A Two Step Fine-tuning Approach for Text Recognition on Identity Documents

Francesco Visalli[1][a], Antonio Patrizio[2] and Massimo Ruffolo[1,2][b]
[1]*High Performance Computing and Networking Institute of the National Research Council (ICAR-CNR), Via Pietro Bucci 8/9C, Rende (CS), 87036, Italy*
[2]*Altilia.ai, Piazza Vermicelli, c/o Technest, University of Calabria, Rende (CS), 87036, Italy*

Keywords: Text Recognition, Identity Documents, Transfer Learning, Fine-tuning, Scene Text Recognition, Deep Learning, Transformer.

Abstract: Manually extracting data from documents for digitization is a long, tedious and error-prone job. In recent years, technologies capable of automating these processes are gaining ground and managing to obtain surprising results. Research in this field is driven by the strong interest of organizations that have identified how the automation of data entry leads to a reduction in working time and a speed-up of business processes. Documents of interest are heterogeneous in format and content. These can be natively machine readable or not when they are images obtained by scanning paper. Documents in image format require pre-processing before applying information extraction. A typical pre-processing pipeline consists of two steps: text detection and text recognition. This work proposes a two step fine-tuning approach for text recognition in Italian identity documents based on Scene Text Recognition networks. Experiments show promising results.

1 INTRODUCTION

Feeding enterprise information systems with data coming from documents often requires manual data entry that is a long, tedious, and error-prone job. Nowadays, automatic information extraction can offer a big boost in efficiency, accuracy, and speed in all those business processes where data capture from documents plays an important role. Therefore, automating data entry can save a lot of time and speed-up the execution of business processes allowing employees to focus on core and more valuable aspects of their daily activities.

Organizations in all industries may require to process myriads of documents with a variety of formats and contents. Usual examples of documents of interest are invoices, orders, bills, receipts, payrolls, application forms, passports and other identity documents, and so on. Such documents are often digitized as images, so they require pre-processing before applying information extraction. A typical pre-processing pipeline consists of two steps: text detection and text recognition.

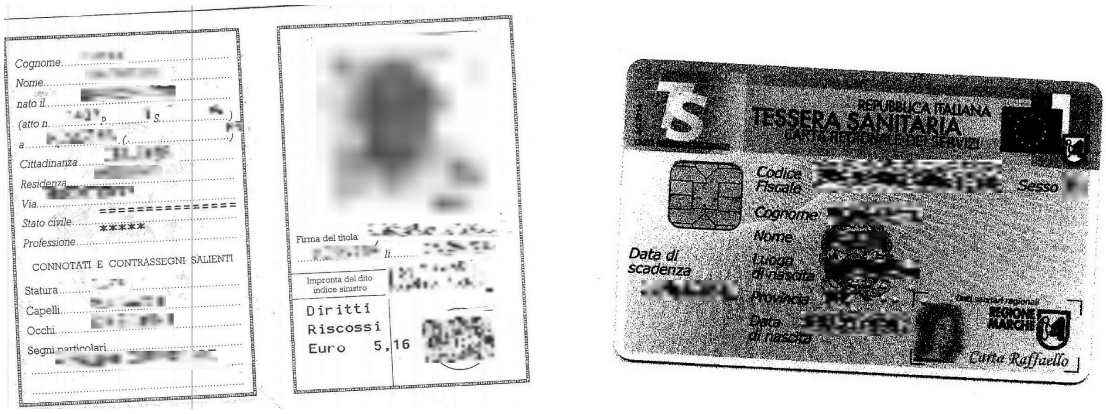
This work addresses the text recognition task for identity documents. This is a particularly difficult problem because while the text detection step can be performed by many different computer vision algorithms (Ren et al., 2017; Baek et al., 2019b; Wang et al., 2019; He et al., 2020), classic Optical Character Recognition (OCR) algorithms (e.g. Textract[1], Tesseract[2] or Calamari[3]), that work well on standard documents such as receipts or invoices, fail in recognizing texts within identity documents. In particular, identity documents are paramount in business processes related to customer subscription and on boarding where users frequently submit photos taken via smartphones, or poor-quality scanned images that are blurred, unclear, and with very complex framing angles.

The text recognition method we present in this work is based on transfer learning and Scene Text Recognition (STR) networks. Transfer learning is a widely used technique within machine learning and, in particular, in the fields of computer vision and natural language processing (Yosinski et al., 2014; Devlin

[a] https://orcid.org/0000-0002-6768-3921
[b] https://orcid.org/0000-0002-4094-4810

[1]https://aws.amazon.com/textract/
[2]https://github.com/tesseract-ocr/tesseract/
[3]https://github.com/Calamari-OCR/calamari/

Visalli, F., Patrizio, A. and Ruffolo, M.
A Two Step Fine-tuning Approach for Text Recognition on Identity Documents.
DOI: 10.5220/0010252208370844
In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) - Volume 2*, pages 837-844
ISBN: 978-989-758-484-8



(a) Italian identity card.

(b) Italian tax code.

Figure 1: Examples of identity documents processed. On the left an Italian identity card, on the right an Italian tax code where sensitive data has been bleared for privacy reasons.

et al., 2019). It relies on the fact that models trained to perform a specific task in a given domain (source) can be applied to do a similar task in another domain (target), thus reducing the need for labeled data in the target domain. STR is a special form of OCR aiming at recognizing text in natural scenes. Text in the wild presents much more complex problems than that found in scanned documents. Some of these problems concern, for example, complex backgrounds, multiple font colors, irregular fonts, different font sizes, different text orientations, imperfect image conditions such as non-uniform illumination, low resolution, motion blurring, and so on. Many of these problems are also found in identity documents. For these reasons, STR networks are ideal candidates to experiment with transfer learning approaches in order to tackle our text recognition problem on identity documents.

To validate the approach we propose we built two datasets, one composed of text extracted from real-world identity documents (i.e. Italian identity cards and tax codes, Figure 1) and another one composed of synthetic text examples randomly generated. We created synthetic texts having same visual and content features of real documents (Figure 2) in order to have more training examples and get better text recognition performances. To train the STR algorithm we adopted a two phase fine-tuning procedure. In the first phase we used synthetic examples, then we used real-world documents to get the final model. In order to prove the effectiveness of our method, we propose a section of ablation studies (Section 4.4). It is important to underline that although we experimented with Italian documents, the approach we propose is general.

The main contributions of this work are:

- the definition of a transfer learning approach for identity documents based on STR algorithms;
- an exhaustive experimental evaluation of the proposed approach that shows the achievement of results that allow the application of the obtained model in real-world scenarios.

The rest of this paper is organized as follows: in Section 2 we introduce and discuss a list of works related to the extraction of text on identity documents; in Section 3 we describe Scene Text Recognition networks and principles underlying our two step fine-tuning approach; in Section 4 we present the creation process of the datasets leveraged for the experiments along with the discussion of the results of these and a series of ablation studies to demonstrate the effectiveness of our method; finally, in Section 5 we draw conclusions and present the future work.

2 RELATED WORK

Information extraction from document images is a research field that is attracting a lot of attention. In the following, we describe works that carry out the whole information extraction process on identity documents.

The work presented by (Lladós et al., 2001) shows that information extraction from identity documents is a well known problem in literature. The authors experimented with Spanish identity cards, Spanish drive licenses and passports achieving 94.50%, 70.00% and 65% of accuracy, respectively. Text regions are preprocessed for background removal and image binarization. OCR is performed as a combination of voting strategy combining several classifiers. The final step of the recognition pipeline consists in a parser designed to consider linguistic and geometric context to overcome typical OCR confusions (e.g. ”D” and ”O” or ”B” and ”8”).

(Bulatov et al., 2017) propose an information extraction pipeline for mobile applications. They focus on Russian driving license and internal Russian passport, achieving an average per-frame accuracy over five field of 89.34%, 92.51% and 94.29% calculated on Russian driving license captured by smartphone, internal Russian passport captured by smartphone and internal Russian passport captured by webcam, respectively. Text recognition is performed by per-character field segmentation (Chernov et al., 2016). Results are very good but multiple frames contain the same document. Moreover, a correction step of the text recognized through language model is performed.

(Viet et al., 2019) present an end-to-end information extraction systems for Vietnamese identity cards. Text bounding boxes are recognized by an encoder-decoder attention model that follows the architecture of STR networks (Shi et al., 2017). Dataset is composed of real examples and data augmentation is performed. The authors reach an average accuracy of 92.00% over fields of interest. However, they only focus on three fields (ID number, name and date of birth), this makes the task easier. Moreover, because of the data augmentation, the distribution of the characters in the fields is almost always the same.

(Attivissimo et al., 2020) propose a method that starts with the classification of the document and ends with the recognition of the text areas detected. They created a synthetic dataset of Italian identity documents composed of paper identity card, electronic identity card, driving license, health insurance card and passport. Text recognition is performed with a typical STR encoder-decoder architecture (Yousef et al., 2018), reaching 92.72% of accuracy on the text fields extracted from the document photos. Although results are good, they are obtained on a synthetic dataset. Typically, these synthetic images can hardly replicate those of real scenarios.

As further evidence of the strong interest in automated information extraction systems from identity documents, in recent years, several commercial systems have emerged[4].

3 METHODOLOGY

In this section we first provide basic concepts about STR networks and their architecture, then we introduce principles underlying our two step fine-tuning method.

[4]https://microblink.com/products/blinkid

3.1 Model

Scene Text Recognition (STR) has a wide range of applications and is attracting a huge interest in the research community. In last years, the use of deep learning methods (Chen and Shao, 2019; Chen et al., 2020), has evolved this field very rapidly. STR networks, due to their nature, are able to recognize text on complex backgrounds while state of the art OCR tools usually fail.

STR deep learning solutions typically follow a standard encoder-decoder architecture (Shi et al., 2019; Baek et al., 2019a). The encoder module performs feature extraction and sequence modeling. The feature extraction is the stage in which the model abstracts an input image and outputs a visual feature map. It is performed by a Convolutional Neural Network or some variant of this (Simonyan and Zisserman, 2015; Lee and Osindero, 2016; He et al., 2016). In order to enlarge the feature context, the extracted features from the previous stage are processed by a sequence model (e.g. Long Short-Term Memory networks).

The decoder module performs the prediction stage. It takes the sequence map output from the encoder and transforms it in a sequence of characters. Predictions are usually made through Connection Temporal Classification (CTC) (Graves et al., 2006) or attention-based mechanism. The former allows predicting a non-fixed number of characters even though a fixed number of features are given, the latter enables a STR model to learn a character-level language model.

In this work we leverage the so called Self-Attention Text Recognition Network (SATRN) (Lee et al., 2019). It follows the encoder-decoder architecture described above. This network is inspired by Transformer (Vaswani et al., 2017), which has deeply revolutionized the NLP field (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019) and also influenced the CV (Computer Vision) world (Parmar et al., 2018). Visual features extracted by the encoder are passed to self-attention modules which are modified version of the original Transformer self-attention modules. The decoder retrieves these enriched features from the encoder and passes them through multi-head attention and point-wise feedforward layers. The output of this process is a sequence of characters.

Due to the full-graph propagation of self-attention mechanism, SATRN is able to recognize text with arbitrary arrangements and large inter-character spacing. This is an important feature for our text recognition task on identity documents. In fact, these kind of documents are often photographed or scanned with

very complex angles. Because of this, text areas result not horizontally aligned unless the document is previously rotated in a pre-processing step. This alignment step is quite common in pipelines for extracting data from document images. Using an STR such as SATRN allows avoiding an explicit phase of text alignment.

3.2 Two Step Fine-tuning Approach

Transfer learning is a widely used technique in the world of deep learning. It allows transferring the knowledge learned by a model from one task to another (or from one domain to another for the same task). Models that leverage the transfer learning need less data than those trained from scratch. It is a must when the architecture is a deep one. In fact, deep learning models usually require lot of data in order to learn a task. Fine-tuning is a transfer learning approach that consists in some minor adjustment of the weights of the network.

SATRN is pretrained on two large datasets which are MJSynth (Jaderberg et al., 2014) and SynthText (Gupta et al., 2016). It is important to note that these two datasets are designed for STR, more details about them are provided in Section 4.1. The key concept here is that SATRN needs to be fine-tuned in order to properly recognize text on identity documents. In fact, unlike problems they share, text areas in identity documents are different from those in natural scenes both from a visual and linguistic point of view.

Starting from real Italian identity documents (Figure 1), our goal is to train SATRN in order to recognize text on them. The language gap is the biggest one to bridge. Character distribution learned on natural text scenes is completely different from that learned on identity documents. Moreover, most of the text in MJSynth and SynthText is in English while the documents we want to recognize are in Italian, thus increasing the language gap. Due to privacy reasons, identity document datasets are not distributed. To provide more examples to the model, we created a big synthetic dataset of text areas. These were created trying to replicate the real ones as much as possible (Figure 2, more about it in Section 4.1). Since synthetic examples can never be like real ones and to avoid overfitting the synthetic distribution, we fine-tuned SATRN in two distinct phases: first with the synthetic dataset and then with the real dataset. In this way we can use more synthetic examples than we could have used by training the network with both synthetic and real ones in a single training set. This is why we talk about two step fine-tuning approach.

4 EXPERIMENTS

In this section we describe experiments carried out in order to train SATRN for the text recognition task on Italian identity documents. First we present datasets used and how we made them. Then, we describe in depth the experiments and the results obtained. The section concludes with some ablation studies demonstrating the effectiveness of our two step fine-tuning approach.

4.1 Datasets

SATRN was originally trained with the combination of MJSynth (Jaderberg et al., 2014) and SynthText (Gupta et al., 2016), which are two very big datasets of Scene Text Recognition. The first contains 9 million text boxes while the second 8 million. As described in Section 3.2, in order to fine-tune the network on identity document texts, our two step approach leverages two different datasets.

The first training step is performed on a synthetic dataset created by us. Our goal was to recognize most of the fields from paper identity cards (front and back) and plastic tax codes (front) for a total of 13 and 8 fields, respectively. Identity cards fields include: name, surname, date of birth, place of birth, citizenship, residence, address, civil status, profession, municipality of issue, date of issue, Identity card code and expiration date. While tax codes fields include: tax code, name, surname, place of birth, county, date of birth, sex and expiration date.

In order to create text areas of interest we leverage a data generator[5]. It takes as input dictionaries of fields and a set of backgrounds. Generality dictionaries were created by taking the fields randomly from the web[6,7], field dictionaries like data were randomly generated, tax codes were calculated by a REST service[8]. In order to reproduce real documents we add some text skewing and text blurring to the boxes (Figure 2). We balanced the examples for each field. We experimented with two synthetic datasets of different sizes: 8886 and 14762.

We further fine-tuned the model on a real-world dataset. Fields are the same of those synthetics, described before. The size of the dataset is 714. To extract text areas from documents we leveraged object detection algorithms (Ren et al., 2017; He et al., 2020). Because we work on documents having a stan-

[5]https://github.com/Belval/TextRecognitionDataGenerator
[6]https://github.com/napolux/paroleitaliane
[7]https://github.com/napolux/italia
[8]http://webservices.dotnethell.it/codicefiscale.asmx

dard layout, the object detection model can be trained with few examples.

(a) Real example cropped from an Italian identity card.

(b) Real example cropped from an Italian tax code.

(c) Synthetic example similar to those cropped from real Italian identity cards.

(d) Synthetic example similar to those cropped from real Italian tax codes.

Figure 2: Comparison between real and synthetic text boxes.

4.2 Experimental Setting

We leveraged the implementation of SATRN proposed in vedastr[9]. SATRN was originally tested with three different architectures: SATRN-small, SATRN-middle and SATRN-big. Differences are in the channel dimension per layer, encoder and decoder layers. We chose SATRN-small which guarantees a good trade-off between accuracy and efficiency. The number of channel dimension per layer, encoder and decoder layers in SATRN-small is 256, 9 and 3, respectively. The model was trained using default hyperparameters: Adam (Kingma and Ba, 2015) as optimizer, 16 as batch size dimension and $1e-4$ as initial learning rate.

We trained our models for 6 epochs taking each time the best one in terms of validation loss and metrics. In order to evaluate the models we used two different metrics: the case sensitive accuracy at bounding box level (i.e. a bounding box is correct if all the characters inside it are correctly recognized) and the mean character error rate (mCER). The CER is defined in Equation 1, where $ed(s1,\ s2)$ is the edit distance of two sequences $s1$ and $s2$ normalized by the maximum length (Wick et al., 2018). The learning policy is different from that of the original work. Instead of using Cyclic learning rate (Smith, 2017) we used a simple weight decay setting the gamma at 0.1 and the decay at the second and fourth epoch.

$$CER = \frac{ed(s1,\ s2)}{max(|s1|,\ |s2|)} \tag{1}$$

[9]https://github.com/Media-Smart/vedastr

4.3 Results and Discussion

All results presented from now on are calculated through 5-fold cross validation. For each fold, we took the 80% of the original size of the dataset as training set and the remaining 20% as validation set. Unfortunately, a direct comparison on public identity documents is not possible because datasets are not released for privacy reasons.

Table 1 shows the results of the two step fine-tuning approach. On the rows there are pairs of dataset used for each experiment. $SYNTHETIC_{SMALL} \times REAL$ is the model first fine-tuned on the smaller synthetic dataset (the one that contains 8886 examples) and then fine-tuned again on the real dataset, whereas $SYNTHETIC_{LARGE} \times REAL$ is the model first fine-tuned on the larger synthetic dataset (the one that contains 14762 examples) and then fine-tuned again on the real dataset. On the columns there are results expressed in accuracy and mean character error rate (mCER).

Table 1: Results of the two step fine-tuning approach.

	Accuracy	mCER
$SYNTHETIC_{SMALL} \times REAL$	88.44%	2.71%
$SYNTHETIC_{LARGE} \times REAL$	86.98%	3.02%

We can observe that the model trained on the smaller synthetic dataset during the first step performs better than the model trained on the larger, both in terms of accuracy and in terms of mCER. Usually, the larger is the dataset the better are results. This conflicting behaviour could be due to the fact that during the first fine-tuning the model overfits the synthetic distribution when trained on more examples. It could be a first clue that shows us how synthetic datasets are different from real ones even if very similar.

(a) Example of poor quality text box.

(b) Example of outlier.

Figure 3: Examples on which the model makes mistakes.

Studying the predictions of the model we noticed that most of the errors are made on very noisy examples or on examples with character distributions completely different from all those seen in both the synthetic and the real dataset (3). We are confident

that by increasing the size of the real dataset results can be further improved. It is noteworthy that results presented here are obtained without any kind of image pre-processing to reduce noise. Results can be further improved by applying post-processing techniques such as OCR correction or dictionaries matching. Some of the errors made by the model could be easily corrected applying some of the techniques just mentioned. For example it could be very easy and effective match some fields (i.e. name of people) on dictionaries.

4.4 Ablation Studies

In this section we present a series of ablation studies to demonstrate the effectiveness of our approach. First we carried out experiments to demonstrate how synthetic datasets are different from those real. Table 2 shows the evaluation of the models trained on the synthetic datasets without the fine-tuning on the real one. Results are calculated on the five folds of the real dataset. On the rows there are the datasets and on the columns the metrics. We can observe far worse results than those obtained after performing a further fine-tuning on the real dataset. Best results are obtained on the smaller synthetic dataset following the trend of the experiments in Section 4.3.

Table 2: Results of the models fine-tuned only on the synthetic datasets calculated on the validation folds of the real one.

	Accuracy	mCER
$SYNTHETIC_{SMALL}$	77.31%	6.80%
$SYNTHETIC_{LARGE}$	73.95%	7.92%

Table 3 has the same structure of previous ones. The first row shows results of the evaluation of SATRN on the five folds of the real dataset without any fine-tuning. Results confirm the need and the effectiveness of transfer learning from natural scenes to identity documents domain.

The second row shows results of the fine-tuning of SATRN only on the real dataset. They are comparable to those in Table 1. We can observe that the model fine-tuned only on the real dataset performs way better than the models fine-tuned only on the synthetic datasets (Table 2). This once again confirms the importance of having real data. However, best results are those presented in Table 1 obtained performing the training both on the synthetic dataset and then on the real one. This confirms the effectiveness of our approach.

Table 3: Evaluation of SATRN from scratch and results of the model fine-tuned only on the real dataset.

	Accuracy	mCER
$NO\ FINE-TUNING$	39.69%	23.39%
$REAL$	85.15%	3.47%

5 CONCLUSIONS AND FUTURE WORK

In this work we proposed a transfer learning procedure to tackle the problem of text recognition on identity documents. To do this we leveraged Scene Text Recognition (STR) networks. In order to better transfer knowledge from text in natural scenes to those within Italian identity documents we carried out a two step fine-tuning approach.

We created two datasets of text areas, one composed of cropped text extracted from real-world examples through object detection algorithms and another one composed of synthetic text examples randomly generated. We created synthetic texts having the same visual and content features of real documents. Therefore, we first trained the STR network on the synthetic dataset and then on the real one.

We achieved an accuracy of 88.44% (considering a text box correct if all the characters inside it are correctly recognized) and a mean character error of 2.71%. Results reported are obtained directly from the network without any form of pre or post-processing. Finally, we demonstrated the effectiveness of our approach through a section of ablation studies.

We plan to perform further experiments extending datasets with examples on which the network misbehaves. We also intend to experiment with some pre and post-processing techniques in order to improve input images and correct output results. These experiments will include per-field analysis. Afterwards, we want to integrate the work presented here into a complete pipeline of information extraction for identity documents.

REFERENCES

Attivissimo, F., Giaquinto, N., Scarpetta, M., and Spadavecchia, M. (2020). An Automatic Reader of Identity Documents. *CoRR*, abs/2006.14853.

Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S. J., and Lee, H. (2019a). What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis. In *International Conference on Computer Vision (ICCV)*.

Baek, Y., Lee, B., Han, D., Yun, S., and Lee, H. (2019b). Character Region Awareness for Text Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9365–9374. Computer Vision Foundation / IEEE.

Bulatov, K. B., Arlazarov, V. V., Chernov, T. S., Slavin, O., and Nikolaev, D. P. (2017). Smart IDReader: Document Recognition in Video Stream. In *7th International Workshop on Camera-Based Document Analysis and Recognition, 14th IAPR International Conference on Document Analysis and Recognition, CBDAR@ICDAR 2017, Kyoto, Japan, November 9-15, 2017*, pages 39–44. IEEE.

Chen, X., Jin, L., Zhu, Y., Luo, C., and Wang, T. (2020). Text Recognition in the Wild: A Survey. *CoRR*, abs/2005.03492.

Chen, Y. and Shao, Y. (2019). Scene Text Recognition Based on Deep Learning: A Brief Survey. In *2019 IEEE 11th International Conference on Communication Software and Networks (ICCSN)*, pages 688–693.

Chernov, T., Ilin, D., Bezmaternykh, P., Faradzhev, I., and Karpenko, S. (2016). Research of Segmentation Methods for Images of Document Textual Blocks Based on the Structural Analysis and Machine Learning. *Vestnik RFFI*, pages 55–71.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Graves, A., Fernández, S., Gomez, F. J., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Cohen, W. W. and Moore, A. W., editors, *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 369–376. ACM.

Gupta, A., Vedaldi, A., and Zisserman, A. (2016). Synthetic Data for Text Localisation in Natural Images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2315–2324. IEEE Computer Society.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. (2020). Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):386–397.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition. *CoRR*, abs/1406.2227.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Lee, C. and Osindero, S. (2016). Recursive Recurrent Nets with Attention Modeling for OCR in the Wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2231–2239. IEEE Computer Society.

Lee, J., Park, S., Baek, J., Oh, S. J., Kim, S., and Lee, H. (2019). On Recognizing Texts of Arbitrary Shapes with 2D Self-Attention. *CoRR*, abs/1910.04396.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Lladós, J., Lumbreras, F., Chapaprieta, V., and Queralt, J. (2001). ICAR: Identity Card Automatic Reader. In *6th International Conference on Document Analysis and Recognition (ICDAR 2001), 10-13 September 2001, Seattle, WA, USA*, pages 470–475. IEEE Computer Society.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. (2018). Image Transformer. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4052–4061. PMLR.

Ren, S., He, K., Girshick, R. B., and Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.

Shi, B., Bai, X., and Yao, C. (2017). An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304.

Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., and Bai, X. (2019). ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2035–2048.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Smith, L. N. (2017). Cyclical Learning Rates for Training Neural Networks. In *2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017*, pages 464–472. IEEE Computer Society.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R.,

Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Viet, H. T., Hieu Dang, Q., and Vu, T. A. (2019). A Robust End-To-End Information Extraction System for Vietnamese Identity Cards. In *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 483–488.

Wang, X., Jiang, Y., Luo, Z., Liu, C., Choi, H., and Kim, S. (2019). Arbitrary Shape Scene Text Detection With Adaptive Text Region Representation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6449–6458. Computer Vision Foundation / IEEE.

Wick, C., Reul, C., and Puppe, F. (2018). Calamari - A high-performance tensorflow-based deep learning package for optical character recognition. *CoRR*, abs/1807.02004.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 5754–5764.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3320–3328.

Yousef, M., Hussain, K. F., and Mohammed, U. S. (2018). Accurate, Data-Efficient, Unconstrained Text Recognition with Convolutional Neural Networks. *CoRR*, abs/1812.11894.