Ad Lingua: Text Classification Improves Symbolism Prediction in Image Advertisements

Andrey Savchenko[1,3], **Anton Alekseev**[1], **Sejeong Kwon**[2],
Elena Tutubalina[1], **Evgeniy Miasnikov**[1] and **Sergey Nikolenko**[1,4]
[1]Samsung-PDMI Joint AI Center,
Steklov Institute of Mathematics at St. Petersburg / St. Petersburg, Russia
[2]Samsung Research / Seoul, Korea
[3]National Research University Higher School of Economics / Nizhny Novgorod, Russia
[4]Neuromation OU / Tallinn, Estonia
andrey.v.savchenko@gmail.com, anton.m.alexeyev@gmail.com,
sejeong.kwon@samsung.com, tlenusik@gmail.com,
deeprehension@gmail.com, sergey@logic.pdmi.ras.ru

Abstract

Understanding image advertisements is a challenging task, often requiring non-literal interpretation. We argue that standard image-based predictions are insufficient for symbolism prediction. Following the intuition that texts and images are complementary in advertising, we introduce a multimodal ensemble of a state of the art image-based classifier, a classifier based on an object detection architecture, and a fine-tuned language model applied to texts extracted from ads by OCR. The resulting system establishes a new state of the art in symbolism prediction.

1 Introduction

Visual advertisement, both image and video, can efficiently convey persuasive messages to potential customers. Much of the power of visual advertising comes from multiple ways to interact with the user, embedding messages in both literal and symbolic forms. One of the most complex tasks in ad analysis is *symbol interpretation*, a much harder problem than, for instance, object detection. For example, in the LEE denim jeans brand ad shown in Fig. 1 a human figure is depicted with a lion's head, which might arguably symbolize "courage", "strong character", etc. In real life, lions are rarely seen wearing denim jeans, so analyzing this scene with a machine learning model trained on regular photographs would be quite challenging at least in terms of object detection. Moreover, establishing a direct association between the (possibly detected) lion and human "courage" is a hard problem by itself. Both factors make symbol interpretation difficult. But understanding ads computationally has important applications: decoded messages can help improve ad targeting, tools for message understanding can lead to better descriptions of visual content, and they can also inform the users how exactly they are being persuaded.

Hussain et al. (2017) present a crowdsourced dataset of advertisements, including images and videos. They introduced several annotation tasks: topic detection, sentiment detection, symbol recognition, strategy analysis, slogan annotation, and question answering for texts related to the ads' messages and motivation. In this work we focus on the image-based multi-label classification task for symbols. In this problem, each annotated image in the dataset has several bounding boxes and textual description that can be mapped to a limited number of categories of symbols. Related work has mostly concentrated on this dataset, and it often combines text and images, but problem settings vary widely. In particular, the ADVISE framework embeds text and images in a joint embedding space in the context of choosing human statements describing the ad's message (Ye and Kovashka, 2017). Zhang et al. (2018) introduce various feature extraction methods to capture relations between ad images and text. Ahuja et al. (2018) use a co-attention mechanism to align objects and symbols. Dey et al. (2018) suggest a data fusion approach to combine text and images for topic classification. Multi-modal image+text prediction is an important research field usually concerned with other tasks; for surveys of this field see, e.g., (Zhang et al., 2020; Xipeng et al., 2020).

Proceedings of the 28th International Conference on Computational Linguistics, pages 1886–1892
Barcelona, Spain (Online), December 8-13, 2020

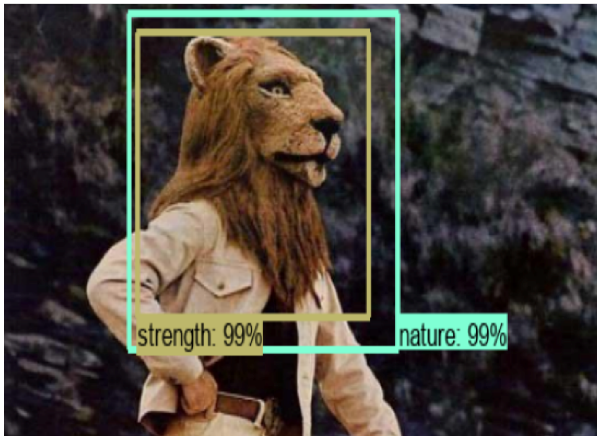


Figure 1: An ad fragment. Symbols annotated by MTurkers: “strength”, “being someone different”, “Courage”, “nature/animal”, “Animal”. Bounding boxes correspond to clustered (53) labels.

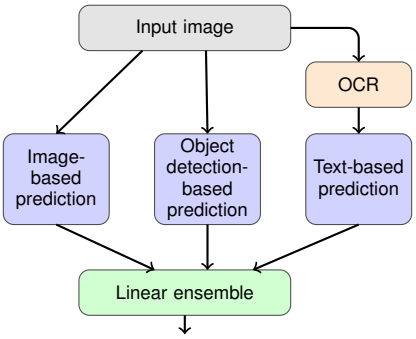


Figure 2: Our approach.

Tesseract (Smith, 2007)	Bite the boredom, TCT am 5 bite-sized treats with a crunchy outside PUR RCo e eB Pree ee Rar ed oe 2 ; RN good
EAST + *Tesseract* (Kopeykina and Savchenko, 2019)	Bite dale boredom, unleash the fun! bite sized treats rh Pe tah outside and delicious aay filled Sita bo ii te STi) me KFc macaroni and cheese. Sl PPPS m)/7-) UC So good
Google Android Vision API	FC ma heese A Bite the boredom, unleash the fun! 5 bite-sized treats with a cr good chy and a delicious center filled with ni and creamy cheese. soft
CloudVision (Otani et al., 2018)	Bite the boredom, unleash the fun! 5 bite-sized treats with a crunchy outside and a delicious center filled with soft macaroni and creamy cheese. KFC ma and heese good KFC
AR-Net Caption	a poster for the film

Figure 3: Sample texts obtained via OCR/captioning. Symbols annotated by MTurkers: “Fun/Party”.

Hussain et al. (2017) remark that “reading the text might be helpful”. Following this and the evidence provided by the winners of the *Towards Automatic Understanding of Visual Advertisements* challenge[1] for a different task on the same dataset (a CVPR 2018 workshop; the task was to choose the correct action-reason statements for a given visual ad), we believe that the textual and visual content of an ad often complement each other to convey the message. In this work we study how to use the text in ads for symbolism prediction along with image-based features.

2 Data

We have used advertisement images from the dataset by Hussain et al. (2017), annotated with the help of MTurkers. In particular, we have used symbolism annotations. In media studies, certain content that stands for a conceptual symbol is called a “signifier” (in the case of this dataset, signifiers are marked with bounding boxes on the image), and the symbols themselves are the “signified” (Williamson, 1978). Hussain et al. (2017) report that the annotators found a total of 13,938 images requiring non-literal interpretation, which was treated as an indicator that the ad contains symbolism. Following the approach of Hussain et al. (2017), we treat symbol prediction as a multi-label classification problem, using the two provided label sets with 221 and 53 labels. In the original paper, the label set has been reduced from 221 to 53 via clustering[2], and the data was split into training/test sets as 80%:20%.

[1] https://people.cs.pitt.edu/~kovashka/ads_workshop/

[2] The latter labels set (“53”): natural, sex, nature, danger, fun, violence, beauty, death, health, adventure, power, love, sexy, sports, environment, speed, fashion, food, energy, injury, strength, safety, youth, travel, hot, entertainment, technology, smoking, family, excitement, healthy, relaxation, art, christmas, refreshing, happiness, fitness, protection, delicious, clean, freedom, vacation, comfort, desire, variety, hunger, strong, humor, party, alcohol, happy, animal cruelty, class.

3 Methods

We propose a composite approach shown in general in Fig. 2, combining features mined from an ad image and from the text extracted on that image. In this section, we elaborate on the methods used on each of these feature extraction steps and in the final ensembling.

Image-Based Prediction. The image-based classifier is a central part of the proposed system. We have trained several state of the art convolutional architectures, namely *MobileNet v1* (Howard et al., 2017), *Inception v3* (Szegedy et al., 2015), and *EfficientNet-B3* (Tan and Le, 2019), for the multi-label classification task with C labels as follows: (1) add a new head with C outputs with sigmoid activations and train this head over 5 epochs with the Adam optimizer (with frozen weights of the baseline model) for the binary cross-entropy loss; (2) train the entire model over 5 epochs with the Adam optimizer; (3) tune the entire model over 3 epochs with the SGD optimizer (learning rate 0.001). The best model, namely *EfficientNet-B3*, was able to obtain F1-score 0.1912 (for the label set of size 221), more than 3% higher when compared to previous state of the art (0.1579) for this task (Hussain et al., 2017). With 53 labels (clusters of symbols), *EfficientNet-B3* with the same training procedure obtains F1-score of 0.2774, also exceeding previous results (0.2684) (Hussain et al., 2017). Thus, even at this level we have already exceeded state of the art by using better convolutional backbones and tuned training schedules.

Object Detection. We have also tried an approach that uses the *location* of signifiers (symbols) on input images to improve symbol recognition. We used the *Faster R-CNN* model (Ren et al., 2015) with *InceptionResNet* backbone trained on *OpenImages v4* with 601 categories (Kuznetsova et al., 2018). Objects were detected in both training and validation sets, and we retained only those objects from the training set that intersect with given symbols with IoU (Intersection over Union) over 0.6 (fixed threshold); the objects were cropped and put in correspondence with symbol labels. We solved multilabel classification for symbol recognition by feeding the images of these objects into pre-trained CNN feature extractors, namely *MobileNets v1/v2* and *EfficientNets B0/B3* (Tan and Le, 2019), with a single dense layer on top of them for classification. The final decision for a test image was made as follows: (1) detect objects using the same *Faster R-CNN*; (2) extract their visual features with pre-trained CNNs; (3) classify the features with the shallow network; (4) unite its outputs (scores or predictions of symbol posterior probabilities) with non-maximal suppression, return only symbols with scores exceeding a given threshold. The F1 measure scores of this pipeline are lower than those of the previous approach. However, the methods are arguably very different, so joining both in a single ensemble might improve the results (see below).

Image-to-text: optical character recognition and captioning. We test our main hypothesis (that text is complementary to visual content) by extracting the text via the following OCR techniques: (1) open source OCR solution *Tesseract* (Smith, 2007), (2) an approach for text processing from (Kopeykina and Savchenko, 2019), in which the bounding boxes of text regions are obtained using the EAST text detector (Zhou et al., 2017) and the words in each region are recognized with Tesseract OCR Engine without text detection ("–oem 1 –psm 13"); (3) Google text recognition library (Google vision API for Android) available at `com.google.android.gms.vision.text`, following the approach by Myasnikov and Savchenko (2019); (4) OCR data from CloudVision used by Otani et al. (2018)[3].

Our experiments indicate that OCR quality has a very significant impact on the output of the predictive system. Unfortunately, the *Tesseract* text detector proved to be far from accurate on the advertisement dataset. It successfully detected text on only half of the images (1182 out of 2084 validation images), much less than commercial solutions: 1739 detected by Google CloudVision API and 1894 provided by Otani et al. (2018). *EAST+Tesseract* approach extracted texts from 1943 validation images, but low recognition quality and arguably random text block order led to inferior prediction results; see Fig. 3 for an illustration. The best-performing texts were those obtained from Otani et al. (2018).

In a different take on extracting text, instead of OCR we have tried to predict symbols based on text obtained via image captioning. The task of generating image captions requires to produce image textual

[3]OCR data on Figshare: `https://figshare.com/articles/dataset/OCR_results/6682709`, GitHub repository: `https://github.com/mayu-ot/ads2018`

descriptions that not only express content information of the input source but also should be naturally coherent. In our case we have chosen this approach for experiments as a completely different way to obtain texts related to the images of interest (Savchenko and Miasnikov, 2020). We have used one of the best-performing image captioning models, namely *AR-Net* (Chen et al., 2018) pre-trained on *Google's Conceptual Captions* (Sharma et al., 2018). *AR-Net* achieves a notable improvement in captioning due to a specific regularization strategy: previous RNN state reconstruction allows the gradient and state information to propagate through time more robustly. We note that the idea of combining captions and recognized texts is potentially fruitful because these texts are usually very different. The DenseCap (Johnson et al., 2016) approach, extracting captioned bounded boxes given an image, might be especially promising. We leave this idea for further study.

Text-based models. We have applied several text classification techniques to OCR results to establish new baselines, training a BERT-based multi-label classification model using the *Simple Transformers* library (Rajapakse, 2020) based on the *Transformers* library (Wolf et al., 2019). We have compared three architectures: (1) BERT (Devlin et al., 2018); we used *bert-base-uncased* from Wolf et al. (2019); (2) RoBERTa (Liu et al., 2019); we used *roberta-base* from Wolf et al. (2019); (3) *Bag-of-Ngrams* baseline, tokenizing extracted texts and preserving only $10{,}000$ most frequent unigrams and bigrams. As a multi-label classification model we have used logistic regression (Pedregosa et al., 2011), one per each label. We have also experimented with multi-output logistic regression trained on SGNS (Mikolov et al., 2013a; Mikolov et al., 2013b) and fastText (Bojanowski et al., 2016) representations, but the prediction quality was clearly worse. In preprocessing, for training we filtered out items without recognized text from the train set and lowercased the text. In the experiments, models were trained in the following settings: 15 training epochs, batch size 16, learning rate 4e-5, and other parameters at default values from Rajapakse (2020). All text-based models perform clearly inferior to image-based ones; this is natural because text is not always present, often short, and even the best OCR methods make quite a lot of mistakes. However, combining image- and text-based approaches can yield significant improvements.

Ensemble. Due to the high risk of overfitting on a small dataset, we have chosen a simple weighted sum of image-based, object detection, and text-based predictions as an ensembling strategy. For an input ad a, each model in the ensemble yields a vector $f(a) \in [0,1]^{|L|}$, where L is a set of labels (classes), for each data point, and the resulting ensemble outputs 1 if $\lambda_{\text{img}} f_{\text{img}}(a) + \lambda_{\text{obj}} f_{\text{obj}}(e) + \lambda_{\text{txt}} f_{\text{txt}}(a) > \theta$, where $f_*(e) \in [0,1]^{|L|}$ are the predictions of individual models described above, and coefficients λ_*, $\sum_* \lambda_* = 1$, and the threshold θ are tunable parameters. For tuning, we have used the predictions of all three "elementary models" (image-, object-, and text-based) on both training and test sets, sampled (5 times) a fraction of the training set (0.1 and 0.05 for 221 and 53 labels respectively; we use the training set due to the small dataset size), then sampled λ_{img}, λ_{obj} and λ_{txt} from a Dirichlet distribution and evaluated the $F1_{\text{macro}}$ (not *micro* as an extra measure against overfitting) on the chosen training set subsample for every θ from the set $\{0.0, 0.05, ...1.0\}$. Then we averaged the 5 sets of parameters. It order to compare against image-only baselines on the whole test set, we also trained a similar blend of image- and object-based classifiers (2 models) and used it as a *backoff* model.

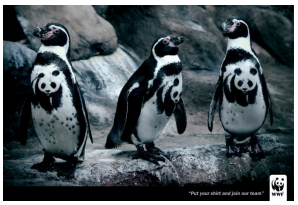
4 Results and Discussion

Table 1 shows the results in terms of F1_{micro}-scores on the test set; text-based and ensemble models are evaluated separately on images where OCR detects something (w/text) and on all ads, falling back on the 2-model ensemble when OCR does not produce anything (all, with backoff).

We see that *Bag-of-Ngrams* is a strong baseline in this task, in some cases even outperforming fine-tuned BERT and RoBERTa. The results confirm our main conclusion: while text-based models alone are hopelessly outmatched even on images with OCR-detected text, they do add a significant boost to image-based models when ensembled together. Another important point is the importance of OCR quality: significant gains are achieved only with the best OCR techniques, while adding text recognized by the basic *Tesseract* fails to achieve meaningful improvements. In general, Table 1 shows that we have significantly improved state of the art results in the symbolism prediction task for both label types.

OCR	Model	Dataset: 221 labels Image-only results: **0.1967**			Dataset: 53 labels Image-only results: **0.2814**		
		Text-based (w/text)	Ensemble (w/text)	Ensemble + Backoff (all)	Text-based (w/text)	Ensemble (w/text)	Ensemble + Backoff (all)
Tesseract	BoN	**0.0881**	**0.1996**	**0.2002**	0.1345	0.2889	0.2865
	BERT	0.0220	0.1932	0.1967	**0.1794**	**0.2935**	**0.2892**
	RoBERTa	0.0220	0.1934	0.1967	0.1765	0.2918	0.2882
EAST + Tesseract	BoN	**0.1198**	**0.2087**	**0.2122**	0.1457	0.2955	0.2957
	BERT	0.0689	0.1949	0.1989	0.1933	**0.3050**	**0.3046**
	RoBERTa	0.0225	0.1975	0.2013	**0.1994**	0.2914	0.2919
AR-Net Captions	BoN	0.1014	0.1918		0.1586	0.2862	
	BERT	**0.1089**	0.1975		0.1615	0.2834	
	RoBERTa	0.0226	**0.2012**		**0.1733**	**0.2871**	
Google Android Vision API	BoN	**0.1407**	**0.2102**	**0.2106**	0.2076	0.3026	0.2994
	BERT	0.0971	0.1975	0.2000	0.2278	0.3054	0.3017
	RoBERTa	0.1201	0.1992	0.2014	**0.2409**	**0.3189**	**0.3128**
CloudVision (Otani et al., 2018)	BoN	**0.1830**	**0.2310**	**0.2292**	0.2420	**0.3186**	**0.3177**
	BERT	0.1520	0.2014	0.2023	0.2653	0.3048	0.3051
	RoBERTa	0.1580	0.2006	0.2016	**0.2898**	0.3074	0.3075

Table 1: $F1_{micro}$-scores; (w/text) — only ads with extracted text; (all) — all ads; BoN — Bag-of-Ngrams.



OCR	CONVERSE There's no lace like home	LIZER BIKE HELMETS	Put your shirt and join our team WWF
Text-based	sex;love;fashion	danger;safety;protection	nature;environment
Image-based	nature;fun;love;sports	danger;violence;death;love;injury	nature;danger;death;environment
Blend	None	danger;violence;death;safety	nature
Ground truth	fun;safety;comfort;humor	injury;safety	environment

Table 2: Three sample results; the text-based model is based on RoBERTa and CloudVision OCR. Listed symbols are separated by semicolons.

Table 2 shows error analysis for sample symbols predicted by three models. First, models predict related symbols due to a vague difference (Examples 2 and 3). Second, advertisers encode general knowledge in images, e.g., a dog might be a metaphorical representation of safety (Example 1). Finally, the ads contain short text fragments that do not describe the picture in detail, e.g., bike helmets prevent head injury, while the image shows a person in hospital (Example 2).

5 Conclusion

We have presented a novel approach to symbol classification on multimodal advertisement data, improving upon state of the art results already with pure image-based approaches and showing further improvements with text-based methods. We have introduced linear ensembles of the developed image-based models and text-based models that operate with OCR-extracted text, demonstrating superior performance for the ensembles and significant improvements over state of the art for both datasets (label types). Possible directions for future work include: (1) OCR postprocessing improvement: spelling correction, noise removal, etc.; (2) enhancing text-based prediction and/or object detection using thesauri, knowledge graphs, e.g., replacing specific entities with hypernyms similar to Ilharco et al. (2019) or word associations datasets, e.g., Wordgame (Louwe, 2020); (3) developing a joint architecture for images and texts: obviously, simple blending might not be the best choice for the task.

Acknowledgements

This research was done at the Samsung-PDMI Joint AI Center at PDMI RAS and supported by Samsung Research. We thank the anonymous referees for insightful comments that helped improve the paper.

References

Karuna Ahuja, Karan Sikka, Anirban Roy, and Ajay Divakaran. 2018. Understanding visual ads by aligning symbols and objects using co-attention. *CoRR*, abs/1807.01448.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and Wei Liu. 2018. Regularizing rnns for caption generation by reconstructing the past with the present. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7995–8003.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Arka Ujjal Dey, Suman K. Ghosh, and Ernest Valveny. 2018. Don't only feel read: Using scene text to understand advertisements. *CoRR*, abs/1806.08279.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1705–1715.

Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. 2019. Large-scale representation learning from visually grounded untranscribed speech. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 55–65.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Lyudmila Kopeykina and Andrey V Savchenko. 2019. Automatic privacy detection in scanned document images based on deep neural networks. In *Proceedings of the International Russian Automation Conference (RusAutoCon)*, pages 1–6. IEEE.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2018. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Anneloes Louwe. 2020. Wordgame: English word associations. Kaggle dataset, `https://www.kaggle.com/anneloes/wordgame/`.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Evgeny Myasnikov and Andrey Savchenko. 2019. Detection of sensitive textual information in user photo albums on mobile devices. In *Proceedings of the International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, pages 0384–0390. IEEE.

Mayu Otani, Yuki Iwazaki, and Kota Yamaguchi. 2018. Unreasonable effectiveness of ocr in visual advertisement understanding. Slides, `http://people.cs.pitt.edu/~kovashka/ads_workshop/otani.pdf`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Thilina Rajapakse. 2020. Simple transformers.

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

Andrey V Savchenko and Evgeniy V Miasnikov. 2020. Event recognition based on classification of generated image captions. In *Proceedings of the International Symposium on Intelligent Data Analysis (IDA)*, pages 418–430. Springer.

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

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Mingxing Tan and Quoc V Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.

Judith Williamson. 1978. *Decoding advertisements: ideology and meaning in advertising.* Marion Boyers.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Qiu Xipeng, Sun TianXiang, Xu Yige, Shao Yunfan, Dai Ning, and Huang Xuanjing. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63:1872–1897.

Keren Ye and Adriana Kovashka. 2017. ADVISE: symbolism and external knowledge for decoding advertisements. *CoRR*, abs/1711.06666.

Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. 2018. Equal but not the same: Understanding the implicit relationship between persuasive images and text. *CoRR*, abs/1807.08205.

Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. 2020. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560.