CCpdf: Building a High Quality Corpus for Visually Rich Documents. . .

15

total. We decided to limit ourselves to one dump only due to computational and

storage limitations. One with enough computing resources may easily reproduce

our pipeline and create a corpus up to 84 times larger.

6

Conclusions

Large corpora of documents are crucial for 2D language model pretraining. Re-

cent approaches to their creation have had limitations in terms of diversity and

multilinguality. Diversity of the dataset is a crucial property, as data used in the

training phase impact the biases of the model. Efficient design of a pipeline for

creating such a corpus has not been studied before. In this work we addressed

those limitations by designing a process of downloading diversified samples of

PDFs and their efficient processing. To obtain documents we used Common

Crawl, which is a popular source of data for language model pretraining, but

has rarely been used in the context of 2D language models. The PDF files used

for this project were balanced across languages and domains, which guarantees

diversity with respect to layouts and topics. To make the processing pipeline

efficient in terms of computing time and data quality, we tested different strate-

gies of OCR processing, i.e. usage of the embedded textual layer for documents

not requiring OCR, and predefining the OCR language. The language detection

step was also carefully analyzed.

The result of this work is an index of PDF files with their URL addresses and

metadata, and the script for downloading it is available at our repository27 . The

supplied data were analyzed in terms of not only document length and layout,

but also metadata connected to the PDF format (i.e., the PDF version and the

creator tool), which can help understand better the dataset itself, but also give

an insight into the content of the Internet.

Acknowledgments

The Smart Growth Operational Programme partially supported this research un-

der projects no. POIR.01.01.01-00-0877/19-00 (A universal platform for robotic

automation of processes requiring text comprehension, with a unique level of

implementation and service automation) and POIR.01.01.01-00-1624/20 (Hiper-

OCR - an innovative solution for information extraction from scanned docu-

ments).

References

1. Abadji, J., Suarez, P.O., Romary, L., Sagot, B.: Towards a cleaner document-

oriented multilingual crawled corpus. ArXiv abs/2201.06642 (2022)

27

https://github.com/applicaai/CCpdf16

M. Turski et al.

2. Allison, T., Burke, W., Constantinou, V., Goh, E., Mattmann, C., Mensikova, A.,

Southam, P., Stonebraker, R., Timmaraju, V.: Research report: Building a wide

reach corpus for secure parser development. In: 2020 IEEE Security and Privacy

Workshops (SPW). pp. 318–326 (2020). https://doi.org/10.1109/SPW50608.

2020.00066

3. Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey,

D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier,

S., Lo, K., Murray, T., Ooi, H.H., Peters, M., Power, J., Skjonsberg, S., Wang,

L.L., Wilhelm, C., Yuan, Z., van Zuylen, M., Etzioni, O.: Construction of the

literature graph in semantic scholar. In: Proceedings of the 2018 Conference of the

North American Chapter of the Association for Computational Linguistics: Human

Language Technologies, Volume 3 (Industry Papers). pp. 84–91. Association for

Computational Linguistics, New Orleans - Louisiana (Jun 2018). https://doi.

org/10.18653/v1/N18-3011, https://aclanthology.org/N18-3011

4. Biten, A.F., Tito, R., Gomez, L., Valveny, E., Karatzas, D.: OCR-IDR: OCR an-

notations for industry document library dataset. arXiv preprint arXiv:2202.12985

(2022)

5. Borchmann, Ł., Pietruszka, M., Stanisławek, T., Jurkiewicz, D., Turski, M., Szyn-

dler, K., Gralinski, F.: DUE: End-to-end document understanding benchmark. In:

NeurIPS Datasets and Benchmarks (2021)

6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Nee-

lakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A.,

Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C.,

Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner,

C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are

few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H.

(eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901.

Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/

file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

7. Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D.,

Mitchell, M., Gardner, M.: Documenting large webtext corpora: A case study

on the colossal clean crawled corpus. In: Proceedings of the 2021 Conference

on Empirical Methods in Natural Language Processing. pp. 1286–1305. Asso-

ciation for Computational Linguistics, Online and Punta Cana, Dominican Re-

public (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.98, https:

//aclanthology.org/2021.emnlp-main.98

8. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He,

H., Thite, A., Nabeshima, N., Presser, S., Leahy, C.: The Pile: An 800gb dataset

of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020)

9. Garncarek, Ł., Powalski, R., Stanisławek, T., Topolski, B., Halama, P., Turski,

M., Graliński, F.: LAMBERT: Layout-aware language modeling for informa-

tion extraction. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) Document Anal-

ysis and Recognition – ICDAR 2021. pp. 532–547. Springer International Pub-

lishing, Cham (2021). https://doi.org/10.1007/978-3-030-86549-8\_34, https:

//link.springer.com/content/pdf/10.1007%2F978-3-030-86549-8\_34.pdf

10. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D.,

Smith, N.A.: Don’t stop pretraining: Adapt language models to domains and tasks.

Proceedings of the 58th Annual Meeting of the Association for Computational

Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.740, http://

dx.doi.org/10.18653/v1/2020.acl-main.740CCpdf: Building a High Quality Corpus for Visually Rich Documents. . .

17

11. Habernal, I., Zayed, O., Gurevych, I.: C4Corpus: Multilingual web-size corpus with

free license. In: Proceedings of the Tenth International Conference on Language

Resources and Evaluation (LREC’16). pp. 914–922. European Language Resources

Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/

L16-1146

12. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for

document image classification and retrieval. In: ICDAR (2015)

13. Huber, P., Aghajanyan, A., Oğuz, B., Okhonko, D., tau Yih, W., Gupta, S., Chen,

X.: CCQA: A new web-scale question answering dataset for model pre-training

(2021)

14. Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., Heard, J.: Building

a test collection for complex document information processing. In: Proceedings of

the 29th Annual International ACM SIGIR Conference on Research and Devel-

opment in Information Retrieval. p. 665–666. SIGIR ’06, Association for Comput-

ing Machinery, New York, NY, USA (2006). https://doi.org/10.1145/1148170.

1148307, https://doi.org/10.1145/1148170.1148307

15. Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: Docbank: A benchmark

dataset for document layout analysis (2020)

16. Liu, V., Curran, J.R.: Web text corpus for natural language processing. In: 11th

Conference of the European Chapter of the Association for Computational Lin-

guistics. Association for Computational Linguistics, Trento, Italy (Apr 2006),

https://www.aclweb.org/anthology/E06-1030

17. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M.,

Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining

approach (2019)

18. Luccioni, A.S., Viviano, J.D.: What’s in the box? An analysis of undesirable content

in the Common Crawl corpus. In: ACL (2021)

19. Masson, C., Paroubek, P.: NLP analytics in finance with DoRe: A French 250M to-

kens corpus of corporate annual reports. In: Proceedings of The 12th Language Re-

sources and Evaluation Conference. pp. 2261–2267. European Language Resources

Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/

2020.lrec-1.275

20. Patterson, D.A., Gonzalez, J., Le, Q.V., Liang, C., Munguía, L.M., Rothchild, D.,

So, D.R., Texier, M., Dean, J.: Carbon emissions and large neural network training.

ArXiv abs/2104.10350 (2021)

21. Powalski, R., Borchmann, Ł., Jurkiewicz, D., Dwojak, T., Pietruszka, M., Pałka,

G.: Going full-TILT boogie on document understanding with text-image-layout

transformer. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) Document Analy-

sis and Recognition – ICDAR 2021. pp. 732–747. Springer International Pub-

lishing, Cham (2021). https://doi.org/10.1007/978-3-030-86331-9\_47, https:

//link.springer.com/content/pdf/10.1007%2F978-3-030-86331-9\_47.pdf

22. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A.: Imagebert: Cross-modal

pre-training with large-scale weak-supervised image-text data (2020)

23. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li,

W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text

transformer (2019)

24. Schwenk, H., Wenzek, G., Edunov, S., Grave, E., Joulin, A.: CCMatrix: Mining

billions of high-quality parallel sentences on the web. In: ACL (2021)

25. Smith, J.R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., Lopez,

A.: Dirt cheap web-scale parallel text from the Common Crawl. In: Proceedings of18

M. Turski et al.

the 51st Annual Meeting of the Association for Computational Linguistics (Volume

1: Long Papers). pp. 1374–1383. Association for Computational Linguistics, Sofia,

Bulgaria (Aug 2013), https://www.aclweb.org/anthology/P13-1135

26. Smith, R.: Tesseract Open Source OCR Engine (2022), https://github.com/

tesseract-ocr/tesseract

27. Turc, I., Chang, M.W., Lee, K., Toutanova, K.: Well-read students learn better: On

the importance of pre-training compact models. arXiv preprint arXiv:1908.08962v2

(2019)

28. Wenzek, G., Lachaux, M.A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A.,

Grave, E.: CCNet: Extracting high quality monolingual datasets from web crawl

data (2019)

29. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C.,

Che, W., Zhang, M., Zhou, L.: LayoutLMv2: Multi-modal pre-training for visually-

rich document understanding. In: ACL-IJCNLP 2021 (January 2021)

30. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-training of

text and layout for document image understanding. In: Proceedings of the 26th

ACM SIGKDD International Conference on Knowledge Discovery & Data Mining

(2020)

31. Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., Wei, F.:

Layoutxlm: Multimodal pre-training for multilingual visually-rich document un-

derstanding. arXiv preprint arXiv:2004.2104.08836 (2021)

32. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua,

A., Raffel, C.: mT5: A massively multilingual pre-trained text-to-text transformer.

In: NAACL (2021)

33. Zhong, X., Tang, J., Yepes, A.J.: Publaynet: largest dataset ever for document lay-

out analysis. In: 2019 International Conference on Document Analysis and Recog-

nition (ICDAR). pp. 1015–1022. IEEE (Sep 2019). ht