UAB
Universitat Autònoma
de Barcelona

Dipòsit digital
de documents
de la UAB

This is the **published version** of the master thesis:

Molina Rodríguez, Adrià; Ramos Terrades, Oriol , dir.; Lladós, Josep , dir. Bridging Cross-Modal Alignment for OCR-Free Content Retrieval in Scanned Historical Documents. 2023. (Màster Universitari en Visió per Computador/Computer Vision)

This version is available at https://ddd.uab.cat/record/284333

under the terms of the CC BY license

Bridging Cross-Modal Alignment for OCR-Free Content Retrieval in Scanned Historical Documents

Adrià Molina Rodríguez

Abstract

In this work, we address the limitations of current approaches to document retrieval by incorporating vision-based topic extraction. While previous methods have primarily focused on visual elements or relied on optical character recognition (OCR) for text extraction, we propose a paradigm shift by directly incorporating vision into the topic space. We demonstrate that recognizing all visual elements within a document is unnecessary for identifying its underlying topic. Visual cues such as icons, writing style, and font can serve as sufficient indicators. By leveraging ranking loss functions and convolutional neural networks (CNNs), we learn complex topological representations that mimic the behavior of text representations. Our approach aims to eliminate the need for OCR and its associated challenges, including efficiency, performance, data-hunger, and expensive annotation. Furthermore, we highlight the significance of incorporating vision in historical documentation, where visually antiquated documents contain valuable cues. Our research contributes to the understanding of topic extraction from a vision perspective and offers insights into annotation-cheap document retrieval systems.

"Truth is like a blanket that always leaves your feet cold.
You push at it, stretch it, it will never be enough."
Dead Poets Society

I. Introduction

CURRENT trends in document retrieval have primarily centered around tasks related to visual elements, such as author, layout-sensitive, or class retrieval when being addressed from the computer vision perspective. When it comes to topic modeling or topic retrieval, the most typical approach involves operating directly within the language domain, assuming that all the semantic nuances are captured there. However, in cases where a document lacks textual analogue, the conventional solution has been to extract the text using optical character recognition (OCR).

In this work, we aim to address the limitations of these existing approaches. Firstly, OCR-based text extraction presents challenges in terms of efficiency, performance, data-hunger as inaccuracies in word recognition can significantly impact the quality of topic retrieval. Moreover, while off-the-shelf reading systems tend to not generalize on historical documents, the process of annotating data for OCRs is expensive, particularly for historical documentation. Note that while nowadays the population of documents exhibits some uniformity because of the increasing globalization and standardization of document generation software around the globe, in a historical scenario population tends to be sparser and more diverse due to differences in technology and tools. This sparsity in the population of historical documents causes the absence of robust tools for information extraction and search in document databases.

Additionally, it is not often trivial to provide an initial proposal or determine the appropriate annotations without the assistance of experts. This issue is further compounded in non-English scenarios and historical setups due to the aforementioned reasons. Therefore, providing solutions that do not rely on text recognition for document retrieval is not only convenient but also more feasible, especially for historical documentation.

In summary, historical documentation introduces two primary challenges in real-world applications. First, the sparsity of historical data, exacerbated by longer time gaps compared to modern documents, complicates the process of transfer learning.

Author: A. Molina, amolina@cvc.uab.cat
O. Ramos, Computer Science Department & Computer Vision Center, Universitat Autònoma de Barcelona
J. Lladós, Computer Science Department & Computer Vision Center, Universitat Autònoma de Barcelona
Thesis dissertation submitted: September 2023

Off-the-shelf methods often struggle to extract information effectively due to the unique and specific nature of the historical data at hand.

Furthermore, annotating historical data presents a unique challenge. In contrast to other document types, historical documents often necessitate expert assistance for annotation, particularly when dealing with low-resource languages and intricate layouts. This added layer of complexity further emphasizes the intricacies involved in working with historical documentation.

Alternatively, it is often unnecessary to recognize all visual elements within a document to identify its underlying topic. Visual cues, such as icons, writing style, or font, can serve as sufficient indicators to narrow down the range of potential topics. This realization highlights a key issue with OCR systems: they require the capacity to localize and recognize every single pattern, regardless of its relevance to the topic extraction process.

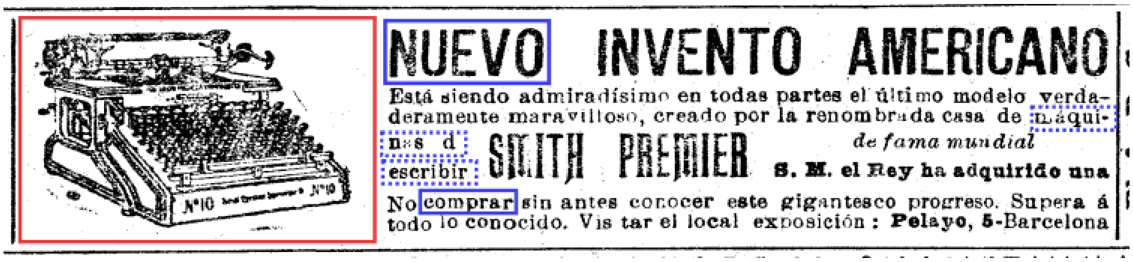
In this work, we present an approach that addresses this challenge by incorporating topic spotting without the need for such wasteful capacity. Our method allows every feature or kernel in the convolutional neural network (CNN) to be dedicated solely to the purpose of topic extraction. By doing so, we create lightweight models that are more efficient in extracting relevant topics.

This consideration is particularly important in the context of historical documentation. Visually antiquated documents often contain cues that automatically exclude topics related to modern technologies. Furthermore, when extracting information from historical newspapers, incorporating vision becomes an intriguing trend. This is because the topic can often be determined not only from the textual information but also from visual elements such as tables, diagrams, charts, photographs, and illustrations (see Figure 1). Leveraging these visual elements becomes particularly valuable in cases where the degradation of the document renders certain words unrecognizable by any means.

Our hypothesis proposes a paradigm shift by directly incorporating vision into the topic space, rather than relying on text as an intermediary representation. We argue that this approach is not only feasible but also offers significant benefits, as the model can discover efficient shortcuts to determine the topic from a given sample.

The main contribution of this work is to address the problem of topic extraction from a vision perspective. Our approach focuses on exploring the capabilities of ranking loss functions to learn complex topological representations that mimic the behavior of text representations. Given a set of textual queries (QbS), we evaluate the performance of the vision encoder for document retrieval in terms of accuracy and rank correlation. Therefore, the main objectives posed in this document are: solving the absence of annotated data for content-based document retrieval, proposing a method that can leverage information without the usage of supervised text recognition, and performing a deep analysis of the behavior of the solution under different use cases.

Through experiments and evaluation of accuracy and rank correlation in these scenarios, we can assess the performance and effectiveness of the vision encoder for document retrieval tasks. This research contributes to the understanding of topic extraction from a vision perspective and provides insights into the potential utilization of visual cues in document retrieval systems in an annotation-cheap manner.



NUEVO INVENTO AMERICANO

Está siendo admiradísimo en todas partes el último modelo verdaderamente maravilloso, creado por la renombrada casa de máquinas de escribir SMITH PREMIER de fama mundial S. M. el Rey ha adquirido una No comprar sin antes conocer este gigantesco progreso. Supera á todo lo conocido. Visitar el local exposición: Pelayo, 5-Barcelona

Fig. 1: Fragment of a document showing textual tokens (blue) indicating the presence of an advertisement. The word ”typewriter” is deteriorated and truncated (dashed), posing challenges for OCR-based approaches. The visual content (red) can potentially help overcome these challenges in the language domain, as identifying the object as a ”typewriter” can fill the missing gap in the textual domain. It is also worth noting that, in this scenario, it is easier to propose a query as an annotation for the document rather than providing the entire transcription. This aligns with our motivation to make document retrieval annotation more cost-effective.

II. State of the art

As previously introduced in Section I, our primary goal is to develop an information extraction method that can operate effectively without the need for explicit character recognition, a common requirement in the realm of historical document analysis. However, this objective presents a significant challenge due to the limited availability of data suitable for such methods.

In this section, we will initially explore how the field of information retrieval has been approached from the perspective of Natural Language Processing (NLP). Additionally, we will discuss the potential strategies for merging vision and language to comprehend multi-modal data.

A. Text-Based Information Retrieval

Before the omnichannel era, the task of information extraction relied solely on textual information. In the absence of advanced technologies and multimodal data sources, retrieving relevant information from a collection of documents was primarily approached from the perspective of Natural Language Processing (NLP) using textual data.

To retrieve information from textual data of variable sizes, traditional topic models such as bag-of-words or Term Frequency Inverse Document Frequency (TF-IDF) rely on the relative occurrence of distinctive elements. However, these frequentist approaches have limitations in capturing semantic relationships and handling polysemy. Latent Semantic Indexing (LSI) [1] addresses these issues by computing the SVD of the frequency matrix, capturing co-occurrence patterns, and disambiguating word meanings. Another approach, Latent Dirichlet Allocation (LDA) [2], models documents as probabilistic distributions over latent topics, providing a flexible representation. These advanced topic models enhance information retrieval by capturing semantic nuances and improving document similarity calculations, facilitating a deeper understanding of variable-sized textual data. Newer approaches such as SentenceBERT [3] try to take advantage of masked text encoding such as BERT, introduced by Delvin *et al.*[4], which creates a word representation by analyzing the context in which the word appears rather than the semantics of the word itself.

B. Vision and Language

Over the past decade, the integration of visual and textual information has become increasingly important in various procedures. However, obtaining the underlying textual content of documents is not always feasible. Vision and language research has responded to this challenge by developing innovative solutions, ranging from Optical Character Recognition (OCR) to Visual Question Answering (VQA) systems. Incorporating textual information into image retrieval pipelines introduces two challenges. Firstly, in ”focused-scene-text” scenarios, where text occupies the main focus of an image [5], state-of-the-art methods achieve approximately 90% accuracy [6], [7]. Secondly, in scenarios with incidental text [8], determining the relevance of textual information for retrieving insights from scenes becomes nontrivial. While document retrieval is typically treated as a focused scene text retrieval problem, documents present additional challenges due to their dense and structured nature. Extracting meaningful information from documents relies on comprehending the combination of detected tokens, posing a performance bottleneck as incomplete understanding leads to poor retrieval performance. Due to the specific nature of document retrieval, the existing research primarily focuses on retrieving documents based on factors such as the author, date, location, or classification label [9]–[12] alternatively, some DocVQA approaches [13] propose obtaining natural language answers based on visual documentation. However, there is a lack of benchmarks that specifically address the need to retrieve documents based on their topic or semantic information. One notable exception is TextTopicNet, introduced by Gomez *et al.*[14] and preceded by Lazebnik *et al.*[15], which emphasizes the alignment between topic vectors derived from Wikipedia captions and their corresponding images. Subsequently, CLIP refined this approach for image-caption retrieval [16]. Although CLIP has achieved remarkable success in various image-related tasks, its visual encoder remains text-agnostic posing a significant challenge when developing a content-based document retrieval system. This raises the question: How have we been retrieving information from scanned documents, given the limited literature in this area?

From a practical standpoint, using OCR as a means to directly perform information retrieval on the textual domain poses certain challenges and limitations. Firstly, there are efficiency issues associated with the process of extracting text using OCR. OCR involves converting image-based or scanned documents into editable text, which can be a computationally intensive task. This extraction step adds an additional processing overhead before the actual topic modeling can take place. Consequently, this can impact the overall efficiency of the information retrieval process. Furthermore, the quality of the extracted text from OCRs may become a performance bottleneck.

In contrast to the uni-modal approach of OCR, which focuses solely on extracting information from textual data, Word Spotting has emerged as a powerful solution for retrieving information in a cross-modal setting. Word Spotting aims to identify visual regions where specific textual queries appear [17]. Unlike OCR approaches, Word Spotting is limited to identifying atomic queries, such as individual words, rather than capturing the broader information contained in the document. However, as queries become more complex, a more challenging problem arises: Visual Question Answering (VQA) operates in a cross-modal scenario, utilizing natural language queries that require the system to understand and respond to questions in human language.

VQA goes beyond simple word spotting by combining visual and textual understanding, enabling machines to comprehend and provide meaningful answers based on the given visual content. This can include various types of visual content, such as images [18], scene-text images [19], or even documents [20], [21].

The goal of this work is to bridge the gap between Word Spotting and DocVQA by proposing a novel saddle-point solution. This solution enables the retrieval of documents based on their content, without the need to provide natural language answers to a given query. By solely relying on image features, similar to how it is done in Natural Language Processing (NLP), a global understanding of the document can be established, as discussed in Section II-A.

This research explores the ranking capabilities of a contrastive optimization regime [16], [22] and evaluates its performance in topic-sensitive document retrieval. It provides a valuable contribution to the field by addressing the limitations of relying solely on OCR or recognition-based solutions in historical documentation. Given the age-related damage and illegibility of historical documents, visual insights become essential for understanding and extracting information. By proposing a novel saddle-point solution that focuses on image features, this work offers a fresh approach to document retrieval in historical archives. It serves as a baseline for leveraging visual understanding and demonstrates the significance of combining visual and textual comprehension in historical document analysis. By bridging the gap between traditional OCR methods and the unique challenges of historical documentation, this research advances the field and paves the way for future advancements in information extraction and preservation.

III. Problem Statement

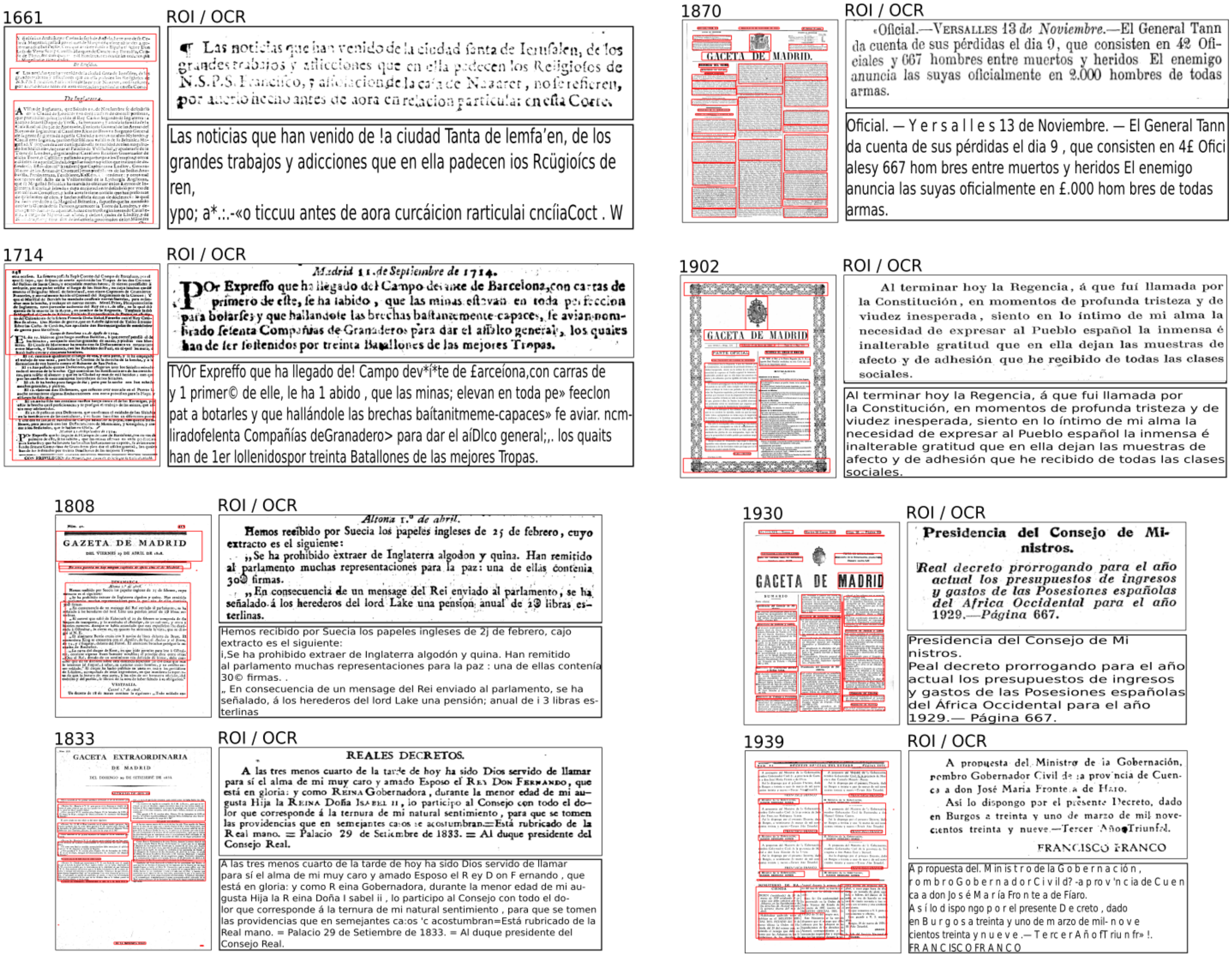
In Section I, we highlighted the unique challenges associated with topic-sensitive retrieval in the analysis of historical documents. The inherent nature of historical documentation renders the reliance solely on text-based approaches for retrieval unfeasible. The damaged condition of the documents, the lack of accurate recognition of ground truth, and the complexities of dealing with multiple languages severely impede the effectiveness of text-based methods. Consequently, it becomes crucial to incorporate visual insights to overcome these limitations and enable topic-sensitive retrieval in historical document analysis.

It is worth noting that the absence of appropriate datasets further exacerbates the challenge. Existing historical document datasets primarily focus on tasks such as classification [12], date and author retrieval [9], or word spotting [23], rather than specifically addressing the retrieval challenges. Although topic-sensitive retrieval is a common task in the NLP field with ample available datasets [24], it remains largely unaddressed in the document analysis field.

From a computer vision perspective, previous works [14] have posed topic-sensitive retrieval as a metric learning problem, clustering LDA embeddings extracted from corresponding Wikipedia articles with their respective images. However, in the case of documents, which heavily rely on textual information, defining a topic space beyond their transcription proves challenging. This discrepancy highlights the impracticality of relying solely on textual information, particularly in the context of historical documents. It may seem hard to define such topic space: except it's not.

A. Topic Retrieval Historical Dataset

The Boletín Oficial del Estado (BOE) Figure 2, also known as the Official State Gazette, holds immense importance in Spain. Serving as the official diary, the BOE is responsible for publishing all laws, decrees, provisions, resolutions, and general regulations approved by the government. Its primary objective is to ensure transparency and public access to the norms and administrative acts that impact citizens. Originally known as ”La Gazeta de Madrid” (Madrid's Gazette), the BOE was initially distributed in newspaper format. These historical documents, along with human summarizations, are available online[1] and date back to the 15th century up until the present day. The main objective for this work is to create a pipeline that will be available to correctly match such sumarizations and queries, to its correspondent documents.


1661
ROI / OCR
1870
ROI / OCR
1714
ROI / OCR
1902
ROI / OCR
1808
ROI / OCR
1930
ROI / OCR
1833
ROI / OCR
1939
ROI / OCR

Las noticias que han venido de la ciudad ſanta de Ieruſalen, de los grandes trabajos y aflicciones que en ella padecen los Religioſos de N.S.P.S. Franciſco, y afiniacion de la caſa de Nazaret, no ſe refieren, por auerlo hecho antes de aora en relacion particular en eſta Corte.

Las noticias que han venido de !a ciudad Tanta de lemfa'en, de los grandes trabajos y adicciones que en ella padecen ios Rcügioícs de ren,
ypo; a*.:.-«o ticcuu antes de aora curcáicion rarticulai cncíiaCoct . W

Madrid 11. de Septiembre de 1714.

POr Expreſſo que ha llegado del Campo delante de Barcelona, con cartas de primero de eſte, ſe ha ſabido, que las minas eſtavan en toda perfeccion para bolarſe; y que hallandoſe las brechas baſtantemente capaces, ſe avian nombrado ſetenta Compañias de Granaderos para dar el aſſalto general, los quales han de ſer ſoſtenidos por treinta Batallones de las mejores Tropas.

TYOr Expreffo que ha llegado de! Campo dev*í*te de £arceíona,con carras de
y 1 primer© de elle, le ha 1 abido , que las minas; elevan en toda pe» feeclon
pat a botarles y que hallándole las brechas baítanitmeme-capaces» fe aviar. ncm-
liradofelenta Compañías deGranadero> para dar el aíDlco general;,. los quaits
han de 1er lollenidospor treinta Batallones de las mejores Tropas.

Altona 1.º de abril.

Hemos recibido por Suecia los papeles ingleses de 25 de febrero, cuyo extracto es el siguiente:

„Se ha prohibido extraer de Inglaterra algodon y quina. Han remitido al parlamento muchas representaciones para la paz: una de ellas contenia 300 firmas.

„En consecuencia de un mensage del Rei enviado al parlamento, se ha señalado á los herederos del lord Lake una pension anual de 3© libras esterlinas.

Hemos recibido por Suecia los papeles ingleses de 2j de febrero, cajo extracto es el siguiente:
¡,Se ha prohibido extraer de Inglaterra algodón y quina. Han remitido al parlamento muchas representaciones para la paz : una de ellas contenía 30© firmas. .
„ En consecuencia de un mensage del Rei enviado al parlamento, se ha señalado, á los herederos del lord Lake una pensión; anual de i 3 libras esterlinas

REALES DECRETOS.

A las tres menos cuarto de la tarde de hoy ha sido Dios servido de llamar para sí el alma de mi muy caro y amado Esposo el Rey Don Fernando, que está en gloria: y como Reina Gobernadora, durante la menor edad de mi augusta Hija la Reina Doña Isabel II, lo participo al Consejo con todo el dolor que corresponde á la ternura de mi natural sentimiento, para que se tomen las providencias que en semejantes casos se acostumbran.=Está rubricado de la Real mano. = Palacio 29 de Setiembre de 1833. = Al duque presidente del Consejo Real.

A las tres menos cuarto de la tarde de hoy ha sido Dios servido de llamar para sí el alma de mi muy caro y amado Esposo el R ey D on F ernando , que está en gloria: y como R eina Gobernadora, durante la menor edad de mi augusta Hija la R eina Doña I sabel ii , lo participo al Consejo con todo el dolor que corresponde á la ternura de mi natural sentimiento , para que se tomen las providencias que en semejantes ca:os 'c acostumbran=Está rubricado de la Real mano. = Palacio 29 de Setiembre de 1833. = Al duque presidente del Consejo Real.

«Oficial.—Versalles 13 de Noviembre.—El General Tann da cuenta de sus pérdidas el dia 9, que consisten en 42 Oficiales y 667 hombres entre muertos y heridos El enemigo anuncia las suyas oficialmente en 2.000 hombres de todas armas.

Oficial. — Y e r s a l l e s 13 de Noviembre. — El General Tann da cuenta de sus pérdidas el dia 9 , que consisten en 4£ Ofici alesy 667 hom bres entre muertos y heridos El enemigo anuncia las suyas oficialmente en £.000 hom bres de todas armas.

Al terminar hoy la Regencia, á que fuí llamada por la Constitución, en momentos de profunda tristeza y de viudez inesperada, siento en lo íntimo de mi alma la necesidad de expresar al Pueblo español la inmensa é inalterable gratitud que en ella dejan las muestras de afecto y de adhesión que he recibido de todas las clases sociales.

Al terminar hoy la Regencia, á que fui llamada por la Constitución, en momentos de profunda tristeza y de viudez inesperada, siento en lo íntimo de mi alma la necesidad de expresar al Pueblo español la inmensa é inalterable gratitud que en ella dejan las muestras de afecto y de adhesión que he recibido de todas las clases sociales.

Presidencia del Consejo de Ministros.

Real decreto prorrogando para el año actual los presupuestos de ingresos y gastos de las Posesiones españolas del Africa Occidental para el año 1929.—Página 667.

Presidencia del Consejo de Mi nistros.
Peal decreto prorrogando para el año actual los presupuestos de ingresos y gastos de las Posesiones españolas del África Occidental para el año 1929.— Página 667.

A propuesta del Ministro de la Gobernación, nombro Gobernador Civil de la provincia de Cuenca a don José María Fontana de Haro.

Así lo dispongo por el presente Decreto, dado en Burgos a treinta y uno de marzo de mil novecientos treinta y nueve.—Tercer Año Triunfal.

FRANCISCO FRANCO

A p ropuesta del. M in is t r o de la G o b e r n a c ió n , r o m b r o G o b e r n a d o r C i v i l d? -a p r o v 'n c ia de C u e n ca a don Jo s é M a r ía Fro n te a de Fáro.
A s í lo d ispo ngo p o r el presente D e creto , dado en B u r g o s a treinta y uno de m arzo de mil- n o v e cientos treinta y n u e v e .— T e r c e r A ñ o fT riu n fr» !.
FR A N C IS CO FR A N C O

Fig. 2: Diverse Examples from the "Boletín Oficial del Estado" across different years. Noteworthy observations arise when examining the layout evolution. Certain documents (1661, 1714, 1808) adhere to a single-column structure, with two of them (1661, 1714) displaying archaic handwriting styles that pose challenges for modern OCR systems. Conversely, other time periods (1833, 1902, 1939) adopt a dual-column layout, occasionally alternating with phases of triple-column formatting (1870, 1930). The dataset's remarkable diversity precludes the existence of a universal segmentation method performing at the cutting edge. The automated detection of Regions of Interest (ROIs) presents a twofold challenge, tied both to the quality of the extracted OCR and the efficacy of the segmentation technique.

The BOE dataset consists of a vast collection of around 1.5 million scanned multipage documents. However, for the purposes explained in the following sections, we will be utilizing a smaller subset of 200,000 documents from the total dataset. It is important to note that the dataset contains OCR transcriptions that are not entirely reliable but are sufficient for establishing certain heuristics, which we will explore as follows.

B. Addressed Limitations

When it comes to createing the annotations for query-document matching, the most direct approach would involve relying on OCR transcriptions. However, it's important to note, as depicted in Figure 2, that there exists an inconsistency in both the

[1] https://www.boe.es/buscar/gazeta.php

quality of the OCR-extracted text and the segmentation of document regions. One potential solution could be to refine these methods specifically for this dataset. Nevertheless, it's worth mentioning that we currently lack annotations for the data. In the context of historical documentation, domains tend to be sparse; this means that the changes resulting from factors like time and location are significantly more pronounced compared to modern documents, which have a more concise representation due to globalization and shorter timestamps.

While annotating this dataset remains a possibility, it's a resource-intensive approach that requires experts, particularly in cases where the text is not straightforward even for native Spanish speakers. This challenge is compounded by the complexity of historical language and context; hence, in this document, we explore an annotation-cheap solution for information extraction in historical documents.

C. Query and Document Matching

One of the primary challenges we encounter when using this dataset is accurately associating queries with the corresponding regions in the document. The newspaper format, which allowed writers to cover various topics on a single page, presents a particular difficulty. Given that a query relates to a specific news item on one of the pages, it becomes crucial to identify which part of the document the query is referring to. It is essential to establish robust methods for disambiguating and pinpointing the relevant content before feeding the data into subsequent models. Assigning a whole page to just one of its topics would be arbitrary and inadequate for effective topic retrieval. In this regard, we leverage the OCR transcriptions and employ a two-step selection criteria Figure 5. First, we utilize a Mask-RCNN model [25] trained on the Prima Layout Analysis Dataset [26] to identify relevant regions in the document. Then, we proceed to match the query (human-annotated summary) with the available OCR text comparing SentenceBert [3][2] encoding with a cosine similarity measure. However, it is important to note that the OCR transcriptions are not entirely reliable, as depicted in Figure 5. To further enhance the accuracy of the associations made between queries and document regions, we employ a filtering step using the Hungarian distance measure. This measure evaluates the similarity of named entities by taking into account the Edit Distance for each possible pair of strings. By applying this additional filtering, we ensure that only matches with a satisfactory Hungarian distance for their named entities are considered.

It is important to note that this procedure assumes the query to be present in the document in all cases. However, failures of the Mask-RCNN model may challenge this assumption. To mitigate potential issues, we eliminate matchings with low similarity, considering only pairs with a similarity greater than 50% as valid. This filtering step helps address cases where the Mask-RCNN model fails to accurately identify the query in the document. This is imposed in the test set in order to provide a fair comparison of valid examples. For training set, is left as a parameter that may be tuned in order to increment performance. During Section VI it will be shown how, up to a point, using more data is more worthwhile than using less data with better quality.

To address the computational expenses and accommodate the diverse nature of the data, we made the deliberate decision to create a smaller subset consisting of more modern documents from the available one million dataset. Specifically, we focused on documents spanning the years 1923 to 1976. This strategic selection allows us to optimize our analysis while still capturing significant temporal diversity. The chosen time range encompasses distinct historical periods, including two military dictatorships (1923-1931, with one being fascist from 1939-1976), a republican period (1931-1936), and a civil war (1936-1939). Despite the reduced dataset size, this subset provides a rich and varied collection of documents and topics for our research.

To address the computational demands of our research, we conducted data processing using four Titan Xp GPUs in 8 parallel threads for each device. This configuration of high-performance hardware allowed us to efficiently process our dataset, which involved intricate tasks such as PDF parsing, mask computation, and OCR extraction for each region. Despite the optimized hardware setup, the sheer volume of data required approximately a month to complete the entire process. This timeframe highlights the significant computational resources needed to handle the dataset efficiently.

It is important to note that while we were able to handle the processing requirements for the subset of data we worked with, addressing the computational expenses of handling the entire dataset, consisting of 1 million documents, will pose a challenging task in the future. As we look ahead, addressing these computational demands will be an important focus area for further research. Finding efficient solutions to handle the extensive dataset will enable us to explore the complete collection of documents and extract valuable insights from the historical data.

[2] Using `distiluse-base-multilingual-cased-v1` weights in `sentence-transformers`'s implementation.

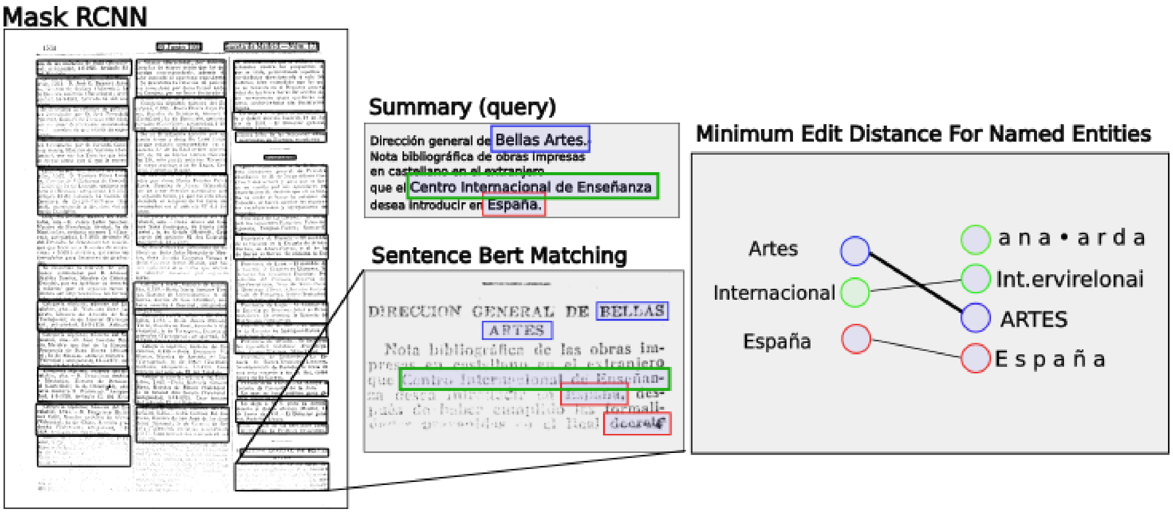

Mask RCNN
Summary (query)
Dirección general de Bellas Artes.
Nota bibliográfica de obras impresas
en castellano en el extranjero
que el Centro Internacional de Enseñanza
desea introducir en España.
Sentence Bert Matching
DIRECCION GENERAL DE BELLAS
ARTES
Minimum Edit Distance For Named Entities
Artes
Internacional
España
a n a • a r d a
Int.ervirelonai
ARTES
E s p a ñ a

Fig. 3: Query-Region Matching system for creating the Ground Truth (GT) for the Proposed Dataset In this system, we begin by extracting the text regions from the document and identifying the most similar region using the available OCR. To ensure the reliability of the OCR, we compare the named entities extracted from it. It is important to note that this procedure assumes the presence of the query in the document, enabling the system to function even when dealing with documents of poor OCR quality. Given the inherent challenges in recognizing such documents, we anticipate minimal variations in similarities. Therefore, using the maximum similarity as a criterion is sufficient to establish the matching.

By focusing on a specific time range, we can better adapt our data curation pipeline and algorithms to the characteristics and challenges present in the chosen historical period. Older documents present unique challenges when applying state-of-the-art layout segmentation algorithms and OCR methods that are predominantly designed for contemporary document standards. Adapting our pipeline to accommodate the diverse requirements of each historical period would have introduced substantial complexity and potentially compromised the accuracy of our results.

While our decision to work with a smaller subset is driven by the computational demands and the need for a more refined segmentation model, we recognize that further research is necessary to develop a comprehensive dataset that encompasses the entire collection of historical documents. This future work will enable us to capture the vast variance of document types and topics throughout history, contributing to a more comprehensive understanding of historical document analysis in the field of computer vision.

D. Topic Retrieval Problem Statement

As discussed in the previous sections (Section I), the primary aim of this research is to establish a visually meaningful representation that effectively preserves semantics within its inherent topic space T. To achieve this goal, we introduce the concept of a ”document” denoted as d and its ”query” q , which has an associated topic represented as $T(d) \in \mathbb{R}^n$. Each document consists of both a visual $d^v \in \mathcal{D}$ and a textual d^t representation. In the case of historical documents, d^t is annotation-expensive and most of the times unfeasible to use; therefore, although we introduce the problem formulation for any modalities s, the textual representation of the multi-modal scenario will be the query rather than the transcription of the document.

To represent these visual and textual aspects, we employ real-valued vectors denoted as v and t, with $\phi^s : \mathcal{D} \rightarrow \mathbb{R}^n$, where $\nu = \phi^v(d^v_n)$ and $\tau = \phi^t(q_n)$ respectively stand for the encoded visual and textual representations of the document d through the neural network encoder.

The challenge of contrastive learning lies in addressing minimization schemes that often lead to trivial solutions [27]. In our multi-modal scenario, the problem can be formulated as creating a representation such that the expression $\langle \tau_n, \nu_n \rangle$ is minimized. However, this formulation can lead to undesired results, such as $\phi^s(d_n) = \phi_s(d_{n+1})\ \forall n$. To deal with this issue, contrastive approaches introduce a maximization constraint $\min\left\{\langle \tau_n, \nu_n \rangle) - \langle \tau_m, \nu_n \rangle\right\}$, seeking a saddle point in the optimization space. This constraint is explicitly satisfied by triplet loss [22], which utilizes a single negative example per query, or more recent approaches such as SimCLR, MoCo, or CLIP [16], [28], [29], which consider every element in the batch, except the positive, as a negative example. These methods aim to address the triviality of the minimization procedure by introducing negative examples that provide a continuous space.

However, this approach is not fully aligned with the concept of topic modeling, where smooth transitions between documents are expected, meaning that it's possible to establish topics as linear combination of other ones. Contrastive learning approaches assume that the summation of hard and soft negatives leads to such a continuous space. In our work, we seek to explicitly define the surface $\phi^s(d)$ in terms of a predefined topic space T, drawing inspiration from Engilberge *et al.*[30]. By doing so, we aim to create a representation that aligns more closely with the inherent structure of the topic space. To conclude, we aim to evaluate two main aspects of the model's performace. First, we define the retrieval capabilities as the capacity of the model to establish a common representation for both d_n and q_n. On the other hand, we pose ranking capabilities as the capacity of the model to behave equally in several modalities. Therefore, in evaluating our model's performance, we must consider both its retrieval and its ranking capabilities.

On one hand, we have to be accountable for traditional retrieval evaluations such as mean average precision (mAP) and top-k accuracy. However, note that there is only one query corresponding to each document, which doesn't align with the idea of mean average precision's monotony. For the sake of simplicity, accuracy will be reported for top-1, 5 and 10 in the ranking withing a support set. This decision is taken partially because of the queries having atomic nuances that wouldn't be distinguished in bigger datasets (using the whole test set as support). While further research is needed in terms of a better evaluation of the problem statement, the behavior shown in the experiments should be enough to address most of the stated hypothesis. Therefore, during the experiments on Table I, a support set of 256 documents will be chosen as a distractors for the correct document d; which will will have to be ranked on the top of the support set. Additionally, we propose the idea of both image and queries spaces being similar, which will be reported as it's mentioned in the following section, where we will introduce the formal definition of the problem and the metrics relating the meaningfulness of the generated embedding space.

To gain a clear understanding of the desired topic modeling behavior, we introduce the following notation and definitions. Let us consider a specified radius, denoted by $\epsilon > 0$, which represents a tolerance threshold. Now, focusing on the document representations in a particular domain, indicated as $\phi^s(d)$, and their corresponding representations in the topic space, denoted as $T(d)$, we define two sets of documents. Firstly, we define $B_s(d_n^s, \epsilon) = \{d : \|\phi^s(d^s) - \phi^s(d_n^s)\| \leq \epsilon\}$ as the set of documents surrounding document d_n in the domain s. This set resembles the concept of a ball in a metric space, capturing documents within a certain distance from the focal document d_n. Analogously, we establish $B_T(d_n, \epsilon')$ as the set of documents surrounding document d_n in the topic space, represented by $T(d)$. In other words, it encompasses documents in the topic space that are close to document d_n based on the same criteria, but now applied to the topic representations.

Now, our objective is to establish that for a given document d_n in domain s, the sets of surrounding documents in both representations, $B_s(d_n^s, \epsilon)$ and $B_T(d_n, \epsilon')$, are equal. In other words, the neighbors of d_n in the domain s representation, denoted as $\text{kNN}_{\phi_s}(d_n)$, should match the neighbors of d_n in the topic space representation, denoted as $\text{kNN}_T(d_n)$. This condition ensures that the neighborhoods of documents in both the original domain and the topic space are preserved, thereby establishing the notion of ”behaving equally”.

For the evaluation of ranking similarity, we propose several metrics, starting with the Spearman rank correlation (Eq 1). This metric assesses the correlation of two sets based on their relative positions in an all-vs-all ranking. The rank correlation is defined using a ranking function R and the respective covariance between two random variables. Unlike Pearson's correlation test, Spearman rank correlation serves as a ”linearizer” for the data, making it robust to non-linear behaviors. It focuses on studying monotonicity rather than linear correlation, which aligns well with the idea of $\text{kNN}_T(d) \approx \text{kNN}_{\phi^s}(d)$, regardless of the specific values for ϵ and ϵ'.

$$\rho = \frac{COV(R(X), R(Y)}{\sigma_{R(X)}, \sigma_{R(Y)}} \tag{1}$$

Additionally, we may be interested in measuring the dissonance between the localities of each document rather than the global behavior of the data distribution. For this purpose, we introduce a measurement of intersection over union (IoU). Given that our notion of neighborhood is defined under different ϵ (Eq 2) for both T and ϕ (which is a requirement when seeking rank correlation rather than regression, as it wouldn't be possible under some constraints on T), there is a need for a discrete definition that does not require manual crafting of how a locality is expressed in each domain.

$$IoU_{\epsilon,\epsilon'} = \frac{|B_s(d_n^s, \epsilon) \cap B_T(d_n, \epsilon')|}{|B_s(d_n^s, \epsilon) \cup B_T(d_n, \epsilon')|} \tag{2}$$

To address this, we define a discrete interpretation (Eq 3) of the IoU measurement that can be applied for each control point k, expressing the number of neighbors being taken into account for both sets. This eliminates the necessity of establishing localities in a Euclidean regime and, instead, looks for discrete topologies that mimic each other. For example, IoU_1 will measure if, given a document d, its top match in both spaces is the same, while IoU_5 would measure the number of common documents within the top-5 ranked ones. By using these discrete IoU metrics, we can ensure that the surrounding documents for each document are the same in both domains, providing valuable insights into the local behavior of the representations. In this way, we can evaluate the model in terms of retrieved content rather than metric regression since in a real world scenario we are interested in the ranking behaving as close to the target T as possible, regardless of the resulting topology.

$$IoU_k^{Rank}(d) = \frac{|\,\mathrm{kNN}_T^k(d) \cap \mathrm{kNN}_\phi^k(d)|}{|\,\mathrm{kNN}_T^k(d) \cup \mathrm{kNN}_\phi^k(d)|} \tag{3}$$

In summary, by employing both the Spearman rank correlation and the discrete IoU metrics, we can comprehensively evaluate the ranking similarity between our model's representations (ϕ^s) and the true topic space (T). These evaluation measures capture both the global and local aspects of ranking behavior, allowing us to validate the effectiveness of our approach in preserving the meaningful relationships and relative importance of documents in the context of the topic space.

By adopting this comprehensive evaluation scenario, which addresses both topic modeling and retrieval requirements, we can thoroughly assess the effectiveness and relevance of our proposed approach. This ensures that the model's representations not only capture meaningful semantics within the topic space but also facilitate accurate and contextually relevant document retrieval. Such a dual evaluation strategy will provide valuable insights into the performance of our method in real-world applications and its potential to serve as a reliable tool for information retrieval tasks that demand both semantic consistency and effective ranking capabilities.

IV. Syntactic Graphs

In addition to the document retrieval perspective, we introduce the concept of a structured solution that aligns with the inherent syntactic complexities found in information-rich historical data. Given that BOE documents primarily focus on legal statements, laws, and legislative regulations, there is a noticeable repetition of patterns in the queries. Often, a specific token indicates whether the document contains a law, royal decree, or resolution. This token is associated with a noun chunk that represents the entity affected by the resolution referred to in the query. As a result, various verbs, attributes, and number modifiers interact within these structures. These patterns tend to recur across different queries. While the tokens' significance may vary, the underlying structures remain consistent over time.

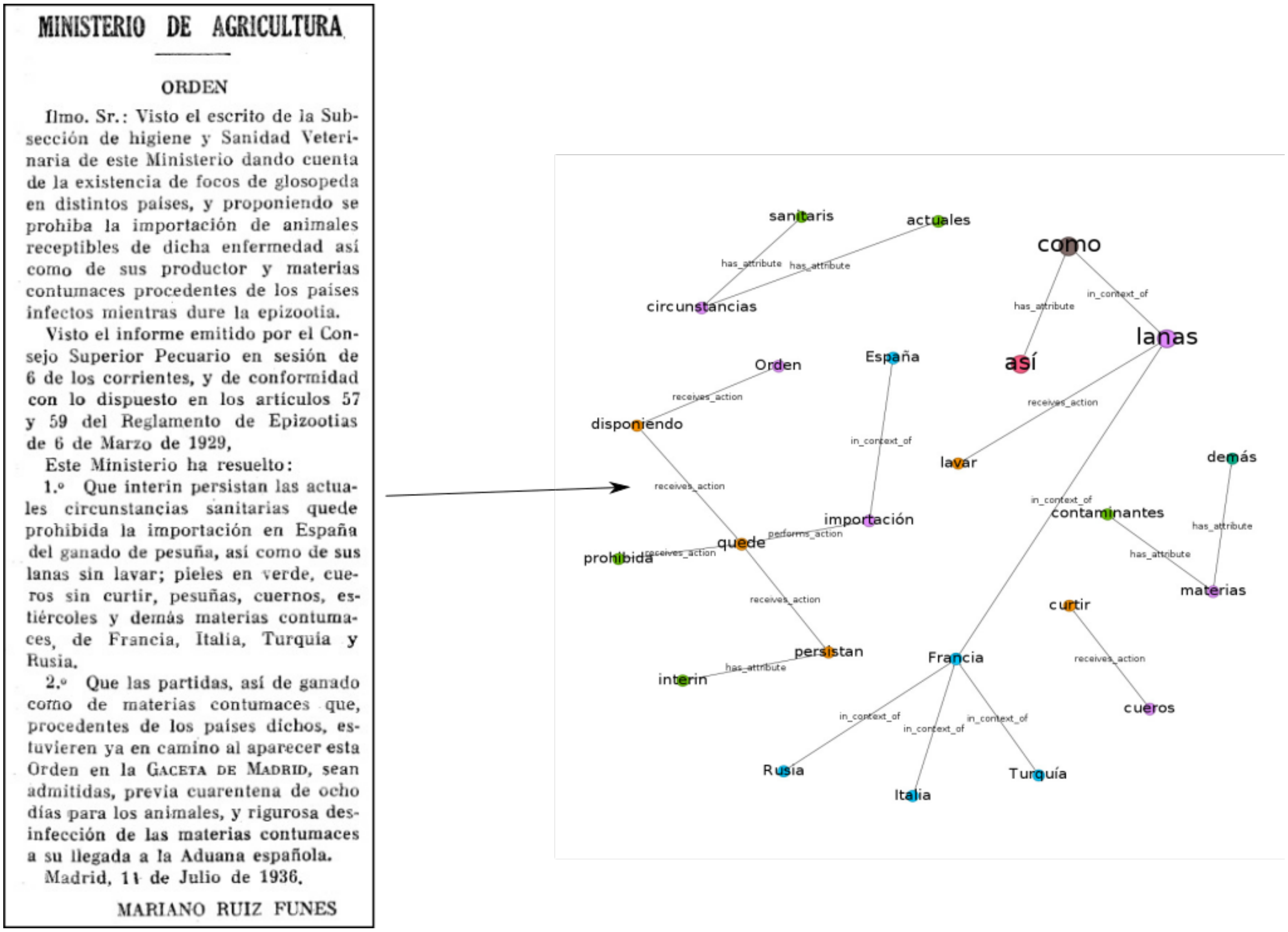
This perspective reveals a graph-like behavior that we can potentially leverage to encode the document's content. By analyzing the relationships and connections between tokens, we can capture the intrinsic structure of the information within the documents. This approach holds promise for effectively representing and understanding the content of BOE documents. In other words, here a structured perspective is tackled in contrast to the unstructured nature of topic models.

For doing so, we made use of Spacy's [31] POS tagger features[3]. Here, we define some syntactical rules that produce this syntactic graph for each query. Most of the rules stated in the following sections are based on well-known syntactic relationships in lingüistics and Natural Language Processing (NLP). While syntactic trees structure information in a hierarchical manner, the addressed challenge here lies in establishing a transformation on such a tree that can capture semantic structures in the text from the functional relationships in language.

1) Subject-Verb-Object: The described relationship between tokens in a query, particularly the Subject-Verb-Object structure, is a crucial aspect of understanding and extracting information from text. To achieve this, all the verbs within the query are extracted. For each verb, its head token (the subject) and its children (typically the object, acting as a direct complement) are extracted. This process allows you to establish a triplet relationship in a graph format, which can be represented as follows:

$$\texttt{<subject>} \xrightarrow[\texttt{action}]{\texttt{performs}} \texttt{<verb>} \xrightarrow[\texttt{action}]{\texttt{receives}} \texttt{<object>}$$

[3] https://github.com/explosion/spacy-models/releases/tag/es_dep_news_trf-3.6.1



MINISTERIO DE AGRICULTURA

ORDEN

Ilmo. Sr.: Visto el escrito de la Subsección de higiene y Sanidad Veterinaria de este Ministerio dando cuenta de la existencia de focos de glosopeda en distintos países, y proponiendo se prohiba la importación de animales receptibles de dicha enfermedad así como de sus productor y materias contumaces procedentes de los países infectos mientras dure la epizootia.

Visto el informe emitido por el Consejo Superior Pecuario en sesión de 6 de los corrientes, y de conformidad con lo dispuesto en los artículos 57 y 59 del Reglamento de Epizootias de 6 de Marzo de 1929,

Este Ministerio ha resuelto:

1.º Que interin persistan las actuales circunstancias sanitarias quede prohibida la importación en España del ganado de pesuña, así como de sus lanas sin lavar; pieles en verde, cueros sin curtir, pesuñas, cuernos, estiércoles y demás materias contumaces, de Francia, Italia, Turquía y Rusia.

2.º Que las partidas, así de ganado como de materias contumaces que, procedentes de los países dichos, estuvieren ya en camino al aparecer esta Orden en la GACETA DE MADRID, sean admitidas, previa cuarentena de ocho días para los animales, y rigurosa desinfección de las materias contumaces a su llegada a la Aduana española.

Madrid, 11 de Julio de 1936.

MARIANO RUIZ FUNES

Fig. 4: Example of a fragment of a document and its respective query ”*Orden disponiendo que interin persistan las actuales circunstancias sanitaris quede prohibida la importación en España de ganado de pesuña, así como de sus lanas sín lavar, pieles en verde, cueros sín curtir, pesuñas, cuernos, estiércoles y demás materias contaminantes, de Francia, Italia, Turquía y Rusia*” parsed through the proposed text to graph parser. Since syntactical relationships rely on functional aspects of language rather than semantical, it's not trivial to yield a proper generalization for each kind of functional relationship.

However, if the `<object>` token isn't a direct complement, but rather part of a prepositional phrase or serves as an attribute of a linking verb, it's assumed that the verb is part of a subordinate infinitive subject or a nominal predicate. In this case, the object is omitted since it functions as the prepositional link for the subordinate phrase. On the other hand, the attribute for the nominal predicate is used as object.

2) Token-Attribute: For the Token-Attribute relation, all adjectives and manner adverbs are parsed along its respective roots. By doing so, the relation is stablished as follows:

$$\texttt{<object>} \xrightarrow[\texttt{attribute}]{\texttt{has}} \texttt{<attribute>}$$

Note that the object must be a noun chunk, which is represented by its root connecting the chunk to its modifiers.

3) Quantity-Token: Quantity modifiers are those numerical tokens related to an object similarly to Token-Attribute structures. All numerical tokens are parsed and the graph is constructed relating the quantity to the object which is being modified. The measurment or the units of such quantity often is indicated through Subject-Verb-Object structure. Therefore, it's not needed to indicated the units which the quantity is referencing, only the modified token.

$$\texttt{<object>} \xrightarrow[\texttt{quantity}]{\texttt{has}} \texttt{<number>}$$

4) Shared Context: In natural language text, numerous tokens are interconnected through the use of prepositions, subordinating clauses, and enumerations, forming cohesive chunks of information. This contextual interplay within chunks, functioning as unified entities, is effectively encoded through the concept of a ”Shared Context” relationship. This mechanism allows us to capture certain relationships that are challenging to generalize using conventional means, particularly when considering the functional aspects of language, such as syntax. Consequently, this approach leads to a notable reduction in the complexity of graphical representations by simplifying each chunk to its root element. This methodology reveals two distinct types of contextual relationships:

$$\texttt{<token>} \xrightarrow[\texttt{context}]{\texttt{in}} \texttt{<root of chunk>}$$

$$\texttt{<root of chunk>} \xrightarrow[\texttt{context}]{\texttt{in}} \texttt{<root of preposition>}$$

Employing this approach provides a powerful means to dissect and illustrate the intricate relationships that exist among various components within text. This technique proves especially valuable in the extraction and comprehension of document structures and content, particularly in contexts such as legal documents like BOE documents. For further implementation details, the parser has been made publicly available via GitHub[4].

V. Methods

The initial strategy we have undertaken involves leveraging the concept of $T(q)$ as a pseudolabel for the text encoder ϕ_t. In order to grasp a deeper understanding of this approach, please refer to Figure 5. Within this framework, a crucial aspect is the imposition of a maximization constraint between two critical elements: $T(q)$ and τ. This constraint is designed with the specific goal of ensuring that the ranking of embeddings, both for $T(q)$ and $\phi_t(q_n)$, exhibit a high degree of ranking similarity within each batch. In this approach, the concept of a ”pseudolabel” is central. Essentially, $T(q)$ is employed as a surrogate label for the text encoder ϕ_t. This implies that $T(q)$ serves as a reference point or a label-like representation for the textual information. This section will address the basic blocks of our pipeline Section V-A, the metric learning approach to solve our re-identification problem Section V-B, the usage of pseudolabels to assist the re-identification problem Section V-C and, lastly, an alternative approach briefly tackled which incorporates structural information into the groundtruth.

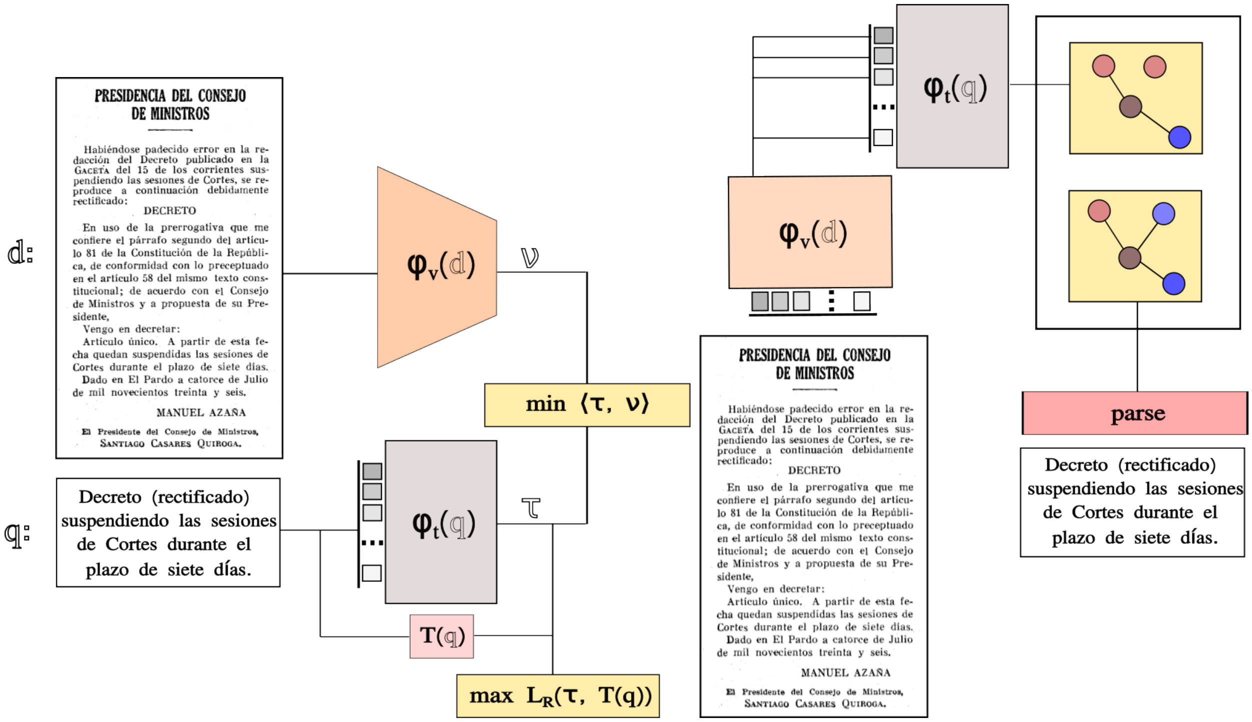
A. Image and Text Encoders

We utilize both Image and Text transformers to define ϕ_v and ϕ_t respectively. For the image encoder, we tokenize each 16x16 patch within the Region of Interest (ROI) using a linear transformation, while textual tokens are embedded. It’s worth noting that a pivotal decision in this setup is the deliberate avoidance of cross-attention between the two modalities. This choice is motivated by the need for the image encoder to capture intrinsic characteristics within its pixel content, allowing attention maps to exclude irrelevant textual tokens (i.e., words in the document) rather than leading to a trivial solution where the topic can be described by zero-ing the visual features and working only on the textual domain. This approach leaves the task of information integration to the attention mechanism, eliminating the need to make explicit decisions about handling heavily damaged visual tokens. In such cases, the visual encoder autonomously zeroes out tokens that lack meaningful information while emphasizing others. For specific implementation and chosen architectures consult Annex B; but, in summary, a ViT 8x12 (heads x layers) is chosen as visual encoder for Table I experiments; while a transformer encoder 2x4 has proven to be enough to capture the semantic nuances in the queries.

B. Learning Objectives with τ and ν

As suggested by Musgrave *et al.*[32], we have observed that among a wide array of loss functions, there are no significant differences in performance (see Annex B). However, contrastive learning plays a pivotal role in accurately deduplicating

[4] https://github.com/CVC-DAG/spacy-graph-parser



PRESIDENCIA DEL CONSEJO DE MINISTROS

Habiéndose padecido error en la redacción del Decreto publicado en la GACETA del 15 de los corrientes suspendiendo las sesiones de Cortes, se reproduce a continuación debidamente rectificado:

DECRETO

En uso de la prerrogativa que me confiere el párrafo segundo del artículo 81 de la Constitución de la República, de conformidad con lo preceptuado en el artículo 58 del mismo texto constitucional; de acuerdo con el Consejo de Ministros y a propuesta de su Presidente,

Vengo en decretar:

Artículo único. A partir de esta fecha quedan suspendidas las sesiones de Cortes durante el plazo de siete dias.

Dado en El Pardo a catorce de Julio de mil novecientos treinta y seis.

MANUEL AZAÑA

El Presidente del Consejo de Ministros,
SANTIAGO CASARES QUIROGA.

Decreto (rectificado) suspendiendo las sesiones de Cortes durante el plazo de siete días.

PRESIDENCIA DEL CONSEJO DE MINISTROS

Habiéndose padecido error en la redacción del Decreto publicado en la GACETA del 15 de los corrientes suspendiendo las sesiones de Cortes, se reproduce a continuación debidamente rectificado:

DECRETO

En uso de la prerrogativa que me confiere el párrafo segundo del artículo 81 de la Constitución de la República, de conformidad con lo preceptuado en el artículo 58 del mismo texto constitucional; de acuerdo con el Consejo de Ministros y a propuesta de su Presidente,

Vengo en decretar:

Artículo único. A partir de esta fecha quedan suspendidas las sesiones de Cortes durante el plazo de siete dias.

Dado en El Pardo a catorce de Julio de mil novecientos treinta y seis.

MANUEL AZAÑA

El Presidente del Consejo de Ministros,
SANTIAGO CASARES QUIROGA.

Decreto (rectificado) suspendiendo las sesiones de Cortes durante el plazo de siete días.

Fig. 5: Optimization scheme for ϕ_v and ϕ_t within the context of a document-query pair (left). And graph encoding using query to graph parser as groundtruth (right). Our hypothesis centers on the notion that $T(q)$ serves as a valuable ”pseudolabel,” aiding in the discernment of key elements crucial for identifying the content within document d.

document and summary pairs. In this context, Circle Loss [33] has proven to be a viable option for re-identifying pairs. While its primary application is in person re-identification, our problem shares similarities since there are no distinct positive and negative labels for samples. Instead, each sample has only one correct pair. Therefore, Circle Loss exhibits a slight performance advantage over other metric learning approaches, such as triplet loss [22]. In summary, circle loss 4 solves the re-identification problem by optimizing all similarity pairs S grouping them between positive and negative matches, P and N respectively, with a margin factor $m = 0.40$ and temperature $\gamma = 80$ in our experiments.

$$\mathcal{L}_{circle} = \log[1 + \sum_{i=0}^{P} \sum_{j=0}^{N} e^{\gamma(S_j - S_i + m)}] \tag{4}$$

Note that this is suited for cases where re-identification is done via several views of a sample. In our case, we only have two views per sample: the query and the document. Therefore, it’s important to note that Equation 4 simplifies to Equation 5 since there’s a single positive pair S_P.

$$\mathcal{L}_{circle} = \log[1 + \sum_{j=0}^{N} e^{\gamma(S_j - S_P + m)}] \tag{5}$$

In summary, the base solution is interpreted as a re-identification probelm with two views on the same document rather than a classic metric learning perspective. Although this is the main perspective of this work, alternative baselines are proposed in Table I capturing several approaches in the literature exposed during Section II.

C. Auxiliar Loss with τ and $T(q)$

It is important to note that establishing a correspondence between the pseudolabel $T(q)$ and the text embedding ν often requires them to have the same shape. While we demonstrate in further experiments the ability to use arbitrarily shaped topic models with surrogate ranking loss functions, for our current approach, we have opted for LDA (Latent Dirichlet Allocation) as pseudolabels for the queries. This decision is motivated by the considerations outlined in Section II-A.

As discussed in Section II-A, TF-IDF does not provide pseudolabels of arbitrary sizes; instead, it generates large vectors corresponding to the dictionary size. On the other hand, LSI (Latent Semantic Indexing) can model a corpus with a specified embedding size but relies on TF-IDF eigenvalues, potentially limiting its variability.

In contrast, the generative and probabilistic nature of LDA offers sufficient variance in the topic space to represent any given number of topics, effectively maximizing the out-class variance. Therefore, LDA serves as a suitable choice for our pseudolabeling approach.

To generate T such that $T : q \rightarrow \mathbb{R}^n$, we make use of the training sample pairs (q, d) that define the topic model that will be used as pseudolabels for distilling the semantical information into an encoder. Once a document is processed through $\phi_{\{t,c\}}$, we consider the embedding under a restriction within $T(q)$ such that its rank similarity is maximized. For doing so, we use traditional metric learning loss function; but we also propose the usage of a ranking surrogate that mimics the behavior of Spearman Rank Correlation defined in Section III.

Although this can be seen as a third view in the re-identification process, we also provide results with a constraint that maximizes the rank correlation rather than the similarity. This is relevant since the solution space containing correlated rankings should be greater than the solution space implying equality (maximum similarity); therefore flexibility is provided through this perspective. Given two similarity pairs, S_i and S_j, contained in the batch, we compare them using a logistic activation function with a sufficiently high temperature, $\gamma = 1 \times 10^4$.

$$\mathrm{I}_{i,j} = \frac{1}{1 + e^{\frac{1}{\gamma}(S_i - S_j)}} \approx \begin{cases} 0 \text{ if } S_i > S_j \\ 1 \text{ if } S_i < S_j \end{cases} \tag{6}$$

Therefore, for a given query $T(q)$, we can determine the position of each batch sample in the ranking by calculating the cumulative sum of the indicators. For example, the i-th document in the batch B is ranked in the N-th position with respect to the query as:

$$\text{Rank}_i \approx |B| - \sum_{j \neq i}^{B} I_{i,j} \tag{7}$$

Here, both S_i and S_j represent similarities with respect to the query, which is excluded from the batch. Once each of the samples is ranked such that $R(T) \approx (Rank_0, Rank_1, ..., Rank_n)$, we can approximate Spearman's Rank Correlation as stated in Section III.

$$R(f) = R((f(x)|x \in B)) \tag{8}$$

The loss function $\mathcal{L}_\rho$ for Spearman's Rank Correlation can be expressed as:

$$\mathcal{L}_\rho = 1 - \frac{COV(R(T), R(\phi))}{\sigma_{R(T)}\sigma_{R(\phi)}} \tag{9}$$

In this way, we construct the loss surrogate for Spearman's Rank Correlation by approximating the binary step function with a logistic function.

D. Graph Encoder Through LLMs

Finally, we utilize a visual encoder-decoder model, as depicted in Figure 5, which undergoes fine-tuning with the assistance of structured queries represented by syntactic graphs. While this approach is currently in its initial stages, we posit that leveraging structured data may enhance information extraction by imposing explicit constraints on the detected visual tokens. For instance, imposing the location of specific nodes can provide valuable guidance in describing the document's content. While the visual

Configuration					Evaluation				
Topic Distillation	Contrastive Loss	Topic Loss	Acceptance	Distilled Encoder	Acc@1	Acc@5	Acc@10	IoU$^{10}_{rank}$	rho
✓	Circle	Circle	0.4	τ	0.596	0.786	0.836	0.224	0.346
				ν	**0.663**	**0.817**	**0.866**	0.255	**0.431**
			0.17	τ	0.608	0.789	0.849	0.22	0.348
				ν	0.602	0.756	0.796	0.255	**0.438**
✗	Circle	-	0.4	-	0.042	0.196	0.317	0.096	0.166
		-	0.17		0.054	0.190	0.304	0.097	0.146
✗	Rank	-	0.4	-	0.0	0.009	0.017	0.0	0.306
✗	Triplet	-	0.4	-	0.019	0.112	0.185	0.061	0.191
			0.17		0.334	0.621	0.710	**0.274**	0.349
✗	SimCLR	-	0.4	-	0.502	0.719	0.786	0.226	0.292
			0.17		0.548	0.744	0.804	0.207	0.266
✗	CLIP	-	0.4	-	0.342	0.624	0.743	0.191	0.294

TABLE I: Table of results comparing the usage of pseudolabels (topic distillation) and whether the pseudolabels have been applied to the vision encoder (ν) or the text encoder (τ).

encoder, discussed in Section V-A, remains the same, we introduce a text decoder in the form of GPT-2 [34]. The decoder plays a pivotal role in generating both the nodes and edges within the graph. It undergoes fine-tuning to capitalize on the inherent knowledge within the decoder, allowing it to leverage the relationships between the identified nodes effectively.

Our rationale behind this approach is as follows: the visual encoder excels at identifying visual regions containing atomic information pertinent to the query, essentially representing them as nodes. Meanwhile, GPT-2 is proficient in establishing meaningful relationships between these nodes through the creation of edges. It's important to note that this end-to-end approach is still in its early stages and may not yield immediately promising results. However, we believe it warrants exploration because it allows us to work at a granular, atomic level of representation rather than relying on unstructured textual representations.

While further research is essential to refine and optimize this approach, it holds the potential to provide more nuanced and structured insights, offering a promising direction for future research on the topic.

VI. Results

A. Quantitative Evaluation

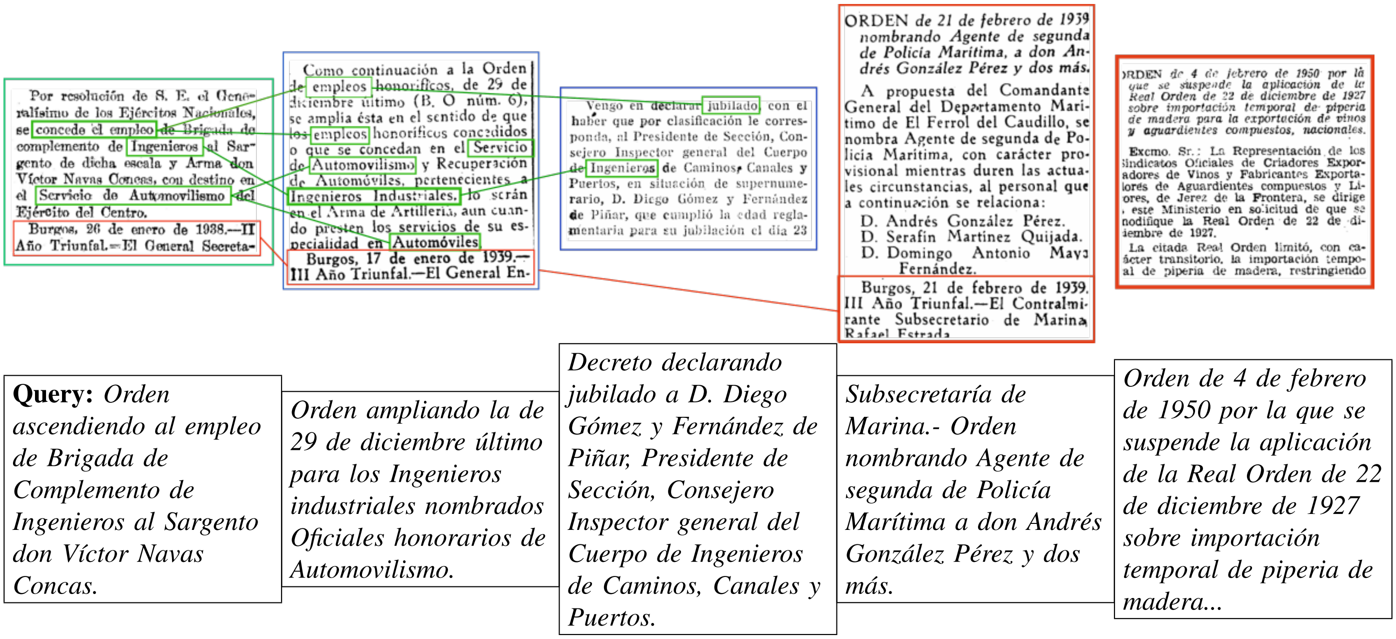
The experiments presented in Table I highlight several key insights.

Firstly, the choice between the textual or visual encoder significantly depends on data quality. It is evident that the use of triplet loss is not the optimal strategy, even though it can yield competitive results in terms of topology intersection over union when a sufficient quantity of data is available, regardless of its quality, performance is far from the optimum obtained through the experiments.

While approaches like SimCLR show promise with increasing data quantity, they fall short of outperforming the effectiveness of Circle Loss. It's essential to note that the benefits of Circle Loss are most pronounced when topic pseudolabels are incorporated. Further research should investigate different combinations of loss functions to gain deeper insights into their behaviors. However, at least in the case of Circle Loss, the inclusion of topic pseudolabels has proven crucial for achieving success in retrieval and rank correlation metrics.

When employing Spearman Rank Correlation as a loss function, it becomes apparent that retrieval is somewhat neglected, as there is no inherent constraint to align the embedding spaces effectively. Although rank correlation exhibits competitiveness with other approaches, it is recommended to use it in conjunction with additional constraints for optimal results.

Future research avenues should explore the maximization of metrics by experimenting with comprehensive loss combinations to enhance ranking distillation. There is evidence that such distillation can benefit the retrieval task, suggesting that numerous



Por resolución de S. E. el Generalísimo de los Ejércitos Nacionales, se concede el empleo de Brigada de complemento de Ingenieros al Sargento de dicha escala y Arma don Víctor Navas Concas, con destino en el Servicio de Automovilismo del Ejército del Centro.

Burgos, 26 de enero de 1938.—II Año Triunfal.—El General Secreta-

Como continuación a la Orden de empleos honoríficos, de 29 de diciembre último (B. O núm. 6), se amplia ésta en el sentido de que los empleos honoríficos concedidos o que se concedan en el Servicio de Automovilismo y Recuperación de Automóviles, pertenecientes a Ingenieros Industriales, lo serán en el Arma de Artillería, aun cuando presten los servicios de su especialidad en Automóviles.

Burgos, 17 de enero de 1939.—III Año Triunfal.—El General En-

Vengo en declarar jubilado, con el haber que por clasificación le corresponda, al Presidente de Sección, Consejero Inspector general del Cuerpo de Ingenieros de Caminos, Canales y Puertos, en situación de supernumerario, D. Diego Gómez y Fernández de Piñar, que cumplió la edad reglamentaria para su jubilación el día 23

ORDEN *de 21 de febrero de 1939 nombrando Agente de segunda de Policía Marítima, a don Andrés González Pérez y dos más.*

A propuesta del Comandante General del Departamento Marítimo de El Ferrol del Caudillo, se nombra Agente de segunda de Policía Marítima, con carácter provisional mientras duren las actuales circunstancias, al personal que a continuación se relaciona:

D. Andrés González Pérez.
D. Serafín Martínez Quijada.
D. Domingo Antonio Mayo Fernández.

Burgos, 21 de febrero de 1939. III Año Triunfal.—El Contralmirante Subsecretario de Marina Rafael Estrada

ORDEN *de 4 de febrero de 1950 por la que se suspende la aplicación de la Real Orden de 22 de diciembre de 1927 sobre importación temporal de pipería de madera para la exportación de vinos y aguardientes compuestos, nacionales.*

Excmo. Sr.: La Representación de los Sindicatos Oficiales de Criadores Exportadores de Vinos y Fabricantes Exportadores de Aguardientes compuestos y Licores, de Jerez de la Frontera, se dirige a este Ministerio en solicitud de que se modifique la Real Orden de 22 de diciembre de 1927.

La citada Real Orden limitó, con carácter transitorio, la importación temporal de pipería de madera, restringiendo

Query: *Orden ascendiendo al empleo de Brigada de Complemento de Ingenieros al Sargento don Víctor Navas Concas.*

Orden ampliando la de 29 de diciembre último para los Ingenieros industriales nombrados Oficiales honorarios de Automovilismo.

Decreto declarando jubilado a D. Diego Gómez y Fernández de Piñar, Presidente de Sección, Consejero Inspector general del Cuerpo de Ingenieros de Caminos, Canales y Puertos.

Subsecretaría de Marina.- Orden nombrando Agente de segunda de Policía Marítima a don Andrés González Pérez y dos más.

Orden de 4 de febrero de 1950 por la que se suspende la aplicación de la Real Orden de 22 de diciembre de 1927 sobre importación temporal de pipería de madera...

Fig. 6: Success case of the retrieval model. We observe the correct document placed at the top (left), two documents (blue) with some relevant semantic features (green), and some signs of bias (red). The last two retrieved elements in the 5-nearest neighbors don't share any apparent semantic meaningfulness with respect to the query.

combinations hold the potential for improving overall performance.

B. Qualitative Analysis

From a qualitative perspective, our focus lies on assessing both retrieval performance and result explainability. In this section, we delve into cases of both failure and success with our model, aiming to gain insights into its behavior.

First, in Figure 6, we observe a document that has been accurately ranked as the top result by our 1-Nearest Neighbor approach. What's even more interesting is the similarity between the second and third-ranked documents, as they both contain significant content related to engineering job status. The second-ranked document even touches upon engineering in the context of the driving sector. All documents present words related to the military context. We will later show strong evidence that this is intended by the model and demonstrates a perfectly fine behavior to expect from such a solution.

However, there are also indications of potential bias in our results. The second and fourth-ranked documents exhibit a shared historical fragment at the bottom, reminiscent of the early years of the military uprising in Spain (1936-1939). Although the first-ranked document, relevant to the query, pertains to the ”II Victorious Year,” and the others contain ”III Victorious Year,” we need to exercise caution when interpreting the model's behavior. There are still contradictory insights when claiming the amount of temporal bias in the model.

In Figure 7, we inspect the corresponding activations within the network for both the visual and text encoders. From a visual standpoint, we observe the tokenizer prioritizing certain words. While there is some activation around ”engineers,” it doesn't seem to strongly correlate with the words shared by the top-ranked documents. However, it's worth noting that the topic may be correlated with several words, and it's expected that LDA, represented as a probability density function, exhibits activations at various points.

On the textual side, understanding what transpires within the transformer is more complex. In the last layer, we observe that the retrieval token isn't a composition of easily interpretable tokens; it receives high activations from seemingly arbitrary tokens. These tokens, however, were previously composed in the first layer, where maximum attention per head appears to be relevant.

The most intriguing insight from our activation maps is that the query-key pair ”Orden” and ”Concediendo” yields the highest

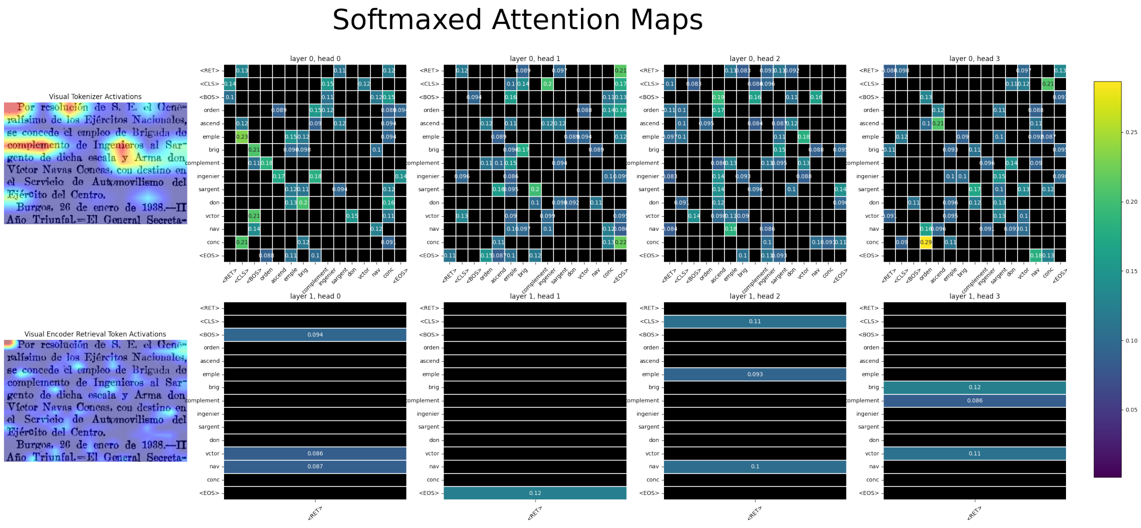


Fig. 7: Activation maps for document visual representation (left) in its tokenizer (top), last layer (bottom) and activations of the different attention heads for query tokens (right) in the first encoder layer (top) and last encoder layer (bottom) relative only to the <RET> retrieval token.

logit. This pair represents a common formulaic sentence structure used to indicate the granting of a position to a named entity in legal resolutions. While these words may not exist in the visual representation, they could be implicitly represented through the broader context of the document, which discusses jobs.

While relying solely on inspection of activation maps may present challenges in achieving explainability, we also explore the perturbations, or ”brain damage,” as introduced by Le Cun et al. [35], inflicted upon the network when masking and removing specific query tokens. This analysis yields critical insights.

Firstly, it validates the earlier hypothesis posited in Section IV, emphasizing the structural importance of queries. As illustrated in Figure 8, preserving some token pairs while masking others results in reasonable similarity scores. In stark contrast, removing these tokens leads to a significant drop in similarity. However, some exceptions are noteworthy. For instance, while eliminating all tokens except ”Orden” results in a state of uncertainty with a 0.5 similarity score, the token ”Empleo” (job) maintains a 64% similarity by itself. When combined with ”Brigada” (brigade), it rises to 70%, followed by ”ingeniero” (engineer) and ”Sargento” (sergeant) with 69% similarity. From this observation, we can confidently conclude that the presence of such tokens among the top-k retrieved documents in Figure 6 is likely not indicative of bias within the image encoder. Instead, it reflects a correct understanding of the topic’s connection to both work and war. Furthermore, it becomes apparent that the token ”engineering” is strongly correlated with military terminology, as illustrated in various examples throughout this document.

Additionally, this analysis highlights that certain tokens can introduce misleading information, while others cause similarity scores to slightly increase. A case in point is the named entity ”Don Victor Navas Concas.” This insight is particularly valuable, emphasizing that topics should not be overly biased towards named entities, as they often contribute minimally to the comprehension of the topic. Surprisingly, the tokens ”orden” and ”concediendo” tend to not be useful in encoding the query’s information, which notably conflicts with the insights enlighted by the analysis on attention maps in Figure 7.

We anticipate further analysis of activations in the subsequent sections to provide deeper insights, where we will address how the visual tokenizer isn’t using whole words and how it may be affected by the used receptive field or a reasonable behavior of the visual transformer. Additionally, other success cases may be found in Annex B.

In Figure 9, we list some of the documents that were placed at the bottom of the ranking to their respective queries. Note that while some of the errors can be attributed to segmentation and cropping issues, some failure cases cannot be attributed to natural noise present in the dataset.

Upon closer inspection of the same figure, we observe that not all query-document mismatches can be attributed to segmentation issues or data noise. Some of these mismatches, albeit intriguing, are not rooted in flaws within the data.

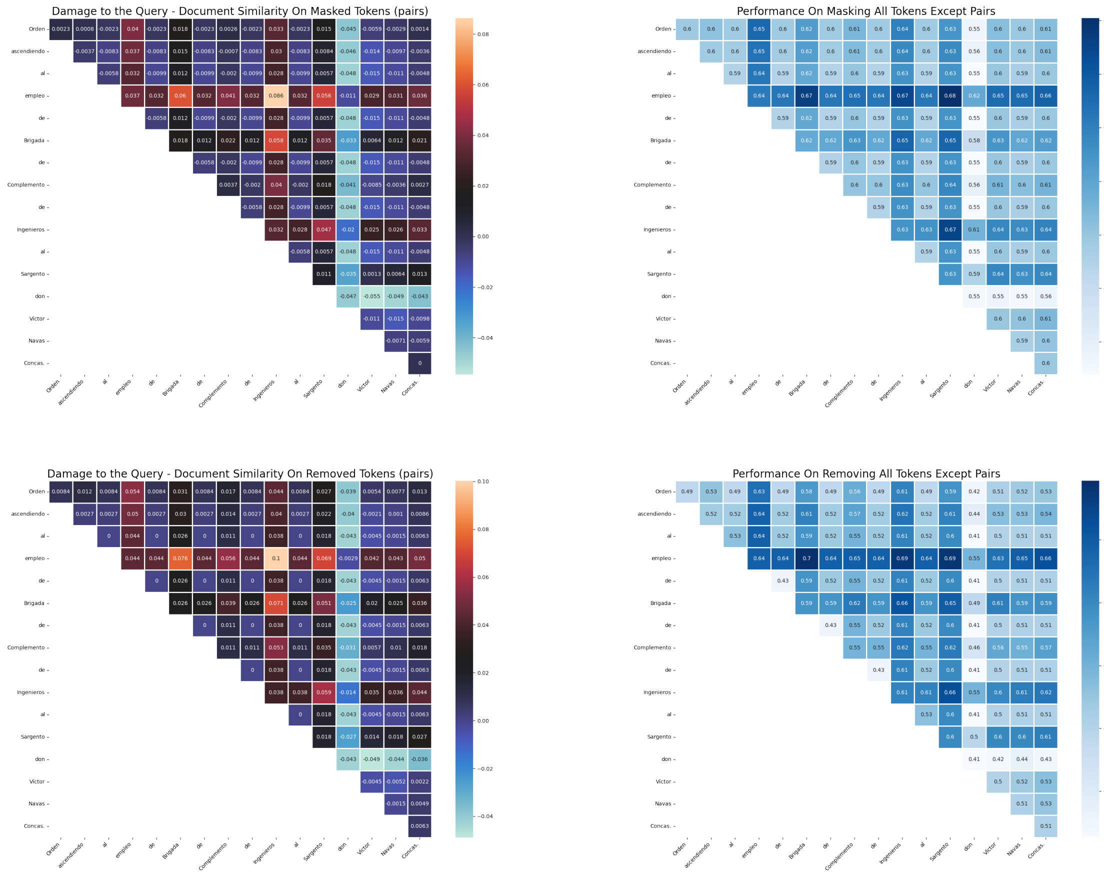
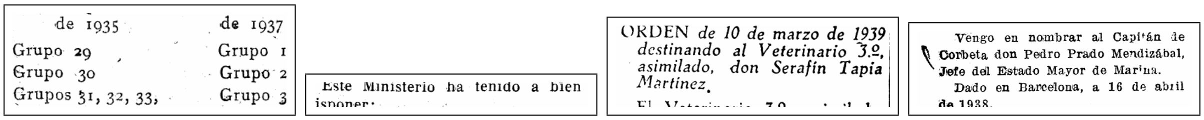


Fig. 8: Drop of query-document similarity (left) when masking pairs of tokens (top), similarity value (top) when masking (top) or removing (bottom) all tokens except pairs. Diagonal expresses only one token is being masked or presserved at a time.

This discovery is further elucidated in Figure 10, which provides a clear illustration of such cases. In Figure 10, we encounter a scenario where two documents while sharing a common theme of ”engineers”, are fundamentally unrelated in their actual content. This divergence in meaning, despite the surface-level connection, underscores the need for a nuanced understanding of the semantic relationships between queries and documents. There’s little evidence that the text encoder is working at ”word level” rather than at ”topic level”; which is also aligned with what was shown in Figure 6.

VII. Hypothesis Discussion

In the course of our work, we introduced a series of hypotheses that serve as fundamental pillars in our research. In this section, we embark on an in-depth discussion of each hypothesis, critically assessing whether they have been met and exploring their implications within the context of our study.



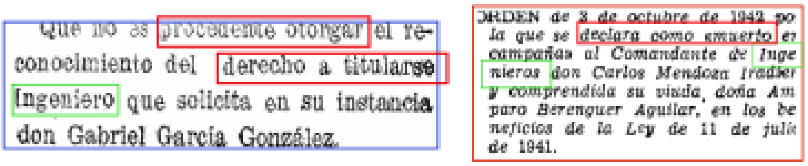
de 1935 de 1937
Grupo 29 Grupo 1
Grupo 30 Grupo 2
Grupos 31, 32, 33, Grupo 3

Este Ministerio ha tenido a bien
isponer:

ORDEN de 10 de marzo de 1939
destinando al Veterinario 3.º,
asimilado, don Serafín Tapia
Martínez.

Vengo en nombrar al Capitán de
Corbeta don Pedro Prado Mendizábal,
Jefe del Estado Mayor de Marina.
Dado en Barcelona, a 16 de abril
de 1938.

Fig. 9: Examples of documents that were not correctly matched with their corresponding queries. The documents vary in quality, ranging from severely missegmented (left) to gradually improving in quality, ultimately resulting in well-structured crops (right) containing all the essential information required to identify the content of the legal resolution.



Que no es procedente otorgar el reconocimiento del derecho a titularse Ingeniero que solicita en su instancia don Gabriel García González.

ORDEN de 3 de octubre de 1942 po la que se declara como «muerto» en campañas al Comandante de Ingenieros don Carlos Mendoza Iradier y comprendida su viuda, doña Amparo Berenguer Aguilar, en los beneficios de la Ley de 11 de julio de 1941.

Query: *Orden denegando el reconocimiento del derecho a titularse Ingeniero, que solicitaba don Gabriel García González.*

Retrieved: *Orden por la que se declara como "muerto en campaña" al Comandante de Ingenieros don Carlos Mendoza Iradier, y comprendida su viuda, doña Amparo Berenguer Aguilar, en los beneficios de la Ley de 11 de julio de 1941.*

Fig. 10: Example of missclassified sample with respect to the query (right). We observe some similarities with the correct document (blue) and the top ranked document (red), but the query talks about some engineer not being able to graduate, the retrieved document talks about the death of a lead eningeer in the army.

A. Visual Cues Should Narrow the Topic Space

Firstly, it's mentioned during Section I that visual cues should help to narrow the posterior topic space; meaning that given the set of all the possible topics a document may belong to, it's heavily narrowed given certain visual cues. This should be the case for some notions of time where heavily damaged documents are related to the most ancient topics in the dataset, whereas some printed documents with modern layouts should be more frequently related to modern topics. However in Figure 11 we observe some bias towards features being related to temporal insights.

The debate regarding whether the model's performance arises from its ability to identify crucial tokens for mapping related topics, as highlighted by Figures 8 and 12, or if this behavior is a consequence of temporal insights present in the documentation, remains an open discussion. However, there exists ample empirical evidence to confidently assert that, to a certain extent, the model is implicitly acquiring the skill of identifying relevant tokens within each query to assess their correspondences in the visual domain. In other words, there is strong evidence supporting the notion that the model is learning how to skim through the documentation to facilitate an understanding of the underlying topics.

Conversely, the activation maps depicted in Figure 7 do not demonstrate a discernible preference for regions indicating the date of the document, despite the potential presence of pertinent information embedded within visual nuances, such as font style. In contrast, when we examine high activations, it becomes apparent that they are often associated with specific words. It is essential to acknowledge that this observation should not be considered a definitive assertion, but rather a tentative observation that suggests visual tokens tend to focus on portions of words rather than the entire word itself.

In some instances, we can discern how activations are more pronounced in segments of words, particularly those that correspond to tokenized representations following lemmatization. Additionally, we observe situations where groups of words, such as "engineering" and "armed conflicts," are aggregated within the visual tokenizer, even though they are ultimately separated in the final representation. This phenomenon could potentially explain the occurrence of the failure case illustrated in Figure 10. In this scenario, a common error arises where terms associated with engineering and military subjects exhibit a notably high degree of correlation. This behavior is further exemplified in the successful scenario depicted in Figure 8, where we demonstrated a significant correlation between engineering and military terminology within the generated topic space. This alignment with the coexistence of engineering and military terms in visual tokens supports the assertion that the model's capability for topic identification relies on keyword recognition and, therefore, it should not be considered an error but a byproduct of the correct optimization of keywords that yielded to situations where words weren't actually correlated.

It is crucial to reiterate that these tendencies are not observable within the textual encoder, underscoring the distinctive behavior of visual tokens in contrast to their textual counterparts.

There is currently insufficient conclusive evidence to substantiate the assertion that visual cues have a constraining effect on the topic space. In order to establish this relationship through empirical experimentation rather than analytical exploration, it is imperative that we undertake the extraction of a defined set of visual and textual tokens, as illustrated in Figure 1. Following the extraction of these tokens, we anticipate observing a situation where the probability of a topic t appearing, denoted as $P(t)$, is influenced significantly by a subset of visual tokens $R \in V$, resulting in a pronounced narrowing of the topic space, i.e., $P(t|r \in R) >> P(t)$. Conversely, certain visual tokens may serve as distractors, showing little to no correlation with the topic space, as expressed by $P(t|r \in V \setminus R) \approx P(t)$.

However, it is imperative to emphasize that such a scenario and dataset are not currently available. Hence, exercising prudence and caution is necessary before making any claims regarding the existence of such behavior within the scope of the current research. Therefore, this hypothesis remains inconclusive as we cannot provide such evidence.

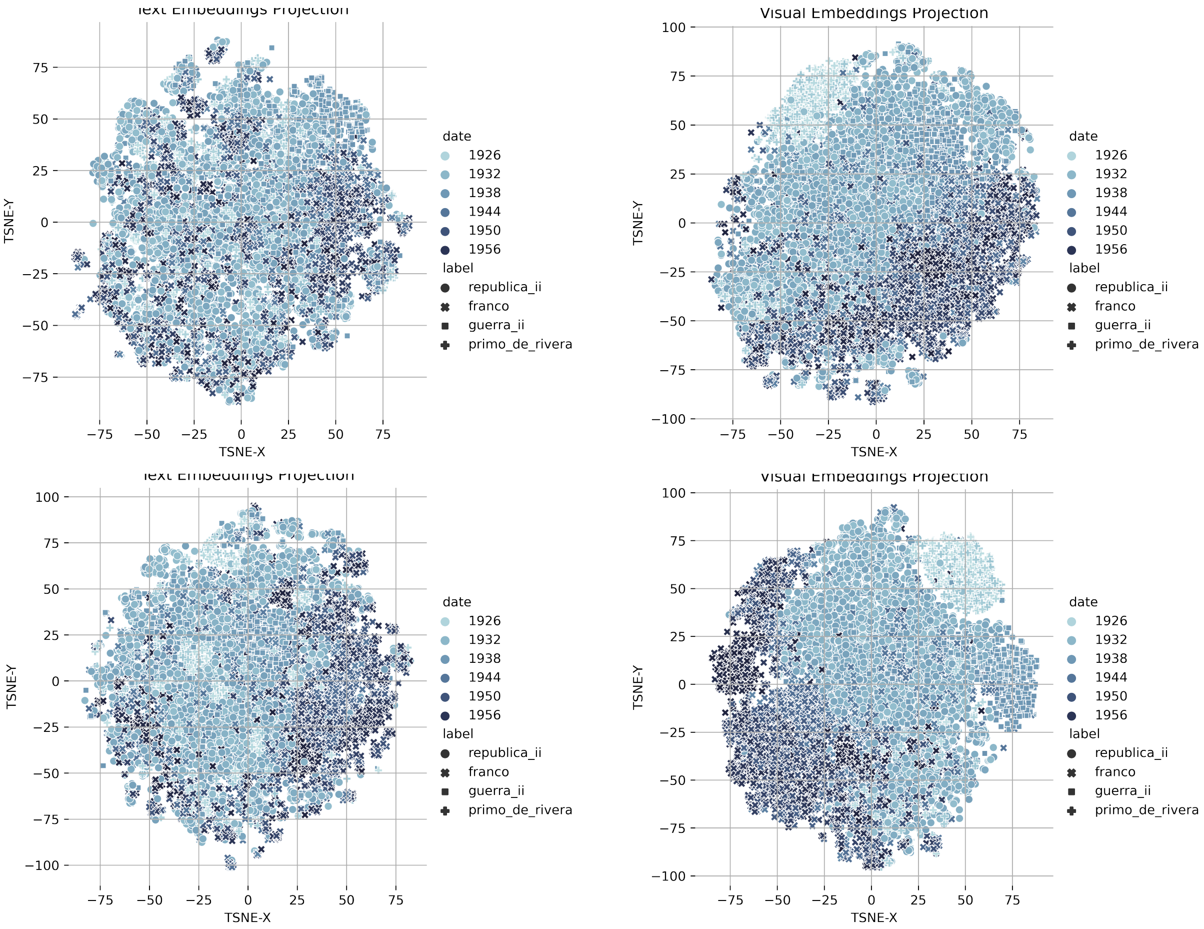


Fig. 11: Text encoder projection (left) and visual encoder projection (right) using T-SNE for the model optimized using $T(q)$ on the text encoder (top) and without it (bottom). Notably, the visual encoder exhibits a distinct bias towards the historical context of the documents, a nuance that is scarcely captured by the textual features. This observation aligns with the notion that visual features may serve as a valuable means to narrow down the potential range of topics, a dimension that the text encoder tends to remain agnostic about. When not using the topic model as pseudolabels, the bias is more clearly shown in the text encoder.

B. A Few Words are Worth A Thousand Pixels

Another recurrent hypothesis addressed through this work is that it's worth noting that topics are often decided by a relatively small set of tokens rather than the whole size of the corpus. It should be learned through the visual encoder to find such tokens in order to perform an efficient semantic retrieval. As it's been previously stated, it's noticeable that the visual tokenizer, sometimes, avoids segmenting the whole word for the sake of extracting only a part. The cause is trivial: images are tokenized in 16x16 patches. Even though the cause is clear, the reason may be interesting.

The attention mechanism's behavior of occasionally combining two tokens when a word is split in half lacks a clear rationale. However, it is worth noting that in certain instances, the attention mechanism seems to focus primarily on the root of the word. This behavior provides compelling evidence supporting the notion that deriving topic information directly from the visual domain, rather than relying on OCR extraction, yields a more efficient representation.

In simpler terms, akin to the way words are tokenized through lemmatization, the visual tokenizer appears to learn how to

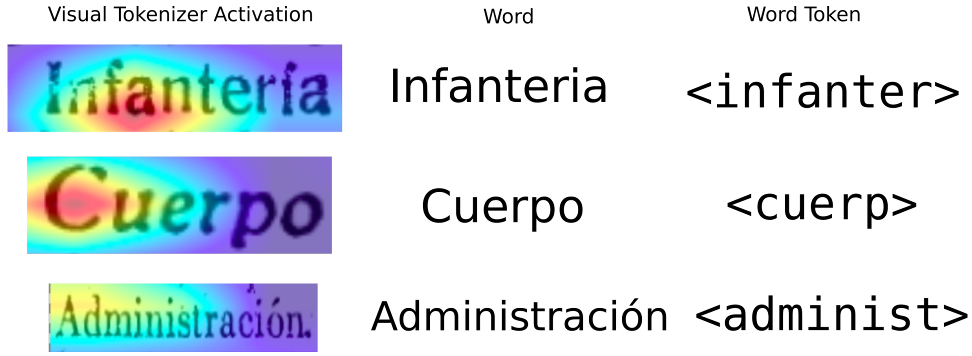


Fig. 12: Examples of visual words (left) where the visual tokenizer prioritizes only a patch of the image rather than combining several patches. The exhibited behavior shows some similarity to how we lemmatize words (right) during tokenization in order to create robust representations.

utilize only the essential components of the image Figure 12. This characteristic holds significant promise from the perspective of historical document analysis, as it introduces robustness when dealing with unreadable or deteriorated portions of the image. Further research in this direction could prove valuable, especially for highly damaged documents, by emphasizing the learning of efficient visual word representations.

In conclusion, our findings strongly support the hypothesis that deriving the topic model directly from visual features results in a more efficient representation capable of effective operation without the need for complete words. This approach enables the representation to focus on specific word components, offering great promise for applications in historical document analysis and presenting a potential solution for addressing challenges associated with image deterioration and incomplete textual information.

This conclusion is reinforced by the observations in Figure 8, where tokens that consistently appeared in the retrieved documents in Figure 6 were influenced by the network's understanding of the importance of these tokens in the query, rather than being indicative of a visual bias towards specific elements. This highlights the accurate behavior of the intrinsic topic spotting facilitated by the proposed solution.

C. Optimizing Among the Topic Space Yields Greater Performance

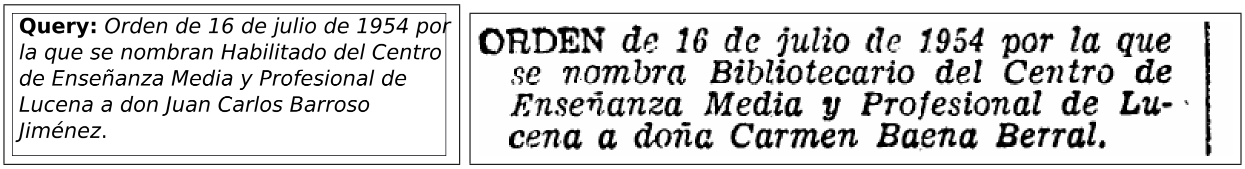
This research places a significant emphasis on investigating the potential enhancement of document retrieval through the utilization of ranking distillation with topic models as pseudolabels. The evidence presented in Table I supports this hypothesis. It not only underscores the superiority of the rank distillation approach compared to other methods but also highlights how it can transform Circle Loss into a viable solution when ranking distillation is employed.

While further experiments could be conducted to maximize these metrics and build robustness for real-world scenarios, it's worth noting that, within the scope of the conducted experiments, none of the alternative approaches demonstrated the same level of performance as the proposed optimization scheme. To establish more robust evidence, future research should aim to facilitate fair comparisons, mitigating potential cherry-picking issues. As mentioned in Appendix B, a substantial number of experiments were conducted to explore various approaches. While none of these approaches contributed significantly to performance when ranking distillation was not utilized (as evidenced in Table I), a more equitable comparison framework needs to be established to provide stronger evidence.

D. Structured Encoding Capture Semantic Nuances

Lastly, the process of encoding data in a structured manner using a graph encoder remains at a preliminary stage, as elaborated further details in Annex B, this preliminary work has yielded seemingly unintelligible results, gibberish, stemming from two primary challenges that warrant future attention.

First, the pursuit of an end-to-end perspective necessitates a high-capacity architecture. However, such models demand an extensive corpus of annotated data. Whether we employ the entire dataset comprising 1.5 million documents or enhance the architectural representation, mitigating overfitting remains a critical task. Research endeavors should focus on the extraction



Query: *Orden de 16 de julio de 1954 por la que se nombran Habilitado del Centro de Enseñanza Media y Profesional de Lucena a don Juan Carlos Barroso Jiménez.*

ORDEN de 16 de julio de 1954 por la que se nombra Bibliotecario del Centro de Enseñanza Media y Profesional de Lucena a doña Carmen Baena Berral.

Fig. 13: Query (top) and document (bottom) on which we observe some important nodes (Juan Carlos Barroso) missing in the fragment of the document.

of meaningful nodes within documents, as previously discussed, guided by a graph-based framework. The establishment of relationships among nodes should be a subsequent step, emphasizing link prediction. This approach offers the potential for addressing the pronounced overfitting, either through model simplification or scaling with more extensive data.

On the other hand, we must address the quality of the underlying data. As illustrated in Figure 13, the visual representation of the graph conspicuously lacks crucial nodes. From a mode-seeking perspective, this deficiency is untenable, particularly as certain elements of the ground truth appear arbitrarily erroneous. Consequently, two distinct strategies warrant consideration: either rectifying the ground truth by encompassing all regions pertinent to the query, rather than solely the top one, to encapsulate comprehensive information, or transitioning the model paradigm to a mode-covering perspective, possibly involving an adversarial setup. This shift in approach aims to generate viable graph representations that may not be exact solutions but are more practical and avoid arbitrary offsets on the loss function.

VIII. Conclusions

In this research, we draw attention to several significant conclusions and areas for future exploration. Firstly, in Section III, we introduced a novel dataset that tackles the challenge of content-based document retrieval from a visual perspective. Surprisingly, this aspect had been largely overlooked, likely due to the predominant focus on digitally native documents where content is primarily text-based but digitally native. In light of this, we formulated the problem as a re-identification task, employing classical retrieval evaluation techniques and an embedding alignment approach to compare the spatial relationships within the resulting topologies.

Additionally, we've provided parsed graphs for each query in the dataset, using a syntactic graph parser that is adaptable to multiple languages thanks to the underlying SpaCy [31] framework. This capability distinguishes our solution from existing repositories, as no other resource currently offers such versatility.

Moving forward, there are several directions for future work. Firstly, we intend to complete the dataset by incorporating the entire available corpus of documents, totaling 1.5 million entries. We also plan to enhance the segmentation algorithm to ensure more robust training and evaluation, addressing the issues identified in Figure 9. Furthermore, we aim to introduce multilingual sources into the dataset to increase its diversity.

A crucial aspect we intend to explore is the inclusion of data sources with a richer visual component, including figures and drawings, as exemplified in Figure 1. This expansion will enable us to further investigate the interplay between visual content and document retrieval, opening up exciting avenues for research and development.

In Section V, we suggest employing topic models as a form of supervision for document re-identification. Subsequently, in Section VI, we presented some evidence demonstrating that utilizing topic models as pseudo-labels for the retrieval task may enhance retrieval metrics. However, it is noteworthy that there is no clear distinction regarding the model's ability to establish a topology that shares local features in both modalities, as measured by the proposed IoU_{rank} metric. In this matter, improving the representation of the generated embedding should be considered a necessary work for expanding the literature in the historical document understanding field.

In the same section, we presented compelling evidence demonstrating that vision encoders can effectively enhance the utilization of visual tokens. This development holds promise for gaining a deeper understanding of historical documents, where words may be partially damaged, occluded, or unreadable, making traditional OCR solutions unfeasible.

On the contrary, the hypothesis positing that visual cues should definitively aid in distinguishing certain topics remains inconclusive. Nevertheless, our results do suggest that the utilization of pseudolabels has a mitigating effect on the temporal bias present in the acquired representation. As previously discussed, the inclusion of data that allows for a comprehensive evaluation of this hypothesis will be a crucial aspect of our future research endeavors. Some promising insights, such as those revealed in Figure 8, indicate that the model may operate with some independence from the temporal bias. It carefully analyzes the queries, identifying crucial tokens to construct its internal topic representation. To recap, we contend that the model can proficiently engage in topic understanding of documents due to its capability to detect important keywords. While the presence of bias cannot be entirely dismissed, there exists empirical evidence to support the assertion that the achieved solution excels in topic identification through implicit learning of word recognition.

Finally, harnessing the potential of graph representations for historical documents holds promise as a direction that could help alleviate some of the challenges encountered in our approach. This includes the explicit identification of relevant nodes for topic extraction, rather than relying on an inconclusive selection of regions.

In summary, this work has addressed various aspects of information extraction from historical document collections, offering a novel and unexplored perspective on a classic task, information retrieval. We have provided both evaluation metrics and data to support our findings. While we consider some of the evidence strong and conclusive, we acknowledge that further research is necessary to fully resolve other critical aspects of document understanding.

Acknowledgment

This work has been partially supported by the Spanish project PID2021-126808OB-I00, and the Culture Department of the Generalitat de Catalunya, and the CERCA Program / Generalitat de Catalunya.

We would like to extend our gratitude to Professors O. Ramos and J. Lladós for their supervision, support and patience throughout this endeavor.

We also extend our thanks to the Document Analysis Group, Computer Vision Center, and the Procrastination Frog.

References

[1] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research 3 (2003) 993-1022*, 2003.

[3] N. Reimers and I. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks*, 2019. arXiv: 1908.10084 [cs.CL].

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019. arXiv: 1810.04805 [cs.CL].

[5] H. Li, P. Wang, and C. Shen, *Towards end-to-end text spotting with convolutional recurrent neural networks*, 2017.

[6] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 67–83.

[7] M. Liao, G. Pang, J. Huang, T. Hassner, and X. Bai, "Mask textspotter v3: Segmentation proposal network for robust scene text spotting," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, Springer, 2020, pp. 706–722.

[8] A. Singh, G. Pang, M. Toh, J. Huang, W. Galuba, and T. Hassner, "Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8802–8812.

[9] V. Christlein, A. Nicolaou, M. Seuret, D. Stutzmann, and A. Maier, *Icdar 2019 competition on image retrieval for historical handwritten documents*, IEEE, 2019.

[10] A. Gordo, F. Perronnin, and E. Valveny, "Large-scale document image retrieval and classification with runlength histograms and binary embeddings," *Pattern Recognition*, vol. 46, no. 7, pp. 1898–1905, 2013.

[11] D. Picard, *Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision*, 2023. arXiv: 2109.08203 [cs.CV].

[12] G. Jaume, H. K. Ekenel, and J.-P. Thiran, "Funsd: A dataset for form understanding in noisy scanned documents," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, IEEE, vol. 2, 2019, pp. 1–6.

[13] R. Tito, D. Karatzas, and E. Valveny, “Document collection visual question answering,” in *Document Analysis and Recognition – ICDAR 2021*, Springer International Publishing, 2021, pp. 778–792. DOI: 10.1007/978-3-030-86331-9_50.

[14] L. Gomez, Y. Patel, M. Rusinol, D. Karatzas, and C. Jawahar, “Self-supervised learning of visual features through embedding images into text topic spaces,” in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2017, pp. 4230–4239.

[15] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, IEEE, vol. 2, 2006, pp. 2169–2178.

[16] A. Radford, J. W. Kim, C. Hallacy, *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[17] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, “A novel word spotting method based on recurrent neural networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 2, pp. 211–224, 2011.

[18] S. Antol, A. Agrawal, J. Lu, *et al.*, “Vqa: Visual question answering,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[19] A. F. Biten, R. Litman, Y. Xie, S. Appalaraju, and R. Manmatha, “Latr: Layout-aware transformer for scene-text vqa,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 548–16 558.

[20] M. Mathew, D. Karatzas, and C. Jawahar, “Docvqa: A dataset for vqa on document images,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2200–2209.

[21] R. Tito, D. Karatzas, and E. Valveny, “Hierarchical multimodal transformers for multipage docvqa,” *Available at SSRN 4466918*,

[22] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, “Learning local feature descriptors with triplets and shallow convolutional neural networks.,” in *Bmvc*, vol. 1, 2016, p. 3.

[23] P. Krishnan and C. Jawahar, “Hwnet v2: An efficient word image representation for handwritten documents,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 22, no. 4, pp. 387–405, 2019.

[24] T. Nguyen, M. Rosenberg, X. Song, *et al.*, “Ms marco: A human-generated machine reading comprehension dataset,” 2016.

[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[26] A. Antonacopoulos, D. Bridson, C. Papadopoulos, and S. Pletschacher, “A realistic dataset for performance evaluation of document layout analysis,” in *2009 10th International Conference on Document Analysis and Recognition*, IEEE, 2009, pp. 296–300.

[27] F. Wang and H. Liu, “Understanding the behaviour of contrastive loss,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2495–2504.

[28] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[29] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[30] M. Engilberge, L. Chevallier, P. Pérez, and M. Cord, “Sodeep: A sorting deep net to learn ranking loss surrogates,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 792–10 801.

[31] M. Honnibal and I. Montani, “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing,” To appear, 2017.

[32] K. Musgrave, S. Belongie, and S.-N. Lim, “A metric learning reality check,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, Springer, 2020, pp. 681–699.

[33] Y. Sun, C. Cheng, Y. Zhang, *et al.*, “Circle loss: A unified perspective of pair similarity optimization,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6398–6407.

[34] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[35] Y. LeCun, J. Denker, and S. Solla, “Optimal brain damage,” *Advances in neural information processing systems*, vol. 2, 1989.

[36] S. Schweter and A. Akbik, *Flert: Document-level features for named entity recognition*, 2020. arXiv: 2011.06993 [cs.CL].

[37] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: 1412.6980 [cs.LG].

Appendix

In this concluding appendix, we endeavor to distill the technical intricacies discussed throughout the document into a coherent and accessible summary. Our objective is to provide a formal yet comprehensible guide that promotes the reproducibility of our experiments while further expanding the qualitative standpoint of the behavior of our solution.

A. Dataset Construction

One crucial decision in our problem statement is the determination of a valid matching point. As detailed in Section III of the problem statement, the required Sentence-BERT similarity threshold for including a sample in the dataset is considered as a hyperparameter in our training regime, as discussed in Annex B. However, for the test set, we cannot allow this flexibility, as it could lead to misleading and overfitted results. Therefore, we rely on a minimum Sentence-BERT similarity threshold based on empirical observations of the distribution, as shown in Figure 14. It's worth noting that a 40% similarity threshold should already filter out all undesirable examples. However, to ensure the exclusion of all misleading documents, we have imposed a minimum similarity threshold of 50%.

Additionally, another challenge arises from the segmentation model, which sometimes crops small patches that yield no meaningful information. Unfortunately, these patches may exhibit high similarity scores due to the presence of important named entities, biasing Sentence-BERT towards maximizing the likelihood of such regions. To address this issue, we have imposed a minimum size requirement of 125x125 for both the test and training datasets. As observed in the histogram Figure 14, there's a low variance in the height of the crops, which is an indicator of consistency within the dataset. This threshold affects entirely to those with small widths.

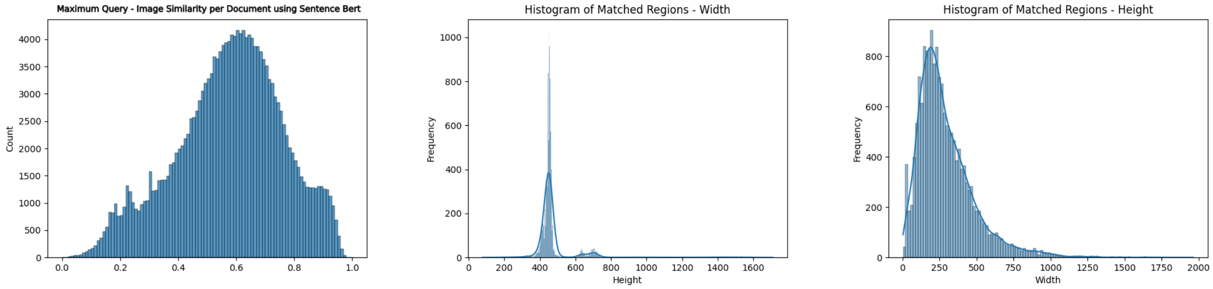


Fig. 14: Sentence-Bert cosine similarity between the query and the top-1 segmented region in the document (left). Histogram for crops height (center) and width (right). There's a noticeable low variance on height, indicating consistency on the data.

In order to extract the named entities from both queries and OCR transcriptions, `flair/ner-spanish-large` [36] has been used, which is available in Huggingface. Once the named entities are extracted, an inexact hungarian algorithm is performed relating each of the named entities from OCR to Query by minimum edit distance as is described in Algorithm 1.

Algorithm 1: Hungarian Algorithm for Sentence Alignment

```
Input  : Two sentences, $s1$ and $s2$, and a word distance function, $word\_distance$
Output: Total minimum edit distance between the aligned words
$s1\_split \leftarrow$ Split sentence $s1$ into words;
$s2\_split \leftarrow$ Split sentence $s2$ into words;
$weights\_matrix \leftarrow$ Initialize a matrix of zeros with dimensions $(\text{len}(s1\_split), \text{len}(s2\_split))$;
for $n$ in range(0, len($s1\_split$)) do
    for $m$ in range(0, len($s2\_split$)) do
        $weights\_matrix[n, m] \leftarrow$ edit_distance($s1\_split[n]$, $s2\_split[m]$);
($row\_ind, col\_ind$) $\leftarrow$ linear_sum_assignment($weights\_matrix$);
return $\sum_{i=0}^{len(row\_ind)} weights\_matrix[row\_ind[i], col\_ind[i]]$
```

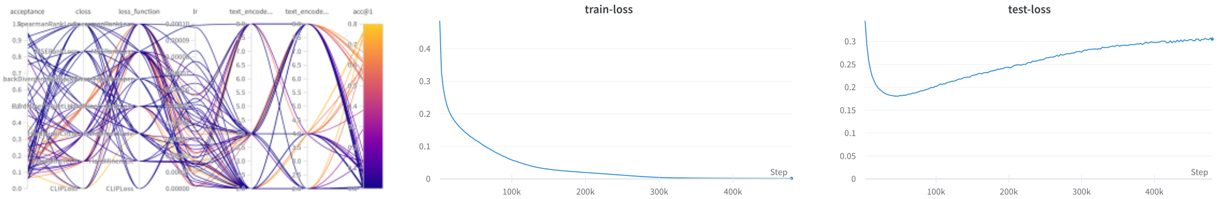


Fig. 15: Flow chart for the experiments ran on W&B sweeps (left) of the unstructured approach. And loss function for train (center) and test (left) for the structured graph approach, showing significant overfitting.

B. Training Regime

The training regime for the various experiments detailed in Table I is as follows: a visual transformer, ViT-B/16, is utilized in conjunction with OpenAI's CLIP implementation [16], which consists of 12 attention layers and 8 heads. Notably, the encoder architecture is maintained as a variable parameter, although our final results are presented in terms of a transformer encoder, implemented in vanilla PyTorch, comprising 2 heads and 4 layers, which has proven to deliver satisfactory performance.

Both encoder models undergo training for 50 epochs, with the number of training samples determined by the acceptance rate, either 17% or 40%. Training is conducted concurrently using the Adam optimizer [37] in PyTorch, with an initial learning rate set at 1×10^{-5}. This learning rate is scheduled to decrease by a factor of 0.1 on a plateau, which occurs when the mean Average Precision (mAP) on the training set no longer increases by at least 0.001 for 10 consecutive epochs. All of the batches are hard mined with the exception of the ones operating on SpearmanRankLoss1 since it doesn't measure negatives for samples.

For testing purposes, a random seed is applied to select a test dataset, ensuring an unbiased and non-sampling method approach, drawn from a 10% subset of the total document collection. Within this subset, 11,000 documents exceed the threshold of a 50% minimum Sentence-BERT similarity score. Inputs of the vision encoder (ROIs) are rescaled to 224x224 image size and normalized $N(0,1)$ with its total mean and variance batched among 128 images. No augmentation is applied in any case.

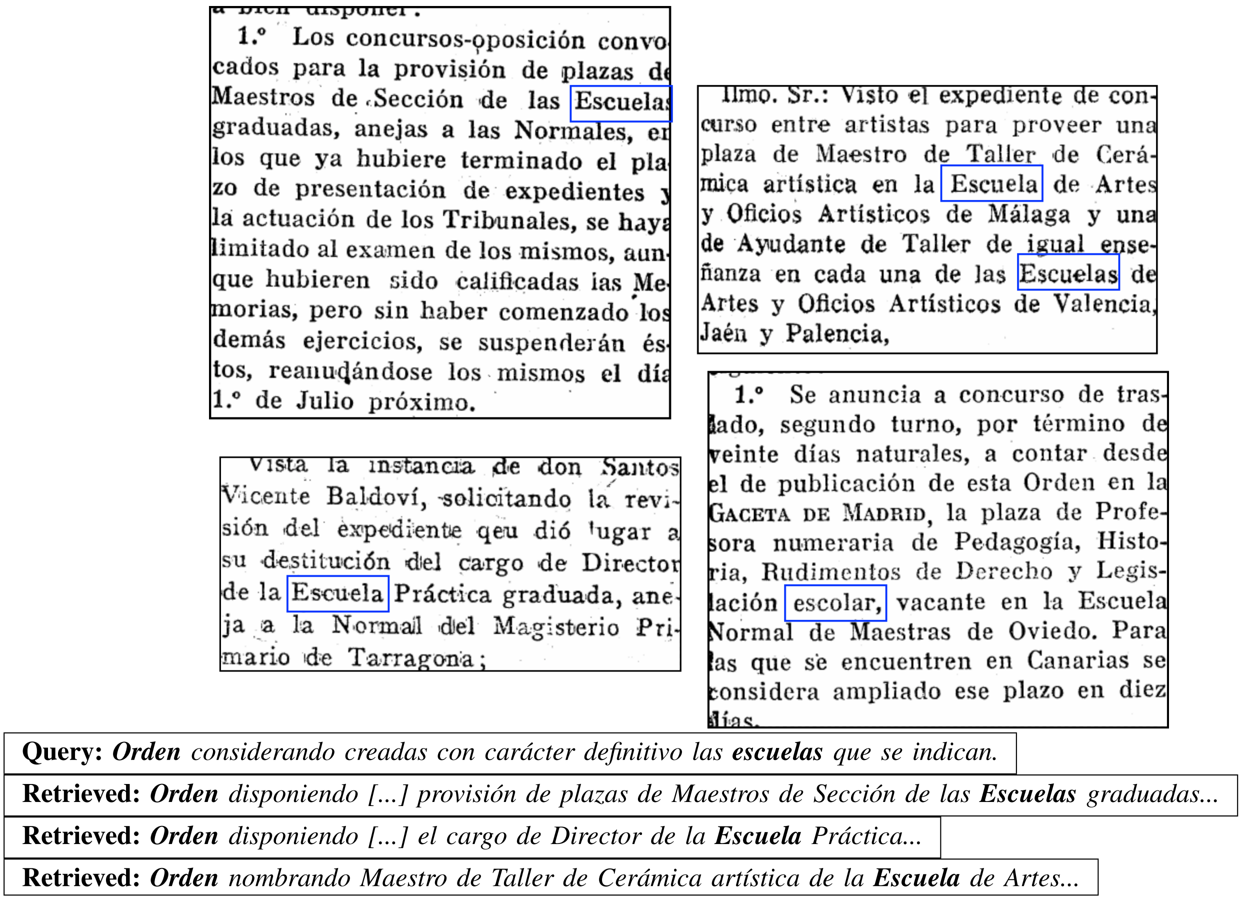
To ensure a fair and unbiased comparison, we leveraged W&B's Sweep implementation. This allowed us to simulate a ”competition” involving a total of 62 different agents and a cumulative computing time spanning 72 days, executed on 5 parallel Titan Xp GPUs. This extensive simulation was conducted prior to generating the results presented in Table I, with the aim of avoiding parameter cherry-picking that might favor our hypothesis. It's worth noting that conducting a full-fledged competition exceeds the scope of a master's thesis. Consequently, we performed the ablation study using globally optimized parameters. While additional research is essential to bolster the confidence of our conclusions, we can draw preliminary insights from this approach.

For the structured approach, an Encoder-Decoder architecture is used with the vision encoder previously introduced, but with GPT-2 as text decoder. The output of the decoder is tokenized with a special token beginning of edge, `<BOE>`, and end of edge, `<EOE>` . We also introduce special tokens for each of the edge types and beginning and end of sequence respectively `<BOS>` and `<EOS>`. Then, the graph is tokenized such that each node token is contained within `<BOE> [node] [class] <EDGE> [node] [class] <EOE>` . Note that the presence of the `[class]` token will indicate whether the token is a generic word, `<WORD>` , or a named entity: `<PER>`, `<ORG>`, `<LOC>` and `<MISC>`, being person, organisation, location and miscelanium respectively.

This pipeline optimizes cross entropy loss with Adam, an initial learning rate of $1E^{-5}$ with a 50 steps warmup and 8 512x512 images as batch. A cosine scheduler is used with a cycle of 10 epoches. As it's been previously stated in Section VI this resulted in a severe overfitting for the regions explained above.

C. Qualitative Examples

In this last section we present additional matterial in terms of qualitative evaluation to support the statements done in the document. We exemplify different cases and behaviors adopted by the architecture.



a bien disponer:

1.° Los concursos-oposición convocados para la provisión de plazas de Maestros de Sección de las Escuelas graduadas, anejas a las Normales, en los que ya hubiere terminado el plazo de presentación de expedientes y la actuación de los Tribunales, se haya limitado al examen de los mismos, aunque hubieren sido calificadas ias Memorias, pero sin haber comenzado los demás ejercicios, se suspenderán éstos, reanudándose los mismos el día 1.° de Julio próximo.

Vista la instancia de don Santos Vicente Baldoví, solicitando la revisión del expediente qeu dió lugar a su destitución del cargo de Director de la Escuela Práctica graduada, aneja a la Normal del Magisterio Primario de Tarragona;

Ilmo. Sr.: Visto el expediente de concurso entre artistas para proveer una plaza de Maestro de Taller de Cerámica artística en la Escuela de Artes y Oficios Artísticos de Málaga y una de Ayudante de Taller de igual enseñanza en cada una de las Escuelas de Artes y Oficios Artísticos de Valencia, Jaén y Palencia,

1.° Se anuncia a concurso de traslado, segundo turno, por término de veinte días naturales, a contar desde el de publicación de esta Orden en la GACETA DE MADRID, la plaza de Profesora numeraria de Pedagogía, Historia, Rudimentos de Derecho y Legislación escolar, vacante en la Escuela Normal de Maestras de Oviedo. Para las que se encuentren en Canarias se considera ampliado ese plazo en diez días.

Query: ***Orden*** *considerando creadas con carácter definitivo las* ***escuelas*** *que se indican.*

Retrieved: ***Orden*** *disponiendo [...] provisión de plazas de Maestros de Sección de las* ***Escuelas*** *graduadas...*

Retrieved: ***Orden*** *disponiendo [...] el cargo de Director de la* ***Escuela*** *Práctica...*

Retrieved: ***Orden*** *nombrando Maestro de Taller de Cerámica artística de la* ***Escuela*** *de Artes...*

Fig. 16: Failure case with retrieved documents (top) to the given query and document's respective true queries (bottom). We observe highlighted similarities ("Escuela", school, appearing several times) among the retrieved content.

1) Errors due to brief queries: Some of the queries are expremely brief and uniformative, this is the case of Figure 16, where the correct document have been placed at the bottom part of the support set (position 130). Nevertheless, note that all of the returned documents contain relevant tokens with respect to the query; showing that the method is working as expected despite the odds. Interestingly, we observe that all of the retrieved documents are "Orders", note that none of the retrieved segments contain such information, which may be inducted from the context or other shortcuts.

The low amount of information provided by this query, gives us to opportunity to further analyze its inner behavior. IN 17, where the word 'school' tends to yield a significant contribution into the tokens that later on construct the retrieval vector. From the visual point of view, there is no clear interpretation as "schools" (escuelas) is significantly less activated in the tokenizer than other words. In the last layer, results are cryptic and inconclusive.

2) Errors due to resolution: Note that through the encoding of the documents, all images are resized to 224x224 for simplicity. This is a terribly bad practice as shown in 18, since unconsistency in the resolution of the crops may create different images where tokens lose structure and become impossible to understand by the visual encoder. In order to deploy a more consistent model, a version with flexible resolution should be trained.

3) Visual Tokenizer Behavior: Throughout the document, we conducted multiple assessments of the activations within the visual tokenizer. Our observations revealed instances where individual words contribute significantly to the overall encoding, as well as cases where groups of words collectively influence the representation, as exemplified in Figures 7 and 18.

Figure 18 also sheds light on the challenge of uninformative token activations, primarily stemming from issues related to aspect ratios and resolution in certain samples. These challenges suggest the need for improved representations in the input space.

In Figure 12, we discussed how, at times, it isn't necessary to consider an entire visual word to contribute to the model's

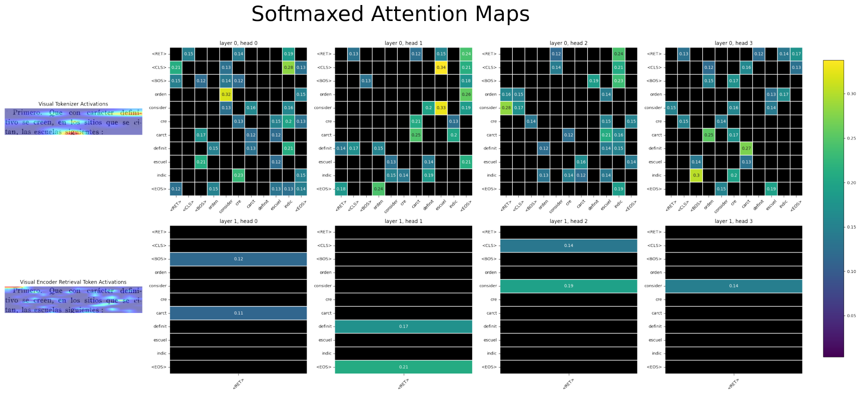
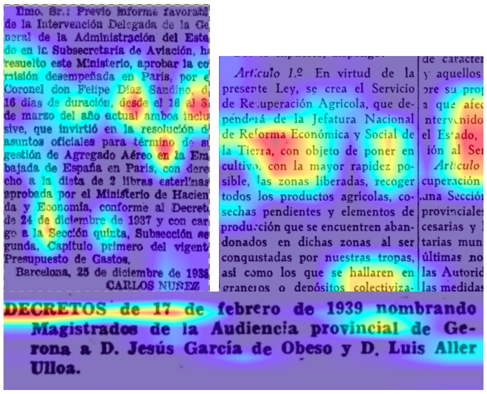


Fig. 17: Activation maps for the case of use where the brief query causes wrong matching of documents, yet good results in terms of usability.



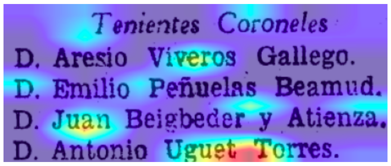
Ilmo. Sr.: Previo informe favorabl
de la Intervención Delegada de la Ge
neral de la Administración del Esta
do en la Subsecretaría de Aviación, h
resuelto este Ministerio, aprobar la co
misión desempeñada en Paris, por e
Coronel don Felipe Díaz Sandino, d
16 días de duración, desde el 16 al 3
de marzo del año actual ambos inclu
sive, que invirtió en la resolución d
asuntos oficiales para término de su
gestión de Agregado Aéreo en la Em
bajada de España en Paris, con dere
cho a la dieta de 2 libras esterlinas
aprobada por el Ministerio de Hacien
da y Economía, conforme al Decret
de 24 de diciembre de 1937 y con car
go a la Sección quinta, Subsección se
gunda, Capítulo primero del vigent
Presupuesto de Gastos.
Barcelona, 25 de diciembre de 1938
CARLOS NUÑEZ

Artículo 1.º En virtud de la presente Ley, se crea el Servicio de Recuperación Agrícola, que dependerá de la Jefatura Nacional de Reforma Económica y Social de la Tierra, con objeto de poner en cultivo, con la mayor rapidez posible, las zonas liberadas, recoger todos los productos agrícolas, cosechas pendientes y elementos de producción que se encuentren abandonados en dichas zonas al ser conquistadas por nuestras tropas, así como los que se hallaren en graneros o depósitos colectiviza-

de caracter
y aquellos
ore su pro
a que afec
ntervenido
el Estado,
ión al Se
Artículo
cuperación
una Secció
provinciale
cesarias y
tarias mun
últimas no
las Autorid
las medidas

DECRETOS de 17 de febrero de 1939 nombrando Magistrados de la Audiencia provincial de Gerona a D. Jesús García de Obeso y D. Luis Aller Ulloa.

Fig. 18: Different aspect ratios may heavily affect the receptive field of the vision tokenizer. Unifying all aspect rations with a padding token strategy should improve the explainability of the activation maps.



Tenientes Coroneles
D. Aresio Viveros Gallego.
D. Emilio Peñuelas Beamud.
D. Juan Beigbeder y Atienza.
D. Antonio Uguet Torres.

Fig. 19: We observe a heavy bias for the tokenizer towards activating on named entities in some cases.

performance effectively. Lastly, Figure 19 underscored the critical role that named entities play in the encoding of images.

In conclusion, achieving explainability in attention-based models is a non-trivial task. However, it's evident that many issues with tokenization arise when fragments exhibit problematic aspect ratios, leading to the grouping of regions rather than the precise attention to important tokens. The primary reason for the challenges seen in Figure 18, particularly in the upper documents, lies in the complexities of extrapolating the concept of a ”word.” This complexity is better handled in other examples, such as the lower document in Figure 18, Figure 19, and those discussed in Section VI.