answer; captioning tasks, which evaluate the ability to describe a scene or an event; and close-ended tasks such as multiple-choice visual question-answering.

In [24], Tsimpoukelli et al. present *Frozen*, a simple-yet-effective approach for transferring the FSL abilities inherent in large auto-regressive language models to a multimodal setting (vision and language). *Frozen* is a multimodal few-shot learner, with the surprising ability to learn a variety of new tasks when conditioned on examples, represented as a sequence of multiple interleaved image and text embeddings. *Frozen* can rapidly learn words for new objects and novel visual categories and do visual question-answering with only a handful of examples. While this work serves as an important baseline for FM3, a key limitation is that it achieves far from SoTA performance on the specific tasks that it learns in a few shot setting. *Frozen* shows that training a visual encoder through a pre-trained and frozen language model results in a system capable of strong out-of-distribution (zero-shot) generalization. Furthermore, *Frozen* confirms that the ability to rapidly adapt to new tasks given appropriate prompts is inherited from the pre-trained language model and transfers directly to multimodal tasks.

While *Flamingo* and *Frozen* are both ICL-based FSL methods, the differentiating factors are: (i) the scale of data used to train these models, and (ii) architectural variations. *Flamingo* is trained on large-scale multimodal web corpora while is *Frozen* is trained on the Conceptual Captions dataset [43]. The architectural design choices differ between the two in using pre-trained modality encoders vs. training them from scratch. Similar to FM3, *Flamingo* uses off-the-shelf pre-trained encoders and only generates adapter components (in the form of Perceiver Resampler blocks) while *Frozen* utilizes a pre-trained LM but trains its own vision encoder that feeds the LM. Inspired by this observation, FM3 borrows the idea of using separate text and vision adapters in the form of hypernetworks so as to offer the model additional degrees of freedom, which in turn, helps render better performance.

4 FM3

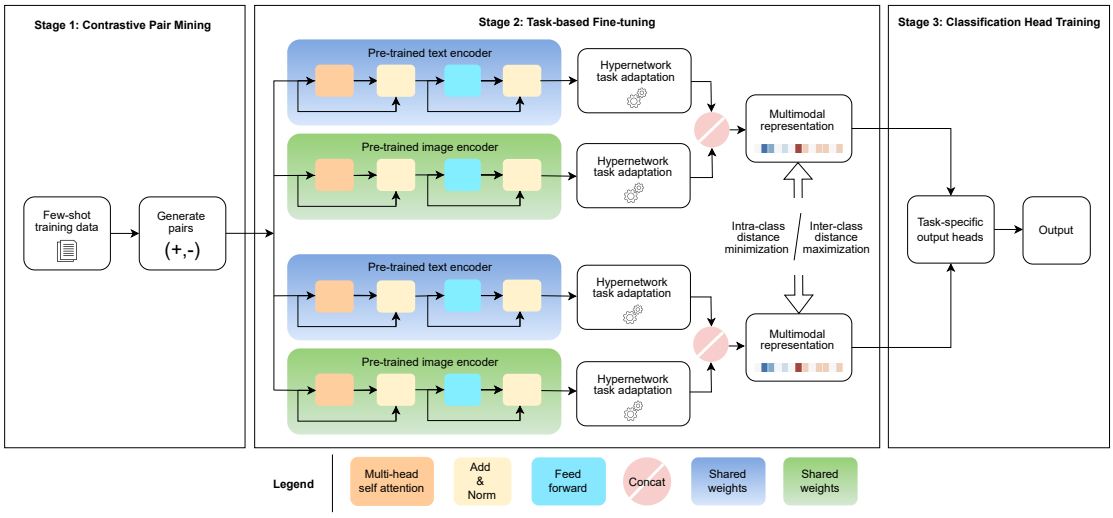

Figure 1: Architectural overview of FM3. FM3 consists of three stages: (i) contrastive pair mining for fine-tuning, which generates positive and negative pairs, (ii) task-based fine-tuning involves adapting the pre-trained text and image encoder models for down-stream tasks using hypernetworks, and (iii) training task-specific classification heads.

4.1 Methods

Figure 1 offers a visual summary of the architectural stages of FM3.

4.1.1 Task and batch sampling

At each iteration during training, we randomly select a task with a sampling probability that can be manually specified based on the dataset size. Once the task list has been sampled, for tasks with multiple datasets, we randomly sample a dataset corresponding to that task to fill a batch of samples.

4.1.2 Contrastive fine-tuning for few-shot learning

Similar to [23], we utilize a contrastive learning approach to FSL. Contrastive learning effectively enlarges the size of training data which is critical in few-shot scenarios and thus fosters effective learning for tasks with limited annotated data. Assuming a small number (k) of labeled examples for a classification task, the potential size of the fine-tuning set T derived from the number of unique positive and negative contrastive pairs that can be generated would be $\frac{k(k-1)}{2}$, which is significantly larger than just k [23]. In this stage, we sample R positive and R negative triplet pairs, where R is a hyperparameter (set to 20, following [23]). We utilize multiple negatives ranking loss [44] for contrastive fine-tuning owing to its superior performance [44] and its ability to randomly sample negative pairs from each batch in an automated fashion.

4.1.3 Task-based fine-tuning using hypernetworks

To our knowledge, FM3 is the first to utilize hypernetworks in a multimodal setting. Using frozen modality encoders has the distinct advantage of preventing catastrophic forgetting (compared to fine-tuning the encoders themselves) [15]. As such, we utilize an independent hypernetwork for each modality.

In this step, we perform task-specific fine-tuning of SoTA pre-trained text and vision models, namely a pre-trained multilingual MPNet [31] Sentence Transformer [22] from Huggingface [45] as our text backbone and CoCa-Base [46] as our vision backbone. We adopt the idea of hypernetworks from [37], which is a parameter-efficient method for multitask fine-tuning. We train shared hypernetworks to generate task-specific adapters conditioned on the task, layer ID, and adapter position embeddings. These shared hypernetworks capture knowledge across tasks, enabling positive transfer to low-resource and related tasks, while task-specific layers allow the model to adapt to each individual task. We optimize a distance function based on cosine similarity, minimizing it for positive pairs and maximizing it for negative pairs.

4.1.4 Task-specific classification head training

Lastly, we train task-specific classification heads on the fine-tuned model obtained from the above step. The generated embeddings corresponding to the data samples for each task, along with their labels, constitute the training set for the respective classification head. We use logistic regression for binary classification tasks such as SST-2, QQP, QNLI, etc. and softmax for multiclass classification tasks such as VQA, SNLI-VE, GLUE, NER, etc.

4.2 Tasks and Datasets

Table 1 delineates the domains, tasks, and datasets for training and evaluating FM3.

Domain	Task	Dataset
Language understanding	Neural entity recognition (NER)	CoNLL-2003 [47]
	GLUE benchmark [3]	QNLI [4], MNLI [5], QQP [6], and SST-2 [7]
Vision-and-language reasoning	Visual entailment	SNLI-VE [1]
	Visual question answering (VQA)	VQAv2 dataset [2] (with questions from Visual Genome [48] as additional data), OK-VQA [49]

Table 1: Datasets for training and evaluation

5 Experiments

5.1 Experimental setup

We finetuned FM3 on the Conceptual Captions dataset (which *Frozen* [24] is trained on) for vis-à-vis comparisons. We use the AdamW optimizer with global norm clipping of 1, no weight decay for the hypernetworks and weight decay of 0.1 for the other trainable parameters. We anneal the learning rate, increasing it linearly from 0 to 10^{-3} up over the first 5000 steps then held constant for the duration of training and then decayed exponentially. Unless specified otherwise we train our models for 500K steps.

All datasets were trained with the same weights. Since the performance of models trained with a contrastive objective is sensitive to the batch size, we use a relatively large batch size of 32.

5.2 Baselines

We utilize *Flamingo* [15] and *Frozen* [24] as our primary baselines since they deal with multimodal multitask learning in the context of FSL. We include UniT [38] as an additional baseline since it deals with multimodal multitask learning of tasks that *Flamingo* [15] and *Frozen* haven't been evaluated on. For each task, we compare FM3 with both task-specific zero-/few-shot and pre-trained/fine-tuned SoTA. Since none of the above baselines support multilingual tasks, we utilize [39] as a baseline to qualify the performance of FM3 on non-English tasks.

5.3 Results

Method	FT	Shot(s)	OKVQA [49]	VQAv2 [2]	Flickr30K [50]	Multi30K [42]	SNLI-VE [1]	CoNLL-2003 [47]	QNLI [4]	MNLI [5]	QQP [6]	SST-2 [7]
Zero/Few shot SoTA	✗	Various	[51] 43.3 (16)	[24] 38.2 (4)	[52] 94.9 (0)	-	[53] 87.3 (0)	[54] 65.4 (5)	-	[53] 86.4 (5)	[55] 67.8 (16)	[55] 93.5 (16)
Flamingo-80B	✗	0	50.6	56.3	67.2	-	-	-	-	-	-	-
		4	57.4	63.1	75.1							
		32	57.8	67.6	75.4							
Frozen	✗	0	5.9	29.5	-	-	-	-	-	-	-	-
		1	9.7	35.7								
		4	12.6	38.2								
UniT	✗	Fully Supervised	-	67.0	-	-	73.2	-	88.0	81.8	90.6	91.5
FM3-5B	✗	0	51.5	57.3	93.2	27.2	86.2	51.3	53.6	58.2	41.4	75.5
		4	55.5	64.6	94.3	35.2	89.9	66.2	67.6	55.5	54.4	86.8
		16	57.8	67.9	96.1	37.1	92.4	77.3	75.4	81.1	66.7	95.7
		64	**58.9**	**71.2**	**99.1**	39.9	**94.4**	**82.7**	83.7	86.4	**78.9**	**99.2**
Pre-trained FT SoTA	✔	Various	54.4 [51] (10K)	80.2 [56] (444K)	98.8 [57] (31K)	49.3 [58] (31K)	91.0 [59] (430K)	94.6 [60] (430K)	99.2 [33] (105K)	92.0 [61] (393K)	89.2 [62] (364K)	97.5 [61] (67K)

Table 2: **Comparison to the state of the art.** A *single* FM3 model reaches the state of the art on a wide array of vision-language understanding tasks with FSL, significantly outperforming previous best zero- and few-shot methods with as few as 16 examples. More importantly, using only 64 examples and without adapting any model weights, FM3 *outperforms* the current best methods – fine-tuned on thousands of annotated examples – on 4 tasks. Best few-shot numbers across all shots are in **bold**, best numbers across both zero/few-shot (prompt based) and fine-tuned models are underlined. For each baseline, we chose the best numbers across spanning all variants/experiments (unless explicitly stated).

Table 2 performs a comparative analysis of FM3 with *Flamingo*, *Frozen*, and the respective zero-/few-shot and fine-tuned SoTA on each task with number of support examples/shots as $k \in \{0, 4, 16, 64\}$. While Flickr30K Image-to-Text uses Recall@1, Mulit30K `en-de` uses BLEU, CoNLL-2003 and QQP use F1, all other tasks utilize accuracy as their performance metric. Note that since *Frozen* is an auto-regressive model/decoder which undergoes prompt-based fine-tuning, k indicates the number of support examples as part of the prompt/prefix passed as input to the model, while FM3 being an encoder-based architecture, k indicates the number of examples we contrastively fine-tune on.

Few-shot results. FM3 outperforms zero-/few-shot baselines on 7 out of the 10 benchmarks considered. This is achieved with as few as 16 examples per task, demonstrating superior adaptation to these tasks. More importantly, FM3 is often competitive with SoTA methods fine-tuned on up to hundreds of thousands of annotated examples. On 4 out of 10 tasks, FM3 even outperforms the fine-tuned SoTA despite using a single set of model weights and only 64 task-specific examples.

Scaling with respect to parameters and shots. As table 2 indicates, more the number of shots, better the few-shot performance, similar to GPT-3 [8]. The performance improvement shows diminishing returns as the number of shots increases.

5.4 Inference runtime analysis

Table 3 summarizes our inference runtime analysis. We measure the time taken to run FM3 and our primary baselines on the test set of VQAv2 [2] and OKVQA [49] and averaging it out by the total number of total samples. These measurements are from a platform with NVIDIA A100 with 32GB VRAM. **Bold** numbers indicate best performance. Underlined numbers indicate the next best baseline on which the % improvements for FM3 are based.

	VQAv2 (sec.)	**OKVQA** (sec.)
FM3	**0.187** (58% ↑)	**0.214** (46% ↑)
Flamingo	0.353	0.371
Frozen	0.295	0.312

Table 3: Inference runtime analysis of FM3 vs. *Flamingo* and *Frozen* on VQAv2 and OKVQA. Measurements are based on wall clock time (sec.) so lower is better.

6 Ablation analysis

Table 4 summarizes the results for FM's ablation experiments. **Bold** numbers indicate best performance. Underlined numbers indicate the baseline on which the % numbers are based. We analyze the impact of the following design decisions on FM3's performance:

- **Direct encoder fine-tuning with no hypernetworks.** While frozen modality encoders prevents catastophic forgeting [15], we adopt a hypernetwork-fee architecture and fine-tune the encoders themselves with data from our target tasks.
- **Hypernetwork size selection.** We perform comparisons with varied parameter size allocations for hypernetworks to quantify the effect of hypernetwork size. The parameter count for hypernetworks is expressed as a percentage of the baseline FM3 model parameter count.
- **Compute/memory vs. performance trade-offs.** We vary the choice of text and vision encoders (which in turn, varies the number of parameters and time complexity of the model). For our text encoder, we choose mulitingual MiniLM [63] with 57% lesser parameters compared to our default choice of the `paraphrase-multilingual-mpnet-base-v2` variant of multilingual MPNet [45]. For our vision encoder, we choose the `vit-base-patch16-224` variant of ViT [64] with 78% fewer parameters compared to our default choice of CoCa-Base [46].

	VQAv2 (acc.)	**OKVQA** (acc.)
FM3 (default)	**71.2**	**58.9**
No hypernetworks	64.3 (90.0% ↓)	52.1 (88.4% ↓)
Hypernetworks with 5% parameters	69.5 (97.6% ↓)	56.7 (88.4% ↓)
Text encoder: MiniLM	68.1 (95.6% ↓)	55.4 (94.0% ↓)
Vision encoder: ViT	66.2 (92.9% ↓)	51.2 (86.9% ↓)
Text/vision encoders: MiniLM/ViT	65.5 (91.9% ↓)	52.3 (88.7% ↓)

Table 4: Ablation analysis of FM3 on VQAv2 and OKVQA with number of shots as 64. Default FM3 uses hypernetworks with 10% parameters, and MPNet/CoCa as text/vision encoders. Measurement units are % accuracy so higher is better.

7 Future Work

While FM3 establishes a new SoTA on several tasks, there are significant opportunities for improvement centered around three major aspects: (i) data, (ii) model architecture, and (iii) loss function. First, *Flamingo*

[15] highlights the importance of a diverse dataset amalgamated from various disparate sources (*Flamingo* uses >2B image-text pairs vs. 3.3M that FM3 was trained on) in training the neural network. Using the publicly available massive LAION-400M dataset [65] would be a great starting point. Second, the model architecture can incorporate other techniques that offer reasonably high performance with a reduced parameter count such as low-rank based adaption methods, for e.g., LoRA [29]. Third, following [66, 25], we can formulate the ranking loss [44] as a binary classification problem. This has reported to lead to an increase in performance [66, 25]. In other words, given an aligned image-text pair, we randomly select a different image or a different caption to form an unaligned pair. Similar to FM3's current framework, the final concatenated multimodal embedding can still be used as the input for classification to predict whether the given pair is aligned or not. Finally, FM3 is easily extendable to other languages, tasks, and modalities.

8 Conclusion

FM3 combines the best of both worlds of in-context learning and fine-tuning as a front-runner in the niche domain of few-short multilingual multimodal multitask learning. It offers a scalable architecture that can span modalities, tasks, and languages, all while setting a new standard with SoTA performance on a plethora of tasks and competitive performance on others. FM3 outperforms zero-/few-shot baselines on 7 out of 10 benchmarks with as few as 16 examples per task. Moreover, FM3 is competitive with fine-tuning a plethora of task-specific SoTA models on fine-tuned on up to hundreds of thousands of annotated examples. On 4 out of 10 tasks, FM3 even outperforms the fine-tuned SoTA despite using a single set of model weights and only 64 task-specific examples. Lastly, FM3 also yields a $\sim$50% latency improvement compared to the next best FSL SoTA baseline on VQA and OKVQA datasets.

References

[1] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[2] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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, 2017.

[3] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[4] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[5] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[6] Adina Williams, Nikita Nangia, and Samuel R Bowman. First quora dataset release: Question pairs. *Quora blog*, 2017.

[7] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[10] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

[11] Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*, 2022.

[12] Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*, 2022.

[13] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

[14] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2022.

[15] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[16] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638*, 2022.

[17] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

[18] Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*, 2021.

[19] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[21] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

[22] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[23] Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*, 2022.

[24] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[25] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[26] Subhojeet Pramanik, Priyanka Agrawal, and Aman Hussain. Omninet: A unified architecture for multi-modal multi-task learning. *arXiv preprint arXiv:1907.07804*, 2019.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[28] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[29] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[30] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[31] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

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

[33] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[34] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*, 2020.

[35] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172, 2013.

[36] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[37] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489*, 2021.

[38] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021.

[39] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3977–3986, 2021.

[40] Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. *arXiv preprint arXiv:2006.06402*, 2020.

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[42] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016.

[43] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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, 2018.

[44] Matthew Henderson, Rami Al-Rfou, Brian Strope, Y Sung, L Lukács, R Guo, S Kumar, B Miklos, and R Kurzweil. Efficient natural language response suggestion for smart reply. arxiv. *Preprint posted online May*, 1, 2017.

[45] HuggingFace. Multilingual mpnet sentence transformer. *HuggingFace*, 2022. Available at `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`. Accessed Nov 13 2022.

[46] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[47] Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

[48] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[49] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[50] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[51] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*, 2021.

[52] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[53] Tu Vu, Minh-Thang Luong, Quoc V Le, Grady Simon, and Mohit Iyyer. Strata: Self-training with task augmentation for better few-shot learning. *arXiv preprint arXiv:2109.06270*, 2021.

[54] Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. Few-shot named entity recognition: An empirical baseline study. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10408–10423, 2021.

[55] Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. Differentiable prompt makes pre-trained language models better few-shot learners. *arXiv preprint arXiv:2108.13161*, 2021.

[56] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[57] Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. X2-vlm: All-in-one pre-trained model for vision-language tasks. *arXiv preprint arXiv:2211.12402*, 2022.

[58] Bin Shan, Yaqian Han, Weichong Yin, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-unix2: A unified cross-lingual cross-modal framework for understanding and generation. *arXiv preprint arXiv:2211.04861*, 2022.

[59] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022.

[60] Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. Automated concatenation of embeddings for structured prediction. *arXiv preprint arXiv:2010.05006*, 2020.

[61] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[62] Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*, 2021.

[63] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

[64] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[65] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[66] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.