Project ID: 10

Text & Vision-Fused Framework for Academic Paper Review

Yichen Yang, Tongan Cai, Shuyang Huang, Jiachen Liu
{yangych,johncta,shuyangh,amberljc}@umich.edu

Abstract

The great boom in Artificial Intelligence these years has been leading to more and more great work, yet high volumes of submissions in academic paper greatly overwhelmed review committee. This Text & Vision-Fused Framework will distinguish papers of lower quality based on contents, vocabulary usage and image quality with a deep-learning-based model. It can serve as an efficient and reasonably accurate filter for academic paper review process, with experiment showing that the framework is superior in acceptance classification on ICLR academic paper submission with 98.7% correctly reject rate and only 4% mis-reject rate.

1 Introduction

The great boom in Artificial Intelligence these years has been providing great chances for researchers to develop novel algorithms and applications, consequently bringing out a great number of research papers. Taking IEEE conference on Computer Vision and Pattern Recognition (CVPR) as an example, as Figure 1 shows, the number of submissions for CVPR 2010 was 1724, while in 2017 it was 2680. It rises to 5000$\sim$ at 2019.

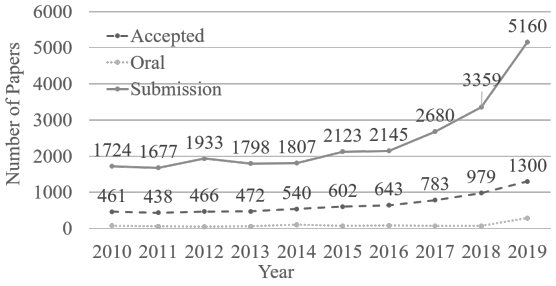


Figure 1: CVPR2010-2019 Number of Submissions & Accepted Papers

While more ideas and methods are being presented, the submissions greatly overwhelmed paper review committee. How to efficiently determine the quality of academic paper remains a demanding question. If some papers of low quality could be judged and ruled out ahead of time, the committee members will be greatly relieved.

Our team proposed a novel way for academic paper review process to save reviewers' labor by adopting deep learning method to judge the quality of academic paper. With the help of this academic paper review framework, we can effectively improve the efficiency of this process and accuracy.

This framework is designed to have printed academic paper in PDF format as input. The information is then extracted to be text sequences, "Gestalt" (the look of first 8 pages) and image set. A fused deep model is constructed with Hierarchical Attention Network [19] for text classification and a ResNet [5] for the "Gestalt" image classification. It also consider the statistics of the image sets and text sequences as biases. To our best knowledge, this is so far the **FIRST** framework to fuse text & vision features of academic papers for acceptance prediction, both of which are proved to be valuable in this classification task.

2 Related Works

2.1 Text Classification

Ever since the 1960s, text classification has been widely investigated [16]. While in the earliest stage, text classification can be regarded as filtering out irrelevant documents from a large-scale corpus, the overall task of text classification was to allocate documents into predefined categories. Well-known methods like Rocchio Relevance Feedback and Decision Tree stems ever then. The vast development in statistical and machine learning techniques in late 1990s gave birth to the new era of text classification, works using SVM [2], Naive Bayes [14], Maximum Entropy [15], ConvNet [20] has been coming out all the time. In recent years, attention-based networks are also giving very ideal performance on text classification, like Recurrent Neural Network [3] and recurrent neural networks based on long short-term memory (LSTM) [6]. In this project, we utilized a Hierarchical Attention Network with two levels of attention [19].

2.2 Deep-Network & Image Classification

Doing classification or recognition on images is just human nature. Among research works decades ago, LeCun [12] introduced a backprop method for convolutional networks, and opened up the door of contemporary research in image classification with deep neural network. Famous models like VGG (Very Deep Convolutional Network) [17], DenseNet (Densely Connected Convolutional Networks) [7], AlexNet [10] are widely used as image classification modules. In this project, we would like to borrow the idea of Residual Learning and construct ResNet (Residual Network) [4]. Contemporary image classification mainly focuses on the recognition of objects in the image. Well-used image databases like MNIST [13], CIFAR [9] and ImageNet [1] all have images that would be easily recognize by human. In this project, we would like to classify on images that human may not be able to easily distinguish–the "Gestalt" of academic papers [8].

2.3 Computer-Aided Paper Review

Paper review process could be time-consuming and tough, especially in the recent years, when the number of papers in general computer science is at a vast increase. Computer-aided paper review may be of auxiliary help to paper review committee, yet there would be large bias in deploying such system. There are already some research work in Computer-Aided essay grading, either with text classification method [11] or with convolutional recurrent neural network [18], but few would stretch out to generalize it on paper review process. One inspiring work will be "Deep Paper Gestalt" [8] that considers a "Gestalt" method as imitation of human reading for pre-filtering of academic paper submissions. In this project, we borrow Huang's idea and try to go deeper with fused image & text model.

3 Preliminary

3.1 Residual Network

For "Gestalt" image Based Classifier, we use ResNet-18 (pretrained on ImageNet) [5] [8] (Figure 2), as ResNet is one of the state-of-the-art neural network on image classification task. The smallest version was used because of the limited computational resources and relatively small dataset. The flow of "residue" are indicated in Figure 2. Comparing to a full ResNet, the last fully-connect layer was removed in this project for building the fused network. This image based classifier reads the low-resolution image "Gestalt" of the PDF and generate the feature vector.

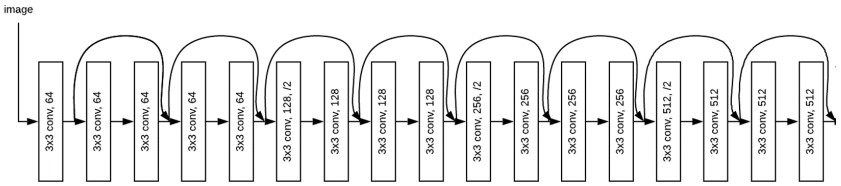


Figure 2: ResNet-18 Structure

3.2 Hierarchical Attention Network

For text sequences, we made a simple implementation of HAN (Hierarchical Attention Network), as in *Hierarchical Attention Networks for Document Classification* [19]. Figure 3 shows the basic 2-layer structure of HAN, and reveals how it catches the attention of a whole paragraph. The model consists of two attention-based RNN layers, with each layer consisting of two sub layers, one for encoding and another one for attention catching. Words of sentences after padding are used as input. After they are embedding with `word2vec`, word attention sub layer returns the attention of words in one sentence. This return results will be used as inputs of another attention-based RNN network, which uses the same structure to catch attention of sentence attention. Int this way, our deep model may return the attention of whole paper, and correspondingly check the logicality.

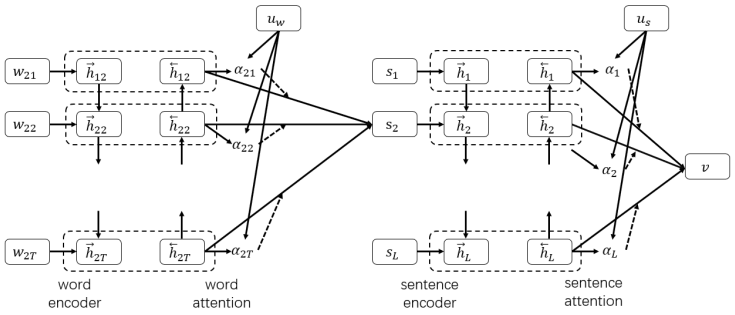


Figure 3: HAN Structure

4 Proposed Method

4.1 Overall Structure

The overall structure is shown as the Figure 4:

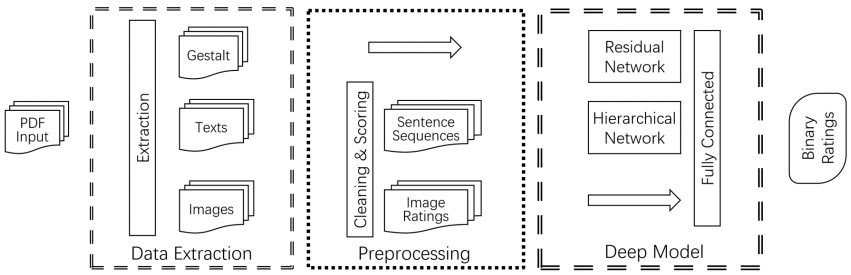


Figure 4: Framework Overall Structure

4.2 Data Extraction

The preprocessing mainly focuses on getting usable and formatted features from individual PDF file. We are developing program in Python to convert full PDF to image "Gestalt", text sequences, image sets and biases.

Gestalt Package `pdf2image` is used to convert PDF input to images, first 8 pages are stacked and down-scaled to the size of 680 $\times$ 440 as "Gestalts" [8].

Sequence of Sentences From the PDF input, a basic parse is performed by `pdfminer.layout`. Text features are extracted and stored as raw strings and processed to calculate vocabulary size. Breaklines, stop words and brackets are removed into a big string. The big string is then splitted by comma to be considered as sequences, after which words are stemmed. To improve efficiency and rule out noise, sentences with more than 50 characters and 13 words are sampled.

Images The PDF input is also processed by `pdfimages` utils to extract images. The images inside PDF files can be extracted as raw inputs (preserving their size and format) and saved into a separate directory. Images that are in single color or too small sized are excluded in calculation.

Biases For individual PDF input, number of pages, number of sequences, mean and variance in length of sequence and vocabulary size are calculated. Number of useful images, size and color of images are examined.

4.3 Preprocessing

We used the following notation for the manipulation of images:

Table 1: Notations and Definitions

Notation	Definition
$\|\cdot\|$	Normalize by Maximum value
Φ	Golden ratio
n_i	Number images
$\mathbf{s_i}$	List of image size of the i^{th} dimension
$\mathbf{s_i}[k]$	Size of image in the i^{th} dimension
$\mathbf{c_j}$	List of image color information in the j^{th} channel
$\mathbf{c_j}[k]$	The k^{th} image color information (matrix) in the j^{th} channel

We follow the convention that for $i = 0, 1$ (representing the image resolution),

$$\bar{\mathbf{s_i}} = \frac{1}{n_p}\sum_{k=0}^{n_i}\mathbf{s_i}[k], \quad \sigma^2_{\mathbf{s_i}} = \frac{1}{n_p}\sum_{k=0}^{n_i}(\|\mathbf{s_i}[k] - \bar{\mathbf{s_i}}\|)^2$$

And for $j = 0, 1, 2$ (the image RGB color channel), the mean & variance value over the channel:

$$\bar{\mathbf{c_j}} = \frac{1}{n_p}\sum_{k=0}^{n_i}\frac{1}{\mathbf{s_1}\mathbf{s_2}}\sum_{\mathbf{s_i}}\mathbf{c_j}[k] \qquad \sigma^2_{\mathbf{c_j}} = \frac{1}{n_p}\sum_{k=0}^{n_i}(\|\frac{1}{\mathbf{s_1}\mathbf{s_2}}\sum_{\mathbf{s_i}}\mathbf{c_j}[k] - \bar{\mathbf{c_j}}\|)^2$$

For the overall rating of size, we define

$$\mathbf{R}_{img} = \frac{1}{\sigma_{\mathbf{s}}}\cdot\frac{(\min_i\{\bar{\mathbf{s_i}}\}/\max_i\{\bar{\mathbf{s_i}}\})}{\Phi}\cdot\frac{n_i}{\sigma_{\mathbf{c}}} \quad \text{where} \quad \sigma_{\mathbf{s}} = \sqrt{\sum_i \sigma^2_{\mathbf{s_i}}} \quad \sigma_{\mathbf{c}} = \sqrt{\sum_i \sigma^2_{\mathbf{c_j}}}$$

The rating considers the golden cut of images and bias the variance, as well as the number of images and uniformity analysis of color.

So far, we already have number of pages, number of sequences, mean and variance in length of sequence and vocabulary size for text and an overal rating for images. They will collaborate with the result of deep model for the use of the final fully connected layer.

4.4 Deep Model

Our design of deep model can be separated into two sub-models: Residual Network (Section 3.1) for "Gestalt" images and Hierarchical Attention Network (Section 3.2) for sentence sequences. The models are customized and a meta fully connected layer wraps both networks up, which allows for an all-inclusive backpropagation through the whole network.

5 Experiment & Result

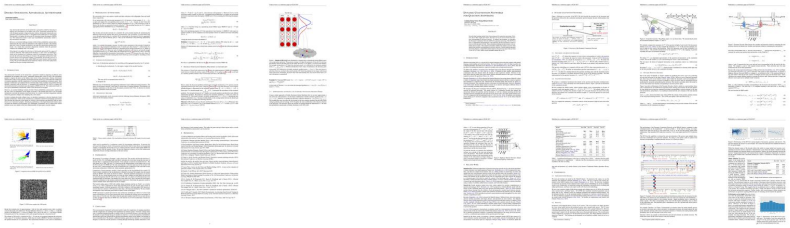
5.1 Data

For this project, we used PDF files of academic papers as input. We considered submissions to *International Conference on Learning Representations (ICLR)* from 2017 to 2019 as our dataset. The files are crawled from OpenReview website, and the statics is shown in Table 2:

Table 2: Dataset statistics

Year \ Type	Oral	Poster	Reject
2017	18	183	245
2018	23	313	486
2019	24	478	1077
Total	1039		1808
Total after cleaning	895		1519

Actual usable dataset after running cleaning (ruling out blank PDF/ignoring paper less than 4 pages/removing files unable to parse) turned out to have 895 accept and 1519 reject samples. One sample from the accept category and one sample from the reject category (Converted as "Gestalt" image) is shown as Figure 5.



(a) Rejected (b) Accepted

Figure 5: Gestalt Samples of ICLR Paper

In previous work by Huang [2018], the dataset used was from CVPR and ICCV from 2013 to 2018, with conference papers as positive class and workshop paper as negative class. However, it's insufficient to distinguish the difference between workshop and conference papers. Many low quality papers are rejected before getting recommended to workshops.

One other limitation come with other major conferences is that they hardly ever release rejected submissions as OpenReview does. This is why our dataset is of relatively small size and we do not generalize the results to other conferences i.e., we should not apply it to a two-column formatted academic paper if we train the framework on ICLR data.

5.2 Baselines

For reference purpose, we also build baseline classifiers in contrast with the framework. The baselines we're choosing come from decomposing our deep model. For image baselinle, we simply take a one layer convolutional network with structure of `Conv-ReLU-Pool-FC`.

For text baseline, we choose to take the basic RNN model with structure `RNN-FC-Sigmoid`. The input sequences are flattened and regarded as single sentence. Besides the basic RNN model, CNN model, as long as model with RNN on word level and CNN on sentence level are implemented. In following parts, the baseline statistics of text come from the basic RNN model.

Ples, we also fused the CNN baseline model and a RNN model to construct a simple fused model.

5.3 Results

For result, we define $Correct\_Accept$ and $Miss\_Accept$ to be the predicted accept class, and All_Accept to be the sum of $Correct\_Accept$ & $Miss\_Accept$. $Should\_Accept$ and $Should\_Reject$ are the ground truth. The metrics we are using are defined in Table 3[1].

Table 3: Metrics Defined for Results

Precision Rate (PR)	$Correct\_Accept/All\_Accept$
Correct Reject Rate (CRR)	$Correct\_Reject/Should\_Reject$
Miss Reject Rate (MRR)	$Miss\_Reject/Should\_Accept$

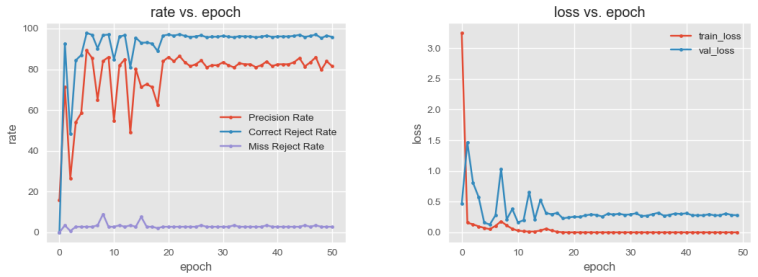
[1]For the metric, $Recall = Correct\_Accept\_Rate = Correct\_Accept/Should\_Accept = 1 - MRR$

The best results we discover for each classifier are shown in Table 4

Table 4: Best Results for Classifiers & Baselines

	CNN(Image)	RNN(Text)	Fused CNN+RNN	Proposed Framework
PR	81.548%	60.317 %	88.158 %	**93.289%**
CRR	95.969%	64.286 %	97.659 %	**98.700%**
MRR	5.517%	45.714 %	7.586 %	**4.138%**

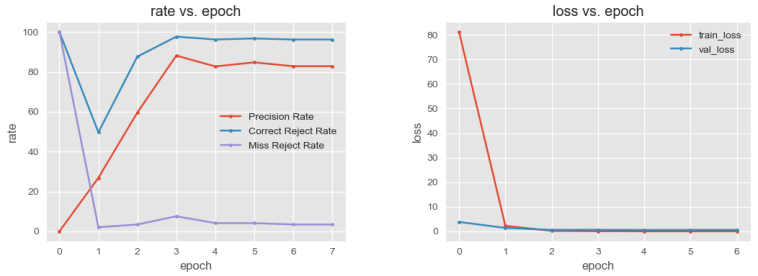
Figure 6 and Figure 7 show the learning curve for proposed model and simple fused structure respectively.



(a) Proposed Model Accuracy Curve for 50 Epoch

(b) Proposed Model Loss Curve for 50 Epoch

Figure 6: Proposed Model Learning Curves for 50 Epoch



(a) Fused Baseline Accuracy Curve for 8 Epoch

(b) Fused Baseline Loss Curve for 8 Epoch

Figure 7: Fused Baseline Learning Curves for 8 Epoch

From the above results, it can be concluded that the proposed framework is proved to be functional and useful to some extent, considering it will be able to correctly reject 98.7% paper of low quality and only sacrifice 4% good papers. There is a small overfit for the train and validation loss. With more training on accepted data sample, the framework is expected to achieve even lower Miss Reject Rate and therefore would be suitable to work as a pre-filter of academic paper submissions.

Data Visualization It's not hard to discover that the the main contribution to the final decision may come from image-based classifier. We would like to go into detail and find out what feature the deep classifier is looking at and therefore adopted the idea of Class Activation Mapping [21], which considers weighted sum of spatial visual features on the "Gestalt" images.

The results on our ResNet are presented for both positive and negative classes:

Figure 8: Heatmap for Accepted Papers

Notice that there's kind of a similar pattern and the network is "looking" at the top left part of the image, yet it does not make much sense.

Figure 9: Heatmap for Rejected Papers

For the negative samples, the patterns are more diverse, and the seemingly common problem comes from the figures. Still, it's not very intuitive enough to gain much useful information regarding what pattern the good paper will have.

6 Conclusion

In this project, we proposed a Text & Vision Fused Framework for academic paper review cooperating with Residual Network and Hierarchical Attention Network. The experiment on paper collections submitted to International Conference on Learning Representations (ICLR) proved that our framework is effective i.e., 98.7% accuracy in filtering out paper of low quality with minimum sacrifice (4%) as mis-reject. Comparing to baseline classifiers, our framework shows its value in improving the final result by considering text and image features in a fused pattern.

As the **FIRST** framework to fuse text & vision features of academic papers for acceptance prediction, our framework demonstrates its novelty as previous works hardly ever address the problem of computer-aided academic paper review process, and a few of the similar works take only "Gestalt" image features and only applied simple computer vision method in classifying images, without addressing their actual contents. Comparing to previous works on text method for essay grading, our framework considers the appearance of the paper as a whole, and is more specialized for the classification of academic papers. This framework also suggests the influence of images and vocabulary set the author adopted in the work, including their quantity and diversity (or uniformity).

In the mean time, we also recognized that this framework is suffering from low text extraction quality, and we have insufficient data samples for deeper models. Hierarchical Attention model is strong in structural analysis of text, but suffers from gradient vanishing and may not be the best suit for this framework. Future work may focus on improving the data extraction quality and keep the structural feature of academic paper, and try to find out a more suitable text classification model.

It's proposed that this framework may also have a online interface for file transfer and tear down each of the model for a detailed scoring factor, which may improve the interpretability of the framework. A generative model may also help to generate what an accept paper would "look like" based on the current discovery on "Gestalt" images.

Author Contribution Statement

We equally distributed our work with specific focus individually. Y. Yang and T. Cai implemented the proposed model (ResNet & HAN), with J. Liu and S. Huang also contributed to HAN implementation; S. Huang implemented several text baselines (RNN/CNN/CRNN) together with J. Liu. Y. Yang created the Fused (CNN+RNN) Baseline with Tongan, who implemented the Vision part (CNN). Tongan implemented the crawler, extraction and preprocessing modules with suggestions from others. All members tested and optimized multiple models. All members contributed to the Progress Report together. T. Cai & Y. Yang designed the poster and wrote the final report draft; J. Liu studied for related works and provided valuable suggestions; S. Huang provided great helps and improvements in formatting.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[2] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, CIKM '98, pages 148–155, New York, NY, USA, 1998. ACM.

[3] J. L. Elman. Finding structure in time. *Cognitive Science*, 14:213–252, 1990.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Sepp Hochreither and Jürgen Schmidhuber. Long short-term memory. 1997.

[7] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[8] Jia-Bin Huang. Deep Paper Gestalt. *arXiv e-prints*, page arXiv:1812.08775, Dec 2018.

[9] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[11] Leah S. Larkey. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 90–95, New York, NY, USA, 1998. ACM.

[12] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989.

[13] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[14] Andrew McCallum and Kamal Nigam. A comparison of event models for naive Bayes text classification. In *Learning for Text Categorization: Papers from the 1998 AAAI Workshop*, pages 41–48, 1998.

[15] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, 1999.

[16] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, March 2002.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[18] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *EMNLP*, pages 1882–1891. The Association for Computational Linguistics, 2016.

[19] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

[20] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 649–657. Curran Associates, Inc., 2015.

[21] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition*, 2016.