3-2023

Concept-oriented transformers for visual sentiment analysis

Quoc Tuan TRUONG
Singapore Management University, qttruong.2017@phdis.smu.edu.sg

Hady Wirawan LAUW
Singapore Management University, hadywlauw@smu.edu.sg

Concept-Oriented Transformers for Visual Sentiment Analysis

Quoc-Tuan Truong*
Amazon
truquoc@amazon.com

Hady W. Lauw
Singapore Management University
hadywlauw@smu.edu.sg

ABSTRACT

In the richly multimedia Web, detecting sentiment signals expressed in images would support multiple applications, e.g., measuring customer satisfaction from online reviews, analyzing trends and opinions from social media. Given an image, visual sentiment analysis aims at recognizing positive or negative sentiment, and occasionally neutral sentiment as well. A nascent yet promising direction is Transformer-based models applied to image data, whereby Vision Transformer (ViT) establishes remarkable performance on large-scale vision benchmarks. In addition to investigating the fitness of ViT for visual sentiment analysis, we further incorporate concept orientation into the self-attention mechanism, which is the core component of Transformer. The proposed model captures the relationships between image features and specific concepts. We conduct extensive experiments on Visual Sentiment Ontology (VSO) and Yelp.com online review datasets, showing that not only does the proposed model significantly improve upon the base model ViT in detecting visual sentiment but it also outperforms previous visual sentiment analysis models with narrowly-defined orientations. Additional analyses yield insightful results and better understanding of the concept-oriented self-attention mechanism.

CCS CONCEPTS

• **Information systems** → **Web mining**; *Multimedia information systems*; • **Computing methodologies** → *Computer vision.*

KEYWORDS

visual sentiment analysis, concept orientation, transformers

ACM Reference Format:
Quoc-Tuan Truong and Hady W. Lauw. 2023. Concept-Oriented Transformers for Visual Sentiment Analysis. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3539597.3570437

1 INTRODUCTION

Visual imagery is fast becoming the preferred mode of expressing oneself on the Web, as attested to by the ostentatious use of photos on social media such as Instagram and Facebook, the curative collections on Pinterest, as well as the illustrative images within product reviews. While a part of these are objectively descriptive in nature, many are also subjectively expressive, meant to convey a certain sentiment across. For example, a review that negatively describes one's experience with a particular establishment may be embellished with photos that spotlight causative issues (e.g., dirty conditions, faulty mechanisms).

Given the increase in both importance and volume of visual imagery, detecting whether an image expresses positive or negative sentiment has extensive applications. For one, we could extract visual signals regarding customer satisfaction with items (products or services) from online reviews. For another, from abundant social media postings, we could glean either signals of individual preferences for items, or population-wide trending opinions regarding various issues.

Visual sentiment analysis is an emerging and active research area. Early approaches rely on low-level image features combined with classification methods [6, 36, 52]. With the advent of Convolutional Neural Networks (CNN), more recent approaches rely on deep neural networks [8, 9, 20, 33, 50], exploring the utility of various CNN architectures [18, 26, 37, 38].

Pertinently, these previous works presume that an image is associated with a single sentiment. We posit that an image, being a rich form of expression, may exhibit multiple sentiments simultaneously, either synchronously or contrastively, depending on the foci of interest within the said image. For example, a photo of a smiling person in a dirty restaurant may express positivity towards the person and negativity towards the locale. For another example, a well-taken snapshot of an insect may inspire wonders (positive) for an entomologist who studies insects yet may grip an entomophobic with fear (negative). Yet another example is how an otherwise non-descript house (negative) may have beautiful windows (positive). In such cases, whether the sentiment is positive or negative depends on the aspect or concept of interest.

In this work, we address the problem of *conceptual* visual sentiment analysis. Given an image and a concept, we seek to detect positive or negative sentiment. Concept here could capture various notions. For instance, it could mark the time of day (e.g., morning, evening), location (e.g., downtown, suburb, neighborhood), type of place or product (e.g., barber, restaurant, nail salon), or generally keywords associated with images. The idea is that the same image may be construed positively or negatively under different concepts.

Contributions. To our awareness, this problem has not been widely-studied in the literature (see Section 2). Thus, formulating a *generalized* notion of conceptual visual sentiment analysis is one of our contributions. A related notion is the user- or item-orientation in VS-CNN [41], which is a narrowly construed notion of concept that would not generalize to new users or items. Nevertheless, we will compare the proposed approach to a version of VS-CNN that orients the parameters according to concepts for parity.

*Work done prior to Amazon

WSDM '23, February 27-March 3, 2023, Singapore, Singapore

ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3570437

Another contribution is to bring in the Transformer architecture to the realm of visual sentiment analysis. Transformer [45] achieves remarkable efficiency in training on large-scale Natural Language Processing (NLP) datasets. In turn, it is successfully applied to image classification by Vision Transformer or ViT [15]. We investigate the efficacy of ViT on visual sentiment analysis (see Section 3.1).

A third contribution is to propose SentiViT, a *concept-oriented* Transformer that could detect different sentiments on an image, depending on concept. A key principle is to introduce concept orientation into the self-attention mechanism. This gives rise to two variants: *feature fusion* and *attention fusion* (see Section 3.2).

The final contribution is to systematically evaluate the efficacy of the proposed model on a number of real-world visual sentiment analysis datasets (see Section 4). Not only do we evaluate this quantitatively across multiple datasets, but we also bolster the understanding of the models through illustrative case studies that shine some light on the inner workings of the models.

2 RELATED WORK

In this work, we focus squarely on *visual* sentiment analysis, i.e., with images the sole source of features. Other forms of sentiment analysis may consider other types of data. The classical ones include textual sentiment analysis [4, 5, 13, 19, 43] formulated mostly as text classification and occasionally as regression on polarity scores [32]. Another flavor is multi-modal methods that seek to bridge multiple modalities for sentiment analysis [10, 24, 27, 42, 48, 51].

Within visual sentiment analysis, as previously alluded to, our Transformer-based approach is distinguished from those based on low-level image features [6, 36, 52] or deep learning with CNN [8, 9, 20, 33, 41, 50]. Meanwhile, [49] considers yet more external information in the form of friends interactions on social network. Yet, while building on transformers [45], particularly on Vision Transformer (ViT) [15], we introduce novel components to alter the self-attention mechanism so as to reflect the concept orientation.

To encourage research on visual sentiment analysis, there exist several public datasets and resources such as VSO [7] based on Flickr photos and T4SA [44] based on Twitter photos. In this paper, we work with VSO, as well as with other datasets based on Yelp online reviews, which we will also release publicly upon publication of the work. As labeling is labor intensive, one approach to product sentiment labels for images is to derive them the text associated with those images, for which SentiWordNet [2, 16] lexicon is helpful.

There are other formulations that may be incidentally related, but are not visual sentiment analysis per se. Some analyze the image aesthetics [12, 23, 30, 34], interestingness [21], or popularity [17, 25, 31, 40]; each is a distinct notion from sentiment. In turn, [3] seeks to recognize facial expression (a very specific type of image), while [22, 29] looks into affective classification which covers a wide range of human emotions. Finally, a class of works seek to establish connections between low-level image features and emotions with the objective of performing automatic image retrieval [11, 35, 47].

3 CONCEPT-ORIENTED TRANSFORMERS

In this section, we describe the proposed model SentiViT, which is a concept-oriented Transformer for visual sentiment analysis. For ease of reference, we first revisit the base model ViT before discussing two variants for introducing concept orientation into the self-attention mechanism.

3.1 Base Model: ViT

Let $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ be an input image, where H, W, and C are the height, width and the number of channels, respectively. To handle a 2D image, ViT splits the image into a sequence of N patches of size $(P \times P)$ pixels and flattens them, yielding $\mathbf{X} = [\mathbf{x}_p^1; \ldots; \mathbf{x}_p^N]$, where $\mathbf{x}_p^i \in \mathbb{R}^{P^2 C}$ is the i-th patch of the input image and $N = HW/P^2$. Each of the patches are then mapped to D dimensions with a trainable linear projection $\mathbf{E} \in \mathbb{R}^{P^2 C \times D}$. Together with position embeddings $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$, it forms the input sequence:

$$\mathbf{z}_0 = [\mathbf{x}_{class}; \mathbf{x}_p^1 \mathbf{E}; \ldots; \mathbf{x}_p^N \mathbf{E}] + \mathbf{E}_{pos}, \tag{1}$$

where $\mathbf{x}_{class}$ is a learnable embedding prepended to the input sequence as similar to BERT [14].

Suppose that there are L blocks in the Transformer. Each block consists of a multi-head self-attention (MSA) layer and a multi-layer perceptron (MLP). Inputs to those layers are normalized by Layernorm (LN), and residual connections are employed. Mathematically, each block $l \in [1, \ldots, L]$ can be formulated as follows:

$$\mathbf{z}'_l = \text{MSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \tag{2}$$

$$\mathbf{z}_l = \text{MLP}(\text{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l \tag{3}$$

Output from the L^{th} layer of the first token (input $\mathbf{x}_{class}$) is normalized and put through a fully-connected (FC) layer to compute prediction $\mathbf{y}$ of classes as:

$$\mathbf{y} = \text{FC}(\text{LN}(\mathbf{z}_L^0)) \tag{4}$$

Zooming in on a multi-head self-attention (MSA) module, its input sequence $\mathbf{H} \in \mathbb{R}^{(N+1) \times D}$ (after LN) is linearly transformed into different spaces of queries $\mathbf{Q} \in \mathbb{R}^{(N+1) \times D}$, keys $\mathbf{K} \in \mathbb{R}^{(N+1) \times D}$, and values $\mathbf{V} \in \mathbb{R}^{(N+1) \times D}$:

$$\left[\mathbf{Q}; \mathbf{K}; \mathbf{V}\right]^T = \mathbf{H} \left[\mathbf{W}_Q; \mathbf{W}_K; \mathbf{W}_V\right]^T \tag{5}$$

Self-attention, defined as a weighted sum over all values in the sequence, is computed through:

$$\textsc{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \textsc{Softmax}(\mathbf{Q}\mathbf{K}^T / \sqrt{D})\mathbf{V} \tag{6}$$

MSA is simply a process of splitting queries, keys, and values for M times (number of heads) and performing the Attention function in parallel, then projecting their concatenated outputs:

$$\text{MSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \textsc{Concat}(\mathbf{h}_1, \ldots, \mathbf{h}_M)\mathbf{W}, \tag{7}$$

$$\text{where } \mathbf{h}_i = \textsc{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i),\ i \in [1, \ldots, M] \tag{8}$$

3.2 Proposed Model: SentiViT

Our focus is on investigating principled ways to introduce concept orientation into the base ViT model. Arguably, MSA is considered the most consequential component in the Transformer block. And by visualizing the attention scores, [46] shows that different heads in the MSA are capable of behaving differently given the same input sequence, which we hypothesize to be especially salient for concept orientation. Therefore, we propose two approaches to incorporate the notion of concepts into the self-attention mechanism, namely *feature fusion* and *attention fusion*. Without loss of generality, in

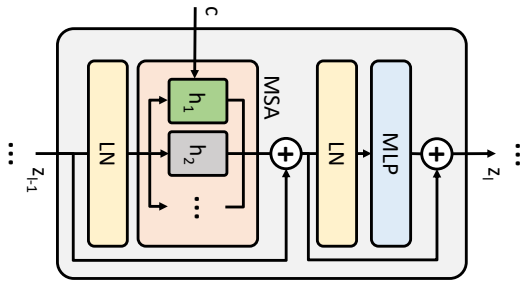


Figure 1: Illustration of a Transformer block with MSA module of one concept-oriented head $\mathbf{h}_1$ (green), and other heads $\mathbf{h}_2, \ldots, \mathbf{h}_M$ (grey) are shared among the concepts.

the following exposition (Figure 1), we take the very first head, subsequently called *concept-oriented head*, of a Transformer block. Beyond one head, the techniques can be applied to multiple heads, as we shall explore in the experiments.

Feature Fusion: SentiViT-F. One hypothesis is that concepts can act on feature-level input to the self-attention layer. Different dimensions in the input features might carry varying visual semantics. In different concepts the model needs to look for important features related to such concepts to better detect the sentiment. For a particular concept c, we model it as a learnable vector $\mathbf{m}_c \in \mathbb{R}^D$. The input sequence $\mathbf{H}$ to the concept-oriented head is transformed via Hadamard product $\odot$ with $\mathbf{m}_c^T$, right before self-attention is performed:

$$\left[\mathbf{Q}_c; \mathbf{K}_c; \mathbf{V}_c\right]^T = \mathbf{H} \odot \mathbf{m}_c^T \left[\mathbf{W}_Q^c; \mathbf{W}_K^c; \mathbf{W}_V^c\right]^T, \quad \mathbf{m}_c^T = \mathbf{c}^T \mathbf{W}_C \quad (9)$$

where $\mathbf{c}^{|C|}$ is one-hot encoded vector of the concept c, $\mathbf{W}_C \in \mathbb{R}^{|C| \times D}$ is a matrix of learnable parameters for all of the concepts, $\mathbf{W}_Q^c \in \mathbb{R}^{D \times d}$, $\mathbf{W}_K^c \in \mathbb{R}^{D \times d}$, and $\mathbf{W}_V^c \in \mathbb{R}^{D \times d}$ are matrices of learnable parameters of the concept-oriented head, with $d = D/M$ being the number of hidden dimensions of one head.

One perspective of this is to view $\mathbf{m}_c$ as a mask of features input to the self-attention. The vector $\mathbf{m}_c$ is essentially a set of scale factors, which are capable of intensifying important features and suppressing less relevant features to the concept c. With this fusion between features and concept, we obtain concept orientation at the input level of the self-attention.

Attention Fusion: SentiViT-A. Another approach is to directly introduce concept orientation into the self-attention mechanism, by learning concept-specific transformations of queries, keys, and values. Intuitively, this is more general and expressive than *feature fusion* technique, as values are transformed features while combinations of queries and keys opt for feature selection in the self-attention process. This fusion of the concept at the attention level can be expressed mathematically as follows:

$$\left[\mathbf{Q}_c; \mathbf{K}_c; \mathbf{V}_c\right]^T = \mathbf{H} \left[\mathbf{c}^T \mathbf{W}_{CQ}; \mathbf{c}^T \mathbf{W}_{CK}; \mathbf{c}^T \mathbf{W}_{CV}\right]^T \quad (10)$$

where $\mathbf{c}^{|C|}$ is the same one-hot encoded vector of the concept c, $\mathbf{W}_{CQ} \in \mathbb{R}^{|C| \times D \times d}$, $\mathbf{W}_{CK} \in \mathbb{R}^{|C| \times D \times d}$, and $\mathbf{W}_{CV} \in \mathbb{R}^{|C| \times D \times d}$ are tensors of learnable parameters for all concepts. With all three transformations learned, we have a full treatment of concept-oriented self-attention.

Algorithm 1 Optimization with Mini-Batch Gradient Descent

Input: image corpus $\mathcal{I}$, set of concepts C, distribution of the concepts π_C
Output: learned model parameters

parameter initialization
repeat
  sample concept $c \sim \pi_C$
  sample images uniformly from $\mathcal{I}_c \rightarrow$ mini-batch $\mathcal{B}$
  calculate loss based on the mini-batch $\mathcal{B}$
  take a gradient descent step
  update the shared parameters
  update the concept-oriented parameters
until converged

3.3 Implementation Details

As our focus is on testing the effectiveness of concept orientation, all of our experiments are conducted with the Base configuration of ViT (12 blocks, 12 heads), which takes in input images of size 224×224 and splits into patches of size 16×16. It would then form input sequence of 197 flatted tokens (1 class token and 196 visual patches). Parameters are initialized with the weights pre-trained on *imagenet21k* and fine-tuned on *imagenet2012*, except for the last linear layer of sentiment classification. It is worth mentioning that the multiplication between one-hot encoded vector with matrix/tensor in SentiViT-F (Eqn. 9) and SentiViT-A (Eqn. 10) can be implemented efficiently with *embedding lookup* and *reshape* operators in any deep learning framework, e.g., PyTorch, TensorFlow, JAX.

For training the models, we use SGD with momentum of 0.9 and train for 20 epochs (almost always converge after 10-15 epochs for our datasets) with batch size of 16. The learning rate starts at 0.0 and is linearly increased to 0.01 in the first 2 epochs of training, after which it is decreased to 0.0 in the subsequent epochs using a cosine decay schedule [28]. For each mini-batch, we first sample a concept with the probability proportional to its number of images, then images belonging to that concept will be sampled uniformly to construct the mini-batch. This sampling procedure helps to ensure all images of a mini-batch coming from the same concept, and also all of the images having approximately the same chance of being observed during training. The former is required to develop an efficient implementation for stable optimization, while the latter is important to avoid sampling biases. Alg. 1 sketches the optimization procedure based on the mini-batch gradient descent algorithm.

4 EXPERIMENTS

The objectives are to investigate several research questions on the effectiveness of SentiViT for visual sentiment analysis. First, we consider how modeling of SentiViT with concept orientation can further improve the task of visual sentiment analysis as compared to multiple baselines. Second, we analyze the contributions of putting concept orientation on three main components, namely query (Q), key (K) and value (V), of the self-attention mechanism. In addition, we study the effects of which layer to place, as well as incremental number of self-attention heads used for concept orientation. Last but not least, we look into case study to get a better understanding of the models.

4.1 Experimental Setup

Here we describe the setup of the experiments, including the datasets as well as the metrics used for model evaluation.

Datasets. There are four datasets derived from two sources: *Visual Sentiment Ontology (VSO)* [7] and online reviews from *Yelp.com*.

VSO dataset is constructed by initial adjective-noun pairs (ANPs), e.g., *delicious drink* or *angry face*, associated with sentiment scores from -2.0 (most negative) to 2.0 (most positive). Images are retrieved from Flickr when using these ANPs as queries. First, we define the nouns (e.g., drink, face) to be the concepts. Images belonging to the same noun are merged together. Sentiment is binarized based on the sign of the scores. Second, to remove potential biases, we balance the number of images between two sentiments within each concept via uniform sampling. For each concept, 80% of the images are used for training while 20% are used for testing. We randomly split the concepts into 5 folds for a more holistic view of the model evaluation. Statistics of the dataset after being processed are shown in Table 1. On average, we have about 1333 images per concept.

Yelp data consists of images found within reviews of businesses in 5 US cities: Boston (BO), Chicago (CH), Los Angeles (LA), New York (NY), and San Francisco (SF). Each review has a rating from 1 to 5, together with one or multiple images. Ratings 1 and 2 are considered negative, 3 neutral, while 4 and 5 positive. Here, we concentrate on discriminating between positive and negative only. All images in a review are assigned the same sentiment label. From this Yelp data, we construct three different datasets:

- *Yelp-User* defines each user as a concept, i.e., images belonging to the same user are considered the same concept.
- *Yelp-Business* defines each business as a concept.
- *Yelp-Category* defines each category as a concept. In this case, images from a particular business will be assigned to the category which that business belongs to. If a business belongs to more than one category, we create a new category as an alphabetically-sorted concatenation of those categories.

Statistics of these datasets are shown in Table 1. On average, we have more images per concept for the *Yelp-Category* than *Yelp-User* and *Yelp-Business*, as it has a more general definition. We follow the same procedure as described for the VSO dataset, i.e., 80% for training and 20% for evaluation for each concept.

Metrics. Each model outputs the probability of an image being positive sentiment. This probability can also be viewed as sentiment score while comparing two images.

The first metric is classification accuracy. For a test image i, the model outputs probability $p_i \in [0, 1]$ of being positive. The predicted class $\hat{y}_i = 1$ (positive) iff $p_i > 0.5$, and $\hat{y}_i = 0$ (negative) otherwise. This metric evaluates the number of correct predictions over the total number of test instances, defined as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i}^{N} \mathbb{1}(\hat{y}_i = y_i)$$

where N is the number of instances, $\hat{y}_i \in \{0, 1\}$ is the predicted label of image i, $y_i \in \{0, 1\}$ is the corresponding ground-truth label, and $\mathbb{1}(.)$ is the indicator function.

The second metric is AUC (area under the ROC curve), which allows us to test the ability of a model to assign higher probabilities for true positives than for true negatives, within the same concept.

Table 1: Dataset statistics.

VSO	Fold1	Fold2	Fold3	Fold4	Fold5	Total
#images	38,726	25,164	32,560	15,432	37,448	149,330
#concepts	26	23	18	18	27	112
img/concept	1,490	1,094	1,809	857	1,387	1,333
Yelp-User	BO	CH	LA	NY	SF	Total
#images	11,676	8,890	135,726	65,878	21,510	243,680
#concepts	243	164	2,176	1,258	377	4,218
img/concept	48	54	62	52	57	58
Yelp-Business	BO	CH	LA	NY	SF	Total
#images	15,686	14,034	228,302	93,854	47,902	399,778
#concepts	438	347	4,726	2,259	917	8,687
img/concept	36	40	48	42	52	46
Yelp-Category	BO	CH	LA	NY	SF	Total
#images	18,684	16,928	267,352	109,876	41,924	454,764
#concepts	69	67	675	344	131	1,286
img/concept	271	253	396	319	320	354

For a concept c, the metric is defined:

$$\text{AUC}(c) = \frac{1}{|\mathcal{I}_c^+||\mathcal{I}_c^-|}\sum_{i\in\mathcal{I}_c^+}\sum_{j\in\mathcal{I}_c^-}\mathbb{1}(p_i > p_j)$$

where $\mathcal{I}_c^+$ and $\mathcal{I}_c^-$ are the sets of positive and negative images of concept c, p_i and p_j are the corresponding probabilities of those images. The average AUC is computed across all concepts C as: $\text{AUC} = \frac{1}{|C|}\sum_{c\in C}\text{AUC}(c)$.

All metrics reported in our experiments are the averaged numbers across 10 independent runs owning different random seeds. Thus, we also indicate the standard deviations for all the average numbers reported.

4.2 Quantitative Analyses

We organize discussions along four research questions.

RQ#1: How accurately could SentiViT models detect visual sentiment? We now see the sentiment classification results on the datasets (Table 2). With balanced datasets, the accuracy would be 0.5 for a Random classifier. That also holds for AUC as we have equal numbers of positive and negative images per concept.

As reference baselines, we include CNN models ResNet-152 [18] and EfficientNet-B7 [39], which have shown competitive performances on various vision benchmarks. While EfficientNet-B7 could detect sentiment better than ResNet-152, there is a noticeable gap between EfficientNet-B7 and ViT accross all datasets, both in terms of Accuracy and AUC. The gains for ViT can be attributed to its superior and more efficient architecture. This observation reinforces the effectiveness of ViT for image classification in general.

VS-CNN [41] is a CNN model developed with the notion of user- and item-orientation for visual sentiment analysis on review images. For parity, we adapt VS-CNN to orient the parameters according to concepts (more broadly construed to give it the benefit of doubt). We refer to the VS-CNN version that puts orientation in the last layer (FC7), which is shown to achieve the best performance in the original paper. For *Yelp-User* (Table 2b) and *Yelp-Business* (Table 2c) datasets, VS-CNN with orientation can detect visual sentiment better than non-conceptualized ViT. It is somehow expected as

Table 2: Performance comparison between SentiViT and comparative baselines (RQ#1).

		Method	Fold1	Fold2	Fold3	Fold4	Fold5	Avg.
(a) VSO	Accuracy	ResNet-152	60.2±0.3	64.2±0.2	59.9±0.3	60.3±0.3	62.4±0.1	61.4±0.1
		EffiNet-B7	62.5±0.3	65.3±0.5	61.6±0.6	64.2±0.6	63.4±0.4	63.2±0.4
		ViT	64.0±0.1	67.4±0.3	62.7±0.2	64.6±0.2	66.3±0.2	64.9±0.1
		VS-CNN	63.7±0.3	66.6±0.7	61.0±0.5	65.8±0.5	65.4±0.4	64.3±0.3
		SentiViT-F	68.3±0.2	74.5±0.5	66.8±0.3	70.9±0.3	72.2±0.3	70.3±0.1
		SentiViT-A	**68.6**±0.3	**74.7**±0.4	**66.9**±0.3	**71.3**±0.5	**72.4**±0.5	**70.5**±0.3
	AUC	ResNet-152	62.5±0.9	64.7±0.7	64.6±0.5	65.4±0.5	64.4±0.6	64.2±0.4
		EffiNet-B7	67.6±1.2	64.2±0.8	67.3±1.0	68.7±0.5	64.7±0.9	66.3±0.5
		ViT	66.5±0.5	65.0±0.8	68.6±0.5	66.1±0.4	69.3±0.5	67.1±0.3
		VS-CNN	72.2±0.8	75.2±0.7	69.6±0.5	74.9±0.9	70.6±0.4	72.4±0.4
		SentiViT-F	76.4±0.6	**84.8**±0.5	**79.8**±0.5	80.5±1.3	78.9±1.1	79.9±0.6
		SentiViT-A	**76.8**±0.6	84.1±0.7	79.5±0.9	**82.2**±0.8	**79.9**±0.5	**80.4**±0.4

		Method	BO	CH	LA	NY	SF	Avg.
(b) Yelp-User	Accuracy	ResNet-152	54.1±0.1	55.0±0.1	53.0±0.3	53.5±0.2	55.2±0.4	53.5±0.2
		EffiNet-B7	56.3±0.4	56.7±0.7	55.1±0.9	55.3±0.5	57.8±0.7	55.5±0.4
		ViT	57.2±0.6	57.9±0.6	56.5±0.2	57.3±0.1	56.5±0.5	56.8±0.1
		VS-CNN	62.2±0.8	63.5±0.7	63.0±0.2	62.1±0.4	65.4±0.3	63.0±0.3
		SentiViT-F	67.5±1.0	68.0±0.7	66.8±0.5	66.8±0.4	69.5±0.5	67.1±0.4
		SentiViT-A	**69.0**±0.5	**70.5**±0.2	**68.0**±0.1	**68.3**±0.3	**72.0**±0.6	**68.6**±0.2
	AUC	ResNet-152	55.1±0.6	60.4±1.0	54.6±0.8	55.2±0.6	56.7±1.1	55.2±0.6
		EffiNet-B7	60.3±0.7	56.5±0.8	56.5±0.7	56.9±0.6	58.1±0.5	57.0±0.4
		ViT	60.2±0.7	61.8±0.7	60.0±0.2	59.9±0.2	58.1±0.2	59.9±0.2
		VS-CNN	68.0±1.6	67.8±1.2	68.4±0.8	67.1±0.9	70.4±0.9	68.1±0.9
		SentiViT-F	75.0±1.4	75.0±0.8	74.7±0.6	73.9±0.5	76.7±0.7	74.7±0.5
		SentiViT-A	**76.0**±0.7	**76.3**±0.6	**75.8**±0.2	**75.0**±0.2	**78.2**±0.8	**75.8**±0.2

		Method	BO	CH	LA	NY	SF	Avg.
(c) Yelp-Business	Accuracy	ResNet-152	56.1±0.2	55.5±0.3	57.2±0.3	55.6±0.2	55.9±0.5	56.6±0.1
		EffiNet-B7	56.5±0.8	56.9±0.7	59.0±0.3	57.6±0.3	57.5±0.4	58.3±0.3
		ViT	56.7±0.4	58.6±0.7	60.0±0.1	58.0±0.1	57.9±0.3	59.1±0.1
		VS-CNN	61.5±0.7	62.6±0.4	61.7±0.2	60.8±0.1	61.3±0.2	61.5±0.2
		SentiViT-F	62.8±0.9	65.3±0.7	64.0±0.2	62.9±0.3	62.7±0.4	63.6±0.3
		SentiViT-A	**64.6**±0.9	**66.7**±0.5	**65.7**±0.2	**64.7**±0.5	**64.5**±0.7	**65.3**±0.3
	AUC	ResNet-152	58.2±0.8	58.9±0.4	61.8±0.3	59.8±0.4	59.9±0.7	60.8±0.2
		EffiNet-B7	59.2±0.5	61.7±0.9	64.4±0.9	61.6±0.8	61.6±0.8	63.0±0.5
		ViT	60.6±0.4	62.8±0.2	66.1±0.1	62.4±0.2	62.3±0.2	64.3±0.1
		VS-CNN	66.0±0.3	68.1±0.9	67.5±0.2	65.9±0.2	66.6±0.5	67.0±0.2
		SentiViT-F	69.2±0.8	70.8±0.9	70.5±0.4	68.0±0.5	68.8±0.5	69.6±0.4
		SentiViT-A	**70.1**±0.9	**73.4**±0.8	**72.3**±0.3	**70.7**±0.7	**70.7**±0.8	**71.6**±0.3

		Method	BO	CH	LA	NY	SF	Avg.
(d) Yelp-Category	Accuracy	ResNet-152	59.8±0.2	61.2±0.2	59.4±0.2	59.0±0.1	57.6±0.2	59.2±0.1
		EffiNet-B7	61.4±0.5	61.2±0.4	61.9±0.2	61.3±0.5	61.4±0.6	61.7±0.3
		ViT	66.0±0.3	64.6±0.2	63.4±0.1	62.8±0.2	60.2±0.1	63.1±0.1
		VS-CNN	61.8±0.3	62.2±0.2	59.2±0.2	59.8±0.3	59.3±0.2	59.6±0.2
		SentiViT-F	65.8±0.5	66.1±0.4	64.6±0.1	64.5±0.2	62.7±0.5	64.5±0.1
		SentiViT-A	**66.9**±0.6	**66.9**±0.4	**65.3**±0.1	**65.5**±0.3	**64.2**±0.4	**65.4**±0.2
	AUC	ResNet-152	63.1±0.4	63.3±0.3	61.8±0.2	61.1±0.3	58.1±0.4	61.4±0.2
		EffiNet-B7	63.8±0.4	64.1±0.6	64.8±0.6	64.9±0.8	64.1±0.5	64.7±0.3
		ViT	70.7±0.4	67.2±0.4	66.6±0.1	65.7±0.1	62.7±0.4	66.2±0.1
		VS-CNN	66.5±0.4	66.7±0.7	64.1±0.2	64.2±0.2	63.5±0.4	64.3±0.2
		SentiViT-F	71.5±0.7	71.3±0.6	68.9±0.1	69.1±0.1	66.5±0.4	69.0±0.1
		SentiViT-A	**73.1**±0.6	**72.3**±0.2	**70.1**±0.2	**70.5**±0.3	**68.3**±0.7	**70.3**±0.2

these are two specific concepts (users and items/businesses) that the VS-CNN model has been designed for. However, for more general concepts, *VSO* (Table 2a) and *Yelp-Category* (Table 2d), VS-CNN is not as accurate as ViT, though the former still achieves better sentiment ranking within concept (AUC) on VSO.

Both SentiViT-F and SentiViT-A models perform better than VS-CNN. Importantly, they significantly outperform the base model ViT. These results highlight the success of our modeling based on the advancement of Transformer married with concept orientation for visual sentiment analysis. Between the two models, on average SentiViT-A is consistently better than SentiViT-F across all datasets. It is expected as we have discussed earlier that *attention fusion* is more general and expressive than *feature fusion* due to the direct modeling of concept-oriented self-attention.

Table 3: The impacts of the self-attention components with concept orientation (RQ#2).

		Method	Fold1	Fold2	Fold3	Fold4	Fold5	Avg.
(a) VSO	Accuracy	SentiViT-Q	68.0±0.4	73.9±0.4	66.1±0.3	**71.6**±0.9	71.8±0.5	69.9±0.2
		SentiViT-K	68.1±0.1	73.7±0.6	66.0±0.6	71.3±0.6	73.0±0.5	70.1±0.2
		SentiViT-V	**68.9**±0.4	74.5±0.3	**67.0**±0.2	**71.6**±0.7	**73.0**±0.2	**70.7**±0.1
		SentiViT-A	68.6±0.3	**74.7**±0.4	66.9±0.3	71.3±0.5	72.4±0.5	70.5±0.3
	AUC	SentiViT-Q	74.1±0.2	82.4±0.8	76.9±0.7	80.9±1.2	78.2±0.9	78.3±0.5
		SentiViT-K	75.5±0.7	83.6±0.7	78.3±0.5	80.8±0.7	78.6±0.6	79.2±0.3
		SentiViT-V	76.6±0.4	83.4±0.6	**79.8**±0.2	81.4±0.8	80.1±0.5	80.1±0.3
		SentiViT-A	**76.8**±0.6	**84.1**±0.7	79.5±0.9	**82.2**±0.8	**79.9**±0.5	**80.4**±0.4

		Method	BO	CH	LA	NY	SF	Avg.
(b) Yelp-User	Accuracy	SentiViT-Q	67.5±0.5	70.3±0.4	67.7±0.2	68.0±0.1	71.6±0.4	68.2±0.1
		SentiViT-K	68.0±0.8	69.9±0.6	67.3±0.2	67.4±0.3	70.8±0.9	67.8±0.2
		SentiViT-V	66.9±0.6	67.8±0.9	66.4±0.2	66.4±0.2	69.0±1.0	66.7±0.3
		SentiViT-A	**69.0**±0.5	**70.5**±0.2	**68.0**±0.1	**68.3**±0.3	**72.0**±0.6	**68.6**±0.2
	AUC	SentiViT-Q	74.6±0.6	76.1±1.0	75.0±0.4	74.6±0.4	76.9±0.7	75.1±0.2
		SentiViT-K	74.8±0.7	73.8±1.0	74.3±0.4	73.8±0.5	76.4±0.5	74.4±0.4
		SentiViT-V	74.2±0.9	74.9±0.7	74.6±0.3	73.9±0.3	76.1±0.8	74.5±0.3
		SentiViT-A	**76.0**±0.7	**76.3**±0.6	**75.8**±0.2	**75.0**±0.2	**78.2**±0.8	**75.8**±0.2

		Method	BO	CH	LA	NY	SF	Avg
(c) Yelp-Business	Accuracy	SentiViT-Q	**64.6**±0.7	**67.3**±0.3	65.6±0.2	**65.1**±0.3	**64.7**±0.6	**65.4**±0.2
		SentiViT-K	63.4±0.5	65.4±0.9	64.5±0.2	63.4±0.3	63.1±0.7	64.1±0.2
		SentiViT-V	62.5±0.4	64.6±0.7	63.5±0.2	62.8±0.1	62.0±0.2	63.2±0.1
		SentiViT-A	**64.6**±0.9	66.7±0.5	**65.7**±0.2	64.7±0.5	64.5±0.7	65.3±0.3
	AUC	SentiViT-Q	**71.0**±0.5	**73.4**±0.5	**72.3**±0.2	**70.7**±0.5	71.2±0.6	**71.7**±0.2
		SentiViT-K	68.7±0.9	70.7±0.9	71.2±0.2	68.9±0.5	68.9±1.0	70.2±0.3
		SentiViT-V	68.3±0.8	70.7±0.4	70.4±0.2	68.2±0.1	68.0±0.4	69.5±0.1
		SentiViT-A	70.1±0.9	**73.4**±0.8	**72.3**±0.3	**70.7**±0.7	70.7±0.8	71.6±0.3

		Method	BO	CH	LA	NY	SF	Avg.
(d) Yelp-Category	Accuracy	SentiViT-Q	66.8±0.5	66.6±1.0	65.0±0.2	65.1±0.4	63.2±0.4	65.0±0.2
		SentiViT-K	65.8±0.4	66.3±0.7	64.8±0.2	65.3±0.2	62.8±0.4	64.9±0.2
		SentiViT-V	66.2±0.3	66.1±0.5	64.7±0.2	64.6±0.1	62.8±0.6	64.6±0.1
		SentiViT-A	**66.9**±0.6	**66.9**±0.4	**65.3**±0.1	**65.5**±0.3	**64.2**±0.4	**65.4**±0.2
	AUC	SentiViT-Q	72.9±1.0	71.9±0.5	**70.3**±0.2	70.4±0.3	67.6±0.6	**70.3**±0.2
		SentiViT-K	72.1±0.3	71.5±0.5	69.9±0.1	70.2±0.5	66.5±0.7	69.9±0.2
		SentiViT-V	72.4±0.6	72.3±0.2	69.6±0.2	69.9±0.2	67.1±0.5	69.7±0.1
		SentiViT-A	**73.1**±0.6	**72.3**±0.2	70.1±0.2	**70.5**±0.3	**68.3**±0.7	**70.3**±0.2

RQ#2: What are the impacts of self-attention query (Q), key (K), and value (V) on concept orientation? Having established the effectiveness of *attention fusion* for the task, we now turn to its components. As there are three main parts (Q, K, and V), we would like to understand the contribution of each part towards the full orientation (which will be referred to as just A for short). Table 3 reports the experiments when concept orientation is introduced to only a specific part, as well as the full model. On *VSO* dataset, we observe that V achieves the best performance in terms of classification accuracy, while Q is significantly better on the other three datasets. K somehow finds a balance in the middle among the three components. Overall, the full orientation A is always better or comparable with any of the ablation Q, K, or V,

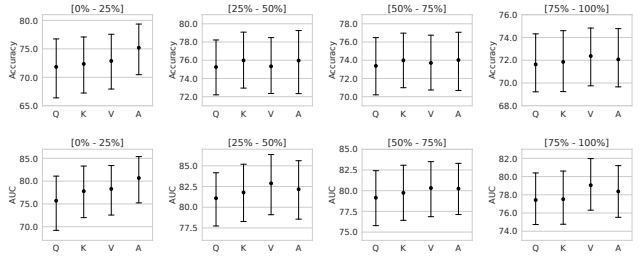


Figure 2: Performance of SentiViT breakdown for concepts with increasing number of images (RQ#2).

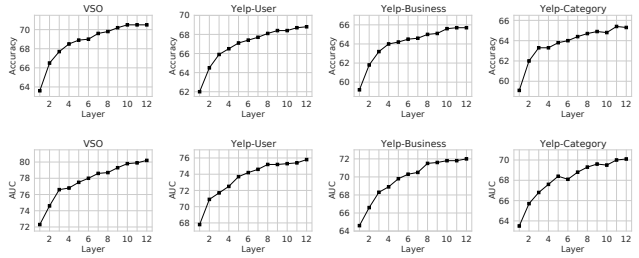


Figure 3: Performance of SentiViT w.r.t. the layer where concept orientation is placed (RQ#3).

especially when it comes to AUC measure in which the ability of contrasting sentiment within concept matters more.

One interesting observation is that neither of Q, K, or V is universally dominant in our experiments. Hypothetically, this phenomenon might be related to the number of images per concept, as on average *VSO* dataset has higher image/concept than the others. In Figure 2, we break down the concepts on *VSO* dataset into quartiles with increasing number of images. It is not clear that Q would perform better than V with low image/concept as V is consistently better regardless of quartile. To better understand this phenomenon, it would require further investigation on the behaviors of Q, K, and V in the self-attention mechanism, we leave it for future work. Importantly, a standing observation from this analysis is that the full orientation A is always the best (or the second best). This agreement with results on other datasets suggests that the full orientation A is a robust choice. It is able to adapt to different scenarios of the data due to flexible control over all spaces of Q, K, and V of the self-attention for concept orientation.

RQ#3: Where should concept orientation be placed in the layer hierarchy of the Transformer? For SentiViT models, concept orientation can be placed arbitrarily in any layer of the Transformer, as illustrated in Figure 1. We would like to examine whether there are differences among the choices of layers, and if the answer is yes, which choice is the most effective one. In our base architecture, there are in total 12 layers, all with the same composition of transformation. We systematically place the concept orientation in each of the layer in the hierarchy, and provide a visualization of the results in Figure 3. At a first glance, a consistent observation is that model performances are increasing while going from lower layers (close to the input space) to higher layers (close to the output space). The performances are boosted sharply for the first few layers, especially in terms of Accuracy as depicted in the first row, and peaked around the layer 12. From the first layer to the last layer, there are notable gaps of improvements. These results provide supporting evidence to two points. First, there are indeed significant impacts in choosing which layer in the Transformer to place the concept orientation. Second, the concepts being modeled are high-level of semantics, therefore the orientation should be done at higher levels of feature abstraction. With the last layer appearing to be the most promising choice to place the orientation, the experimental results reported are conducted under this setting, unless mentioned otherwise.

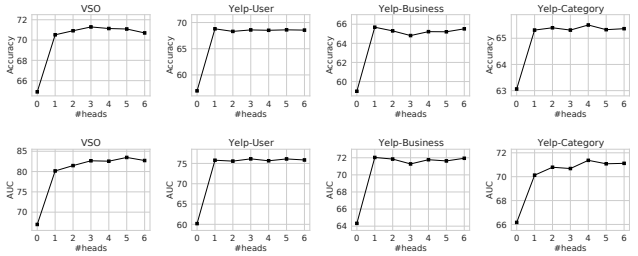


Figure 4: Performance of SentiViT w.r.t. the number of heads used for concept orientation (RQ#4).

RQ#4: Is there benefit in going beyond single-head self-attention for concept orientation? This is a natural question to ask when we observe notable improvements of SentiViT-A over ViT with only one concept-oriented head. We gradually increase the number of self-attention heads used for orientation, and the results are illustrated in Figure 4. On *VSO* dataset, we observe improvements going from one to three heads for classification accuracy, and up until five heads for AUC, before plateauing or slightly decreasing. Similar trends are observed on the *Yelp-Category* dataset, but with smaller magnitudes. On the other two datasets of *Yelp-User* and *Yelp-Business*, the performances are not improving (even slightly decreasing for the latter) while adding number of concept-oriented heads. This is understandable as we have more examples per concept on *VSO* and *Yelp-Category* as compared to *Yelp-User* and *Yelp-Business*. Thus, the model requires higher degree of freedom (more heads) to capture the concepts in the former datasets, while one head is enough for the latter datasets.

4.3 Case Study

Visualizing Attention. To gain insights on how the concept-oriented head helps SentiViT to detect sentiments better than the base model ViT, in Figure 5 we visualize the attention in the input space. These examples from the VSO dataset are correctly classified by SentiViT-A but misclassified by ViT. To compute maps of the attention from the output token to the input space, we use Attention Rollout [1]. For each example, the four images from left to right correspond to the input image, attention map of ViT, attention maps of the shared heads and concept-oriented head of SentiViT-A.

In the first example where the concept is *animals*, an image of a spider with negative label is misclassified as positive by ViT. Visualizing the attention, we see that ViT focuses not only on the spider but also on the bright spot in the input image. This could be a factor in ViT's positive prediction as positive images tend to be brighter than negative ones. For SentiViT, with the concept being *animals*, the concept-oriented head only pays attention to the spider and ignores the bright spot (the last visualized image). With the

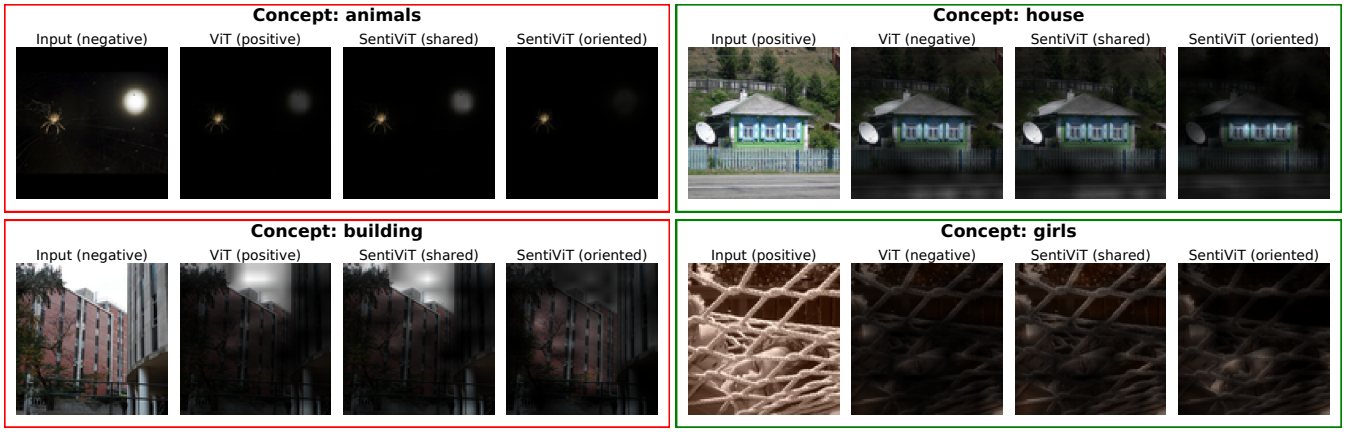


Figure 5: Visualization of attention from the output token to the input space.

relevant signals provided by the concept-oriented head, SentiViT correctly detects the true negative sentiment.

In the second example of another negative image (with *building* concept), a similar observation emerges where ViT almost completely ignores the building and only pays attention to the bright sky. In contrast, the concept-oriented head of SentiViT only focuses on the downgraded building, which we hypothesize might be a clue for SentiViT to predict the negative sentiment correctly. In both cases, the shared heads in SentiViT have similar attention maps as compared to ViT, which means that they might just carry the same information leading to the misclassifications. This observation reinforces the important contribution of the concept-oriented head to predicting the correct sentiments.

On the right hand side of Figure 5 are two examples of positive images, which ViT misclassifies as negative. The top example is of *house* concept, and what is striking in the image are the colorful windows. That is the focus area of SentiViT concept-oriented head, as compared to ViT's more widely-spread attention. Therefore, SentiViT correctly classifies the image as positive. The second example underneath is a difficult case for ViT. Visibly, the attention map is blackout with only a few spots in the corners. It means that ViT could not detect useful information. With the concept being *girls*, SentiViT focuses on smiling face of the baby, which could be a clue to the image's positive sentiment.

Sentiment Reversal Due to Concepts. Another attempt to gain insights is to look at images generally considered to be positive (resp. negative), and then see if other visually similar images would express the reverse sentiment, i.e., negative (resp. positive) if they appear in certain concepts.

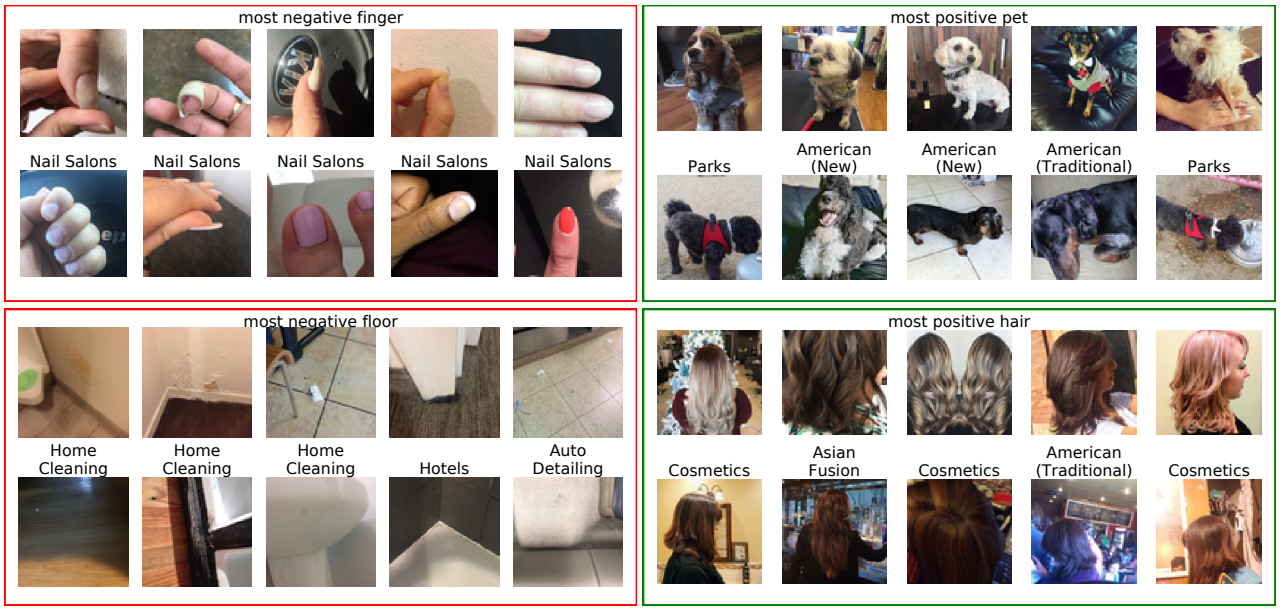
Figure 6a shows four clusters of images on the VSO dataset. In each cluster (bounded by a box), the first rows contain images among those assigned highest probabilities of being negative (resp. positive) by ViT. For the second rows, they are training images from concepts that *reverse* the sentiments of all images in the corresponding first rows, but are *visually similar* to them (in terms of cosine similarity of the output embedding by ViT model). In the first cluster, the first row depicts *flower* images considered the most negative by ViT. They seem to be dried/dying flowers with yellowish-brown colors, which may be considered negative. However, in some other concepts, *evening*, *leaves*, images of similar patterns might well be positive, as they may reflect how an attractive flower may appear in the evening or when the focus is on the leaves rather than on the flower itself. We can observe a similar phenomenon in the second cluster of negative *lake* images. Images of *frozen lakes* are considered negative, though they share a lot of similarities with positive images of *winter*, due to the colors and patterns of ice and snow. On the right hand side with the most positive images, in the top cluster, we have the first row of architectural and colorful buildings. Their counterparts of negative images appear in *crowd* concept as images of crowds usually have buildings as backgrounds. Similarly, the last cluster shows positive images of *shoes* in the first row, and the second row covers concepts of *star*, *island* and *hands*, with some unusual looking fingers or flashy objects.

In turn, Figure 6b shows clusters of images obtained from *Yelp-Category* dataset. On the left hand side, the first cluster shows negative images containing patterns of fingers. It is not surprising that the concept would reverse the sentiment of those *finger* images is the concept of *Nail Salons*, as there are many positive review images about nail salons with such patterns. For the second cluster underneath, it depicts negative images of dirty *floors* possibly in some restaurants, while the same kind of images shown in the second row with clean floors would be positive in *Home Cleaning* concept expressing satisfaction of customers with the cleaning services. Looking at the first cluster on the right hand side, it carries positive images of cute dogs. Those images came from positive reviews of *pet services* (e.g., veterinarian). Although, the second row shows similar images of dogs, they are the negative ones as being parts of complaining reviews in some other concepts, i.e., complaints of dirty/noisy 'pets' being in some *restaurants* or public *parks*. In the bottom-right cluster, the first row presents positive images coming from reviews of *hair* salons. They express satisfaction of users with beautiful hair shapes and colors. Meanwhile, the same kind of images could be complaints about some *cosmetic* products, shown in the second row underneath.

5 CONCLUSION

In this work, we address the problem of *conceptual* visual sentiment analysis, formulating the notion of concept being generalizable to

(a) Examples from VSO dataset.



(b) Examples from Yelp-Category dataset.

Figure 6: Most negative/positive images (by ViT) and images of the concepts that reverse their sentiments (by SentiViT).

various scenarios of images from the Web. Our contributions include systematic investigation of the efficacy of Vision Transformer or ViT for the task, and two proposed formulations for incorporating concept orientation, yielding SentiViT. We conduct experiments on different Web image datasets showing that the proposed SentiViT models perform better than ViT as well as other existing methods for visual sentiment analysis. Ablation study reveals that the *attention fusion* technique is robust to different image population sizes of the concepts. We also provide qualitative analyses with case studies yielding insights on the model predictions.

Future work includes investigating behaviors of the self-attention components influencing sentiment prediction of the SentiViT models. Understanding of such behaviors would lead to more efficient and effective approaches for doing concept orientation, as well as new designs of model architectures for visual sentiment analysis.

ACKNOWLEDGMENTS

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-020).

REFERENCES

[1] Samira Abnar and Willem H. Zuidema. 2020. Quantifying Attention Flow in Transformers. In *ACL*. Association for Computational Linguistics, 4190–4197.

[2] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC*. European Language Resources Association.

[3] Marian Stewart Bartlett, Gwen Littlewort, Mark G. Frank, Claudia Lainscsek, Ian R. Fasel, and Javier R. Movellan. 2005. Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. In *CVPR*. 568–573.

[4] Adam Bermingham and Alan F Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage?. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 1833–1836.

[5] Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In *ICWSM*.

[6] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. 2013. SentiBank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Multimedia*. ACM, 459–460.

[7] Damian Borth, Rongrong Ji, Tao Chen, Thomas M. Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Multimedia*. ACM, 223–232.

[8] Victor Campos, Brendan Jou, and Xavier Giró-i-Nieto. 2017. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. *Image Vis. Comput.* 65 (2017), 15–22.

[9] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks. *CoRR* abs/1410.8586 (2014). arXiv:1410.8586

[10] Yan-Ying Chen, Tao Chen, Winston H. Hsu, Hong-Yuan Mark Liao, and Shih-Fu Chang. 2014. Predicting Viewer Affective Comments Based on Image Content in Social Media. In *ICMR*. ACM, 233.

[11] Carlo Colombo, Alberto Del Bimbo, and Pietro Pala. 1999. Semantics in Visual Information Retrieval. *IEEE Multim.* 6, 3 (1999), 38–53.

[12] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Ze Wang. 2006. Studying Aesthetics in Photographic Images Using a Computational Approach. In *ECCV (Lecture Notes in Computer Science, Vol. 3953)*. Springer, 288–301.

[13] Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In *COLING, Posters Volume*. Chinese Information Processing Society of China, 241–249.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. Association for Computational Linguistics, 4171–4186.

[15] 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. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[16] Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *LREC*. European Language Resources Association (ELRA), 417–422.

[17] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. 2015. Image Popularity Prediction in Social Media Using Sentiment and Context Features. In *Multimedia*. ACM, 907–910.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE Computer Society, 770–778.

[19] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. 2013. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*. 607–618.

[20] Jyoti Islam and Yanqing Zhang. 2016. Visual Sentiment Analysis for Social Images Using Transfer Learning Approach. In *(BDCloud), (SocialCom), (SustainCom)*. IEEE Computer Society, 124–130.

[21] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable?. In *CVPR*. IEEE Computer Society, 145–152.

[22] Jia Jia, Sen Wu, Xiaohui Wang, Peiyun Hu, Lianhong Cai, and Jie Tang. 2012. Can we understand van gogh's mood?: learning to infer affects from images in social networks. In *Multimedia*. ACM, 857–860.

[23] Dhiraj Joshi, Ritendra Datta, Elena A. Fedorovskaya, Quang-Tuan Luong, James Ze Wang, Jia Li, and Jiebo Luo. 2011. Aesthetics and Emotions in Images. *IEEE Signal Process. Mag.* 28, 5 (2011), 94–115.

[24] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. 2015. Visual Affect Around the World: A Large-scale Multilingual Visual Sentiment Ontology. In *Multimedia*. ACM, 159–168.

[25] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. What makes an image popular?. In *Proceedings of the 23rd international conference on World wide web*. 867–876.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (Eds.). 1106–1114.

[27] Zuhe Li, Yangyu Fan, Weihua Liu, and Fengqin Wang. 2018. Image sentiment prediction based on textual descriptions with adjective noun pairs. *Multim. Tools Appl.* 77, 1 (2018), 1115–1132.

[28] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*.

[29] Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Multimedia*. ACM, 83–92.

[30] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*. IEEE Computer Society, 1784–1791.

[31] Philip J McParlane, Yashar Moshfeghi, and Joemon M Jose. 2014. "Nobody comes here anymore, it's too crowded"; Predicting Image Popularity on Flickr. In *Proceedings of international conference on multimedia retrieval*. 385–391.

[32] Bo Pang and Lillian Lee. 2007. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2, 1-2 (2007), 1–135.

[33] Tianrong Rao, Xiaoxu Li, and Min Xu. 2020. Learning Multi-level Deep Representations for Image Emotion Classification. *Neural Process. Lett.* 51, 3 (2020), 2043–2061.

[34] Fabrizio Ravì and Sebastiano Battiato. 2012. A Novel Computational Tool for Aesthetic Scoring of Digital Photography. In *CGIV*. IS&T - The Society for Imaging Science and Technology, 349–354.

[35] Stefanie Schmidt and Wolfgang G. Stock. 2009. Collective indexing of emotions in images. A study in emotional information retrieval. *J. Assoc. Inf. Sci. Technol.* 60, 5 (2009), 863–876.

[36] Stefan Siersdorfer, Enrico Minack, Fan Deng, and Jonathon S. Hare. 2010. Analyzing and predicting sentiment of images on the social web. In *MM*. ACM.

[37] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[38] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[39] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International conference on machine learning*, Vol. 97. PMLR, 6105–6114.

[40] Luam Catao Totti, Felipe Almeida Costa, Sandra Eliza Fontes de Avila, Eduardo Valle, Wagner Meira Jr., and Virgílio A. F. Almeida. 2014. The impact of visual attributes on online image diffusion. In *Proceedings of the 2014 ACM conference on Web science*. ACM, 42–51.

[41] Quoc-Tuan Truong and Hady W. Lauw. 2017. Visual Sentiment Analysis for Review Images with Item-Oriented and User-Oriented CNN. In *Multimedia*. ACM, 1274–1282.

[42] Quoc-Tuan Truong and Hady W. Lauw. 2019. VistaNet: Visual Aspect Attention Network for Multimodal Sentiment Analysis. In *AAAI*. AAAI Press, 305–312.

[43] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *ICWSM*.

[44] Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. 2017. Cross-Media Learning for Image Sentiment Analysis in the Wild. In *ICCV*. IEEE Computer Society, 308–317.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[46] Jesse Vig. 2019. A Multiscale Visualization of Attention in the Transformer Model. In *ACL*. Association for Computational Linguistics, 37–42.

[47] Wei-Ning Wang, Ying-Lin Yu, and Shengming Jiang. 2006. Image Retrieval by Emotional Semantics: A Study of Emotional Space and Feature Extraction. In *SMC*. IEEE, 3534–3539.

[48] Yilin Wang, Yuheng Hu, Subbarao Kambhampati, and Baoxin Li. 2015. Inferring Sentiment from Web Images with Joint Inference on Visual and Social Cues: A Regulated Matrix Factorization Approach. In *ICWSM*. AAAI Press, 473–482.

[49] Yang Yang, Jia Jia, Shumei Zhang, Boya Wu, Qicong Chen, Juanzi Li, Chunxiao Xing, and Jie Tang. 2014. How Do Your Friends on Social Media Disclose Your Emotions?. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI Press, 306–312.

[50] Quanzeng You, Hailin Jin, and Jiebo Luo. 2017. Visual Sentiment Analysis by Attending on Local Image Regions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. AAAI Press, 231–237.

[51] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Cross-modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia. In *WSDM*. ACM, 13–22.

[52] Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. 2013. Sentribute: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM, 10:1–10:8.