Graph Neural Networks for End-to-End Information Extraction from Handwritten Documents

Yessine Khanfir [1] Marwa Dhiaf [1,2,3] Emna Ghodhbani [1] Ahmed Cheikh Rouhou [1]
Yousri Kessentini [2,3]

[1]InstaDeep
[2]Digital Research Centre of Sfax, Tunisia
[3]SM@RTS: Laboratory of Signals, Systems, Artificial Intelligence and Networks
{y.khanfir, m.dhiaf, e.ghodhbani, a.cheikhrouhou}@instadeep.com
yousri.kessentini@crns.rnrt.tn

Abstract

Automating Information Extraction (IE) from handwritten documents is a challenging task due to the wide variety of handwriting styles, the presence of noise, and the lack of labeled data. In this work, we propose an end-to-end encoder-decoder model, that incorporates transformers and Graph Convolutional Networks (GCN), to jointly perform Handwritten Text Recognition (HTR) and Named Entity Recognition (NER). The proposed architecture is mainly composed of two parts: a Sparse Graph Transformer Encoder (SGTE), to capture efficient representations of input text images while controlling the propagation of information through the model. The SGTE is followed by a transformer decoder enhanced with a GCN that combines the outputs of the last SGTE layer and the Multi-Head Attention (MHA) block to reinforce the alignment of visual features to characters and Named Entity (NE) tags, resulting in more robust learned representations. The proposed model shows promising results and achieves state-of-the-art performance on the IAM dataset, and in the ICDAR 2017 Information Extraction competition using the Esposalles database.

1. Introduction

Paper documents exist in different forms and frequently hold valuable information. Historical records may be used to determine ethnic origins or even to glean important historical information. Business and administrative documents, in turn, can be used to carry out statistical analyses. However, the large volume of data makes manual transformation impractical. Therefore, the adoption of automated Information Extraction (IE) systems is necessary to process these documents.

In the literature, information extraction approaches from document images are either based on a two-stage [17, 18] or an end-to-end architecture [4, 5, 16]. A two-stage approach transforms the document image into a textual representation, and then, Natural Language Processing (NLP) techniques are applied to parse the output text and extract the named entity tags. On the other hand, the end-to-end method, also known as the joint learning approach, involves the simultaneous recognition of text and Named Entity (NE) annotations, or the direct identification of NEs on the image level without requiring an explicit recognition step at the text level.

The advancements in NLP over the past decade, have motivated many researchers to tackle IE tasks from scanned documents using deep learning architectures, showing higher performance compared to traditional methods. Recurrent Neural Networks (RNN) have at some point become the most successful in this context. In [17, 18], authors have proposed a two-stage model based on a Long Short-Term Memory (LSTM) architecture, to perform NER on machine-printed and handwritten documents. Although these LSTM-based approaches produced competitive results, their performance depended on the quality of the text recognition stage. In [5, 19], the authors proposed an end-to-end model comprising a Convolutional Neural Network (CNN) and a Bidirectional LSTM (BLSTM) network, to jointly perform HTR and NE recognition on handwritten document images avoiding the explicit transcription step. However, using such recurrent architectures increases the computational cost at the training stage, since their sequential pipelines prevent parallelization.

With the rise of attention mechanism [1], attention-based models have taken over the field of NLP, and have shown unprecedented capabilities in maximizing their performance in context modeling. In [21] a combination of an attention-based model with a BLSTM and a Conditional Random Field (CRF) is introduced to perform NER on handwritten text images. Despite outperforming previous works, it still operated at the line-level, and requires an explicit HTR step, making it sensitive to transcription errors.

In [16] the authors present a transformer model performing joint HTR and NER from historical handwritten text images. The proposed model works at paragraph level, surpassing the line segmentation problems, which allows the model to exploit larger bi-dimensional contextual information to identify the semantic NE tags.

These previously cited approaches are based on a sequential relational inductive bias, that consists in making relational assumptions to produce a model able to make correct predictions. Driven by this observation, authors in [3] propose a method to perform Named Entity Recognition (NER) and relation prediction in semi structured documents with Graph Neural Networks (GNN). Their approach demonstrated a good generalization ability, but still depended on a third-party text recognition tool.

Some works have explored the application of GNNs in the context of document understanding, including table detection [14], table structure recognition [12] and visual question answering [11]. Nevertheless, to the best of our knowledge, no research applied GNNs to propose an end-to-end information extraction model from handwritten documents, where named entities are directly identified on image level, without the need of an explicit recognition step.

Motivated by the capacity of GNNs in understanding the semantic correlations between elements in the same input, thanks to the flexibility that graph structures offer, and the way a GNN represents each element by the context it belongs to, we propose in this work an encoder-decoder model, that combines the advantages of transformers and Graph Convolutional Networks (GCN), for an efficient, end-to-end NER, with improved context modeling capabilities. The proposed model jointly performs text and named entity recognition at paragraph-level, allowing it to avoid unrecoverable early errors due to line segmentation, and to exploit larger contextual information to identify the semantic relations between the named entities. Our approach achieves state-of-the-art performance on the IAM database manually annotated with NE tags and in the ICDAR 2017 IEHHR competition using the Esposalles database. Our contributions can be stated as follows:

- Sparse Graph Transformer Encoder (SGTE): We propose a variant of the Graph Transformer [7] to encode input sequences of visual features, where we leverage graph structures to flexibly define a dynamic scope of attention, that changes according to the position of the indexed feature vector and consequently reduce the computational cost of the encoding step.

- Cross-GCN-based Decoder: We propose a cross graph convolutional network (Cross-GCN) to reinforce the alignment of visual features to characters and NE tags. The Cross-GCN operates on cross graphs that combine the outputs of the last SGTE layer and the Multi-Head Attention (MHA) block, resulting in a significant improvement in representation learning

- We achieve new state-of-the-art performance in single-stage HTR-NER on two benchmark datasets.

The rest of this paper is organized as follows. Section 2 provides a review of the related works. In section 3, we describe in detail our proposed model and contributions. Section 4 provides experimental validation of our approach. Finally, section 5 concludes the work, highlighting its future scope and benefits.

2. Related works

According to the literature, extracting Named Entities (NE) from document images can be performed following two approaches: the first one consists in applying text recognition on text images, then recognizing NEs as a separate NLP application [6, 8, 9, 13, 21]. However, the second approach combines handwritten text and named entity recognition in a joint procedure. As our method aims to jointly recognize text and named entities using graph neural networks, we introduce a detailed study of the joint recognition approaches, and GNN-based works for document understanding.

2.1. Joint Learning approaches for HTR and NE Recognition

Performing information extraction using a joint learning approach makes it more efficient, as it helps avoid unrecoverable early transcription errors [4, 5, 16]. In this case, the joint learning method simultaneously transcribes the text image and extracts the entities in a single stage. A CNN based model is proposed in [20] in order to extract semantically meaningful entities from handwritten word images, bypassing the recognition step. However, this method fails to consider the context surrounding the word, which can lead to incorrect predictions.

In [17], a CNN is combined with an LSTM network to perform IE tasks directly on visual features, and avoid the transcription step. The main drawback of this approach is that it requires a pre-processing step where the input document is segmented into word images.

In [5,19], the authors proposed a CNN-BLSTM architecture that was trained on line-level handwritten text images to integrate a larger context. Still, in this work, the context is limited to the line level, which affects the performance of the extraction of semantic named entity tags. The authors confirm that by integrating a curriculum learning strategy, consisting in training the model first on text lines and then on records, the model reaches a higher final prediction accuracy.

To incorporate a larger context, the authors in [2] propose an end-to-end model, performing handwriting text detection, transcription, and named entity recognition simultaneously at the page level by leveraging shared features for these tasks. However, this approach requires a manual annotation of word bounding boxes, which may be very expensive in real-world applications. In addition, the performance of the multitask model may be decreased if one task is particularly challenging and unrelated to the others.

The transformer architecture [22] came as a better alternative in context understanding, and learning robust representations, as it mitigates the problems of LSTM models by avoiding recursion in order to allow parallel computation, and also to avoid the drops in performance that are due to long range dependencies. In [16], a sequence-to-sequence transformer architecture is proposed to jointly perform HTR and NER from images of handwritten historical marriage records. First, input images are passed through a CNN, to extract visual features, then fed to the transformer encoder to compute hidden representations. The latter is translated by the decoder into a sequence of transcribed characters and NE tags. The authors show that the proposed transformer model outperforms state-of-the-art approaches on Esposalles dataset. Although self-attention mechanism has been demonstrated to be a more reliable substitute to RNNs, it learns the dependencies between all tokens without regard to their distance, and does not offer the option to define the scope of attention in a flexible way. For this aim, we believe that combining a transformer with a GNN can help the model capture the inherent structure in the given graph and learn good representations for generating the target transcription and NE tags.

2.2. Graph Neural Network for Document Understanding

Graph structures can be used to alleviate self-attention's rigidity in defining a model's information-sharing logic. In fact, graph neural networks have become ubiquitous in various fields that deal with graph data, where the topological structure of the input is highly relevant. In [14], a graph-based approach was introduced to detect tables in document images. The proposed model is trained to detect tables in different types of business documents, predicting relationships between table elements. Instead of using the recognized text, they make use of the position, context, and content type. Carbonell et al. [3] tackled information extraction from semi structured business documents (i.e. forms, invoices, ID documents, etc.) using a GNN-based approach. Each input document is turned into a graph, where each node is a word connected to its K nearest neighbors (KNN) where K is a hyperparameter. Words and their bounding boxes are extracted using a third-party Optical Character Recognition (OCR) system. These words were used to compute distances between word pairs in order to determine each element's neighborhood. GNNs are then used for word grouping, entity labeling and entity linking. In both previous works [3, 14], a prior stage of OCR was necessary in order to solicit graph structures to represent and encode document images while preserving their topological structures.

Previous works in the literature have demonstrated the ability of deep neural network architectures to model contextual information, in order to perform IE tasks on document images. In addition, recent studies have shown that GNNs present a robust alternative in modeling the semantic relationships within graph structures, and can be very useful in representation learning. Also, it has been shown that the transformer self-attention mechanism can indeed be generalized to learn graph representations thanks to the Graph Transformer [7], offering the flexibility to control the scope of information propagation. This observation has motivated us to explore the combination of transformer models and GNNs to propose an end-to-end model that jointly performs HTR and NER on paragraph level handwritten document images.

3. Proposed Approach

In this work, we propose an end-to-end encoder-decoder model, that combines transformers and Graph Convolutional Networks, to jointly perform handwritten text and named entity recognition. We simultaneously take advantage of the self-attention mechanism and GNNs in representation learning and relation extraction. Furthermore, our method benefits from the flexibility of graph structures to control the scope of information propagation.

The proposed model is illustrated in Figure 1. We preserve the encoder-decoder form, as it is suitable for our sequence-to-sequence learning task. Input images are fed into a pre-trained ResNet-50 [10] for feature extraction, followed by a 2D-convolutional layer with a kernel size of 1×1 to match the number of features from the backbone network and the encoder input. For the encoding part, the SGTE is used as described in section 3.1. The decoder of the traditional transformer model [22] includes a Masked MHA (MMHA) block to model relationships within the ground truth, and an MHA block responsible for the alignment of the visual features to characters and NEs through the self-attention mechanism. In our proposed model, we extend this specific operation of alignment by introducing a Cross-GCN in the decoder part, built from the output of the SGTE and the decoder MHA block as presented in section 3.2.

3.1. Sparse Graph Transformer Encoder

For the encoding part, we adopt the generalization of transformers to graph structures proposed in [7]. Knowing that the Graph Transformer [7] operates on graphs, using it as an encoder requires a prior graph construction step. Figure 2 provides a minimized illustration of how the initial graph is constructed from feature maps relative to each input document. To simplify the illustration, Figure 2 shows a 4×3 graph generated from a 4×3 stack of feature maps (F). In practice, the feature extraction step yields 256 feature maps, each of size 8×32, so the real size of the (F) is $8 \times 32 \times 256$ that we use to create an initial graph of 8×32 nodes, each initially represented by a vector of size 256. Indeed, elements that share the same spatial position in all feature maps are stacked and assigned as a node representation in the initial graph.

In step (1) of Figure 1, to reduce the complexity of fully-connected graphs, we build a sparse graph that customizes the scope of attention of each element of the feature maps. Our strategy is to select each node's neighborhood according to its original position in (F). We categorize visual features into two types: initially positioned in the first or last line of (F), and initially positioned in between. Elements on the first line, are connected to all neighbors on the same and the next line. Elements on the last line, are connected to all neighbors on the same and the previous line. The remaining feature vectors are connected to the elements on the same, previous, and following line (note that self-connections are also included). Edges are only used to define the attention scope and connections between nodes, thus it is unnecessary to attribute edge representations. Representing the feature maps using sparse graphs instead of sequential structures, allows controlling the feature propagation process.

After the graph is built, it goes through the SGTE layers to update the node representations over the attention heads. For each layer l of the SGTE, the representation h_i^l of the i_{th} node is updated as follows:

$$\hat{h}_i^{l+1} = O_h^l \,\|_{k=1}^{H} \, (\sum_{j \epsilon N_i} w_{ij}^{k,l} V^{k,l} h_j^l), \tag{1}$$

where,

$$w_{ij}^{k,l} = softmax_j(\frac{Q^{k,l} h_i^l \cdot K^{k,l} h_j^l}{\sqrt{d_k}}), \tag{2}$$

and $Q^{k,l}, K^{k,l}, V^{k,l} \in R^{d_k * d}, O_h^l \in R^{d * d}$, $k \in [1, H]$ denotes the number of attention heads, $\|$ denotes concatenation. N_i refers to the set of nodes directly connected with edges to the i^{th} node.

3.2. Cross-GCN-based Decoder

In the decoding step, first, the MMHA block models the semantic relations between elements of the ground truth sequence. The output of this operation is then fed to the MHA block, to align input elements to the target sequence. In this part, we explore the effect of reinforcing the representation learning during the decoding step, with a two layers GCN, as detailed in step (7) of Figure 1. The goal is to jointly benefit from the attention mechanism, and message-passing principles of graph convolutions, to learn robust representations and align visual features to characters and NE tags. To this end, as shown in step (6) of Figure 1, we construct a directed graph of nodes emerging from the output of the last SGTE layer (N_1), and nodes emerging from the output of the MHA block (N_2). In order not to cancel the masking effect of the MMHA block, N_2 nodes are not connected to each other and are only linked to all the nodes in N_1.

The decoder's initial graph is then fed to a two-layers GCN, referred to as the Cross-GCN as it operates on nodes emerging from different components. In this block, each node coming from the MHA component will be represented, by the weighted sum over N_1 nodes. For each layer l of the Cross-GCN, the update of the representation h_i^l of the i_{th} node is computed as follows:

$$\hat{h}_i^{l+1} = \sigma\left(w \cdot h_i^l + W \cdot \sum_{j \epsilon N_1} \frac{h_j^l}{\sqrt{N_{(i)} + N_{(j)}}}\right) \tag{3}$$

Where σ is a non-linearity, w is a weight coefficient multiplied with the initial representation of the i_{th} node before aggregation, N_1 denotes the set of nodes originating from the last layer of the encoder, and W is a weight matrix, $N_{(i)}$ and $N_{(j)}$ refer to the degrees of the i_{th}, j_{th} nodes respectively, h_j^l is the representation l of the j_{th} node and $\frac{1}{\sqrt{N_{(i)}+N_{(j)}}}$ is added as a regularization term.

The representations learned by the Cross-GCN are then combined with the output of the MHA. We apply a direct sum between both signals for a more robust alignment of the visual features to characters and NE tags.

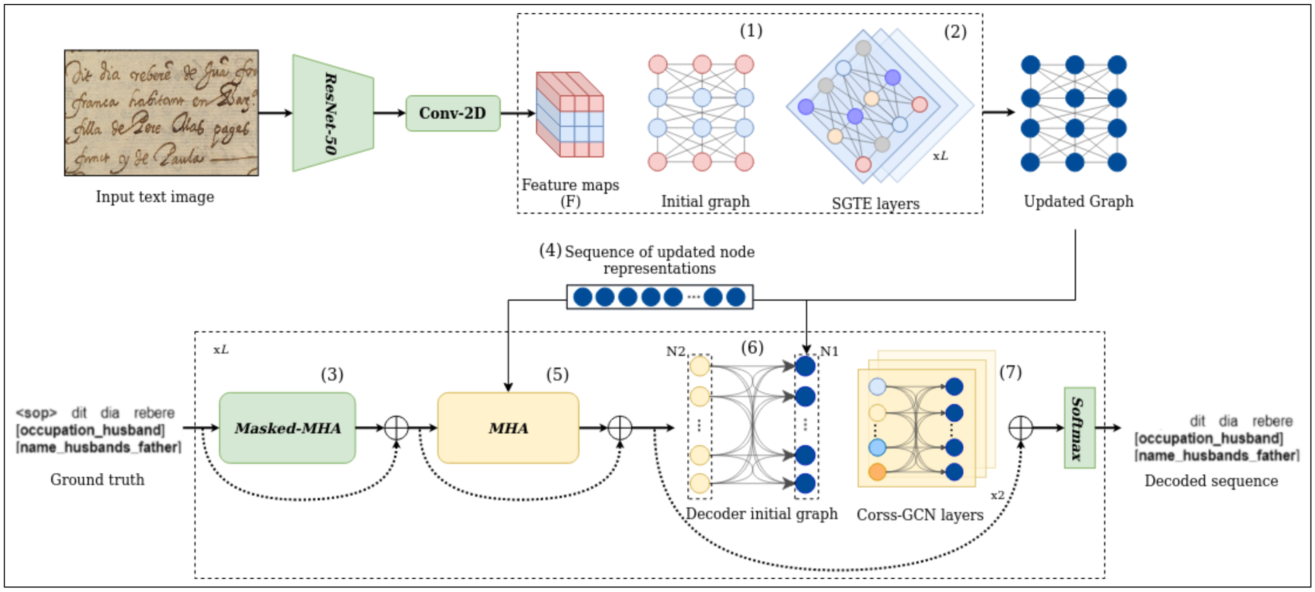


Figure 1. Overview of the proposed architecture: The backbone produces a stack of feature maps from the input image, that will be used to construct the encoder's initial graph (1), the initialization of the graph nodes and edges is illustrated in Figure 2. The graph is then fed to the SGTE to output an updated representation of the feature vectors (2). The ground truth sequence goes into the MMHA block (3). The updated representations of the visual features are stacked back in a sequential form while preserving their initial order (4), then fed as input to the MHA block along with the output of the MMHA (5). The output of the MHA (N_2) and the output of the last encoder layer (N_1) are used to construct the decoder initial graph (6), which will be fed to the Cross-GCN, to reinforce the alignment operation of visual features to characters and NEs (7). Both signals emerging from the GCN and the MHA are combined with a direct sum, then passed to a Softmax layer to perform the predictions.

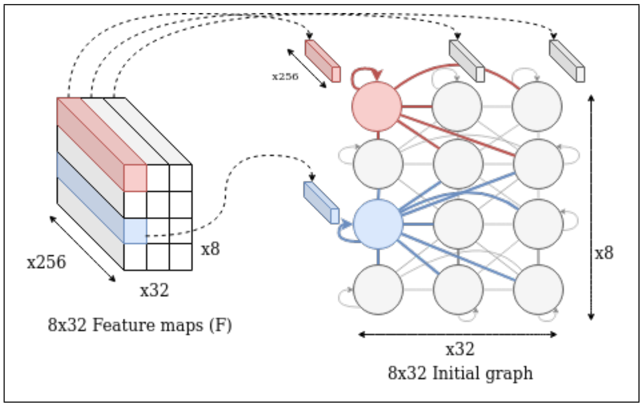


Figure 2. Minimized illustration of the graph construction method: The Backbone outputs, for each image, 256 superposed feature maps, each of size 8 × 32. Feature vectors are retrieved from the 3rd dimension of the collection of feature maps, to preserve their original position in the image and their correspondence to characters. Each feature vector becomes a node representation in an undirected graph that includes self-connections

The SGTE and the Cross-GCN-based Decoder are jointly trained and supervised with the categorical cross-entropy loss computed based on the sequence predicted by the Decoder and the target sequence. We denote $\hat{Y} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_T)$ as the predicted sequence and $Y = (y_1, y_2, ..., y_T)$ as the target sequence, where T represents the sequence length. The loss function L adopted to train our model can be formulated as follows:

$$L\left(\hat{Y}, Y\right) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{C} y_{t,i} log\left(p_{t,i}\right) \tag{4}$$

where,

$$p_t = softmax(\hat{y}_t) \tag{5}$$

Here C refers to the number of unique words in the vocabulary. $y_t\,(y_{t,1}, y_{t,2}, ..., y_{t,C})$ denotes the one-hot encoded vector where $y_{t,i}$ is equal to 1 if it corresponds to the true class at step t, and 0 otherwise. $p_{t,i}$ corresponds to the predicted probability of the i_{th} word in the vocabulary at step t. $p_t\,(p_{t,1}, p_{t,2}, ..., p_{t,C})$ is obtained by applying a $softmax$ scaling to the model's predicted logits over the vocabulary $\hat{y}_t$.

4. Experiments

In this section, we conduct a series of experiments on the Esposalles and IAM datasets. We first present an ablation study to ascertain the contribution of each component within our model. Additionally, we provide a comparison of the results obtained from our method with those from other existing methods.

4.1. Datasets

Esposalles:

We experimented with a subset of the Esposalles database [15], labeled for information extraction. It collects 125 handwritten pages, containing 1221 marriage records. Each record is composed of several text lines giving information on the husband, wife, and their parent's names, occupations, locations, and civil states. We have used 872 records for training, 96 records for validation, and 253 records for the test. For the evaluation, two scenarios are proposed: the basic track where only 4 named entities are considered (name, surname, occupation, location). The complete track is more challenging and comprises 26 named entities (husband, wife, husband's mother, wife's father, etc.).

The manually annotated IAM dataset:

The IAM Database serves as a significant benchmark for IE systems on handwritten documents. It consists of 13353 text lines and 115320 words spread across 1539 scanned text pages. For our experiments on the IAM dataset, we follow the evaluation process proposed by Tuselmann et al. [21], using the traditional RWTH split of the IAM dataset, which was manually annotated with NE labels in [21], to avoid the errors caused by automatic taggers. The split is partitioned into writer-independent training, validation and test partitions composed respectively of 6161, 966, and 2915 lines. The used tag set comprises 6 categories: Location, Time, Cardinal, Nationalities or religious or political groupings (NORP), Person, and Organization.

4.2. Configuration and hyperparameters

The best performance is achieved using an architecture taking as input an image of size 256×1024 for the Esposalles dataset, and 64×1024 at line level for the IAM dataset. The architecture consists of a RestNet-50 as a backbone and a single attention head in the encoding part. The decoder uses one attention head and two GCN layers. We note that we have conducted several experiments to optimize the size of the model (number of layers and heads), and we found that given the size of the training set of Esposalles dataset, using a larger model size decreases the performance of the model as confirmed in [16]. We use Adam optimizer and the optimal learning rate is set to 0.0001.

4.3. Metrics

A first round of experiments is conducted on the Esposalles dataset, to perform ablation studies and to compare with previous approaches that were submitted in the ICDAR 2017 IEHHR competition. For the evaluation on the Esposalles dataset, we use the competition evaluation score computed as follows: For each NE, if the tag is predicted correctly, then the score is given by: $1 - CER$ where CER is the relative Character Error Rate for that word, else the score is 0. Note that the provided evaluation method computes a complete and basic score, for the complete and basic tasks respectively. More details about the metrics can be found in [8]. In the second round of experiments, to fairly compare our system to the state-of-the-art approaches performing NER on the IAM database, we evaluate our model using the same metric adopted in [21] based on the F1-score given by:

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{6}$$

Encoder	Decoder	Complete score
Full Graph Transformer	Baseline Decoder	95.12%
SGTE	Baseline Decoder	95.05%
Baseline Encoder	GCN-based Decoder	92.97%
Ours: SGTE	**GCN-based Decoder**	**96.24%**
Baseline Encoder	Baseline Decoder	95.54%

Table 1. Comparison of different model architectures on the Esposalles dataset

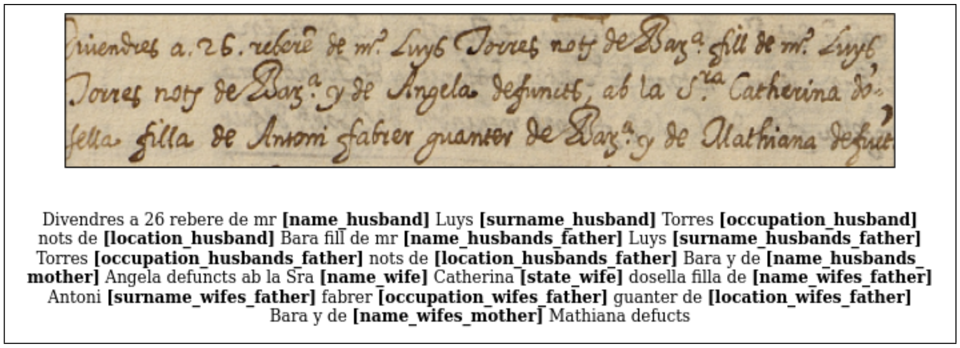
4.4. Results

We start by an ablation study to validate the effectiveness of the SGTE and the Cross-GCN-based decoder. We also investigate the effect of graph sparsity on the model's performance. We then compare our results to state-of-the-art systems on both datasets.

Ablation study

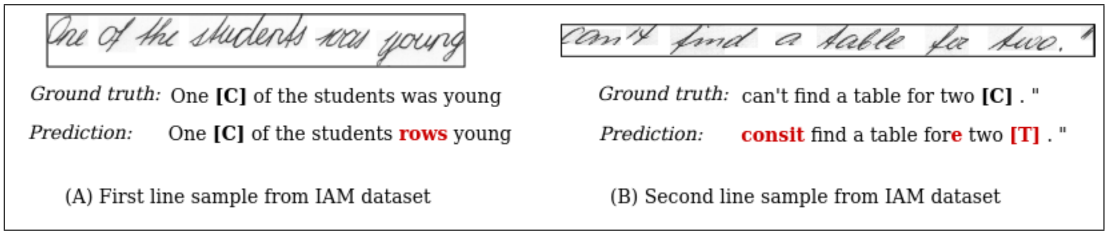
The goal of this section is to validate the effectiveness of the main components of our approach: the SGTE and the Cross-GCN-based decoder. We compare the performance of the proposed model (row 4 of Table 1) to the baseline transformer [16] (row 5 of Table 1), and to two hybrid encoder-decoder architectures where 1) the SGTE is followed by a baseline transformer decoder (row 2 of Table 1), and 2) the Cross-GCN-based decoder is preceded by a baseline transformer encoder (row 3 of Table 1).

Looking at Table 1, we notice that the combination of the sparse graph transformer in the encoder and the Cross-GCN in the decoder achieves the best performance across all tested architectures, with an improvement of 0.7% compared to the baseline transformer [16]. This proves that with this setup, graph convolutions are in fact beneficial in representation learning, and combining their output with that of the MHA, produces a better alignment



Divendres a 26 rebere de mr **[name_husband]** Luys **[surname_husband]** Torres **[occupation_husband]** nots de **[location_husband]** Bara fill de mr **[name_husbands_father]** Luys **[surname_husbands_father]** Torres **[occupation_husbands_father]** nots de **[location_husbands_father]** Bara y de **[name_husbands_mother]** Angela defuncts ab la Sra **[name_wife]** Catherina **[state_wife]** dosella filla de **[name_wifes_father]** Antoni **[surname_wifes_father]** fabrer **[occupation_wifes_father]** guanter de **[location_wifes_father]** Bara y de **[name_wifes_mother]** Mathiana defucts

Figure 3. Joint HTR-NER inference example on a paragraph sample from Esposalles dataset



Ground truth: One **[C]** of the students was young

Prediction: One **[C]** of the students **rows** young

(A) First line sample from IAM dataset

Ground truth: can't find a table for two **[C]** . "

Prediction: **consit** find a table fore two **[T]** . "

(B) Second line sample from IAM dataset

Figure 4. Joint HTR-NER inference example on a line sample from IAM dataset

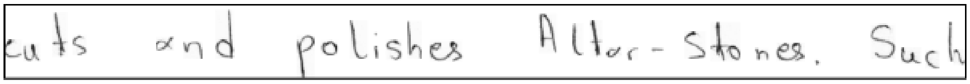
of visual features to characters and NEs, thanks to the effectiveness of message passing iterations in semantic modeling.

However, preserving the transformer encoder and incorporating the Cross-GCN in the decoder, leads to a drop of 2.57% in the complete score. This indicates that when full-attention is employed to encode input visual features, the addition of graph convolutions becomes an excessive learning step that leads to the early overfitting of the model. Besides, replacing the transformer encoder with the Graph Transformer, and using a fully-connected graph, instead of a sparse graph, results in a drop of 0.42% compared to the baseline transformer. We believe that both encoders are theoretically equivalent, thus we assume this slight difference to be due to the use of different implementation frameworks. Furthermore, we observe that the absolute gap between the scores reached in the complete and basic tasks (column 4 of Table 2) varies between 0.02% and 26.45% across the methods, suggesting that some approaches are more sensitive to the length of the inputted context, and the number of classes among which the model has to predict. Our method scores the lowest gap, with a higher performance on the complete task, meaning that in our case, the added classes serve as additional information that the model was able to use to its advantage to make more precise predictions.

Comparison with state-of-the-art approaches

We compare our proposed approach with other methods that participated in the IEHHR, with the same experimental protocol used in the competition. As reported in table 2, our model achieves the best performance in the complete track, which is the most challenging and comprises 26 named entities. We notice an improvement of 4.27% compared to the best system. Against the baseline transformer [16], we notice improvements of 3.14% and 0.7% at line and block levels respectively. We note that for the basic track, which is an easier NER task, our model and the baseline transformer have approximately equal performance. This proves that the combination of transformer and GNNs has a more significant impact for information extraction tasks with increased complexity.

Besides testing our approach on historical handwritten records, we also want to investigate its versatility on another dataset including different NE tags. Hence, we evaluate our model on the IAM dataset, following the same evaluation protocol used by Tuselmann et al. [21]. As presented in Table 3, our model achieves new state-of-the-art performance, with an improvement of 5.6% in terms of F1-score compared to the two-stage NER approach presented in [21]. We also compare our results to the performance of the baseline transformer [16], and an improvement of 3.1% in terms of F1-score is achieved, which confirms the capacity of our model to successfully solve the linking of entities thanks to



*Prediction using GNN-based model **(ours)***	cuts and polishes Aller-stones [L] . Such [L]
Prediction using the baseline transformer in [16]	ents and polishes Afror-stones . Buch
Ground truth	cuts and polishes Altar-Stones [L] . Such

Figure 5. Comparing the inference quality of the baseline transformer and our GNN-based transformer on samples from IAM dataset

System	Basic	Complete	Gap	Level
Hitsz-ICRC-2*	94.16	91.97	2.19	Word
Baseline-CNN*	79.40	70.18	9.22	Word
CITLab-Argus-1*	89.53	63.08	26.45	Line
CITLab-Argus-2*	91.93	91.56	0.37	Line
CITLab-Argus-3*	91.61	91.17	0.44	Line
Carbonell et al. [4]	90.58	89.39	1.19	Line
HMM-MGGI*	80.28	63.11	17.17	Line
Transformer [16]	95.16	93.3	1.86	Line
Transformer [16]	96.25	95.54	0.71	Record
Ours	**96.22**	**96.24**	0.02	Record

* System mentioned in [8]

Table 2. Comparison with IEHHR Competition systems.

System	Precision	Recall	F1-score
Toledo et al. [19]	45.3	28.8	34.0
Rowtula et al. [17]	58.8	41.3	47.4
HTR-NER*	77.3	65.9	70.7
HTR-D-NER*	78.6	73.0	75.4
Annotation-NER [21]	83.8	77.5	80.1
Transformer [16]	98.1	71.4	82.6
Ours	**98.2**	**76.1**	**85.7**

* System mentioned in [21]

Table 3. Results on the IAM dataset

the integration of GNNs.

4.5. Results analysis

This section presents a concrete application of our proposed model on samples from the Esposalles and IAM datasets. Figure 3 shows how the model jointly predicts a sequence of characters and tags from an input paragraph of the Esposalles dataset. Even though the handwriting style is historical and barely human-readable, the model succeeds in almost flawlessly recognizing the characters, and assigning the correct categories to the words.

Figure 4 illustrates two examples of the model's predictions on samples from IAM dataset. It is shown that despite reaching new state-of-the-art performance, the model's accuracy still varies with respect to the difficulty level of the input, and that some flaws are noticed.

In order to examine the impact of substituting the transformer encoder with the SGTE, and incorporating graph convolutions in the decoding step, we compare the performance of our retained model to the standard transformer model. In Figure 5, it is shown that using the combination of the SGTE and Cross-GCN-based decoder, leads to less transcription errors, and results in an enhanced context understanding capabilities. The presented examples demonstrate that our model has successfully recognized NE tags that were overlooked by the baseline transformer. Nevertheless, we notice the occurrence of a few errors, for example, the model sometimes tends to annotate words starting with upper-case characters even if they are not true named entities.

5. Conclusion

In this paper, we introduced an end-to-end encoder-decoder architecture, to take part in the contentious debate comparing multi-stage versus single-stage NER from handwritten documents. It achieves new state-of-the-art performance on the Esposalles dataset and the manually annotated IAM dataset. These results were achieved thanks to two main contributions: in the first place, the combination of two powerful representation learning components, GNNs and attention mechanism, led to an improved adaptation of textual to visual information. In the second place, the use of a sparse version of the Graph Transformer as encoder, allowed the definition of a dynamic attention scope, to avoid unnecessary attention computations. Through this research, we are proposing a widely applicable approach, that could potentially be extended to other sequence-to-sequence applications. We believe that the combination of the Sparse Graph Transformer Encoder with the Cross-GCN-based decoder, can be generalized to other domains, to offer an enhanced alignment of elements of the input to those of the ground truth, regardless of the nature of the task or data.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[2] Manuel Carbonell, Alicia Fornés, Mauricio Villegas, and Josep Lladós. A neural model for text localization, transcription and named entity recognition in full pages. *Pattern Recognition Letters*, 136:219–227, 2020. 3

[3] Manuel Carbonell, Pau Riba, Mauricio Villegas, Alicia Fornés, and Josep Lladós. Named entity recognition and relation extraction with graph neural networks in semi structured documents. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9622–9627. IEEE, 2021. 2, 3

[4] M. Carbonell, M. Villegas, A. Fornés, and J. Lladós. Joint recognition of handwritten text and named entities with a neural end-to-end model. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 399–404, 2018. 1, 2, 8

[5] Marwa Dhiaf, Sana Khamekhem Jemni, and Yousri Kessentini. Docner: A deep learning system for named entity recognition in handwritten document images. In *Neural Information Processing*, pages 239–246, Cham, 2021. Springer International Publishing. 1, 2, 3

[6] Marco Dinarelli and Sophie Rosset. Tree-structured named entity recognition on OCR data: Analysis, processing and results. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 1266–1272, Istanbul, Turkey, 2012. European Language Resources Association. 2

[7] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *AAAI 2021 Workshop on Deep Learning on Graphs: Methods and Applications*, abs/2012.09699, 2020. 2, 3, 4

[8] A. Fornés, V. Romero, A. Baró, J. I. Toledo, J. A. Sánchez, E. Vidal, and J. Lladós. Icdar2017 competition on information extraction in historical handwritten records. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1389–1394, 2017. 2, 6, 8

[9] Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. An analysis of the performance of named entity recognition over ocred documents. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 333–334, 2019. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[11] Yaoyuan Liang, Xin Wang, Xuguang Duan, and Wenwu Zhu. Multi-modal contextual graph neural network for text visual question answering. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3491–3498, 2021. 2

[12] Hao Liu, Xin Li, Bing Liu, Deqiang Jiang, Yinsong Liu, and Bo Ren. Neural collaborative graph machines for table structure recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4523–4532, 2022. 2

[13] Claire Bizon Monroc, Blanche Miret, Marie-Laurence Bonhomme, and Christopher Kermorvant. A comprehensive study of open-source libraries for named entity recognition on handwritten historical documents. In *Document Analysis Systems: 15th IAPR International Workshop, DAS 2022, La Rochelle, France, May 22–25, 2022, Proceedings*, page 429–444, 2022. 2

[14] Pau Riba, Anjan Dutta, Lutz Goldmann, Alicia Fornés, Oriol Ramos, and Josep Lladós. Table detection in invoice documents by graph neural networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 122–127. IEEE, 2019. 2, 3

[15] Verónica Romero, Alicia Fornés, Nicolás Serrano, Joan Andreu Sánchez, Alejandro H. Toselli, Volkmar Frinken, Enrique Vidal, and Josep Lladós. The esposalles database: An ancient marriage license corpus for off-line handwriting recognition. *Pattern Recognition*, 46(6):1658 – 1669, 2013. 6

[16] Ahmed Cheikh Rouhou, Marwa Dhiaf, Yousri Kessentini, and Sinda Ben Salem. Transformer-based approach for joint handwriting and named entity recognition in historical document. *Pattern Recognition Letters*, 155:128–134, 2022. 1, 2, 3, 6, 7, 8

[17] Vijay Rowtula, Praveen Krishnan, C Jawahar, and IIIT CVIT. Pos tagging and named entity recognition on handwritten documents. In *Proceedings of the 15th International Conference on Natural Language Processing*, 2018. 1, 2, 8

[18] Teemu Ruokolainen and Kimmo Kettunen. Named entity recognition in ocred 19th and early 20th century finnish newspaper and journal collection data. In *DHN*, pages 137–156, 2020. 1

[19] J Ignacio Toledo, Manuel Carbonell, Alicia Fornés, and Josep Lladós. Information extraction from historical handwritten document images with a context-aware neural model. *Pattern Recognition*, 86:27–36, 2019. 1, 3, 8

[20] J. Ignacio Toledo, Sebastian Sudholt, Alicia Fornés, Jordi Cucurull, Gernot A. Fink, and Josep Lladós. Handwritten word image categorization with convolutional neural networks and spatial pyramid pooling. In Antonio Robles-Kelly, Marco Loog, Battista Biggio, Francisco Escolano, and Richard Wilson, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, pages 543–552, Cham, 2016. Springer International Publishing. 2

[21] Oliver Tüselmann, Fabian Wolf, and Gernot A Fink. Are end-to-end systems really necessary for ner on handwritten document images? In *International Conference on Document Analysis and Recognition*, pages 808–822. Springer, 2021. 2, 6, 7, 8

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4