S5 Framework: A Review of Self-Supervised Shared Semantic Space Optimization for Multimodal Zero-Shot Learning

Anonymous ACL submission

Abstract

In this review, we aim to inspire research into **S**elf-**S**upervised **S**hared **S**emantic **S**pace (**S5**) multimodal learning problems. We equip non-expert researchers with a framework of informed modeling decisions via an extensive literature review, an actionable modeling checklist, as well as a series of novel zero-shot evaluation tasks. The core idea for our S5 checklist lies in learning contextual multimodal interactions at various granularity levels via a shared Transformer encoder with a denoising loss term, which is also regularized by a contrastive loss term to induce a semantic alignment prior on the contextual embedding space. Essentially, we aim to model human concept understanding and thus learn to "put a name to a face". This ultimately enables interpretable zero-shot S5 generalization on a variety of novel downstream tasks. In summary, this review provides sufficient background and actionable strategies for training cutting-edge S5 multimodal networks.

1 Introduction

There has been a lot of recent advancement in Transformer (Vaswani et al., 2017) architectures for language/vision fusion tasks (Khan et al., 2021). Transformer architectures can implicitly model interactions between visual and textual entities via the attention mechanism (Li et al., 2019b,a; Chen et al., 2019; Li et al., 2020b).

Transformer-based architectures, however, have an inherent property of collapsed representation spaces. Some recent language/vision fusion models address that with self-supervised contrastive pretraining objectives (Radford et al., 2021; Jia et al., 2021). Their most notable limitation is that the dual-encoder architecture prevents them from modeling local context interactions at multiple levels of abstraction.

In another line of work, some models can learn semantic embeddings aligned across modalities at various contextual levels (Li et al., 2019a; Chen et al., 2019), but depend on object detection for visual embedding. That creates additional computation overhead, as well as out-of-domain generalization issues, given that object detectors use human-annotated data.

Some Transformer architectures (Kim et al., 2021; Wang et al., 2021; Ramesh et al., 2021) take the best of both worlds, by using self-supervised multimodal objectives for large-scale pretraining, but neglect the semantic space collapse resulting from Transformer architecture. That makes their semantic embeddings difficult to use on a broad range of tasks without extensive finetuning.

In this work, we review the necessary background literature for a non-expert research practitioner to gain understanding of recent multimodal practices and their shortcomings. We also propose a checklist of informed decisions for training S5 encoder that addresses *all* of the outlined issues. We additionally propose novel experiments to prove effectiveness of S5 models in zero-shot settings. Thus, we equip the reader with actionable self-supervised multimodal training strategies for cutting-edge zero-shot generalization.

More specifically, our framework aims to produce and evaluate a model fulfilling the following goals, in the order of importance:

1. encode image and text into a well-aligned semantic representation space, also optimized for uniformity;

2. capture both global and local context semantic interactions (i.e., interactions on both image/sentence and patch/token levels);

3. adapt well to both single and multiple modality downstream tasks in a zero-shot setting;

4. be easy to use in terms of both architecture versatility and inference computation requirement.

2 S5 Background

In this section, we survey the recent literature related to Transformer model embedding collapse as well as multimodal representation learning approaches. We aim to highlight the recent common practices and their limitations, while introducing the reader to the field.

2.1 Transformer Representation Collapse

Large-scale text Transformers are typically optimized with language modeling likelihood loss, either autoencoding, e.g. BERT (Devlin et al., 2018), or autoregressive, e.g. GPT2 (Radford et al., 2019). Then, for any given hidden state, the ground truth token embedding is pushed toward that hidden state, while all other token embeddings are pushed away from it. Thus, every training example in effect becomes a negative example for the tokens that are not ground truth. Given the Zipf distribution of words in natural languages, majority of the words will appear with a low term frequency. As a result, most of the representations are pushed in a similar direction in the embedding space and degrade into a narrow cone (Gao et al., 2019). The degeneration happens when the convex hull of the representations does not contain the origin. Gao et al. (2019) theoretically shows that this does tend to be the case when layer normalization is used, which happens a lot in Transformers. This way, words with similar frequency are likely to end up close to each other in the representation space, despite not necessarily having similar meaning.

Empirical studies of large-scale Transformers have corroborated this, finding that the singular values of their embedding spaces tend to drop off rapidly, with only first few components accounting for majority of the embedding space variance (Wang et al., 2020a). It has also been shown that BERT embeddings are not uniformly distributed, as they have average cosine similarity greater than 0 (Ethayarajh, 2019).

More than that, it has been empirically shown that word frequency biases BERT embedding space (Li et al., 2020a), as mean l_2-norm of BERT embeddings is correlated with word frequency. The high frequency words are arranged closer to both the origin and each other than the lower-frequency words. That creates "holes" in the outer edges of the embedding space, where semantic meaning is defined poorly.

Semantic embedding spaces can be described in terms of alignment and uniformity metrics (Wang and Isola, 2020). Alignment is the quality of similar examples mapped nearby to each other in the embedding space, whereas uniformity assesses how evenly the representations are distributed on a unit hypershpere. Language Transformer embeddings, then, suffer from excessive alignment: unrelated terms are close to each other in space just due to term frequency.

Contrastive learning objective optimizes asymptotically for the alignment and uniformity of the embedding space (Wang and Isola, 2020). It has been empirically shown effective for both textual (Gao et al., 2021; Yan et al., 2021; Su et al., 2021) and visual (Chen et al., 2020) representation learning. Contrastive tension loss optimizes for the similarity of representations of two augmented views of the same data (positive examples), while increasing their distance with the representations of other data (negative examples) in the training batch (Gao et al., 2021). Even without any augmentation on positive examples – without optimizing for additional alignment – contrastive loss can still improve BERT embedding space by making it more uniform (Yan et al., 2021). While previous approaches (Gao et al., 2021; Yan et al., 2021) use contrastive learning as a finetuning objective, TaCL (Su et al., 2021) successfully uses token-wise contrastive learning as an additional pretraining objective together with Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) denoising objectives (Devlin et al., 2018).

2.2 Advances in Multi-modal Learning

Here, we discuss the recent advances made in vision/language pretraining. This section will address object detection based approaches, contrastive embedding approaches, as well as current mutlimodal unified Transformers. We make a special effort to identify the existing limitations in those approaches, with respect to our goals.

2.2.1 Object Detection Based Approaches

A lot of success in Transformer pretraining for language/vision fusion relates to object detection visual embedding approaches. They typically use an object detector, like Faster R-CNN (Ren et al., 2015), trained on Visual Genome data (Krishna et al., 2016).

This line of work can be traced back to Unicoder-VL (Li et al., 2019a), which pretrains a large-scale denoising autoencoder using Masked Language Modeling, Masked Object Classification, and Visual-Linguistic Matching objectives, on the Conceptual Captions (Sharma et al., 2018) and SBU Captions (Ordonez et al., 2011) datasets.

The Unified Vision-Language Pre-training (VLP) (Zhou et al., 2019) improves on that by accommodating a decoder in the same Transformer stack by the means of a masking mechanism reminiscent to that of UniLM (Dong et al., 2019), thus supporting both discriminative and generative applications.

UNITER (Chen et al., 2019), then, adds MS-COCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2016) to the pretraining, as well as devising novel optimization objectives, including Masked Language and Region Modeling, conditioned on the opposite modality (i.e., if text is corrupted, the image is left intact and vice versa). UNITER also uses Visual-Linguistic Matching and a novel Word-Region Alignment objective with Optimal Transport, which explicitly encourages fine-grained alignment between words and image regions during pretraining.

Oscar (Li et al., 2020b) uses the object detector text output to model each image/text pair in terms of $[w, q, v]$ triplets, with w - textual modality word token embeddings, q - object detector's textual class tag embeddings, and v - object detector's vision region embeddings. VinVL (Zhang et al., 2021) improves upon that with a more robust object detector and a modified contrastive loss.

Despite the impressive improvements made in this area, the models are still constrained by using an explicit object detection model, trained on a human-labeled data. This limits the object detector's zero-shot generalizability and thus undermines zero-shot performance of the whole vision/language model. Another constraint on this is the inference costs of object detection, in terms of both difficulty of use and computation overhead. These architectures fall short on our goals (3): versatility and (4): ease of use.

2.2.2 Contrastive Embedding Approaches

A more recent line of work in text/image fusion does away with the object detector, instead opting to split text and image modalities into separate encoders. The representations of the two encoders are then optimized contrastively in a shared embedding space at the final layers.

CLIP (Radford et al., 2021) is a dual-encoder image/text architecture. It uses ResNet-50 (He et al., 2016) or ViT-L/14@336px (Dosovitskiy et al., 2020) as image encoder and GPT-2 (Radford et al., 2019) style text encoder. Both encoders then project the modality embeddings into a common representation space via a linear projection layer. The model is initialized randomly and then trained contrastively by optimizing the similarity of the corresponding image/text representations and penalizing the similarity of the mismatching ones within each batch. CLIP also uses a novel WebImageText (WIT) dataset, consisting of 400M (image, text) pairs collected from the Internet, which has a similar total word count to the GPT-2 WebText (Radford et al., 2019) dataset. CLIP showcases a strong zero-shot performance on the datasets with a small number of labeled examples. It does not, however, generalize well to the data not likely to be present in WIT, such as MNIST (LeCun and Cortes, 2010) data.

ALIGN (Jia et al., 2021) also uses the dual-encoder architecture with a contrastive representation alignment objective. ALIGN uses EfficientNet (Tan and Le, 2019) with global pooling as the image encoder and BERT (Devlin et al., 2018) with [CLS] token embedding as the text encoder. Both encoders are trained from scratch. The major contribution of that work is to scale up the image/text pretraining dataset. To that end, they propose an improvement over Conceptual Captions (Sharma et al., 2018) data by disabling most of the filtering and postprocessing of Conceptual Captions, until they are left with 1.8B noisy image/text pairs. ALIGN empirically shows that pretraining on the large-scale noisy cross-modal data can still yield strong performance on image/text matching and retrieval.

The dual-encoder choice of the architecture may be problematic. According to a Iki and Aizawa (2021), the dual-encoder architectures, aside from drastically increasing the number of parameters, may also be detrimental for language modeling performance. This may be explained by late-fusion multimodal networks' tendency to overfit due to difference in per-modality optimization rates per Wang et al. (2019). These models are suboptimal with respect to our goals (1): semantic space, and (2): multi-context capture.

2.2.3 Multimodal Transformer Approaches

Yet more recent works propose to use a shared Transformer encoder for both text and image modalities, as do we in S5 Checklist. We will borrow and recombine ideas from these approaches in formulating our checklist in later sections.

ViLT (Kim et al., 2021) is a Transformer encoder. It uses ViT (Dosovitskiy et al., 2020) weights for Transformer initialization. From there, it continues to train the model on concatenated visual and text embeddings. Both modality embedding sequences get their own learnable [class] embedding as a prefix. Following ViT, it uses a simple linear projection for 16x16 pixel image patches to produce visual embeddings. It optimizes with Image Text Matching using encoded text [class] embedding, MLM on text embeddings, as well as Word Patch Alignment with Optimal Transport on stacked encoded text embeddings to encoded visual embeddings, following the Word Region Alignment objective from UNITER (Chen et al., 2019). ViLT uses MS-COCO, Visual Genome, SBU Captions, and Conceptual Captions datasets for pretraining. It produces strong results on VQA (Goyal et al., 2017) and image retrieval (Karpathy and Fei-Fei, 2014) tasks. It is also qualitatively shown to learn semantic alignment between text tokens and image patches.

SimVLM (Wang et al., 2021) is a Transformer architecture for image-to-text tasks. It combines BERT (Devlin et al., 2018) and ViT (Dosovitskiy et al., 2020) approaches in a PrefixLM architecture to model both text/image and text-only data. It uses SentencePiece (Kudo and Richardson, 2018) subword tokenization for text data, and first three blocks of ResNet-152 for image patch embedding, similar to CoAtNet (Dai et al., 2021). The model is optimized with autoencoding loss on the prefix sequence of image and text, as well as autoregressive loss on the remaining text sequence. SimVLM uses ALIGN data for image/text pretraining and C4 (Raffel et al., 2019) dataset for text-only examples. SimVLM establishes a strong performance on visual question answering (Goyal et al., 2017), visual entailment (Xie et al., 2018), and visual reasoning (Suhr et al., 2018). SimVLM also performs better than other vision/language approaches on the GLUE (Wang et al., 2018) benchmark, and is even competitive with BERT, but still falls behind the more recent text-only models (He et al., 2020).

DALL-E (Ramesh et al., 2021) is a massive 12B parameter architecture, aimed at autoregressively modeling text and image tokens as a single stream of data. It consists of a modified VQ-VAE (van den Oord et al., 2017) visual encoder to produce a 32x32 grid of image tokens and an autoregressive Transformer that concatenates up to 256 text embeddings with the 32x32 = 1024 image tokens. The model is optimized with Expected Lowed Bound loss in two stages, first by optimizing the image token encoder, and then learning the prior distribution over text and image tokens. DALL-E also proposes a novel dataset of 250M image/text pairs from the Internet. A lot of engineering work is reported with regard to model scale and mixed-precision training stability. The model shows strong generalization for combination (e.g., to display a specific text within generated image) as well as text-guided image translation, although it does struggle with zero-shot performance on out-of-distribution datasets.

While these models do show impressive results, it can be difficult to extract embeddings representing each modality. One possible way to do this is to take an average-pooled embedding of all encoder representations for the two modalities. However, since the models' representation spaces are not explicitly optimized for semantic alignment, the representations may suffer from the space collapse issue outlined in § 2.1. These approaches do not fit our goals (1): semantic space and (3): versatility.

3 S5 Checklist

In this section, we present the checklist of promising research directions for shared semantic space multimodal learning. We will discuss the model architecture, training objectives and datasets, as well as specific considerations related to multimodal learning with a unified encoder. We aim to empower the reader to produce cutting-edge modeling results.

3.1 Architecture Choice

The first item on the checklist is the underlying model architecture. Kaiser et al. (2017) find that multimodal architectures benefit from parameter sharing via a unified encoder. Following that, we meet the goals (2): multi-context capture and (4): ease of use by choosing a unified Transformer encoder with modality and relative position embeddings. The data streams are separately tokenized and embedded. The input data format is as follows: [CLST], $t_1, \cdots, t_L$, [CLSV], $v_1, \cdots, v_N$, [EOS], for $t_1, \cdots, t_L$ - text tokens, $v_1, \cdots, v_N$ - visual tokens, and [CLST], [CLSV] - learnable class embeddings for contrastive optimization. The embedding and optimization procedure closely follows

ViLT (Kim et al., 2021), except the visual embedding layer is not necessarily linear (more on this in § 3.4.1). Following SimVLM (Wang et al., 2021), we use 2D relative attention (Parmar et al., 2018) between the image stream embeddings.

Indeed, this S5 encoder architecture choice is beneficial not just for simplicity. It also lets the model learn cross-modal interactions at multiple contextual levels (different Transformer layers). It also alleviates overfitting issues related to potential difference in per-modality optimization rates associated with late-fusion multimodal networks (Wang et al., 2019). This architecture choice also has the advantage of preventing the model parameter count from getting too large.

3.2 Training Objective

The second item we need to check off is optimization objective. In order to meet goal (3): versatility, we use both image-only, text-only, and text-image data examples in training. In order to meet goals (1): semantic space, and (2): multi-context capture, the model optimizes for both contextual information as well as the shared representation space alignment. Our learning objective is as follows:

$$\mathcal{L} = \lambda \mathcal{L}_{Denoise} + (1 - \lambda)\mathcal{L}_{Contrast}$$

, with $\mathcal{L}_{Denoise}$ being the modality-specific denoising term to capture contextual information, and $\mathcal{L}_{Contrast}$ - contrastive term to optimize the semantic representation space, with hyperparameter λ. Such a combination of contrastive and denoising terms has been used successfully in pretraining of TaCL (Su et al., 2021), albeit for text only. More details on potential denoising and contrastive task fusion mechanisms are in § 3.4.2. We explore additional S5 checklist options for the two loss terms in more details below.

3.2.1 Denoising Term

When learning the contextual information from the data, we have several checklist options to consider for different modalities.

Text-Only data has been shown to benefit from following denoising objectives: Masked Language Modeling (Devlin et al., 2018), Span Corruption (Joshi et al., 2019; Raffel et al., 2019), Sequence Permutation (Lewis et al., 2019), as well as Gap-Sentence Generation (Zhang et al., 2019), which is essentially Span Corruption directed by the ROUGE1-F1 (Lin, 2004) score. The expectation is that Span Corruption will outperform others, since it has been shown to do well on both generation and understanding tasks (Raffel et al., 2019), and that task-agnostic objective is sufficient for most cases (Rothe et al., 2021), which reduces the incremental value offered by the GSG objective. We thus select Span Corruption as the promising candidate for text-only data.

Image-Only data has less literature, with current trends (Dosovitskiy et al., 2020; Chen et al., 2021) focusing on Masked Patch Prediction (MPP). MPP corrupts 50% of the patch embeddings by either replacing them with a learnable [MASK] embedding (80%), a random other patch embedding (10%), or keeps them the same (10%). This reconstruction loss mimics the MLM objective for text data. We select MPP as a promising candidate. We also use Span Corruption success in NLP to motivate further research into masking spans of multiple image patches, perhaps with more sophisticated visual embedding strategies as presented in § 3.4.1.

Text-Image data has been shown to benefit from conditional Masked Language and Patch Modeling per UNITER (Chen et al., 2019). We don't expect to use Span Corruption here, since we shall remove the assumption that image description must have significant syntactic structure; that will allow us to use large-scale noisy data for pretraining (Jia et al., 2021). Since we do not have class labels for image patches, we use UNITER's Masked Region Feature Regression (MRFR) variant of Masked Region Modeling. Per UNITER and ViLT we also add a Word-Region Alignment with Optimal Transport objective on text embeddings in relation to image embeddings, further encouraging the alignment at patch/token level.

3.2.2 Contrastive Term

This term aims to optimize the Transformer model's semantic space properties by improving alignment between positive examples, while increasing distance between negative examples in each batch. We overview a few checklist options with regard to this objective.

Modality Embedding is the first choice in multimodal contrastive learning. That can be done either as an average-pool of all encoded modality token embeddings, per SBERT (Reimers and Gurevych, 2019), or as [CLSV] and [CLST] embedding outputs. The modality embedding will be used for calculating contrastive loss between positive/negative example pairs.

Augmentation Strategies are used to form positive data examples. Positive text-only pairs are formed with Dropout, Token Deletion, and Feature Deletion (Yan et al., 2021). Positive image-only pairs are formed with Cropping and Color Distortion (Chen et al., 2020). Text-image data is already paired; the positive examples are then implicitly formed by modality transfer without any explicit augmentation.

Contrastive Tension is an extensively used method of contrastive optimization. The NT-Xent (Chen et al., 2020) is used to compute the batch loss for unimodal pairs:

$$\mathcal{L}_{CUniMode} = -\mathbb{E}\left[\log \frac{\exp(\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_i^{+})/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j^{+})/\tau)}\right]$$

, with $\boldsymbol{h}_i, \boldsymbol{h}_i^{+}$ - positive pair embeddings, and τ - hyperprameter, N - batch size. For multimodal pairs, the loss is adapted (Radford et al., 2021; Jia et al., 2021) as follows:

$$\mathcal{L}_{CCrossMode} = \mathcal{L}_{Img2Txt} + \mathcal{L}_{Txt2Img}$$

, such that

$$\mathcal{L}_{Img2Txt} = -\mathbb{E}\left[\log \frac{\exp(\boldsymbol{x}_i^{\top} \boldsymbol{y}_i/\sigma)}{\sum_{j=1}^{N} \exp(\boldsymbol{x}_i^{\top} \boldsymbol{y}_j/\sigma)}\right]$$

$$\mathcal{L}_{Txt2Img} = -\mathbb{E}\left[\log \frac{\exp(\boldsymbol{y}_i^{\top} \boldsymbol{x}_i/\sigma)}{\sum_{j=1}^{N} \exp(\boldsymbol{y}_i^{\top} \boldsymbol{x}_j/\sigma)}\right]$$

, with $\boldsymbol{x}_i, \boldsymbol{y}_i$ - matching text/image pairs, σ - hyperparameter, and N - batch size.

Alignment and Uniformity metrics can also be directly used to optimize the semantic embedding space per Wang and Isola (2020). Thus, we also check the following contrastive loss formulation:

$$\mathcal{L}_{Contrast} = \alpha \mathcal{L}_{Align} + \beta \mathcal{L}_{Uniform}$$

, with α, β - hyperparameters, and

$$\mathcal{L}_{Align} = \mathbb{E}[\|\boldsymbol{h}_i - \boldsymbol{h}_i^{+}\|_2^2]$$

$$\mathcal{L}_{Uniform} = \log \mathbb{E}[\exp(-t\,\|\boldsymbol{h}_i - \boldsymbol{h}_j\|_2^2)], \quad t > 0$$

, with $\boldsymbol{h}_i, \boldsymbol{h}_i^{+}$ - embeddings for positive examples, t - hyperparameter. This way, $\mathcal{L}_{Align}$ directly minimizes l_2 distance between matching examples ensuring alignment, and $\mathcal{L}_{Uniform}$ reduces the Gaussian potential of the batch ensuring uniformity (Wang and Isola, 2020).

Dataset	Scale
Objects365 (Shao et al., 2019)	2M
ImageNet-21K (Ridnik et al., 2021)	14.2M
Graph-RISE (Juan et al., 2019)	260M
JFT-300M (Sun et al., 2017)	300M
3.5B Instagram (Mahajan et al., 2018)	3.5B
MS-COCO (Lin et al., 2014)	120K
SBU Captions (Ordonez et al., 2011)	1M
Conceptual Captions (Sharma et al., 2018)	3.3M
Conceptual 12M (Changpinyo et al., 2021)	12M
DALL-E (Ramesh et al., 2021)	250M
WebImageText (Radford et al., 2021)	400M
ALIGN (Jia et al., 2021)	1.8B

Table 1: A summary of the datasets from the literature for image-only (top portion) and text-image (bottom portion) tasks. Note that Conceptual 12M, DALL-E, and ALIGN are all supersets of Conceptual Captions. Scale denotes the number of examples.

3.3 Training Data

To optimize for goals (1): semantic space and (3): versatility, we check using both single and multimodal data examples in pretraining. We expect the pretraining dataset to be large enough to avoid overfitting, but also unbiased enough to allow for a generalizable semantic embedding space.

For text-only data, C4 (Raffel et al., 2019) has been the prevalent choice. However, the C4 data, as it was originally presented, contains multiple levels of bias (Dodge et al., 2021). For one, it has an ethnic negative sentiment bias (most notably against Arab identities), which may lead to direct negative bias against ethnic identities on downstream tasks. C4 also contains exclusion bias based on race (against Black and Hispanic authors) and sexual identity (against LGBTQ+ communities), brought on by the block-word filtering applied to Common Crawl data. This exclusion is a form of allocation harms, and may exacerbate the current racial inequality as well as stigmatization of LGBTQ+ identities, depriving those groups of benefits of technology and handicapping real-world downstream performance of the model.

While subdomain sampling and filter relaxation can be useful for de-biasing C4, those approaches are not sufficient to meet the task by itself per Dodge et al. (2021). This issue, although unknown, may also be present in image-only and text-image datasets, summarized in Table 1. To train an unbiased S5 encoder and, more importantly, to ensure an equitable distribution of technological benefits in society, we urge the need for additional data cleaning research to produce unbiased large-scale pretraining corpora.

3.4 Special Considerations

Along with introducing multiple data modes and a mix of denoising and contrastive objectives, there will be additional challenges associated with the work. Here, we identify a checklist of choices related to visual embedding construction and multi-domain multi-task optimization.

3.4.1 Visual Embeddings

In the previous vision/language works, visual embeddings were constructed by encoding the result of an object detector. For our checklist, we optimize for goals (3): versatility and (4): ease of use by moving away from object detectors. We discuss the potential embedding approaches below, in the order of increasing complexity.

Linear Projection is the most straightforward way to obtain image pixel patch embeddings (Dosovitskiy et al., 2020; Kim et al., 2021; Chen et al., 2021). This results in quicker image embedding computation, and an overall increased inference performance. This may be a viable option, but it may not capture the necessary image contexts. Also, a linear patch embedding layer may lead to subtle but detrimental instability and heightened sensitivity to optimizer choice during training; that can be alleviated, but not completely resolved for large learning rates, with freezing the linear visual patch embedding layer (Chen et al., 2021).

Convolutional Layers is another choice of image encoder. This encoder could be the first 3 layers of a ResNet (He et al., 2016), per Wang et al. (2021), or 3 MBConv (Sandler et al., 2018) layers, per Dai et al. (2021). This approach will allow the model to leverage the translation equivalence property of convolutional embeddings. This is shown to improve generalization under datasets of limited size (Mohamed et al., 2020), as well as increasing training stability and peak performance of vision transformers (Xiao et al., 2021).

VQ-VAE (van den Oord et al., 2017) visual embeddings rely on learning a codebook of discrete tokens for each image patch, using an encoder, a quantizer, and a decoder. The codebook is optimized by minimizing the original image reconstruction loss from decoding the quantized representations. DALL-E (Ramesh et al., 2021) learns such codebook with pixel-wise reconstruction loss on a very large dataset. PeCo (Dong et al., 2021) instead demonstrates effectiveness of *perceptual* image reconstruction loss on a much smaller ImageNet-1K (Deng et al., 2009) data, using ViT-B as per Chen et al. (2021) to minimize the l_2 distance of original versus reconstructed image representations at different layers of the ViT-B model. Such perceptual codebook tokens are demonstrated to carry high semantic meaning, as opposed to the low-level contents learned with pixel-wise reconstruction loss.

3.4.2 Multi-Domain Multi-Task Optimization

The proposed model aims to optimize both denoising and contrastive loss functions simultaneously, a multi-task learning (MTL) problem as defined by Ruder (2017). Although a long-standing paradigm, the static gradient weighting mentioned in § 3.2 is not necessarily optimal, due to dynamic nature of the multimodal gradient imbalance during training (Wang et al., 2019). Another concern is the diversity of data domains we aim to optimize: text-only, image-only, and text/image, as motivated by goal (3): versatility. We will use this section to discuss alternative optimization strategies for our multi-task and multi-domain problem.

Cascaded Learning is an MTL strategy where each model layer is optimized for a distinct, progressively more complex task. Søgaard and Goldberg (2016) show that syntactic chunking supervised at higher layers of BiLSTM benefits from POS tagging supervised at the lower layers. Hashimoto et al. (2016) extend that result to more tasks and show that *both* low- and high- level tasks benefit from cascaded learning at the corresponding layers. This strategy may be especially applicable to Transformer models, as they tend to learn increasingly complex structural properties of both text (Jawahar et al., 2019) and image (Raghu et al., 2021) data in deeper layers. However, this strategy also requires domain expertise to devise the correct hierarchy of tasks in multimodal context.

Dynamic Gradient Reweighting is an MTL strategy where weights for linear combination of gradients change dynamically throughout the training process. Kendall et al. (2017) propose an uncertainty-based method to assign lower weights to noisier task gradients. Since it is not always clear which task is primary and which is auxiliary (as is the case with our objective), Sener and Koltun (2018) propose to instead find a Pareto-optimal weighting of gradients, such that no other weighting improves all tasks. In multimodal context, Wang et al. (2019) propose introducing and updating an overfitting prior to model optimization. They do so by training for several iterations, estimating the overfitting on each modality, and then re-training the iterations this time using the overfitting prior to weigh the gradient combination.

Differentiable Data Selection is a data resampling approach to overcome data imbalance. The general paradigm is to use a held-out set to train and update a scorer network, which selects the data to be sampled for main model training. The main model and scorer are trained with bilevel optimization. Reinforcement learning is used to update the scorer weights. Wang et al. (2020b) propose a scorer update reward using cosine similarity of main model's gradients on held-out set. Wu et al. (2021) expose a flaw in gradient similarity approach, as it breaks down on highly self-correlated domains, and instead propose an uncertainty-based scorer reward, achieving performance improvements on multilingual and multi-domain tasks. This strategy can be particularly useful to determine the batch data composition for our multi-modal task.

4 S5 Evaluation

S5 Framework aims to produce an encoder optimized for an aligned and uniform semantic space. Cosine similarity, measuring representation proximity in linear space, then directly measures semantic alignment as well as reflects embedding interpretability. Thus, a strong performance on zero-shot tasks with cosine similarity is a direct success criterion of this work. We will review zero-shot evaluation methods from the multimodal literature and further adapt them to be suitable for cosine similarity, thus proposing novel tasks.

Zero-Shot Retrieval is used by ViLT (Kim et al., 2021) on MSCOCO and Flickr30k (Karpathy and Fei-Fei, 2014). We further constrain the task to formulate a novel experiment: *interpretable zero-shot retrieval* with Approximate Nearest Neighbor (ANN) search (Andoni et al., 2018) and cosine similarity. Without any additional training, we use ANN + cosine similarity to evaluate multimodal retrieval with S5 embeddings. This task better simulates the real-world applications of the system: both ANN and cosine similarity are common retrieval optimizations used in the industry, and the domain-specific data is often in short supply. Cosine similarity also allows us to perform retrieval in an interpretable way.

Zero-Shot Cross-Modal Transfer is used by SimVLM (Wang et al., 2021) on SNLI-VE (Xie et al., 2018). We further constrain the task by training an MLP layer on top of the *frozen* S5 encoder on text-only NLI data. This way, if our text and image embeddings are properly aligned in the shared space, we should achieve zero-shot transfer by evaluating the model on text-image entailment data without a drop in performance.

Zero-Shot Semantic Similarity is used by Reimers and Gurevych (2019), Gao et al. (2021), and Yan et al. (2021) to evaluate performance of self-supervised language training on SentEval (Conneau and Kiela, 2018) datasets. We propose to crowdsource a novel text/image dataset comprised of text-image pairs with semantic similarity score on the scale 0-5. Following Reimers and Gurevych (2019), we use Spearman coefficient to calculate the correlation between cosine similarity of data embeddings and ground truth similarity score. Assuming an equivalent data quality, we compare results to similar zero-shot evaluation on SentEval, aiming to not see a significant performance discrepancy on uni-modal versus cross-modal data.

5 Conclusion

Textual concept representation, as robust as it may become from Web-scale data, is still incomplete on its own. Only via incorporating mutlimodal information can machine intelligence advance towards human intelligence (Bisk et al., 2020), the ultimate goal of our work. In this review, we present a framework consisting of the necessary background literature, an informed research decisions checklist, and a few novel zero-shot experiments aimed at improving S5 multimodal learning. We hope this work is useful in cultivating interest regarding the promising multi-modal learning directions.

A limitation of our unified-encoder approach is the computation cost associated with increasing input lengths and batch sizes. The input sequence length increase will come from concatenating the per-modality tokens. The batch size increase will come from contrastive learning objective; current Transformer vision/language models use batch size of 4,096 (Wang et al., 2021; Kim et al., 2021). These issues warrant further exploration of efficient attention mechanisms (Tay et al., 2020).

References

Alexandr Andoni, Piotr Indyk, and Ilya P. Razenshteyn. 2018. Approximate nearest neighbor search in high dimensions. *CoRR*, abs/1806.09823.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph P. Turian. 2020. Experience grounds language. *CoRR*, abs/2004.10151.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *CoRR*, abs/2102.08981.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709.

Xinlei Chen, Saining Xie, and Kaiming He. 2021. An empirical study of training self-supervised vision transformers. *CoRR*, abs/2104.02057.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: learning universal image-text representations. *CoRR*, abs/1909.11740.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. 2021. Coatnet: Marrying convolution and attention for all data sizes. *CoRR*, abs/2106.04803.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. 2021. Documenting the english colossal clean crawled corpus. *CoRR*, abs/2104.08758.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *CoRR*, abs/1905.03197.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. 2021. Peco: Perceptual codebook for BERT pre-training of vision transformers. *CoRR*, abs/2111.12710.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and GPT-2 embeddings. *CoRR*, abs/1909.00512.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. *CoRR*, abs/1907.12009.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *CoRR*, abs/2104.08821.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2016. A joint many-task model: Growing a neural network for multiple NLP tasks. *CoRR*, abs/1611.01587.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Taichi Iki and Akiko Aizawa. 2021. Effect of vision-and-language extensions on natural language understanding in vision-and-language models. *CoRR*, abs/2104.08066.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *CoRR*, abs/2102.05918.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. *CoRR*, abs/1907.10529.

Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. 2019. Graph-rise: Graph-regularized image semantic embedding. *CoRR*, abs/1902.10814.

Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. *CoRR*, abs/1706.05137.

Andrej Karpathy and Li Fei-Fei. 2014. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2017. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *CoRR*, abs/1705.07115.

Salman H. Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2021. Transformers in vision: A survey. *CoRR*, abs/2101.01169.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Yann LeCun and Corinna Cortes. 2010. MNIST handwritten digit database.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020a. On the sentence embeddings from pre-trained language models. *CoRR*, abs/2011.05864.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *CoRR*, abs/1908.06066.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. *CoRR*, abs/2004.06165.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Dhruv Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the limits of weakly supervised pretraining. *CoRR*, abs/1805.00932.

Mirgahney Mohamed, Gabriele Cesa, Taco S. Cohen, and Max Welling. 2020. A data and compute efficient design for limited-resources deep learning. *CoRR*, abs/2004.09691.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, and Alexander Ku. 2018. Image transformer. *CoRR*, abs/1802.05751.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do vision transformers see like convolutional neural networks? *CoRR*, abs/2108.08810.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *CoRR*, abs/2102.12092.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497.

Tal Ridnik, Emanuel Ben Baruch, Asaf Noy, and Lihi Zelnik-Manor. 2021. Imagenet-21k pretraining for the masses. *CoRR*, abs/2104.10972.

Sascha Rothe, Joshua Maynez, and Shashi Narayan. 2021. A thorough evaluation of task-specific pretraining for summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 140–145, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098.

Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. *CoRR*, abs/1810.04650.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany. Association for Computational Linguistics.

Yixuan Su, Fangyu Liu, Zaiqiao Meng, Lei Shu, Ehsan Shareghi, and Nigel Collier. 2021. Tacl: Improving bert pre-training with token-aware contrastive learning. *ArXiv*, abs/2111.04198.

Alane Suhr, Stephanie Zhou, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *CoRR*, abs/1811.00491.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. *CoRR*, abs/1707.02968.

Mingxing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient transformers: A survey. *CoRR*, abs/2009.06732.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. *CoRR*, abs/1711.00937.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. 2020a. Improving neural language generation with spectrum control. In *ICLR*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *CoRR*, abs/2005.10242.

Weiyao Wang, Du Tran, and Matt Feiszli. 2019. What makes training multi-modal networks hard? *CoRR*, abs/1905.12681.

Xinyi Wang, Hieu Pham, Paul Michel, Antonios Anastasopoulos, Jaime Carbonell, and Graham Neubig. 2020b. Optimizing data usage via differentiable rewards. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9983–9995. PMLR.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904.

Minghao Wu, Yitong Li, Meng Zhang, Liangyou Li, Gholamreza Haffari, and Qun Liu. 2021. Uncertainty-aware balancing for multilingual and multi-domain neural machine translation training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7291–7305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross B. Girshick. 2021. Early convolutions help transformers see better. *CoRR*, abs/2106.14881.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2018. Visual entailment task for visually-grounded language learning. *CoRR*, abs/1811.10582.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *CoRR*, abs/2105.11741.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CoRR*, abs/2101.00529.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2019. Unified vision-language pre-training for image captioning and VQA. *CoRR*, abs/1909.11059.