VAuLT: Augmenting the Vision-and-Language Transformer with the Propagation of Deep Language Representations

Georgios Chochlakis, Tejas Srinivasan, Jesse Thomason, Shrikanth Narayanan
University of Southern California
{chochlak, tejass55, jessetho}@usc.edu, shri@ee.usc.edu

Abstract
We propose the Vision-and-Augmented-Language Transformer (VAuLT). VAuLT is an extension of the popular Vision-and-Language Transformer (ViLT), and improves performance on vision-and-language tasks that involve more complex text inputs than image captions while having minimal impact on training and inference efficiency. ViLT, importantly, enables efficient training and inference in vision-and-language tasks, achieved by using a shallow image encoder. However, it is pretrained on captioning and similar datasets, where the language input is simple, literal, and descriptive, therefore lacking linguistic diversity. So, when working with multimedia data in the wild, such as multimodal social media data (in our work, Twitter), there is a notable shift from captioning language data, as well as diversity of tasks, and we indeed find evidence that the language capacity of ViLT is lacking instead. The key insight of VAuLT is to propagate the output representations of a large language model like BERT to the language input of ViLT. We show that such a strategy significantly improves over ViLT on vision-and-language tasks involving richer language inputs and affective constructs, such as TWITTER-2015, TWITTER-2017, MVSA-Single and MVSA-Multiple, but lags behind pure reasoning tasks such as the Bloomberg Twitter Text-Image Relationship dataset. We have released the code for all our experiments at https://github.com/gchochla/VAuLT.

Introduction
The study of social media can have positive societal outcomes. Interest in the importance of emotion has been growing in various sciences (Dukes et al., 2021), law (Bandes and Blumenthal, 2012), Wistrich, Rachlinski, and Guthrie (2014), and politics (Wahl-Jorgensen, 2019). Social media users readily express their emotions about their experiences, and with regards to specific entities and events. They do so through multimedia content such as videos, images, memes, and raw text. This allows researchers to use machine learning (ML) methods to study, for instance, the dynamics of public perception and opinion with respect to on-going events and policy changes and to provide feedback to policy makers, or even look at past events to inform future actions (Hoover et al., 2018). Unfortunately, social media are also ripe with opportunities for social manipulation (Bradshaw and Howard, 2018). This manipulation can be personalized based on the data users, voluntarily or not, supply (Alfano, Sullivan, and Fard, 2022), or take advantage of network effect inherent in human relations and social media (Hoover et al., 2021). Mitigating these risks and identifying widespread manipulation attempts hence becomes pivotal.

Due to the massive volume of data and the dynamic nature of both social media platforms and world events, constant streams of data are required, hence model efficiency is of the essence during training and inference. In addition, since users share multimedia content, multimodal modeling is required, as users can use these additional degrees of freedom in a way that, when expressing irony for example, one modality can be in conflict with the other, and flip the overall appraisal of a post. Moreover, a multimodal model should be able to handle one modality at a time when a post is missing one of the two, as was illustrated by Srinivasan et al. (2022).

Many state-of-the-art vision-and-language models, e.g., VL-BERT (Su et al., 2019) and ViLBERT (Lu et al., 2019), contain an object detector, such as Faster-RCNN (Ren et al., 2015), or take advantage of network effect inherent in human relations and social media (Hoover et al., 2021). Mitigating these risks and identifying widespread manipulation attempts hence becomes pivotal.
to compute input visual features for the main Transformer model (Vaswani et al. 2017) that does the joint vision-and-image modeling; this can be costly to run relative to the rest of the pipeline, even in inference mode (Kim, Son, and Kim 2021). Hence, we may have to resort to a more efficient alternative, such as the Vision-and-Language Transformer (ViLT) (Kim, Son, and Kim 2021), whose visual input consists of non-overlapping image patches projected to token-like embeddings with an linear projection.

Regardless, the aforementioned ML models, designed to understand and extract information from multimedia content such as vision and language, are frequently trained only on paired image and language caption data crawled from the web. Image captions are a weak form of language supervision, often describing literal content with simple syntax and little subtlety, deviating substantially from the text of social media posts.

In this paper, we propose the Vision-and-Augmented-Language Transformer (Vault), which directly addresses the impoverished language representations of the pretrained vision-and-language ViLT model by processing language input through a pretrained large language model (Figure 1). We find that Vault is able to outperform ViLT on a variety of tasks involving affective constructs common in social media, including TWITTER-2015, TWITTER-2017, and MVSA-Multiple.

Our key contributions can be summarized as follows:

- In Vault, we stack Transformer architectures, a language model and a vision-and-language model trained on different kinds of data, and show substantial gains in performance from relatively little tuning data.
- We show that Vault competes with state-of-the-art architectures for the tasks examined.
- We provide evidence that the performance trade-offs of ViLT and Vault depend on the complexity of the image and language model components commensurate with the relative richness of, and variety within, the modalities in the source and target domains, as well as the tasks themselves.

### Related Work

#### Language Transformers

Because of the efficient forward propagation afforded by Transformers (Vaswani et al. 2017), they have been scaled to billions of parameters and trained on vast amounts of crawled, general-purpose text, which grants them the ability to be used as backbones for other downstream tasks with fine-tuning e.g., BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), or without e.g., GPT-type models (Radford et al. 2019). Such models can also be pretrained on more specialized corpora to improve their transferability to specific domains such as the BERTweet (Nguyen, Vu, and Nguyen 2020), a RoBERTa model trained on tweets.

#### Vision-and-Language Transformers

Similar to language transformers, vision-and-language transformers (Su et al. 2019) Tan and Bansal 2019 Chen et al. 2019) are trained with multiple input modalities on web-crawled captions (Sharma et al. 2018a). These models yield state-of-the-art results on many multimodal tasks, such as VQA (Antol et al. 2015), Visual Dialog (Das et al. 2017), and VCR (Zellers et al. 2019). These models frequently use region-of-interest (ROI) features extracted from a pretrained object detector (Ren et al. 2015) as their Transformer’s visual inputs. However, the visual embedding step is computationally expensive during inference and, more so, during training (Kim, Son, and Kim 2021), requiring caching of features in either case. The ViLT vision-language model (Kim, Son, and Kim 2021), based on the Vision Transformer (ViT) (Dosovitskiy et al. 2020), proposed using image patches as visual input to their Transformer model instead of ROI features requiring a heavy visual embedding step. ViLT’s shallow image embedding step results in slightly worse model performance, but is much more computationally efficient. However, ViLT also has limited language understanding capacity due to being initialized from ViT and pretrained solely on image caption data. Vault overcomes this issue by replacing ViLT’s language embedding inputs with features extracted from a large language model pretrained on greater linguistic diversity, which can also be selected to more closely suit the needs of the downstream task of interest.

### Affective Analysis on Social Media

Researchers tend to focus on single modalities when analyzing emotion and sentiment in social media posts, which in the majority of cases is text (Mohammad et al. 2018; Sailunaz and Alhajj 2019; Liew, Turtle, and Liddy 2016; Demszy et al. 2020; Abdul-Mageed and Ungar 2017). There is also a significant amount of work that focuses exclusively on images (Zhao et al. 2016), or even more specifically, Internet memes (Sharma et al. 2020; Kiela et al. 2020; Mathias et al. 2021). However, there are situations where images are treated as nuisance during the data collection phase and are readily filtered out e.g., Sailunaz and Alhajj 2019.

Emotion recognition from text has long relied on word-counting techniques, such as LIWC (Pennebaker, Francis, and Booth 2001) and DDR (Garten et al. 2018), which can be deployed at scale but that disregard context. LSTMs (Baziotis et al. 2018) and Transformers (e.g., SpanEmo by Alhuzali and Ananiadou 2021) model context, provide superior performance but offer less interpretability.

On the other hand, multimodal affect modeling tasks with social media have predominantly focused on sentiment (Yu and Jiang 2019) or related constructs, such as detecting hateful memes (Mathias et al. 2021). Yu and Jiang 2019 have annotated tweets with their corresponding images for sentiment expressed by the author of the tweet towards entities in the text, resulting in the TWITTER-15 and TWITTER-17 datasets, and a task described as Target-oriented Sentiment Classification (TMSC). Similarly, Niu et al. 2016 have released MVSA, where they have annotated the sentiment expressed in a tweet for image and text separately (given both of them, though). The existence of pairs of labels offers the ability to filter MVSA and arrive at a single label for each tweet in different ways depending on the corresponding con-
text MVSA is used in [Hu and Flaxman 2013] have scraped multimodal data from Tumblr and then used user-provided tags as “self-reports” of emotions to automatically generate a “distantly-annotated” dataset.

The proposed approaches for the aforementioned tasks typically involve dedicated architectures or targeted modifications of Vision-and-Language Transformers. An example of the former is TomBERT [Yu and Jiang 2019] for TMSC. TomBERT uses components of BERT to query visual features from ResNet [He et al. 2016] using the standalone sentiment targets, and these are concatenated with the tweet and fed to a regular BERT model for the final prediction. For modifications of multimodal transformers, we can look at VL-BERT+ [Zhu 2020], where data such as race and gender extracted from the images as well as web-recommended tags are fed to VL-BERT.

**Background**

**Task Definitions**

In this section, we define the tasks we use in this work to evaluate our approach.

**Sentiment Classification** For our purposes, sentiment classification corresponds to the task of assigning an image and text pair to one of three categories, positive, neutral, or negative.

**Target-oriented Sentiment Classification (TMSC)** extends the basic sentiment classification formulation to opinion expressed toward given targets within the text of multimodal tweets. In our setting, sentiment targets were repurposed from types of named entities, and an input tweet can contain multiple targets.

**Text Representation** This is a binary task, where multimodal tweets are annotated for whether the text of the tweet is represented in the image. This deparfs from the affective tasks discussed above.

**Multimodal Encoders**

**Vision-and-Language Transformer (ViLT)** is a multimodal transformer. On the language side, ViLT follows conventional practices and embeds input tokens with an embedding matrix, which is essentially a look-up table. On the vision side, it follows ViT [Dosovitskiy et al. 2020], breaking each image into several non-overlapping square patches of constant dimensions, flattening the resulting values, and mapping them to the input space using a linear projection, in a token-like fashion.

ViLT is pretrained on the image captioning datasets Conceptual Captions [Sharma et al. 2018b], MSCOCO [Lin et al. 2014], Visual Genome [Krishna et al. 2017], and SBU Captions [Ordonez, Kulkarni, and Berg 2011], and has been fine-tuned by the original authors for VQAv2 [Goyal et al. 2017], NLVR2 [Suhr et al. 2018], or Flickr30k [Plummer et al. 2015]. Its pretraining objectives include: image-text matching and word-patch alignment, and masked language modeling and whole word masking. The former group focuses on enforcing joint modeling of the two modalities, while the latter is centered on the language understanding of the model. Explicit language training is necessary since the model is initialized from ViT weights, in contrast to BERT or similar models used in previous works [Su et al. 2019; Lu et al. 2019], meaning that the aforementioned pretraining on image-captioning datasets is the only language training ViLT is exposed to.

**Proposed Method**

The proposed Vision-and-Augmented-Language Transformer (VAlLT) has a straightforward architecture. We can use any language model that produces token embeddings for tokenized text, irrespective of the specific tokenizer. We have this requirement because we use the sequence of embeddings produced by the language model to replace the canonical language input of ViLT, namely embedding language tokens with a lookup table. In particular, Transformer-based language models produce contextual token representations by virtue of their attention mechanism and pretraining [Devlin et al. 2018]. Therefore, we are essentially substituting the context-free representations from the lookup table of ViLT, i.e., its learned language embeddings, with context-aware representations from the output layer of a language-only Transformer. Images are embedded in the canonical fashion with ViLT’s trained linear projection of patches.

Concretely, given an input pair $(t, v)$, where $t$ is the text sequence and $v$ the image, we select a language model $L$ and use its corresponding tokenizer $T$ to tokenize $t$ into $(t_{cls}, t_1, t_2, \ldots, t_n) = T(t)$, where $t_{cls}$ is a learnable “class” embedding for text inputs. We pass this input sequence through $L$ to produce the output contextual embeddings $(t_{cls}, t_1, t_2, \ldots, t_n) = L(t_{cls}, t_1, t_2, \ldots, t_n)$. Given flattened image patches from operator $P$, $(v_{cls}, v_1, v_2, \ldots, v_m) = P(v)$, the visual learnable “class” embedding $v_{cls}$, and the trained linear projection $V$, the input to ViLT becomes:

$$
\text{ViLT}((t_{cls}, \hat{t}_1, \hat{t}_2, \ldots, \hat{t}_n) + t_{type} + T_{pos};
(v_{cls}, V v_1, V v_2, \ldots, V v_m) + v_{type} + V_{pos}),
$$

where $:$ denotes concatenation, $t_{type}$ is ViLT’s modality embedding for text, which is constant for all tokens, $T_{pos}$ are ViLT’s position embeddings, a separate one for each position in the sequence (up to 40 in the original implementation), and similarly for the visual input for $v_{type}$, $V_{pos}$. Note that if the language model is transformer-based, as in our work, it adds position embeddings and potentially token-type embeddings to each token $t_i$, which are independent from ViLT’s $T_{pos}$ and $v_{type}$, respectively. In effect, for
VAuLT, we have:

\[ VAuLT(t, v) = \begin{cases} 
\text{ViLT}(L(T(t)) + t^{\text{type}} + T^{\text{pos}}; v_{cls}, V \cdot P(v)) + v^{\text{type}} + V^{\text{pos}}. 
\end{cases} \quad (2) \]

We use language models like BERT and BERTweet that have not been necessarily trained on similar datasets with ViLT or do tokenize text in the same manner as ViLT does. That is to say, the language tokens might not correspond to the tokens ViLT has been trained with.

**Experiments**

Overall, our goal is to show that VAuLT improves upon ViLT on multimodal, small-scale, affective datasets derived from social media, in our case Twitter. More than that, we also show that both ViLT and VAuLT can achieve competitive performance over dedicated architectures for such datasets. We use 5 datasets in total but essentially 3 distinct groups of data collections: TWITTER-15 and TWITTER-17 (Yu and Jiang 2019), MVSA-Single and MVSA-Multiple (Niu et al. 2016), and the Bloomberg Twitter Text-Image relationship dataset (Vempala and Preotuc-Pietro 2019).

**Datasets**

We now discuss further details about the datasets we used in our experiments.

We use TWITTER-15 and TWITTER-17 (Yu and Jiang 2019) annotated for the task of Target-oriented Sentiment Classification. As discussed previously, this involves predicting the sentiment expressed toward a given target in the text of the tweet such as a named entity, with potentially multiple targets per tweet. The number of examples contained in each can be seen in Table 1 with the splits provided by the authors. Accuracy and macro F1-score are typical evaluation criteria. We predict for one target at a time, even if the tweet contains multiple. We follow Yu and Jiang (2019) in replacing the target of interest within the tweet with the placeholder $T$, and appending the actual target string as a second sequence separated by the corresponding separation token of the model’s tokenizer.

We also evaluate our model on MVSA-Multiple and MVSA-Single (Niu et al. 2016), pre-processed for Sentiment Classification. Initially, both datasets contain annotations for the sentiment of both the image and the text of a tweet. The differences between the two datasets are the number of annotators and the number of annotated examples. MVSA-Single has annotations from solely one annotator, while MVSA-Multiple utilizes three separate annotators. It is customary (e.g., Xu and Mar 2017, Zhu et al. 2022) to merge the image and text annotations into a single annotation and aggregate across multiple annotators in the following way. First, for MVSA-Multiple, we perform aggregation by majority vote for each modality, discarding samples with ties. Then, after discarding the examples where the modalities have opposing annotations—one is positive and the other negative—we consider a tweet as positive if it contains at least one positive annotation, negative if it contains at least one negative annotation, and neutral otherwise. Since we do not find well-defined evaluation criteria in the literature, we use only accuracy to compare with existing models, but also use weighted F1-score for our experiments. The cardinality of the remaining tweets is presented in Table 2. Beyond that, however, we had to remove 3 additional samples from MVSA-Multiple because of corrupted images (example IDs: 3151, 3910 and 5995). Since canonical splits are not provided, we follow previous work in randomly splitting the data 8:1:1 (Zhu et al. 2022, Yang et al. 2020, although different splitting ratios have been used, e.g., by Zhang et al. 2022).

Finally, we consider the Bloomberg Twitter Image-Text relationship dataset (Vempala and Preotuc-Pietro 2019), Vempala and Preotuc-Pietro annotate for multiple binary tasks per multimodal tweet, such as if the text is represented in the image, whether the image adds to the textual information, etc. This is more of a pure reasoning task, used as a control for the aforementioned datasets. We only use the first task to evaluate our model for compatibility purposes with the literature, e.g. Khan and Fu (2021). Accuracy and weighted F1-score are used. We do not find any published splits, so based on Khan and Fu (2021), we randomly split 564 examples for validation and 704 for testing. The overall number of examples can be seen in Table 3.

**Implementation details**

We use Python (v3.7.4), the PyTorch (v1.11.0; Paszke et al. 2019) implementations of Transformer models from the Hugging Face transformers library (v4.19.2; Wolf et al. 2020), as well as pretrained models from PyTorch itself.

We keep our learning rate at $2 \cdot 10^{-5}$ unless specified. We use linear warm-up of the learning rate for 10% of the training steps and linear decay to 0 after that. In re-implementations of published models, we use the specified hyperparameter configuration except where noted.

We train and evaluate the model between 4 and 10 times on the development set, depending on resources required and available, to select the hyperparameters, whereas test performance (mean and standard deviation) is reported after 3 different training runs with fixed hyperparameters. The main tuning of all of the hyperparameters, but the number of training epochs, was performed on TWITTER-15 and TWITTER-17 (an arbitrary choice based on the trajectory of our efforts). We searched for whether to use bias correction in AdamW (Loshchilov and Hutter 2017) (no bias correction performed favorably), the number of epochs $\{5, 8, 10\}$ for MVSA-Multiple, $\{8, 15\}$ for Bloomberg and MVSA-Single, and $\{8, 15, 25\}$ for TWITTER-15 and TWITTER-17, and whether to integrate the placeholder $T$ to the tokenizer in TWITTER-15 and TWITTER-17 (yes for VAuLT, no for ViLT and existing baselines or variants thereof).

We find that divergence in the training of VAuLT can rarely occur, but those runs can be filtered by observing increases in the training metrics. While we had to correct for such artifacts during development, no such instances occurred for the models evaluated on the test sets.

In terms of image augmentation, we use simple random cropping instead of the augmentation utilized in ViLT for simplicity. We use augmentation in all cases where ViLT's
which are also substantial improvements but still lag behind compared to the BERTweet ones, and only find overlapping range of values across the two versions of V AuLT on TWITTER-17's accuracy. Additionally, both models provide increases in performance on MVSA-Multiple, with the BERT version of V AuLT improving over ViLT by 6.0% in accuracy and 4.2% in weighted F1 score, respectively, but insignificant improvements are observed from V AuLT with BERTweet. In MVSA-Single, where we have both less data and less reliable annotations, we observe a large drop in performance compared to MVSA-Multiple both in accuracy and weighted F1 score, and overall no improvements from V AuLT models over ViLT. Finally, ViLT does slightly better in the Bloomberg Twitter Image-Text relationship dataset. We see, therefore, that for reasoning tasks, ViLT performs equivalently (even perhaps favorably) to its V AuLT alternatives, but evaluation metrics on affective tasks are evidently improved by both versions of V AuLT.

Comparisons with State of the Art

For completeness, we also compare V AuLT and ViLT with state-of-the-art models for the examined benchmarks. We do not elaborate on the state-of-the-art models because of the variety of tasks, as well as them being not central to the main goals of our work. We find that both V AuLT and ViLT can be competitive with dedicated state-of-the-art models in terms of our evaluation criteria.

We observe, in Table 5, that V AuLT with BERTweet is on par with the state-of-the-art model, EF-CaTr-BERT, on TWITTER-15 and TWITTER-17, except for accuracy in TWITTER-17. V AuLT also performs favorably compared to the rest of the models, where we see that its mean scores improve upon the competition.

For MVSA-Multiple and MVSA-Single, our results when compared to the state-of-the-art are underwhelming, as can be seen in Table 6 with the BERT version of V AuLT barely competing in performance in MVSA-Multiple, while its performance is subpar in MVSA-Single. Perhaps that is unsurprising, since random splits are used in the literature, often with different cardinalities in the different splits, while evaluation criteria remain opaque and pre-processing and normalization techniques can differ between different works.

With regards to the Bloomberg Twitter Image-Text relationship dataset, results given in Table 7 ViLT and the state-of-the-art EF-CaTr-BERT performing better on average.

**Comparison with ViLT**

First, we directly compare the performance of ViLT and V AuLT (with BERT or BERTweet) in terms of the aforementioned evaluation criteria, and we indeed find that V AuLT exceeds ViLT’s performance on affective social media data. Results can be seen in Table 4. We can see that V AuLT with BERTtweet improves accuracy and macro F1 score over ViLT in TWITTER-15 and TWITTER-17, with relative improvements between 9.9% and 19.6%. V AuLT with BERT also provides improvements between 7.2% and 16.0%, which are also substantial improvements but still lag behind compared to the BERTweet ones, and only find overlapping range of values across the two versions of V AuLT on TWITTER-17’s accuracy. Additionally, both models provide increases in performance on MVSA-Multiple, with the BERT version of V AuLT improving over ViLT by 6.0% in accuracy and 4.2% in weighted F1 score, respectively, but insignificant improvements are observed from V AuLT with BERTweet. In MVSA-Single, where we have both less data and less reliable annotations, we observe a large drop in performance compared to MVSA-Multiple both in accuracy and weighted F1 score, and overall no improvements from V AuLT models over ViLT. Finally, ViLT does slightly better in the Bloomberg Twitter Image-Text relationship dataset. We see, therefore, that for reasoning tasks, ViLT performs equivalently (even perhaps favorably) to its V AuLT alternatives, but evaluation metrics on affective tasks are evidently improved by both versions of V AuLT.

**Comparisons with State of the Art**

For completeness, we also compare V AuLT and ViLT with state-of-the-art models for the examined benchmarks. We do not elaborate on the state-of-the-art models because of the variety of tasks, as well as them being not central to the main goals of our work. We find that both V AuLT and ViLT can be competitive with dedicated state-of-the-art models in terms of our evaluation criteria.

We observe, in Table 5 that V AuLT with BERTweet is on par with the state-of-the-art model, EF-CaTr-BERT, on TWITTER-15 and TWITTER-17, except for accuracy in TWITTER-17. V AuLT also performs favorably compared to the rest of the models, where we see that its mean scores improve upon the competition.

For MVSA-Multiple and MVSA-Single, our results when compared to the state-of-the-art are underwhelming, as can be seen in Table 6 with the BERT version of V AuLT barely competing in performance in MVSA-Multiple, while its performance is subpar in MVSA-Single. Perhaps that is unsurprising, since random splits are used in the literature, often with different cardinalities in the different splits, while evaluation criteria remain opaque and pre-processing and normalization techniques can differ between different works.

With regards to the Bloomberg Twitter Image-Text relationship dataset, results given in Table 7 ViLT and the state-of-the-art EF-CaTr-BERT performing better on average.
Table 4: Direct comparison between ViLT and V AuLT (either using BERT†, bert-base-uncased, or BERTweet*, vinai/bertweet-base) on each benchmark.

| Model     | TWITTER-15 | TWITTER-17 | MVSA-M | MVSA-S | Bloomberg |
|-----------|------------|------------|--------|--------|-----------|
|           | Acc mac-F1 | Acc mac-F1 | Acc w-F1 | Acc w-F1 | Acc w-F1 |
| ViLT      | 70.5 62.6  | 62.6 58.1  | 63.2 61.5 | 52.4 49.3 | 63.5 62.9 |
|           | ±1.3 ±2.5  | ±0.1 ±0.7  | ±0.6 ±0.5 | ±1.5 ±1.2 | ±0.6 ±0.6 |
| V AuLT†   | 75.6 70.0  | 70.2 67.8  | 67.0 64.1 | 51.7 49.1 | 61.2 61.0 |
|           | ±0.8 ±1.7  | ±0.4 ±0.1  | ±1.2 ±1.1 | ±1.6 ±1.1 | ±1.2 ±1.2 |
| V AuLT*   | **77.5** 72.9 | **71.0** 69.5 | **64.9** 62.3 | 47.5 45.4 | 61.1 60.6 |
|           | ±0.4 ±0.5  | ±0.5 ±0.7  | ±2.2 ±1.8 | ±1.3 ±0.9 | ±1.9 ±2.3 |

Table 5: Comparing V AuLT (using BERTweet, vinai/bertweet-base) with state-of-the-art models on TWITTER-15 and TWITTER-17 (BERT, mBERT and TomBERT from Yu and Jiang [2019], SaliencyBERT from Wang et al. [2021], and EF-CaTr-BERT from Khan and Fu [2021]).

| Model          | TWITTER-15 | TWITTER-17 | MVSA-M | MVSA-S |
|----------------|------------|------------|--------|--------|
|                | Acc mac-F1 | Acc mac-F1 | Acc    | Acc    |
| BERT (text only) | 74.3 70.0  | 68.9 66.1  |        |        |
| mBERT          | 75.8 71.8  | 68.8 67.1  |        |        |
| TomBERT        | 77.2 71.8  | 70.3 67.5  |        |        |
| SaliencyBERT   | 77.0 72.4  | 69.7 67.2  |        |        |
| EF-CaTr-BERT   | **77.9** 73.9 | **72.3** 70.2 | ±0.8  | ±0.3  |
| V AuLT         | ±0.4  | ±0.5  | ±0.5  | ±0.7  |
| V AuLT†        | 75.6 70.0  | 70.2 67.8  | ±0.8  | ±1.7  |
| V AuLT*        | 68.8 58.8  | 64.0 61.0  | ±0.6  | ±1.5  |
|                | **77.5** 72.9 | **71.0** 69.5 | ±0.4  | ±0.5  |

Table 6: Comparison of V AuLT (using BERT, bert-base-uncased) with state-of-the-art models on MVSA-Multiple and MVSA-Single (ITIN from Zhu et al. [2022], MVAN from Yang et al. [2020]).

| Model | TWITTER-15 | TWITTER-17 |
|-------|------------|------------|
|       | Acc        | Acc        |
| MVAN  | 72.4       | 73.0       |
| ITIN  | **73.5**   | **75.2**   |
| V AuLT| 67.0       | 51.7       |
|       | ±1.2       | ±1.6       |

Table 7: Comparison of ViLT and the state-of-the-art models on the Bloomberg Twitter Image-Text relationship corpus (EF-CaTr-BERT from Khan and Fu [2021]).

| Model          | TWITTER-15 | TWITTER-17 | MVSA-M | MVSA-S | Bloomberg |
|----------------|------------|------------|--------|--------|-----------|
|                | Acc        | Acc        | Acc w-F1 | Acc w-F1 | Acc w-F1 |
| EF-CaTr-BERT   | 64.8       | 64.0       | ±1.1   | ±1.3   |           |
| ViLT           | **63.5**   | **62.9**   | ±0.6   | ±0.6   |           |

Table 8: Comparing ViLT and V AuLT (either using BERT†, bert-base-uncased, or BERTweet*, vinai/bertweet-base, and using a frozen language model) on TWITTER-15 and TWITTER-17.

| Model          | TWITTER-15 | TWITTER-17 | MVSA-M | MVSA-S |
|----------------|------------|------------|--------|--------|
|                | Acc        | Acc        | Acc    | Acc    |
| V AuLT†        | 75.6       | 70.0       | 70.2   | 67.8   |
|                | ±0.8       | ±1.7       | ±0.4   | ±0.1   |
| V AuLT*        | 68.8       | 58.8       | 64.0   | 61.0   |
|                | ±0.6       | ±1.5       | ±0.9   | ±0.6   |
|                | **77.5**   | **72.9**   | **71.0** | 69.5   |
|                | ±0.4       | ±0.5       | ±0.5   | ±0.7   |

Ablation Studies

We present some additional experiments to support our arguments and findings, illuminate properties of the dynamics of stacked Transformer-based models and examine the dynamics between the complexity of the modalities and the capacity of the modality-specific components.

Frozen LM We first examine whether ViLT can “read” the outputs of a language model without any adjustments on the part of the latter. That is to say, we keep the language models frozen and solely train the ViLT parts of V AuLT to the specific task. In this way, we move one step closer to studying whether the language embeddings align a priori, and closer to the efficiency of ViLT, while also allowing for fine-tuning of a large part of our network. We also note the result that early layers of transformers change significantly less compared to later layers during fine-tuning (Merchant et al. 2020), potentially implying that fine-tuning of the language model in V AuLT may be unnecessary. We find that this is not the case.

Results can be seen in Tables 8 and 9. Overall, we find that the fine-tuning of the language model is essential, since in TWITTER-15 and TWITTER-17 we see such notable degradation in performance that TWITTER-15 results drop even below ViLT’s performance. Similar results can be observed also in the Bloomberg Twitter Image-Text relationship dataset, although the effects are not so severe and ranges can be overlapping. Note that we keep every other hyperparameter fixed and only disable the language model’s training in the aforementioned experiments. These experiments
Table 9: Comparing ViLT and V AuLT (either using BERT\textsuperscript{†}, bert-base-uncased, or BERTweet\textsuperscript{∗}, vinai/bertweet-base, and using a frozen\textsuperscript{†} language model) on the Bloomberg Twitter Image-Text relationship dataset.

| Model                   | Acc | w-F1 |
|-------------------------|-----|------|
| V AuLT\textsuperscript{†/f} | 59.3 | 58.7 |
| V AuLT\textsuperscript{†}    | ±0.9 | ±0.8 |
| V AuLT\textsuperscript{*/f}   | ±1.2 | ±1.2 |
| V AuLT\textsuperscript{*}     | ±1.5 | ±1.5 |
|                           | ±1.9 | ±2.3 |

Table 10: Comparing ViLT, V AuLT (with BERT, bert-base-uncased), TomViLT and TomV AuLT on TWITTER-15 and TWITTER-17. Bold assigned only based on means.

| Model            | Acc | mac-F1 |
|------------------|-----|--------|
| TomV AuLT        | 75.0 | 69.0 |
| V AuLT\textsuperscript{†} | ±0.6 | ±0.9 |
| V AuLT\textsuperscript{†/f} | - | - |
| TomViLT          | 73.2 | 67.1 |
| ViLT             | 69.6 | 60.3 |
| TomV AuLT        | ±0.5 | ±0.4 |
| ViLT             | ±1.0 | ±1.5 |
| TomV AuLT        | ±0.5 | ±0.4 |
| ViLT             | ±1.2 | ±1.5 |
| TomV AuLT        | ±0.5 | ±0.4 |
| ViLT             | ±1.0 | ±0.9 |
| TomV AuLT        | ±0.5 | ±0.4 |
| ViLT             | ±1.0 | ±0.9 |

**Deep Vision vs. Deep Language** In this section, we present an experiment supporting the hypothesis that the complexity of the modality-specific components and their effect on the final performance depends on the domain addressed, and that the linear projections used by ViLT and ViT do not inherently restrict the model’s performance. We do so by studying two variants of TomBERT\textsuperscript{†} we call TomV AuLT and TomViLT. TomV AuLT extends TomBERT by replacing the multimodal BERT responsible for the final encoding and prediction with a ViLT. In this way, we replace the linear embeddings of the visual domain of ViLT with ResNet-101’s features (ResNet remains frozen) queried by the target’s embeddings. By disposing of the BERT model encoding the tweet from TomV AuLT, we arrive at the other variant, TomViLT, where only ViLT is used to process the text inputs. We compare these with ViLT and V AuLT (with BERT for fairer comparison) on the development sets of TWITTER-15 and TWITTER-17. In this manner, we present all possible configurations: i ViLT: no deep encoder, ii V AuLT: deep language encoder, iii TomViLT: deep visual encoder, iv TomV AuLT: deep visual and language encoder.

Our results suggest that for the current setting, having a deep visual encoder actually hurts performance in the presence of a deep language encoder. Results are shown in Table 10. Notice that a deep visual encoder performs favorably to no encoder at all (TomViLT vs. ViLT), but, overall, V AuLT has the better performance, and in particular compared to TomV AuLT that has a deep visual encoder in addition to V AuLT’s deep language encoder.

**Conclusion**

In this work, we introduce the Vision-and-Augmented-Language Transformer (V AuLT). V AuLT utilizes a large pretrained language model, such as BERT, to “pre-process” text inputs before propagating these enhanced language representations to ViLT in order to perform multimodal tasks. We find that this is integral, as ViLT’s impoverished language representations, owing to its limited language exposure during pretraining, cannot be easily fine-tuned to perform out-of-distribution tasks, such as affective analysis on multimodal social media data. Our approach obviates the requirements for extensive pretraining of ViLT on the desired domains, as the inclusion of the proper language models, pretrained on the domain of interest, can bridge the distribution shift between source and target domains. However, we also show that ViLT performs on par or even favorably to V AuLT when the tasks involve pure reasoning. Importantly, we demonstrate competitive performance of V AuLT and even ViLT on several tasks when compared to dedicated, state-of-the-art architectures. We also show that joint training of the two models in V AuLT is essential to achieve proper alignment of the language model’s output space with ViLT’s input space. Finally, we study the dynamics between the modality components of ViLT and V AuLT and how performance trade-offs can depend on the complexity and the variety of the modalities in the addressed domains.

Our work opens up several research directions. First, we showed how ViLT can still slightly outperform V AuLT on specific tasks. Anecdotal evidence from experiments in VQA v2 (not presented in this work) further bolster this claim. Extending V AuLT to be strictly superior to ViLT, for example, by improving the alignment of our proposed Transformer stack is a promising direction. Second, we observe rare divergences during training, which again, speak to the need for better alignment mechanisms between the two Transformers. Third, we offer some evidence that shallow visual encoders but deep language encoders can achieve better classification performance to alternatives, a paradigm that could be further investigated in greater detail.

Ethical concerns include the usage of such models for purposes in direct contrast to the applications mentioned previously, such as the suppression of dissent and, thereby, free speech, as well as integration in social manipulation tools.
Acknowledgements

We thank Kristina Lerman for her valuable guidance. This project was funded in part by DARPA under contract HR001121C0168.

References

Abdul-Mageed, M.; and Ungar, L. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers), 718–728.

Alfano, M.; Sullivan, E.; and Fard, A. E. 2022. Ethical pitfalls for natural language processing in psychology. The Atlas of Language Analysis in Psychology.

Alhuzali, H.; and Ananiadou, S. 2021. SpanEmo: Casting Multi-label Emotion Classification as Span-prediction. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 1573–1584. Online: Association for Computational Linguistics.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, 2425–2433.

Bandes, S. A.; and Blumenthal, J. A. 2012. Emotion and the Law. Annual Review of Law and Social Science, 8: 161–181.

Baziotis, C.; Nikolaos, A.; Chronopoulou, A.; Kolovou, A.; Paraskevopoulos, G.; Ellinas, N.; Narayanan, S.; and Potamianos, A. 2018. NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive RNNs and Transfer Learning. In Proceedings of The 12th International Workshop on Semantic Evaluation, 245–255.

Bradshaw, S.; and Howard, P. N. 2018. The global organization of social media disinformation campaigns. Journal of International Affairs, 71(1.5): 23–32.

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Uniter: Learning universal image-text representations. arXiv preprint arXiv:1909.11740.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual dialog. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 326–335.

Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemad, G.; and Ravi, S. 2020. GoEmotions: A dataset of fine-grained emotions. arXiv preprint arXiv:2005.00547.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dukes, D.; Abrams, K.; Adolphs, R.; Ahmed, M. E.; Beatty, A.; Berridge, K. C.; Broomhall, S.; Brosch, T.; Campos, J. J.; Clay, Z.; et al. 2021. The rise of affectivism. Nature human behaviour, 5(7): 816–820.

Garten, J.; Hoover, J.; Johnson, K. M.; Boghrati, R.; Iskivitch, C.; and Dehghani, M. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. Behavior research methods, 50(1): 344–361.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, 6904–6913.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778.

Hoover, J.; Atari, M.; Mostafazadeh Davani, A.; Kennedy, B.; Portillo-Wightman, G.; Yeh, L.; and Dehghani, M. 2021. Investigating the role of group-based morality in extreme behavioral expressions of prejudice. Nature Communications, 12(1): 1–13.

Hoover, J.; Johnson, K.; Boghrati, R.; Graham, J.; and Dehghani, M. 2018. Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation. Collabra: Psychology, 4(1).

Hu, A.; and Flaxman, S. 2018. Multimodal sentiment analysis to explore the structure of emotions. In proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining, 350–358.

Khan, Z.; and Fu, Y. 2021. Exploiting BERT for multimodal target sentiment classification through input space translation. In Proceedings of the 29th ACM International Conference on Multimedia, 3034–3042.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in Neural Information Processing Systems, 33: 2611–2624.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In International Conference on Machine Learning, 5583–5594. PMLR.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 123(1): 32–73.

Liew, J. S. Y.; Turtle, H. R.; and Liddy, E. D. 2016. EmoTweet-28: a fine-grained emotion corpus for sentiment analysis. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), 1149–1156.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft
coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Mathias, L.; Nie, S.; Davani, A. M.; Kiela, D.; Prabhakaran, V.; Vidgen, B.; and Waseem, Z. 2021. Findings of the WOAH 5 Shared Task on Fine Grained Hateful Memes Detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*.

Merchant, A.; Rahimtoroghi, E.; Pavlick, E.; and Tenney, I. 2020. What Happens To BERT Embeddings During Fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 33–44.

Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, 1–17.

Nguyen, D. Q.; Vu, T.; and Nguyen, A. T. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200*.

Niu, T.; Zhu, S.; Pang, L.; and El-Saddik, A. 2016. Sentiment Analysis on Multi-View Social Data. In *MultiMedia Modeling*, 15–27.

Ordonez, V.; Kulkarni, G.; and Berg, T. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001): 2001.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2641–2649.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Sailunaz, K.; and Alhajr, R. 2019. Emotion and sentiment analysis from Twitter text. *Journal of Computational Science*, 36: 101003.

Sharma, C.; Bhageria, D.; Scott, W.; Pykl, S.; Das, A.; Chakraborty, T.; Pulabagari, V.; and Gamback, B. 2020. SemEval-2020 Task 8: Memotion Analysis–The Visuolinguinal Metaphor! *arXiv preprint arXiv:2008.03781*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018a. 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)*, 2556–2565.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018b. 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)*, 2556–2565.

Srinivasan, T.; Chang, T.-Y.; Alva, L. L. P.; Chochlakis, G.; Rostami, M.; and Thomason, J. 2022. CliMB: A Continual Learning Benchmark for Vision-and-Language Tasks. *arXiv preprint arXiv:2206.09059*.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5103–5114.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vempala, A.; and Preot˘iu-Pietro, D. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, 2830–2840.

Wahl-Jorgensen, K. 2019. Emotions, media and politics. John Wiley & Sons.

Wang, J.; Liu, Z.; Sheng, V.; Song, Y.; and Qiu, C. 2021. SaliencyBERT: Recurrent Attention Network for Target-Oriented Multimodal Sentiment Classification. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 3–15. Springer.

Wistrich, A. J.; Rachlinski, J. J.; and Guthrie, C. 2014. Heart versus head: Do judges follow the law or follow their feelings. *Tex. L. Rev.*, 93: 855.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural
Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. Online: Association for Computational Linguistics.

Xu, N.; and Mao, W. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2399–2402.

Yang, X.; Feng, S.; Wang, D.; and Zhang, Y. 2020. Image-text multimodal emotion classification via multi-view attentional network. IEEE Transactions on Multimedia, 23: 4014–4026.

Yu, J.; and Jiang, J. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In IJCAI.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6720–6731.

Zhang, Y.; Zhang, Y.; Guo, W.; Cai, X.; and Yuan, X. 2022. Learning Disentangled Representation for Multimodal Cross-Domain Sentiment Analysis. IEEE Transactions on Neural Networks and Learning Systems.

Zhao, S.; Yao, H.; Gao, Y.; Ding, G.; and Chua, T.-S. 2016. Predicting personalized image emotion perceptions in social networks. IEEE transactions on affective computing, 9(4): 526–540.

Zhu, R. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. arXiv preprint arXiv:2012.08290.

Zhu, T.; Li, L.; Yang, J.; Zhao, S.; Liu, H.; and Qian, J. 2022. Multimodal sentiment analysis with image-text interaction network. IEEE Transactions on Multimedia.