VistaNet: Visual Aspect Attention Network for Multimodal Sentiment Analysis

Quoc-Tuan Truong, Hady W. Lauw
School of Information Systems
Singapore Management University
qttruong.2017@smu.edu.sg hadywlauw@smu.edu.sg

Abstract

Detecting the sentiment expressed by a document is a key task for many applications, e.g., modeling user preferences, monitoring consumer behaviors, assessing product quality. Traditionally, the sentiment analysis task primarily relies on textual content. Fueled by the rise of mobile phones that are often the only cameras on hand, documents on the Web (e.g., reviews, blog posts, tweets) are increasingly multimodal in nature, with photos in addition to textual content. A question arises whether the visual component could be useful for sentiment analysis as well. In this work, we propose *Visual Aspect Attention Network* or *VistaNet*, leveraging both textual and visual components. We observe that in many cases, with respect to sentiment detection, images play a supporting role to text, highlighting the salient aspects of an entity, rather than expressing sentiments independently of the text. Therefore, instead of using visual information as features, *VistaNet* relies on visual information as alignment for pointing out the important sentences of a document using attention. Experiments on restaurant reviews showcase the effectiveness of visual aspect attention, vis-à-vis visual features or textual attention.

Introduction

In this age of the *participative Web*, user-generated content (e.g., reviews) forms a greater part of the Web. It was reported[1] that 90% of consumers would read reviews before visiting a business, and 88% trust those reviews as much as recommendations from acquaintances. Businesses want to learn user preferences for recommendations, or monitor consumer perceptions for marketing and product design.

Key to feeling the pulse of user-generated content is *sentiment analysis*. The common formulation is text classification (Pang and Lee 2007). Given a document (e.g., review, blog post, tweet), we classify it into sentiment classes, which could be binary (positive vs. negative) or ordinal along some rating scale (e.g., 1 to 5). Various textual features and supervised learning techniques have been proposed (Liu and Zhang 2012). More current methods based on deep neural networks (Kim 2014; Zhang, Zhao, and LeCun 2015; Tang, Qin, and Liu 2015) are especially effective.

[1]https://www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic

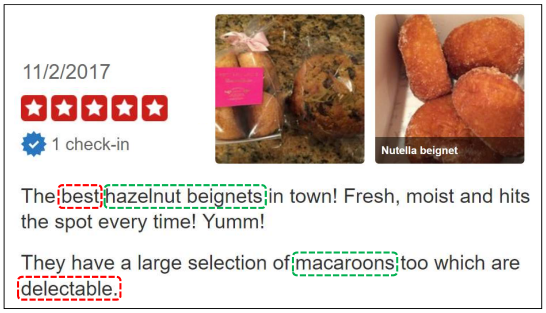

Figure 1: Example of *Yelp* review for Bottega Louie

An emerging concept in sentiment analysis is how different parts of a document are differentially informative. A sentence that expresses sentimentality (e.g., *"The salad was fresh and delicious and the soufle was pure perfection."*) is likely more important than a neutral sentence (e.g., *"I had the Cobb Salad and Chocolate Soufle for lunch."*). Correspondingly, some words (e.g., *"delicious"*) are more influential. These differences in level of informativeness could be captured via *attention* (Yang et al. 2016), which assigns more consequential sentences (or words) higher weights.

Problem. Today's documents contain *more* than text. With smartphones and tablets, it is very convenient to take pictures anytime, anywhere. As a result, many documents are now multimodal. While "multimodal" could refer to image, audio, or video, here we focus on images. Blog posts and reviews often include photos to achieve more vivid descriptions of the authors' experiences. For instance, *Bottega Louie*[2], the most reviewed restaurant in Los Angeles on *Yelp* has 15 thousand reviews (as of the time of writing), with 26 thousand images within. Beyond reviews, it has also been reported[3] that 42% of tweets include an image.

There are synergies between the visual and textual components of reviews. Figure 1 shows a *Yelp* review about *Bottega Louie*, with two images and several sentences describing cakes. We make a couple of observations. First, a sentence within a review tends to focus on one thing

[2]https://www.yelp.com/biz/bottega-louie-los-angeles

[3]https://thenextweb.com/socialmedia/2015/11/03/what-analyzing-1-million-tweets-taught-us/

(e.g., *"beignets"*, *"macaroons"*), and features sentiment-laden words (e.g., *"best"*, *"delectable"*). Second, a photo within a review also tends to focus only on one thing, which tends to be a point mentioned within the review text.

Expectedly, reviews include pictures of things or "aspects" that are especially memorable or important, as such pictures serve to place greater emphasis on those things. The primary means for conveying information, especially on the sentiment, remains the text. Photos play an augmentative role, rather than an independent role; they do not tell the whole story on their own. With this insight, instead of incorporating photos directly as features into the sentiment classification, we propose that they may be better placed for a different role, i.e., as a visual means to direct attention to the most salient sentences or "aspects" within a review.

Contributions. We postulate that there is potential in incorporating visual information into text-based sentiment analysis, and make the following contributions in this paper. *First*, to our best knowledge, this work is the first to incorporate images as attention for review-based sentiment analysis.

Second, we develop a neural network model called *Visual Aspect Attention Network* or *VistaNet*, which considers visual information as a source of alignment at the sentence level. Each sentence in a review could embody some "aspect" (though we do not presume or prescribe a prespecified list of aspects). An image would help identify important sentences within the review that the model should pay more attention when classifying its sentiment.

Third, we conduct comprehensive experiments on *Yelp* restaurant reviews from five major US cities against comparable baselines. While reviews provide a good test case due to the presence of ratings for supervision in training and ground-truth in testing, the model could potentially generalize to other types of Web documents such as blog posts, tweets, or any document containing images.

Related Work

Previous works on sentiment analysis mainly focus on text (Pang and Lee 2007; Tumasjan et al. 2010; Bollen, Mao, and Pepe 2011; Hu et al. 2013; Kiritchenko, Zhu, and Mohammad 2014). Recently, deep learning has made significant inroads in text classification (Kim 2014; Tang, Qin, and Liu 2015; Socher et al. 2013; Lai et al. 2015). The success of RNN with attention (Bahdanau, Cho, and Bengio 2014) gives rise to text classification with hierachical levels (Yang et al. 2016). Where they rely on textual clues alone, our innovation is to also rely on images as visual aspect attention.

Some works study *aspect-level* sentiment analysis (Nguyen and Shirai 2015; Tang, Qin, and Liu 2016). In contrast, we focus on sentiment of the *whole* document.

Visual sentiment analysis (Truong and Lauw 2017; You et al. 2015; Borth et al. 2013) is formulated as image classification. An anachronistic approach is extracting low-level features (e.g. SIFT features) from images, followed by learning a classifier (e.g. SVM, Naive Bayes) (Siersdorfer et al. 2010; Borth et al. 2013). Recent approaches leverage on representation learning using deep learning. Image features are extracted by pre-trained CNN (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014; Szegedy et al. 2015) and fed into a classifier or trained by an end-to-end network (Chen et al. 2014; You et al. 2015). These unimodal approaches rely on only images as features.

A different approach from ours to multimodal sentiment analysis is to derive a joint representation by combining those obtained from the respective components. This could be applicable when there is only one image to one document, both referring to the same meaning, such as for image tweets (You et al. 2016a; You et al. 2016b). In our context, a document may have several images (e.g., blog post, review), each referring to a specific part of the document. Therefore, we assume that images are augmentative rather than representative, and are more suitable as attention rather than features. Experiments will compare *VistaNet* to multimodal baselines that combine the representational features obtained from text and images.

Finding alignment between textual and visual data is also explored by multimodal learning, especially for image captioning (Karpathy and Li 2015; Xu et al. 2015) and visual question answering (Yu et al. 2017; Lu et al. 2016). Different from those problems where there is strong 1-to-1 alignment between image and text, in our problem a document may be associated with several images, each relevant to a specific part of the document, and there is no ground-truth alignment for supervision. We learn the alignment that would help sentiment classification, by paying more attention to image-related sentences (hypothetically more important).

In a different context, "visual attention" is a phrase used to describe which part of an image attracts the attention of human subjects (Itti, Koch, and Niebur 1998; Desimone and Duncan 1995), which is useful for tasks such as scene understanding from images. This is an orthogonal concept to what we are modeling in this work, i.e., which sentence within a text review to pay attention to based on review images.

Visual Aspect Attention Network

We now define the problem, and describe our proposed *Visual Aspect Attention Network* or *VistaNet* model.

Problem. We are given a set of documents $\boldsymbol{\mathcal{C}}$, e.g., reviews. For each document $\boldsymbol{c} \in \boldsymbol{\mathcal{C}}$, its textual component is a sequence of L sentences, $\boldsymbol{s}_i$, $i \in [1, L]$. Let $\boldsymbol{s}_i$ denote a sentence constructed by a sequence of T words $\boldsymbol{w}_{i,t}, t \in [1, T]$. Its visual component is a set of M images $\boldsymbol{a}_j \in \{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_M\}$. L and M may vary between documents. Each document $\boldsymbol{c}$ is also associated with a sentiment label. The problem can thus be stated as follows: given $\boldsymbol{\mathcal{C}}$ as training corpus, the objective is to learn a classification function, to predict sentiment labels for unseen documents.

VistaNet is a hierarchical three-layered architecture, as illustrated in Figure 2. The bottom layer is the word encoding layer with soft attention, where we transform word representations into a sentence representation. The middle layer is the sentence encoding layer where we transform the sentence representations into a document-level representation, with the help of the visual aspect attention. The top layer is the classification layer to assign the document a sentiment label. We now discuss each layer respectively in more detail.

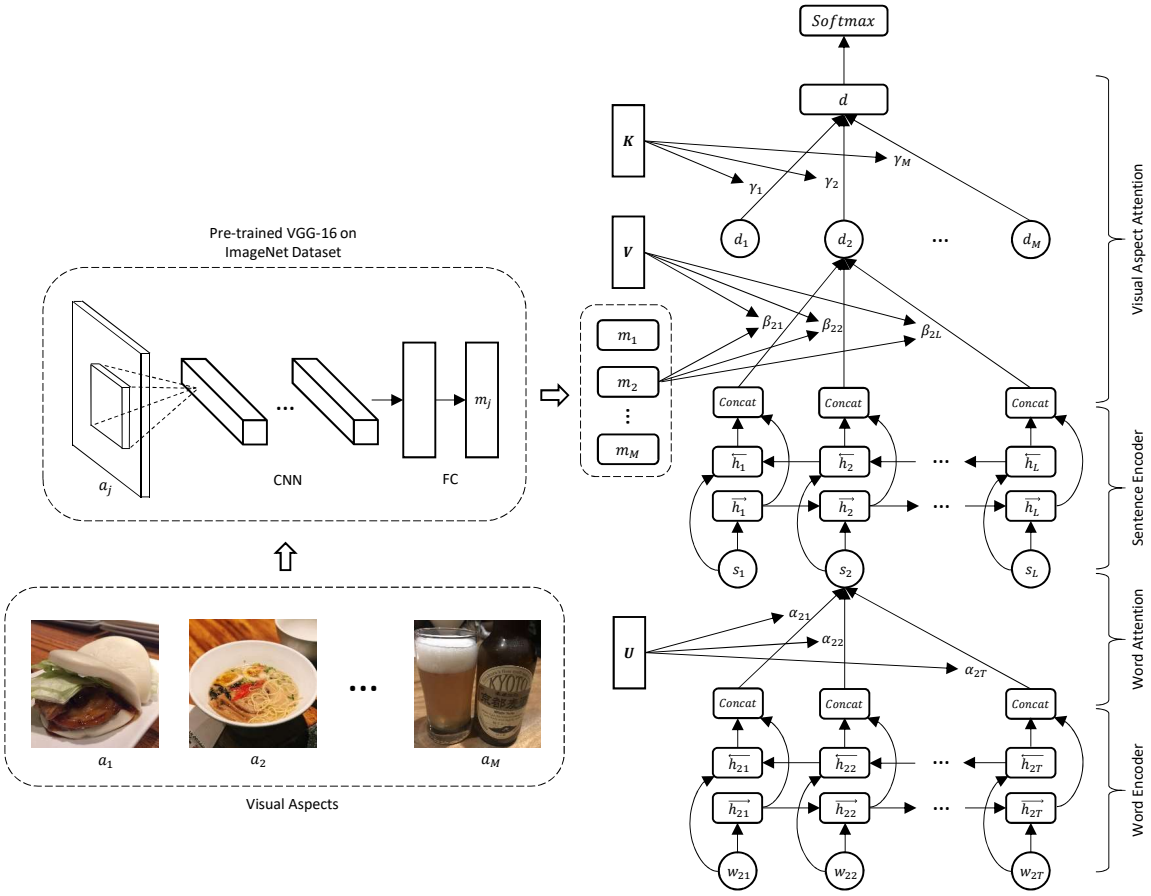

Figure 2: Overall Architecture of *VistaNet*

Word Encoder with Soft Attention

For each word $\boldsymbol{w}_{i,t}$, we derive its embedding $\boldsymbol{x}_{i,t}$ with a learned embedding matrix $\boldsymbol{W}_e$, which can be initialized from pre-trained word embedding models (Mikolov et al. 2013; Pennington, Socher, and Manning 2014) and subsequently adjusted during training.

$$\boldsymbol{x}_{i,t} = \boldsymbol{W}_e \boldsymbol{w}_{i,t}, t \in [1, T] \tag{1}$$

To encode the entire sequence of word embeddings, we use bidirectional recurrent neural network (Bi-RNN) with GRU cell (Cho et al. 2014; Yang et al. 2016), which takes in embedding input $\boldsymbol{x}_{i,t}$ and outputs a new vector of hidden states $\boldsymbol{h}_{i,t} = [\overrightarrow{\boldsymbol{h}}_{i,t}, \overleftarrow{\boldsymbol{h}}_{i,t}]$, which is the concatenation of $\overrightarrow{\boldsymbol{h}}_{i,t}$ generated by the forward RNN and $\overleftarrow{\boldsymbol{h}}_{i,t}$ generated by the backward RNN.

$$\boldsymbol{h}_{i,t} = \text{Bi-RNN}(\boldsymbol{x}_{i,t})$$

All words in a sentence are *not* equal. Some words are more informative and meaningful towards sentiment detection. Hence, when deriving the representation of a sentence from those of its words, each word will be assigned a weight corresponding to its "importance" in the sentence representation. To learn and distribute these weights among words, we employ a soft attention mechanism.

$$u_{i,t} = \boldsymbol{U}^T \tanh(\boldsymbol{W}_w \boldsymbol{h}_{i,t} + \boldsymbol{b}_w) \tag{2}$$

$$\alpha_{i,t} = \frac{\exp(u_{i,t})}{\sum_t \exp(u_{i,t})} \tag{3}$$

$$\boldsymbol{s}_i = \sum_t \alpha_{i,t} \boldsymbol{h}_{i,t} \tag{4}$$

We project $\boldsymbol{h}_{i,t}$, the representation of word $\boldsymbol{w}_{i,t}$, through a layer of neurons with a non-linear activation function $\tanh$, to have its representation in the attention space. Then we multiply the projection with a context vector $\boldsymbol{U}$ (randomly initialized and learned during training) to obtain scalar $\boldsymbol{u}_{i,t}$ that indicates the relative importance of $\boldsymbol{w}_{i,t}$. This is normalized using softmax to produce its attention weight $\alpha_{i,t}$. Finally, the vector representation of the sentence $\boldsymbol{s}_i$ is produced by a weighted summation over all its word representations $\boldsymbol{h}_{i,t}$'s and their attention weights $\alpha_{i,t}$'s.

Sentence Encoder with Visual Aspect Attention

The middle layer aggregates the sentence-level representations from the bottom layer, and aggregates them into a document-level representation using visual aspect attention, assigning greater weight to the more salient sentences. Bi-

RNN outputs hidden states vector $\boldsymbol{h}_i = [\overrightarrow{\boldsymbol{h}}_i, \overleftarrow{\boldsymbol{h}}_i]$ at each position of input sentence $\boldsymbol{s}_i$.

$$\boldsymbol{h}_i = \text{Bi-RNN}(\boldsymbol{s}_i)$$

To get the final representation of the document $\boldsymbol{d}$, one option is to employ text-based soft attention pooling scheme (Yang et al. 2016). In contrast, we advocate using visual information to augment the attention mechanism. A document may be associated with several images, which may be pertinent to different "aspects". Given an image, sentences are differentially informative. In other words, images would highlight different yet important parts of a document. We seek to develop a soft attention mechanism using visual information to improve the quality of learned document representation. We refer to this as *visual aspect attention*.

We first need to encode the input images. VGG convolutional neural networks (Simonyan and Zisserman 2014) has proven effective in learning representations for images for many image-related tasks (Xu et al. 2015; Karpathy and Li 2015). We employ the VGG-16 to get the representation $\boldsymbol{m}_j$ of image $\boldsymbol{a}_j$ by feeding it through the model and geting the output of the last fully-connected layer (FC7) before the classification layer. The image representation $\boldsymbol{m}_j$ is a 4096-dimensional vector encoded from image $\boldsymbol{a}_j$.

$$\boldsymbol{m}_j = \text{VGG}(\boldsymbol{a}_j)$$

With respect to each image representation $\boldsymbol{m}_j$, we learn the attention weights $\beta_{j,i}$'s for sentence representations $\boldsymbol{h}_i$'s.

$$\boldsymbol{p}_j = \tanh(\boldsymbol{W}_p \boldsymbol{m}_j + \boldsymbol{b}_p) \quad (5)$$

$$\boldsymbol{q}_i = \tanh(\boldsymbol{W}_q \boldsymbol{h}_i + \boldsymbol{b}_q) \quad (6)$$

$$v_{j,i} = \boldsymbol{V}^T(\boldsymbol{p}_j \odot \boldsymbol{q}_i + \boldsymbol{q}_i) \quad (7)$$

$$\beta_{j,i} = \frac{\exp(v_{j,i})}{\sum_i \exp(v_{j,i})} \quad (8)$$

To learn these attention weights, first we project both image representation $\boldsymbol{m}_j$ and sentence representation $\boldsymbol{h}_i$ onto an attention space followed by a non-linear activation function; the outputs are $\boldsymbol{p}_j$ and $\boldsymbol{q}_i$ respectively. For the activation function, we use $\tanh$ to scale $\boldsymbol{m}_j$ and $\boldsymbol{h}_i$ into the same range of values, so that neither component dominates the other. To learn the image-specific attention weight of a sentence, we let the image projection $\boldsymbol{p}_j$ interact with the sentence projection $\boldsymbol{q}_i$ in two ways: *element-wise multiplication* and *summation*, for a reason to be discussed shortly. The learned vector $\boldsymbol{V}$ plays the role of global attention context similar to $\boldsymbol{U}$ at word level. This produces an attention value $v_{j,i}$, which is normalized using softmax to obtain $\beta_{j,i}$.

Both element-wise multiplication and summation are needed to compute $v_{j,i}$ to ensure that there is a meaningful interaction between the image and the sentence. Without the element-wise multiplication, and with only summation, the effects of the visual part would have been cleared out by the softmax function when calculating attention weight $\beta_{j,i}$. Without the summation, and with only the element-wise multiplication, the effects of the text part would have been significantly weakened because of the sparsity of the visual part. Hence, both are required for an effective visually-informed soft attention. Our proposed mechanism can be seen to generalize over "bilinear" attention (Kim et al. 2016), which provides tighter interactions than "concat-product" attention (Bahdanau, Cho, and Bengio 2014), if we remove the addition of q_i from Eq. 7.

Using the image-specific attention weights $\beta_{j,i}$, we aggregate the sentence representations $\boldsymbol{h}_i$'s into an image-specific document representation $\boldsymbol{d}_j$ as follows.

$$\boldsymbol{d}_j = \sum_i \beta_{j,i} \boldsymbol{h}_i \quad (9)$$

For a document, we apply this visual aspect attention mechanism for each of its images, yielding a set of aspect-specific document representations $\boldsymbol{d}_j, j \in [1, M]$. All the $\boldsymbol{d}_j$'s need to be aggregated into the final document representation $\boldsymbol{d}$ before classification. Given a document, images are differentially informative. Thus, we seek to learn the importance weight γ_j, signifying how each image-specific document representation $\boldsymbol{d}_j$ would contribute to the final document representation $\boldsymbol{d}$.

$$k_j = \boldsymbol{K}^T \tanh(\boldsymbol{W}_d \boldsymbol{d}_j + \boldsymbol{b}_d) \quad (10)$$

$$\gamma_j = \frac{\exp(k_j)}{\sum_j \exp(k_j)} \quad (11)$$

Aspect-specific document representation $\boldsymbol{d}_j$ is projected into attention space through a layer of neurons with non-linear activation function $\tanh$. The scalar k_j indicating the importance of $\boldsymbol{d}_j$ is obtained by multiplying with global attention context vector $\boldsymbol{K}$ (randomly initialized and learned during training). As shown in Figure 2, the document representation $\boldsymbol{d}_j$'s due to the various images are aggregated into the final document representation $\boldsymbol{d}$ using soft attention pooling with document-to-image attention weights γ_j's.

$$\boldsymbol{d} = \sum_j \gamma_j \boldsymbol{d}_j \quad (12)$$

One limitation of relying only on images found within a document is that the sentiment of the whole document may not be captured completely, because some documents do not have sufficient images to cover all its important aspects. As a result, an important sentence may be overlooked because it does not correspond to any image that can focus some attention to it. To overcome this limitation, in addition to the images found in a document, we include one more global "MEAN" image, which allows those "orphaned" yet important sentences to still be aligned. This additional image plays the role of "global" aspect, and also helps our model to potentially generalize to documents without images.

Sentiment Classification

Finally, in the top layer, after obtaining the high-level representation of the document $\boldsymbol{d}$, we treat it as the features for a softmax-based sentiment classifier, producing the probability distribution over classes $\boldsymbol{\rho}$.

$$\boldsymbol{\rho} = \text{softmax}(\boldsymbol{W}_c \boldsymbol{d} + \boldsymbol{b}_c)$$

The model is trained in a supervised manner by minimizing the cross entropy error of sentiment classification:

$$loss = -\sum_d \log \rho_{d,l}$$

where l is the ground truth label of review $\boldsymbol{d}$.

City	#docs	avg. #s	max #s	avg. #w	max #w	#images
BO	2,080	13.4	85	222.3	1115	10,743
CH	2,165	13.5	96	219.0	1107	12,360
LA	24,860	14.4	104	227.2	1134	137,920
NY	11,425	13.4	95	217.5	1129	61,474
SF	3,775	14.8	98	237.3	1145	22,072
Total	44,305	14.8	104	237.3	1145	244,569

Table 1: Data Statistics

Experiments

We investigate several research questions on the effectiveness of the proposed *VistaNet* for sentiment analysis. First, we consider how our modeling of visual information as attention would perform as compared to multimodal baselines that rely on both textual and visual information as features. Second, we analyze the contributions of the various architectural components of our model by performing an ablation analysis. In addition, we also study the effects of incremental addition of images, as well as look into a case study to get a better understanding of the workings of the model.

Setup

We describe the setup of the experiments, including the dataset, the evaluation tasks, as well as the training details.

Dataset We use a dataset of online reviews crawled from the Food and Restaurants categories of *Yelp.com*, covering 5 different major US cities, namely: Boston (BO), Chicago (CH), Los Angeles (LA), New York (NY), and San Francisco (SF). The statistics are shown in Table 1. LA is the largest, with the most documents and images. BO is the smallest. However, the document lengths, in terms of the number of sentences (#s) and the number of words (#w), are quite similar across the five cities. In total, the dataset has more than 44 thousand reviews, including 244 thousand images. Each review has at least 3 images for the purpose of experiment on visual aspect attention effectiveness.

Task Our target application is sentiment analysis. Since *Yelp* reviews include a rating on the scale of 1 to 5 as five sentiment levels, we treat each rating as a class. We keep the number of examples balanced across classes, and split 80% of the data for training, 5% for validation and 15% for test. Because some cities have smaller data, we merge the training and validation sets of the five cities, while test data is kept separate to maintain statistical property when evaluating models. The metric used is classification accuracy.

Training Details For preprocessing, we use NLTK (Loper and Bird 2002) for sentence and word tokenization. We build the vocabulary from words appearing more than 3 times in the training and validation sets, and replace the other infrequent words with special UNK token. We employ the pre-trained word embeddings from GloVe (Pennington, Socher, and Manning 2014) to initialize the embedding matrix $\boldsymbol{W}_e$ with dimensionality $D = 200$. Word embeddings are fine-tuned during training to adapt to the domain at hand.

All models are tuned with hyper-parameters for their best performance on the validation set. GRU cells are 50-dimensional for word and sentence encoding, (100-dimensional due to bidirectional RNN). Context vectors $\boldsymbol{U}$, $\boldsymbol{V}$ and $\boldsymbol{K}$ are also 100-dimensional for the attention spaces of word, sentence, and document. For images, we use VGG-16 CNN for feature extraction. The image representation is the output from FC7 layer right before the classification layer. We initialize the weights of image encoder using the pre-trained VGG-16 model on ImageNet dataset (Deng et al. 2009) and all the weights of image encoder are fixed during training. We also take the MEAN image from this model. For more details about VGG, we refer readers to (Simonyan and Zisserman 2014). In training, we use RMSprop (Tieleman and Hinton 2012) for gradient based optimization with a mini-batch size of 32. We tune model hyper-parameters on validation set and report the average results with statistical tests on the test sets after multiple runs for every compared methods. *VistaNet* is implemented using TensorFlow[4].

Comparison to Multimodal Baselines

Baselines. We compare the proposed *VistaNet* model that uses visual information as *attention*, to the following multimodal baselines that learn from visual and textual *features*.

- *BiGRU-aVGG* and *BiGRU-mVGG* are composites that concatenate the representations learnt by BiGRU from text and by VGG from images, and feed them to a classification layer. BiGRU is shown to be effective for sequential data, such as text (Tang, Qin, and Liu 2015). For images, we use VGG-16 architecture as encoder with pre-trained model from the ImageNet dataset. The image features are taken from FC7 layer before classification layer. Because there are multiple images per review, there is a need to aggregate the image representations. This results in two variants: *BiGRU-aVGG* that employs averaging pooling and *BiGRU-mVGG* that employs max-pooling for all image feature vectors before we concatenate with the feature vector from text. Weights of the image encoder are fixed during training (also done with *VistaNet*).
- *HAN-aVGG* and *HAN-mVGG* are composites of HAN-ATT (state-of-the-art for textual sentiment analysis) for text and VGG for images. HAN-ATT (Yang et al. 2016) exploits hierarchical structure of documents with word encoder and sentence encoder. The primary difference is *VistaNet*'s modeling of visual aspect attention, as opposed to HAN-ATT's text-only soft attention layers. The two variants correspond to averaging pooling (*HAN-aVGG*) and max-pooling (*HAN-mVGG*) respectively.
- *TFN-aVGG* and *TFN-mVGG* are composites of Tensor Fusion Network (Zadeh et al. 2017) (state-of-the-art for multimodal sentiment analysis). The textual features from HAN-ATT are combined with visual features from VGG using Tensor Fusion Layer and fed through Sentiment Inference Subnetwork to get the final sentiment label. We also apply averaging pooling and max-pooling yielding two variants *TFN-aVGG* and *TFN-mVGG* respectively.

[4]https://github.com/PreferredAI/vista-net

Models	Textual Features	Visual Features	Hierarchical Structure	Visual Aspect Attention	BO	CH	LA	NY	SF	Avg.	Improvement
TFN-aVGG	✓	✓	✓		46.35	43.69	43.91	43.79	42.81	43.89	-
TFN-mVGG	✓	✓	✓		48.25	47.08	46.70	46.71	47.54	46.87	6.8%
~~BiGRU-aVGG~~	~~✓~~	~~✓~~			~~51.23~~	~~51.33~~	~~48.99~~	~~49.55~~	~~48.60~~	~~49.32~~	~~12.4%~~
BiGRU-mVGG	✓	✓			53.92	53.51	52.09	52.14	51.36	52.20	18.9%
HAN-aVGG	✓	✓	✓		55.18	54.88	53.11	52.96	51.98	53.16	21.1%
~~HAN-mVGG~~	~~✓~~	~~✓~~	~~✓~~		~~56.77~~	~~57.02~~	~~55.06~~	~~54.66~~	~~53.69~~	~~55.01~~	~~25.3%~~
VistaNet	✓		✓	✓	**63.81**$^{\star\circ}$	**65.74**$^{\star\circ}$	**62.01**$^{\star\circ}$	**61.08**$^{\star\circ}$	**60.14**$^{\star\circ}$	**61.88**$^{\star\circ}$	41.0%

* Statistical tests show that VistaNet performs significantly better than the base model BiGRU-aVGG ($p < 0.05$).
$^{\circ}$ Statistical tests show that VistaNet performs significantly better than the second-best model HAN-mVGG ($p < 0.05$).

Table 2: Performance Comparison to Multimodal Baselines

Comparison. Table 2 lists the results of the comparative methods, as well as the key attributes of the respective methods. In addition to showing the results for the five cities, it shows the average across the cities, and indicates the degree of improvement with respect to the base with lowest performance (the first row in the table).

Interestingly, the TFN models, which provide rich interactions between textual and visual features, turn out to perform the worst among comparative methods with the accuracies of 43.89% and 46.87% for TFN-aVGG and TFN-mVGG respectively. The results support our hypothesis that review photos weakly express sentiment on their own. Combining features through a complex fusion matrix makes it difficult for the models to find useful textual-visual allignments for the sentiment because they do not carry the same sentiment-driven information. In any case, these performances are substantially higher than random, which would be around 20%.

BiGRU-aVGG (averaging pooling) achieves 49.32% accuracy. With max-pooling, BiGRU-mVGG achieves a higher accuracy of 52.2%, which represents a 5.8% improvement upon BiGRU-aVGG and a 18.9% improvement upon TFN-aVGG. These models incorporate features from both the review text as well as images via concatenation.

Hierarchical HAN-aVGG and HAN-mVGG perform better than BiGRU-mVGG and BiGRU-aVGG. The max-pooling variant is a little higher at 55.01% than the averaging pooling one at 53.16%. These improvements come from the hierarchical modeling in the text module (word level, then sentence level), as compared to BiGRU's single-level modeling of text (word level only). The hierarchical model is supported by soft attention based on the text component.

Our proposed model *VistaNet* performs the best consistently across all the cities. The average accuracy of 61.88% represents a 41.0% improvement upon the the base model TFN-aVGG, and 12.5% improvement upon the most competitive baseline HAN-mVGG. These outperformances are statistically significant across the five cities as well.

All the baseline methods are multimodal, employing both textual and visual features. Our key distinction is modeling visual information as attention, rather than features. This underscores the point that the value of visual information within a review is to draw attention to the salient sentences, rather than to express sentiments directly. The results here provide evidence on the effectiveness of visual aspect attention for multimodal sentiment analysis.

Architecture Ablation Analysis

To investigate the respective contributions of the various components of *VistaNet*'s architecture, we conduct an ablation analysis that starts with the most basic configuration, and incrementally adds a component towards constructing the full architecture. The results are summarized in Table 3.

We start with the base model BiRNN relying only on text. As shown in the first row, this achieves 56.83% on average. Exploting the hierarchical structure of documents, by applying max-pooling on sentence representations, we improve the results by 4.8% as compared to the base model, as shown in the second row. This showcases the value of modeling the hierarchical structure of text. If we apply a soft attention layer based on text alone when aggregating the sentence-level representations, we achieve an improvement of 6.6% over the base model, as shown in the third row of Table 3. By further incorporating visual aspect attention, we achieve an improvement of 8.9% over the base model, as shown in the fourth row. The average accuracy is 61.88%.

The outperformances are statistically significant on the average, as well as across the five cities when compared to the base model. When compared to the second best model, the results are still significant on the average, as well as on four cities. These results support the hypothesis that each component in the *VistaNet* architecture makes a contribution to the performance of the full-fledged model.

Visual Aspect Attention

We evaluate how the number of images may affect the visual aspect attention mechanism. Hypothetically when we increase the amount of visual information, the model will have more choices in aligning the sentences, probably painting a slightly clearer picture towards the overall sentiment.

For each document, we vary the number of images from only the MEAN image, then incrementally adding up to 3 more images. In each case, we sample the specified number randomly from among the images of a document. We do not go beyond 3 images as that would exclude too many documents as requiring 4 or more images would lead to a drastic reduction in data size, which is about 40% in our dataset. We also make sure that all examples regardless of classes have the same number of images to remove the bias from data (a review with more images tends to have higher rating). The model always has the global MEAN image as default.

Table 4 shows the results across all cities when the num-

Bi-RNN	Hierarchical Structure	Soft Text Attention	Visual Aspect Attention	BO	CH	LA	NY	SF	Avg.	Improvement
✓				57.70	60.01	56.74	56.59	55.84	56.83	
✓	✓			60.39	64.39	59.08	59.58	59.18	59.54	4.8%
✓	✓	✓		63.38	64.47	60.65	59.85	58.34	60.56	6.6%
✓	✓	✓	✓	**63.81***	**65.74**$^{\star\circ}$	**62.01**$^{\star\circ}$	**61.08**$^{\star\circ}$	**60.14**$^{\star\circ}$	**61.88**$^{\star\circ}$	8.9%

* Statistical tests show that the improvements are significant over the base with texual input ($p < 0.05$).
$^{\circ}$ Statistical tests show that the improvements are significant over the second best with soft attention ($p < 0.05$).

Table 3: Architecture Ablation Analysis of *VistaNet*

No. of images	BO	CH	LA	NY	SF	Avg.	Improvement
MEAN	62.58	64.60	60.79	59.84	58.44	60.61	-
+1	62.62	64.74	61.25	60.69	59.13	61.16	0.9%
+2	62.56	65.18	61.69	**61.14**	59.93	61.61	1.6%
+3	**63.81**$^{\star\circ}$	**65.74***	**62.01**$^{\star\circ}$	61.08*	**60.14***	**61.88**$^{\star\circ}$	2.1%

* Statistical tests show that the improvements are significant over the base with only the MEAN image ($p < 0.05$).
$^{\circ}$ Statistical tests show that the improvements are significant over the second best with 2 images ($p < 0.05$).

Table 4: Visual Aspect Attention

ber of images varies. We observe a general trend that the classification accuracies tend to increase as we increase the number of images. With MEAN + 3 images, the accuracy of the VistaNet model increases by 2.1% on average. This improvement is statistically significant when compared either with the base of only MEAN image used, or with the closest model with MEAN + 2 images. This supports the contribution of the visual aspect attention mechanism.

Illustrative Examples

To lend some intuitive appreciation for how the visual aspect attention may work to improve the effectiveness of *VistaNet*, here we provide a couple of illustrative examples.

Figure 3 shows an example of a review with a ground-truth rating of 5. On the top left are its three images and MEAN (with their γ's on the top that in this particular case are relatively uniform). On the right are the review sentences (in their original sequence). Based on the image-to-sentence weights β's learned by the VistaNet model, for each image, we show the distribution of β's demonstrating the relative importance of each sentence in the document according to a particular visual aspect. Moreover, within each sentence, certain words are highlighted. This is indicative of attention at the word level. The darker the highlight, the higher the attention weight of a word within a sentence.

The first image visually depicts a dish. Based on the attention weights, the image focuses on the fourth sentence *"the food is great"* which expresses a strong sentiment towards the food. The second image depicts a drink and skews towards the sixth sentence *"i would recommend getting one of their lemonades"*. Based on the highlighted words, we notice *"recommend"* and *"lemonades"* are more emphasized than other words in the sentence. That offers another hint towards the positive sentiment of the user toward the restaurant. The third image, which focuses on *"get their tacos"* phrase within the eighth sentence, also visually depicts *taco*.

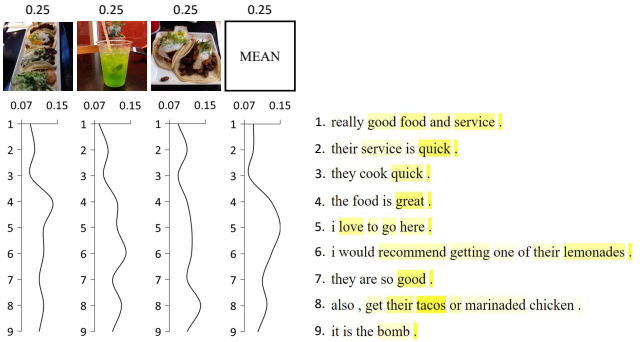

Figure 3: A review with rating 5 of *Komodo* from *Yelp.com* (*best seen in color*)

MEAN increases the probability of document being positive by pointing to the fifth sentence *"i love to go here"*.

Conclusion

We propose a novel approach of using visual information for sentiment analysis called *Visual Aspect Attention Network* or *VistaNet*. The model has a three-layered architecture, aggregating the representations from word to sentence, then to image-specific document representations, and finally to the final document representation. Based on the observation that a sentence tends to focus on something specific, as does each image, we design the model to employ images as alignment to point out the important sentences within a document. Experiments on review datasets from five major US cities show that *VistaNet* outperforms multimodal baselines that uses both textual and visual features on sentiment analysis, supporting our hypothesis that the visual component is more augmentative than representative, and is more effective as an attention mechanism. The datasets and codes used in this submission will be released publicly upon publication.

Acknowledgments

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its NRF Fellowship Programme (Award No. NRF-NRFF2016-07).

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bollen, J.; Mao, H.; and Pepe, A. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*.

Borth, D.; Chen, T.; Ji, R.; and Chang, S. 2013. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *MM*.

Chen, T.; Borth, D.; Darrell, T.; and Chang, S. 2014. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *CoRR*.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

Desimone, R., and Duncan, J. 1995. Neural mechanisms of selective visual attention. *Annual review of neuroscience*.

Hu, X.; Tang, J.; Gao, H.; and Liu, H. 2013. Unsupervised sentiment analysis with emotional signals. In *WWW*.

Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*.

Karpathy, A., and Li, F. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Kim, J.-H.; On, K.-W.; Lim, W.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *CoRR* abs/1408.5882.

Kiritchenko, S.; Zhu, X.; and Mohammad, S. M. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* 50:723–762.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.

Lai, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. Recurrent convolutional neural networks for text classification. In *AAAI*, volume 333, 2267–2273.

Liu, B., and Zhang, L. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*. Springer. 415–463.

Loper, E., and Bird, S. 2002. Nltk: The natural language toolkit. In *ACL Workshop*, ETMTNLP '02, 63–70.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 289–297.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Nguyen, T. H., and Shirai, K. 2015. Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In *EMNLP*, 2509–2514.

Pang, B., and Lee, L. 2007. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Siersdorfer, S.; Minack, E.; Deng, F.; and Hare, J. S. 2010. Analyzing and predicting sentiment of images on the social web. In *MM*, 715–718.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 1631–1642.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *CVPR*, 1–9.

Tang, D.; Qin, B.; and Liu, T. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP*, 1422–1432.

Tang, D.; Qin, B.; and Liu, T. 2016. Aspect level sentiment classification with deep memory network. *CoRR*.

Tieleman, T., and Hinton, G. 2012. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning.

Truong, Q., and Lauw, H. W. 2017. Visual sentiment analysis for review images with item-oriented and user-oriented CNN. In *MM*, 1274–1282.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A. J.; and Hovy, E. H. 2016. Hierarchical attention networks for document classification. In *HLT-NAACL*.

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 381–388.

You, Q.; Cao, L.; Jin, H.; and Luo, J. 2016a. Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks. In *MM*.

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2016b. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *WSDM*.

Yu, D.; Fu, J.; Mei, T.; and Rui, Y. 2017. Multi-level attention networks for visual question answering. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 4187–4195. IEEE.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *NIPS*.