Lacuna Reconstruction: Self-supervised Pre-training for Low-Resource Historical Document Transcription

Anonymous ACL submission

Abstract

We present a self-supervised pre-training approach for learning rich visual language representations for both handwritten and printed historical document transcription. After supervised fine-tuning of our pre-trained encoder representations for low-resource document transcription on two languages, (1) a heterogeneous set of handwritten Islamicate manuscript images and (2) early modern English printed documents, we show a meaningful improvement in recognition accuracy over the same supervised model trained from scratch with as few as 30 line image transcriptions for training. Our masked language model-style pre-training strategy, where the model is trained to be able to identify the true masked visual representation from distractors sampled from *within the same line*, encourages learning robust contextualized language representations invariant to scribal writing style and printing noise present across documents.

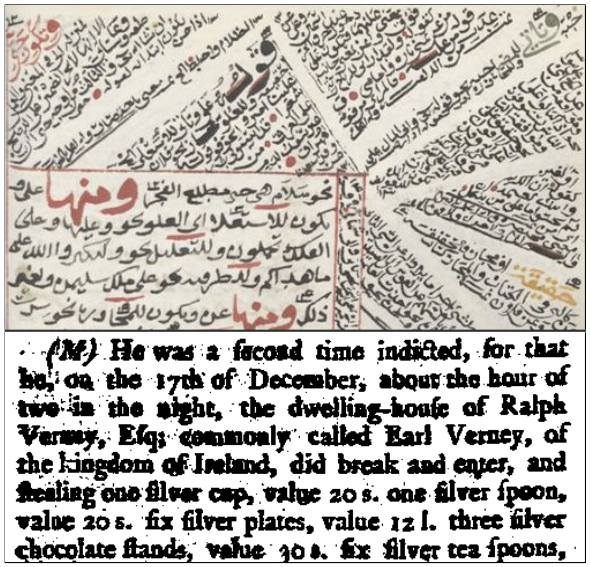

Figure 1: Example page image crops from an Islamicate manuscript dated to 1842 (**Top**, ref: Leiden Or. 669), showcasing its dense, visual complexity with extensive marginalia, and printed proceedings of London's Old Bailey Courthouse (**Bottom**, c. 18th century) (Shoemaker, 2005).

1 Introduction

Document transcription is the task of converting images of handwritten or printed text into a symbolic form suitable for indexing, searching, and computational analysis.[1] Historical documents, whether they were (re)produced via handwriting or the early printing press, confound current statistical document transcription models due to (1) extremely varied style and content across domains, (2) the presence of noise, and (3) a dearth of labeled data.

First, historical printed documents, such as books produced from early modern England (c. 16th–18th centuries; bottom of Fig. 1), use non-standardized spacing and fonts (Shoemaker, 2005) and can contain code-switching that confuses language models (Garrette et al., 2015). However, this variation pales in comparison to their handwritten counterparts. For instance, pre-modern Islamicate manuscripts (i.e., Persian and Arabic

[1] We use the generic term *document transcription* to refer to both the task of optical character recognition (OCR), which is typically reserved for *printed* documents, and handwritten text recognition (HTR) for manuscripts.

handwritten documents from c. 7th–19th centuries; top of Fig. 1), differ in script family, scribal handwriting style, and symbol inventory/vocabulary. As a result, a large degradation in performance is observed when evaluating HTR models on unseen manuscripts (Jaramillo et al., 2018).

Production and imaging noise also present a problem for historical document transcription models. Whether it be uneven inking from a printing press, inconsistent text baselines, or holes resulting from insect damage to ancient pages, techniques must be designed to cope with the noise (Berg-Kirkpatrick and Klein, 2014; Goyal et al., 2020).

While neural networks have a demonstrated capability to model complex data distributions, they typically require large amounts of supervised training data to do so, which is infeasible for historical documents. Unsupervised, non-neural transcription models with fewer parameters alleviate the need to create labeled data (Berg-Kirkpatrick et al., 2013), but struggle with complex handwriting variation. For Islamicate manuscripts, ground truth transcription often requires paleography experts to decipher the ancient writing systems as they appear in each scribal writing style.

In this paper, we propose a self-supervised learning framework designed to overcome these three challenges presented by historical documents. Inspired by the astounding success of self-supervised pre-training techniques for masked language modeling (MLM) in NLP (Devlin et al., 2019), visual models (Chen et al., 2020; Radford et al., 2021), and speech recognition (Baevski et al., 2020), our approach pre-trains a neural text line-image encoder by learning to distinguish masked regions of unlabeled line images from other distractor regions. Specifically, our contribution is the following:

- we show that the recent pre-train/fine-tune paradigm is particularly advantageous for low-resource historical document transcription, obtaining large improvements in both printed and handwritten documents in both English and Arabic-script languages.
- we motivate the self-supervised contrastive loss for document transcription through the lens of "lacuna reconstruction", where blank parts of a document called lacuna must be inferred by human readers.

In doing so, we argue that our approach to pre-training implicitly incentivizes the model to discover and encode discrete character classes in its internal representations, while ignoring style differences occurring in lines using different fonts, languages, or authored by other scribes.

2 Related Work

Masked Pre-training Our approach to self-supervised pre-training follows a growing body of work in both NLP and speech that leverages mask-predict objectives for learning useful, task-agnostic language representations from unlabeled data. In the self-supervised pre-train/supervised fine-tune paradigm, these representations can then be updated on the task of interest using in-domain labeled data. Past work covers learning representations for NLP from monolingual and multilingual text (Devlin et al., 2019; Yang et al., 2019), speech (Baevski et al., 2019; Jiang et al., 2019; Song et al., 2020; Wang et al., 2020), and images grounded with text (Radford et al., 2021). Representations can be learned either through reconstruction objectives (Jiang et al., 2019; Song et al., 2020; Wang et al., 2020) as opposed to a probabilistic contrastive loss (Oord et al., 2018; Baevski et al., 2019, 2020). Most similar to our work is wav2vec2.0 (Baevski et al., 2020), which uses the same two phase training setup with a self-supervised contrastive loss during pre-training and Connectionist Temporal Classification (CTC) loss on transcribed speech data during fine-tuning. Talnikar et al. (2020) presents that the self-supervised loss regularizes the supervised loss during joint learning of both objectives. Follow up work has shown the usefulness of the pre-trained speech representations for exploring speech variation (Bartelds et al., 2020). In this paper, we show that the same learning paradigm can also be successfully applied to very low resource document transcription settings.

Islamicate HTR While machine recognition of handwritten, historic English/German documents can range from 5–12% character error rate (CER) on a sufficient amount of in-sample manuscript training data (Sánchez et al., 2019), performance on Arabic-script languages is much more challenging, leading to substantially higher CER. Pre-modern Islamicate manuscripts (i.e.,

Persian and Arabic handwritten documents from c. 7th–19th centuries), often differ in script family, scribal handwriting style, and symbol inventory/vocabulary. In the top of Figure 1, we present an extreme example of some of the problematic visual variation that can be observed. Even a model trained in a supervised fashion on such a complex document sees a large degradation in performance when evaluating HTR models on unseen manuscripts (Jaramillo et al., 2018) . Until recently, OCR performance on Arabic-script *printed* texts was still poor, typically above 25% CER (Alghamdi and Teahan, 2017), which is too high for downstream users (i.e., researchers and librarians).

Recent studies involving Islamicate manuscripts found that state-of-the-art systems are only able to achieve 40 to mid-20% CER using proprietary software (e.g., Google Cloud Vision, RDI, Transkribus) (Clausner et al., 2018; Keinan-Schoonbaert, 2020, 2019). However, results from these studies only report in-domain performance—an unrealistic scenario where considerable amounts of labeled data can be obtained to enable both training and testing on the same manuscript. In contrast, out-of-domain performance tends to suffer considerably, supported by Romanov et al. (2017)'s study of neural OCR for printed Arabic-script documents. Our work aims to address such performance issues for both in-domain and out-of-domain Islamicate HTR settings by learning general, content-rich pre-trained language representations from large amounts of heterogeneous unlabeled data.

Historical OCR Closely related to manuscript transcription, OCR is another task involving language recognition from images. However, OCR operates on documents that have been printed by a machine with regular, re-used character fonts exhibiting much less superficial glyph variation than human handwriting. OCR is far from a solved problem in the case of documents printed on early modern (c. 16th–18th centuries; see bottom of Fig. 1), movable-type printing presses, where humans would manually set metal type casts with non-standard spacing and fonts (Shoemaker, 2005). In this setting, inking noise and historical font shapes confuse OCR models trained on modern, computer-generated documents (Arlitsch and Herbert, 2004). Berg-Kirkpatrick et al. (2013)'s Ocular explicitly uses a generative probabilistic model inspired by historical printing processes to model such noise. Later work has extended it to handle more typesetting noise (Berg-Kirkpatrick and Klein, 2014), code-switched documents (Garrett, 2014), and produce both diplomatic and normalized transcriptions (Garrette and Alpert-Abrams, 2016). Separately, OCR post-correction models have been proposed to resolve OCR outputs in historical documents (Hämäläinen and Hengchen, 2019; Dong and Smith, 2018) and other low-resource settings (Rijhwani et al., 2020, 2021). In contrast, our approach pre-trains the visual language recognition model's encoder, which produces better contextualized representations in order to reduce the amount of errors the model itself makes. Unlike Ocular, our proposed method does not use a language model and is not fully unsupervised as we require 1-3 pages of transcribed data for learning to transcribe during fine-tuning.

3 Approach

When human readers encounter a lacuna, a blank information gap in a portion of a book or manuscript, they must infer its latent meaning using nearby context like in a cloze test (Taylor, 1953). We argue that the most useful information for inference lies in the ability to reason about the identities of the missing characters in the lacuna using the identities of the surrounding characters. Indeed, MLM-style pre-training techniques are also motivated by the idea of the cloze test, and recent research indicates that language representations learned through the prediction of missing content using surrounding sentential context are useful for many downstream tasks (Devlin et al., 2019; Clark et al., 2019, 2020). Our approach combines the ideas of lacuna inference and masked pre-training to provide a useful learning signal for downstream historical document transcription, a setting with massive digitized collections but few transcribed examples.

Specifically, we introduce a self-supervised pre-training method that randomly masks lacuna-like regions of document line images and learns to reconstruct them by distinguishing them from nearby line image segments, or foils. While lacuna can be reconstructed in a generative way, we find that a discriminative contrastive loss works better in practice. By leveraging a diverse set of unlabeled data

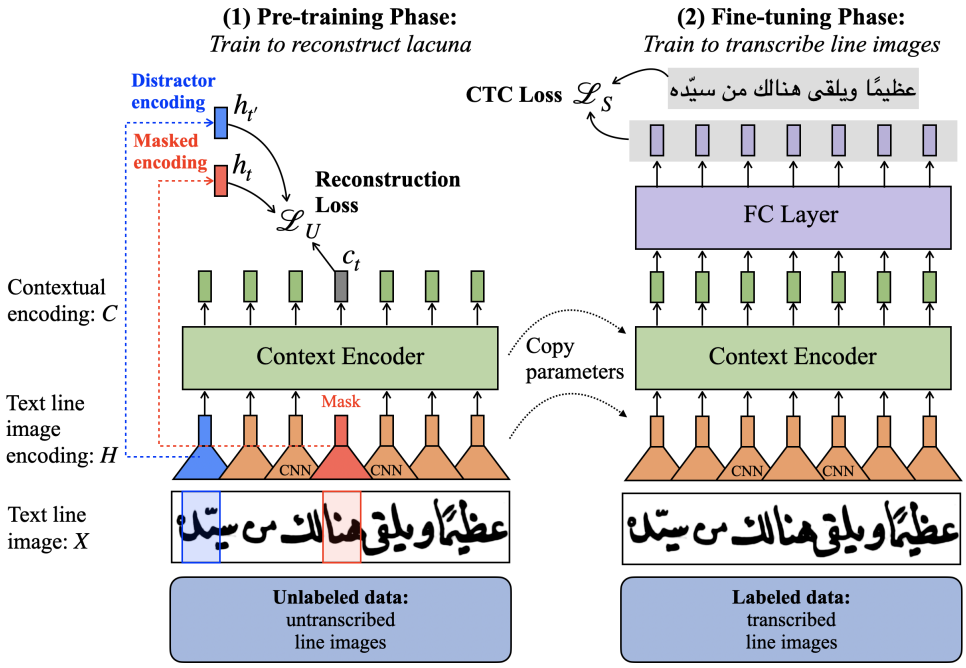

Figure 2: Our proposed two-stage approach for low-resource document transcription first pre-trains a line image encoder using a self-supervised contrastive loss on unlabeled data (left), followed by a fine-tuning phase, in which the pre-trained encoder learns to transcribe 1–3 pages of supervised data using a CTC loss (right).

for pre-training, the model is forced to infer the identities of masked text regions in the presence of scribal writing variation or typesetting noise ubiquitous in historical documents. In the next sections, we describe our model/masking strategy in detail.

3.1 Model

In Figure 2, we show our two-stage pre-train/fine-tune modeling approach. First, we describe the document line image encoder that is shared between stages. For simplicity of description, we assume that each document line image, X, is n pixels tall and m pixels wide, and that pixels are binary-valued. Thus, the space of input text line images can be denoted as $\mathcal{X} = \{0,1\}^{n \times m}$. We first process the input with a **convolutional feature extractor**, $f : \mathcal{X} \mapsto \mathcal{H}$, that maps the input, X, to an encoding matrix, H, using a deep convolutional neural network followed by a reshaping of the image height dimension into the channels dimension. Next, a **contextual encoder**, $g : \mathcal{H} \mapsto \mathcal{C}$, computes a contextualized representation matrix, C, from H using a neural sequence model, parameterized by either an LSTM or Transformer (Hochreiter and Schmidhuber, 1997; Vaswani et al., 2017). We describe both the design of f, which determines the output size of the convolutional encoding space $\mathcal{H}$, and g in Section 5.1. Together, both the convolutional and contextual layers form the encoder of text line images used for downstream document transcription. Ideally, f will capture the underlying visual appearance of distinct character classes, while g will discover linguistic correlations between these classes.

3.2 Masking

During pre-training, we replace randomly sampled, non-overlapping segments of H with a learned mask embedding vector prior to computing contextualized representation matrix C. We train the model to distinguish the masked region from a foil using the contrastive loss presented in Section 3.3.

3.3 Pre-training Objective

We use the following self-supervised contrastive loss whose variants have demonstrated success in self-supervised representation learning (Oord et al., 2018; Baevski et al., 2020):

$$\mathcal{L}_U(c_t) = -\log \frac{\exp\left(s(c_t, h_t)\right)}{\sum_{t'} \exp\left(s(c_t, h_{t'})\right)}$$

Here, c_t (depicted in Figure 2) is the contextual

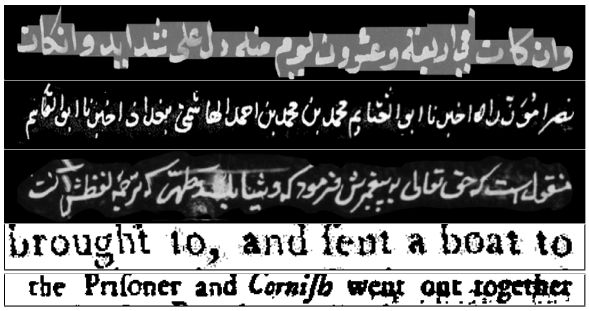

Figure 3: Assortment of cropped, grayscale line images from a selection of our datasets, as extracted by annotators. From top to bottom, RASM 2019 (Keinan-Schoonbaert, 2020), Attar-Mubhij, Ḥuliyya, Trove (Holley, 2010), Old Bailey (Shoemaker, 2005). The Arabic-script line images are shown pre-binarization, while the English line images come binarized.

encoder's output representation of the *masked* line image at position t. Similarly, h_t (also depicted in Figure 2) is the *convolutional* encoder's output representation of the *masked region itself*. Further, $s(c, h)$ represents a scoring function that computes the similarity between representation vectors c and h. We use the cosine similarity similar to Baevski et al. (2020), but compute it only raw vectors, instead of the raw vectors and quantized vectors. The cross-entropy loss requires the model to distinguish the representation of the true masked region, h_t, from distractor representations: the convolutional encodings of other segments, $h_{t'}$ with $t' \neq t$.

3.4 Fine-tuning Objective

After learning pre-trained representations, we add the randomly initialized, fully connected character vocabulary projection layer to the top of our context encoder network (top right of Fig. 2) and perform supervised training using the Connectionist Temporal Classification (CTC) objective (Graves et al., 2006; Graves, 2012; Baevski et al., 2020) with transcribed data. CTC is a commonly used loss function for supervised training in speech and handwriting recognition systems. In this case, CTC is used to marginalize over all monotonic alignments between the sequence of input visual representations and the observed ground truth output sequence of characters.

4 Datasets

In this section, we describe both unlabeled pre-training and labeled fine-tuning/testing datasets used in our experiments.

4.1 Islamicate Manuscripts

First, we introduce a variety of Islamicate manuscript datasets selected for both their uniquely different domain content (e.g., scientific to legal to religious) and their visually distinct scribal handwriting style. All but the first pre-train dataset are professionally transcribed by Islamicate manuscript scholars.

HMML Pre-train Through a collaboration with the Hill Museum and Manuscript Library (HMML), we obtain about 100 early modern, mostly Syrian, naskh[2] manuscripts dating from 1600–1775 with some voweling, but with ornamentally voweled texts excluded (i.e., texts in which every single vowel and orthographic feature is included, usually for ornamental reasons). We filter out manuscripts with extensive marginalia, figures, or tables, though some marginal notes and other elements (e.g., seals, interlinears) are still present. This results in a dataset containing roughly 750,000 *unlabeled* line images.

HMML Fine-tune We obtain transcriptions for 115 line images from a 4-page held-out subset of the HMML Pre-train dataset. This dataset is designed for in-domain fine-tuning/testing experiments with our pre-trained models.

RASM 2019 For the ICDAR 2019 Competition on Recognition of Historical Arabic Scientific Manuscripts, the British Library released 2,164 transcribed line images from scientific manuscripts written in various scribal hands (Keinan-Schoonbaert, 2020). RASM 2019 has become a popular benchmark for Arabic-script handwriting recognition due to its relatively large amount of supervised data for the task.

Attar-Mubhij An Arabic-language legal text with 190 transcribed line images.

[2] https://en.wikipedia.org/wiki/Naskh_(script)

Ḥuliyya A 229-line Persian, nasta'līq[3] devotional text written by an early modern scholar containing mostly Arabic-language prayers.

4.2 Early Modern English Printed Works

Next, we describe several English book and newspaper datasets used in our experiments that were originally printed in early modern England and Australia.

EEBO Pre-train We harvest 750,000 unlabeled line images from a randomly sampled collection of document images from Early English Books Online (EEBO),[4] which contains "almost every work printed in the British Isles and North America, as well as works in English printed elsewhere from 1470-1700."

Trove A dataset of historic Australian newspapers (c. 1803–1954) from the National Library of Australia (Holley, 2010). We use the manually transcribed version totaling 450 lines (Berg-Kirkpatrick et al., 2013).

Old Bailey A manually transcribed set of 20 documents printed 1716–1906, consisting of 30 lines per document, taken from Berg-Kirkpatrick and Klein (2014). Shoemaker (2005) compiled the documents, which describe proceedings of London's Old Bailey Courthouse.

4.3 Line Extraction

Since our model processes individual line images of a document, we use Kiessling (2020)'s line extraction method to automatically segment page images into their component text line images for at-scale collection of the pre-training datasets. We find and discard poorly extracted line images outside an empirically determined pixel width-to-height ratio range of 6–23.

5 Results

In this section, we present document transcription results for both Islamicate manuscripts and early modern English works introduced in Section 4. We compare performance against supervised and unsupervised prior work, and investigate the impact of pre-training/fine-tuning dataset sizes.

[3] https://en.wikipedia.org/wiki/Nastaliq

[4] https://www.proquest.com/eebo

5.1 Experimental Details

Encoder For all experiments, we binarize the line images and scale them to a height of 96 pixels, but allow them to vary in width. We base our CNN architecture on the Kraken OCR system (Kiessling, 2019): two rectangular 4×2 kernels first process the input image, each followed by a Leaky ReLU activation and Group Norm. Two max pooling operations are applied, one before and one after the final 3×3 convolutional layer kernel, with kernel sizes/strides of $4 \times 2/1 \times 2$ for both. The first kernel uses a stride of 4×2 and the final two both use 1×1. The convolutional hidden dimensions are 64, 128, and 256. We use a 3-layer BiLSTM for our contextual encoder with a hidden size of 512. This results in 6,408,000 trainable parameters. Models are implemented in PyTorch (Paszke et al., 2019) and Fairseq (Ott et al., 2019).

Pre-training During pre-training, we perform a non-exhaustive grid search over masking probability and length using 75k lines of data. We determine $p = 0.5/p = 0.65$ to perform best for Islamicate manuscript/English print with a non-overlapping segment length of 12 time steps. We ensure that 8 time steps are between each non-overlapping segment. A maximum of 100 time steps are sampled and used as foils in the denominator of the loss from Sec. 3.3. We use the same learning rate scheduler and Adam optimizer from Baevski et al. (2020) that warms up for the first 8% of updates to a learning rate of 5e-4 and linearly decays it afterwards. Models are pre-trained for 3–5 days on 4 RTX 2080 Ti cards.

Fine-tuning During fine-tuning, we use a tri-stage learning rate schedule with the Adam optimizer, which warms up the learning rate to 5e-4 during the first 10% of updates and decays it linearly by a factor of 0.05 for the final 50% of training. We only update the fully connected layer for the first 200 epochs of training and then proceed to update the contextual encoder as well. These optimization choices are inspired by Baevski et al. (2020). We use a small batch size of 8 and train for a maximum of 700 epochs with the CTC loss (Sec. 3.4). We use greedy decoding after removing the CTC's blank token and do not use any external language model. For Islamicate manuscript experiments we perform NFD unicode normalization.

Baselines				
	Test Dataset CER (↓)			
System	HMML-F	RASM	Attar-Mubhij	Ḥuliyya
Google Cloud OCR	49.0	57.0	61.2	71.4
30 Lines for Supervised Fine-tuning				
	Fine-tune/Test Dataset CER (↓)			
# Lines Pretrain	HMML-F	RASM	Attar-Mubhij	Ḥuliyya
0	51.0	68.9	60.4	70.3
75k	22.7	46.1	30.4	52.9
750k	**14.8**	**36.2**	**23.7**	**45.5**
90 Lines for Supervised Fine-tuning				
	Fine-tune/Test Dataset CER (↓)			
# Lines Pretrain	HMML-F	RASM	Attar-Mubhij	Ḥuliyya
0	36.9	61.7	36.8	52.5
75k	15.2	34.4	20.8	37.5
750k	**10.0**	**25.9**	**15.0**	**28.3**

Table 1: 30 line and 90 line supervised fine-tuning, tested on held-out portion of fine-tuning dataset. Character error rate (CER) is reported.

Character Error Rate (CER) is computed using Kraken OCR (Kiessling, 2019).

Fine-tune/Test Splits For Islamicate manuscript datasets, we hold out 10% of RASM 2019 for testing and the final page each of HMML Fine-tune, Attar-Mubhij, and Ḥuliyya. For English print datasets, we use the same test splits as Berg-Kirkpatrick and Klein (2014) for fair comparison and fine-tune on the validation set of each dataset.

5.2 Islamicate Manuscripts

In Table 1, we present single-run supervised fine-tuning results on in-domain subsets of each dataset limited to 30 and 90 lines (roughly 1 and 3 pages of data, respectively). Each row represents a different set of encoder parameters, which we use to initialize the fine-tuning experiments. Zero lines represents a randomly initialized encoder, while 75k and 750k settings use the encoder parameters pre-trained with our lacuna reconstruction objective on different orders of magnitude of unlabeled HMML Pre-train line images. We also report results obtained from the Google Cloud OCR API as a baseline comparison.

The first thing we can observe is the extremely high character error rates for both the commercial Google Cloud OCR system and the randomly initialized 0k pre-train models, especially in the 30-line setting. Access to about 2 more pages of data (in the 90-line setting) improves results for this setting in the Arabic-language legal text Attar-Mubhij, but does not seem to help much for RASM 2019, a larger collection of scientific manuscripts. This is probably due to the higher amount of diversity in content and style in this benchmark dataset for Arabic-language HTR. Seemingly, without any signal from pre-training and only tens of lines of transcribed data, the model is unable to learn a sufficient visual encoder for the large variety of scribal hands and scripts observed in the manuscripts (examples shown in Fig. 3). Pre-training on just 75k lines halves the error rate for Attar-Mubhij in the 30-line setting. Furthermore, 750k pre-train reduces the Attar-Mubhij CER from 60.4 to 23.7.

The HMML Fine-tune dataset (HMML-F in Table 1) has the largest relative error rate difference between the pre-trained models and models without pre-training. Errors are reduced by about 55% for 75k-30, 70% for 750k-30, 58% for 75k-90, and 73% for 750k-90, which is at least 10 points higher than other datasets on average. Since manuscripts in HMML-F are sourced from the same library as the HMML Pre-train dataset, the results suggest that in-domain pre-training data provides an ad-

Baselines		
	Test Dataset CER (↓)	
System	Trove	Old Bailey
Google Tesseract	37.5	-
ABBYY FineReader	22.9	-
Ocular	14.9	14.9
Ocular Beam	12.9	10.9
Ocular Beam-SV	**11.2**	10.3
Google Cloud OCR	13.3	**8.5**
30 Lines for Supervised Fine-tuning		
	Test Dataset CER (↓)	
# Lines Pretrain	Trove	Old Bailey
0	70.5	60.0
75k	20.3	26.5
750k	**19.6**	**12.2**
90 Lines for Supervised Fine-tuning		
	Test Dataset CER (↓)	
# Lines Pretrain	Trove	Old Bailey
0	38.7	28.6
75k	12.2	9.4
750k	**10.4**	**7.6**

Table 2: 30 line and 90 line supervised fine-tuning, tested on held-out portion of each fine-tuning dataset. Character error rate (CER) is reported (↓). First 5 baselines are taken from Berg-Kirkpatrick and Klein (2014).

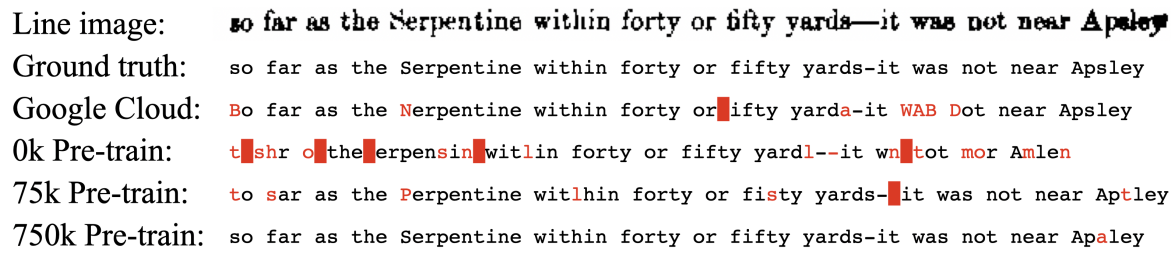

Figure 4: Comparison of results on the Old Bailey test set with errors highlighted. Pre-trained results are from the 90-line fine-tuning setting.

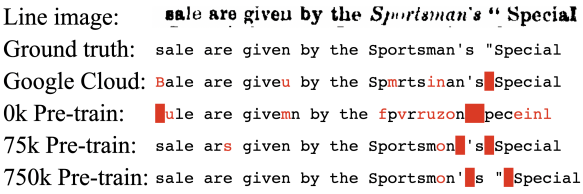

Figure 5: Comparison of results on the Trove test set with errors highlighted. Pre-trained results are from the 90-line fine-tuning setting.

vantage over the other documents from different collections. Regardless, our approach's improved performance on 30-line settings compared to the supervised 90-line results trained from scratch across all datasets is impressive and shows promising generalization ability.

5.3 Early Modern English Printed Works

In Table 2, we present supervised fine-tuning results on in-domain subsets of each dataset limited to the same 30 and 90 line settings as in the Islamicate manuscript experiments. Our first observation is that the randomly initialized encoder from the 0-line pre-train setting sees a much larger improvement from 30 to 90 lines of supervised fine-tuning data than the Islamicate manuscript experiments. We speculate this due to the more similar and repeated glyph shapes on printed data compared to handwritten data, which makes learning of the visual encoder easier. Still, pre-training the visual encoder cuts CER across both datasets, though we do see a slightly bigger relative error rate reduction when fine-tuning on Trove versus Old Bailey.

In Figures 4 & 5, we show comparisons across predicted transcriptions from different systems and datasets for illustrative purposes. First, we observe that Google Cloud OCR, the best baseline system on Old Bailey, consistently struggles with inking variation. For example, the bleeding ink on the initial '`s`' of each line image is mistaken for a '`B`', the '`n`' in '`not`' in Fig. 4 is mistaken for a '`D`' due to the subtle connection of the glyph's legs from over-inking, and the '`m`' in '`Sportsman`' in Fig 5 is confused for the characters '`in`' because of under-inking. However, the 0k pre-train baseline clearly makes the most insertion/deletion/substitution errors since it must learn how to transcribe noisy line images from a randomly initialized encoder using only 90 transcribed line images for supervised parameter learning. Initializing the visual encoder with parameters learned from our self-supervised regime on 75k unlabeled line images from EEBO reduces a lot of these nonsensical errors to only superficial glyph recognition issues. By increasing the pre-training amount by an order of magnitude to 750k, we obtain our best results. Future work could integrate a language model during decoding to address the unlikely sequences of characters/words still output by our best system, like the words '`Apaley`' and '`Sportsmon`'.

5.4 Conclusion

In this paper, we proposed a two-phase pre-train/fine-tune approach for document transcription and applied it to historical documents in low-resource settings. Our pre-training strategy, inspired by reconstructing missing information in documents, or lacuna, uses hundreds of thousands of unlabeled line images to learn rich visual language representations. After supervised fine-tuning on tens of transcribed line images, we showed large character error rate reduction on both Islamicate manuscripts exhibiting major script and style variation and improved over several state-of-the-art OCR systems on early modern English printed works. We estimate that our approach could save human annotators significant amounts of time and enable more distant readings of library collections.

Ethical Considerations

While more accurate transcription of printed and handwritten documents in low-resource settings can expand research access for language and history scholars, it could also potentially facilitate government surveillance of marginalized communities. Separately, bad actors could more easily scan and digitize document images containing sensitive information and use them for nefarious purposes.

References

Mansoor Alghamdi and William Teahan. 2017. Experimental evaluation of arabic ocr systems. *PSU Research Review*, 1(3):229–241.

Kenning Arlitsch and John Herbert. 2004. Microfilm, paper, and ocr: Issues in newspaper digitization. the utah digital newspapers program.

Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. 2019. Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33.

Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2020. Neural representations for modeling variation in english speech. *arXiv preprint arXiv:2011.12649*.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *ACL*.

Taylor Berg-Kirkpatrick and Dan Klein. 2014. Improved typesetting models for historical ocr. In *ACL*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Pre-training transformers as energy-based cloze models. *arXiv preprint arXiv:2012.08561*.

C. Clausner, A. Antonacopoulos, N. Mcgregor, and D. Wilson-Nunn. 2018. Icfhr 2018 competition on recognition of historical arabic scientific manuscripts – rasm2018. In *16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 471–476.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Rui Dong and David A Smith. 2018. Multi-input attention for unsupervised ocr correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2363–2372.

Christopher E. Garrett. 2014. How T. S. Became Known as Thomas Sherman: An Attribution Narrative. *The Papers of the Bibliographical Society of America*, 108(2):191–216.

Dan Garrette and Hannah Alpert-Abrams. 2016. An unsupervised model of orthographic variation for historical document transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 467–472.

Dan Garrette, Hannah Alpert-Abrams, Taylor Berg-Kirkpatrick, and Dan Klein. 2015. Unsupervised code-switching for multilingual historical document transcription. In *NAACL*.

Kartik Goyal, Chris Dyer, Christopher Warren, Max G'Sell, and Taylor Berg-Kirkpatrick. 2020. A probabilistic generative model for typographical analysis of early modern printing. In *Proceedings of 2020 Annual Conference of the Association for Computational Linguistics*.

Alex Graves. 2012. Offline arabic handwriting recognition with multidimensional recurrent neural networks. In *Guide to OCR for Arabic scripts*, pages 297–313. Springer.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Mika Hämäläinen and Simon Hengchen. 2019. From the paft to the fiiture: a fully automatic nmt and word embeddings method for ocr post-correction. *arXiv preprint arXiv:1910.05535*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Rose Holley. 2010. Trove: Innovation in access to information in australia. *Ariadne*, (64).

José Carlos Aradillas Jaramillo, Juan José Murillo-Fuentes, and Pablo M Olmos. 2018. Boosting handwriting text recognition in small databases with transfer learning. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 429–434. IEEE.

Dongwei Jiang, Xiaoning Lei, Wubo Li, Ne Luo, Yuxuan Hu, Wei Zou, and Xiangang Li. 2019. Improving transformer-based speech recognition using unsupervised pre-training. *arXiv e-prints*, pages arXiv–1910.

Adi Keinan-Schoonbaert. 2019. Using transkribus for arabic handwritten text recognition. *British Library Digital Scholarship Blog*.

Adi Keinan-Schoonbaert. 2020. Results of the rasm2019 competition on recognition of historical arabic scientific manuscripts. *British Library Digital Scholarship Blog*.

Benjamin Kiessling. 2019. Kraken-an universal text recognizer for the humanities. *Proceedings of the DH*.

Benjamin Kiessling. 2020. A modular region and text line layout analysis system. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 313–318. IEEE.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. Ocr post correction for endangered language texts. *arXiv preprint arXiv:2011.05402*.

Shruti Rijhwani, Daisy Rosenblum, Antonios Anastasopoulos, and Graham Neubig. 2021. Lexically aware semi-supervised learning for ocr post-correction. *Transactions of the Association for Computational Linguistics*, 9:1285–1302.

Maxim Romanov, Matthew Thomas Miller, Sarah Bowen Savant, and Benjamin Kiessling. 2017. Important new developments in arabographic optical character recognition (ocr). *arXiv preprint arXiv:1703.09550*.

Robert Shoemaker. 2005. Digital london: Creating a searchable web of interlinked sources on eighteenth century london. *Program*.

Xingchen Song, Guangsen Wang, Yiheng Huang, Zhiyong Wu, Dan Su, and Helen Meng. 2020. Speechxlnet: Unsupervised acoustic model pretraining for self-attention networks. *Proc. Interspeech 2020*, pages 3765–3769.

Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, Mauricio Villegas, and Enrique Vidal. 2019. A set of benchmarks for handwritten text recognition on historical documents). *Pattern Recognition*, 94:122–134.

Chaitanya Talnikar, Tatiana Likhomanenko, Ronan Collobert, and Gabriel Synnaeve. 2020. Joint masked cpc and ctc training for asr. *arXiv e-prints*, pages arXiv–2011.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Weiran Wang, Qingming Tang, and Karen Livescu. 2020. Unsupervised pre-training of bidirectional speech encoders via masked reconstruction. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6889–6893. IEEE.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763.