DHQ: Digital Humanities Quarterly
2024
Volume 18 Number 3

Capturing Captions: Using AI to Identify and Analyse Image Captions in a Large Dataset of Historical Book Illustrations

Julia Thomas <ThomasJ1_at_cardiff_dot_ac_dot_uk>, School of English Communication and Philosophy, Cardiff University https://orcid.org/0000-0002-1995-5558
Irene Testini <TestiniI_at_cardiff_dot_ac_dot_uk>, Special Collections and Archives, Cardiff University https://orcid.org/0000-0003-4983-1844

Abstract

This article outlines how AI methods can be used to identify image captions in a large dataset of digitised historical book illustrations. This dataset includes over a million images from 68,000 books published between the eighteenth and early twentieth centuries, covering works of literature, history, geography, and philosophy. The article has two primary objectives. First, it suggests the added value of captions in making digitized illustrations more searchable by picture content in online archives. To further this objective, we describe the methods we have used to identify captions, which can effectively be re-purposed and applied in different contexts. Second, we suggest how this research leads to new understandings of the semantics and significance of the captions of historical book illustrations. The findings discussed here mark a critical intervention in the fields of digital humanities, book history, and illustration studies.

The captions that appear alongside pictures in historical books are situated on a threshold, hanging in their own space, somewhere between the image and the text proper. Like the captions that this paper discusses, the research that we describe here also occupies a threshold: between digital humanities, with its focus on computational tools and remediation to generate research questions, and book history, with its attention to the material narratives of the book as object. It is this liminal space that has come to define the field of illustration studies. In his discussion of “thresholds”, Gérard Genette famously shied away from discussing illustration because it was an “immense continent” that was too large for him to traverse [Genette 1997, 406]. Digital humanities has risen to this challenge both by making illustrated material more accessible than ever before and by engaging with how accessibility and scale can reveal new ways of analysing historical illustrations [Thomas 2017]. 1

From within this critical space, this article sets out to search the “immense continent” of illustration for what is perhaps its most marginalised territory: the caption. Very little work has been undertaken on the significance of captions and how they make their meanings, despite their prevalence in illustrated books. These issues, however, are of some importance for understanding how captions relate to the content of the pictures that they accompany, as well as the wider dialogue between word and image that characterises illustration as a mode of representation. What we outline below is the first attempt to capture the significance of captions by describing the methods and findings of research that identifies and interrogates the captions of historical book illustrations at scale, research that is only possible in a digital environment where images from different books can be viewed alongside each other. 2

The dataset that is central to this study consists of over a million illustrations from 68,000 volumes in the British Library's collection, which were digitised by Microsoft. These illustrations form the basis of the world's largest online resource dedicated to book illustration, The Illustration Archive, which we created on a previous AHRC-funded project (https://illustrationarchive.cf.ac.uk/). The illustrations in this dataset span the sixteenth to the twentieth century, with the majority clustering around the late eighteenth and nineteenth centuries, a period that witnessed a global explosion of illustrated material and was the immediate precursor of our own visual culture. The books are written in different 3 languages, with the vast majority in English, and they cover genres categorised as literature, history, philosophy, and geography, a diverse range that provides insight into the practices and importance of illustration at a time when this mode of representation constituted the “mass image”, the dominant visual form of the day.

The fact that illustrations were so prevalent at this historical moment makes it imperative that we understand their constituents, not least because digitisation has made these illustrations available to view in their hundreds of thousands. Our study suggests that it is the very ambiguity and indeterminacy of captions that opens them up to computational analysis. This proposition might seem counter-intuitive: generally, we think that it is objects that are easier to “describe” and identify computationally that are best suited to analysis. Certainly, some interesting work has been done using historical illustrations in this way. In 2014 an ambitious project was undertaken by Kalev Leetaru to identify the illustrations in books and make them searchable using the captions alongside the paragraphs immediately preceding and following the illustrations [Leetaru n.d.]. More recently, Hoyeol Kim has adapted Victorian illustrations as training sets for deep learning models by creating a dataset containing colourised black-and-white illustrations from the nineteenth century [Kim 2021]. 4

What we propose in this article is that indeterminacy — what is not clearly known, defined, or fixed — also has an analytical value. Captions can be identified computationally precisely because of their indeterminate space between the text and image; the words of the caption are neither part of the image, nor do they fully belong to the body of the text. It is this liminal space that enables the caption to be identified. This indeterminacy can also be understood and exploited on a semantic level (as we describe in the “Analysing Captions” section below). By using the captions of illustrations as a search tool, we can begin to find those instances when the captions seem to capture the visual content of the image, as well as those instances when they do not. The drive for searchability and the retrieval of relevant and accurate “hits” generates an analytical space that reveals new insights into the complex interrelationship between the caption, image, and text. Our study, therefore, sets out to identify captions both as a mechanism for searching illustrations in a big dataset and as a way of exploring how captions signify, an exploration that emerges from the very limits of searchability. 5

What is a Caption?

The function of the *caption* — a piece of text that lies alongside a visual image — can be traced back to the historical use of the *inscription*, *motto*, and *legend*. The word *subscription* was used from the sixteenth century to describe the words placed below a picture or portrait. *Caption* itself is a more recent term. Deriving from the Latin for *taking*, *caption* was originally used in the late eighteenth century in a textual context: to describe the heading of a chapter, section, or newspaper article. Its specific meaning as the “title below an illustration” emanates from America in the early decades of the twentieth century. 6

The Early Modern Graphic Literacies project (University of Turku, Finland), which is working to create an historical taxonomy of visual devices in early English print from the late 1400s to 1800, draws attention to the use of the caption in these forms, noting that the distinction between captions and other image-related texts (for example, titles and running titles) is not always obvious [Ruokkeinen, S. et al. 2023, 9]. Historically, our dataset picks up at the point when the Early Modern Graphic Literacies project ends, but even in this later period, the caption is beset with ambiguity and a startling variety of practices. The difficulty of converting some captions into a readable form is a consequence of the fact that the caption is generally printed using the same reproductive method as the illustration, and numerous printing methods are represented in our dataset: intaglio illustrations (e.g., etchings, steel engravings) include an etched or engraved caption as part of the image, lithographed images have lithographed captions, and wood-engraved and photomechanically produced images usually come with letterpress captions that connect the captions visually to the rest of the text. Captions were sometimes reproduced in the front matter of the book in the form of a “list of illustrations”, with the caption explicitly adopting the role of image title. The choice of words in the caption could be formulated by the author, the artist (or engraver), or by the publisher. 7

The very presence of captions varies considerably across historical book illustrations. There are four main types of illustration represented in our dataset: embellishments that are primarily positioned as the headers or footers of chapters and books (also known as *ornaments* or *decorations*); pictorial capital letters; illustrations that are inset 8 anywhere on a page of text; and full-page plates, where the image occupies its own page. Of these types of illustration, the first two (embellishments and pictorial letters) do not include captions. The second two (inset and full-page illustrations) do, although this is not uniform across all inset and full-page illustrations. In our calculations, and with “embellishments” excluded, we have identified 513,914 captioned illustrations from a total of 665,684 illustrations (where overlapping illustrations are counted as a single image). The majority of captions in our dataset are placed below rather than above the image and occupy their own space at a slight distance from the image (see Figures 3, 5, 7, 8, and 9 below). In the case of those illustrations that are set alongside the text on the page, the caption is also positioned at a slight distance from the text proper (see Figures 1, 2, and 4). These physical features allow us to identify the caption computationally in the ways we outline in the following section.

The positioning of the caption also has semantic implications, suggesting its status on, and between, both sides of the threshold between word and image. Little critical attention has been devoted to the complexities of the caption and how it signifies in relation to the image and to the rest of the text. Roland Barthes is one of the few critics to have engaged with this interaction, suggesting that captions can work to “anchor” the meanings of an image by indicating how the image should be read and directing the viewer's interpretation. Barthes's discussion is based on captions in twentieth-century advertising [Barthes 1977, 32–51], but his ideas are borne out in recent analyses of the captions of historical illustrations. Thomas Smits has discovered that captions were often added and amended when illustrations were reused in nineteenth-century European newspapers [Smits 2020]. Sioned Davies gives an account of the possible reasons why captions were added to the illustrations in the second edition of Charlotte Guest's translation of *Geraint the Son of Erbin* (1849), arguing that they add a specific geographical and historical dimension to the settings of the images that are otherwise indistinct [Davies 2019]. Valentina Abbatelli has also drawn attention to the importance of captions in remediating the racial politics of the early twentieth-century Italian editions of Harriet Beecher Stowe's novel *Uncle Tom's Cabin* [Abbatelli 2018]. According to Abbatelli, the same image of Uncle Tom learning to write is recast in different ways by the captions, with one caption suggesting that he is “failing” to write and another suggesting that he is making an effort. 9

In Davies's and Abbatelli's strikingly different examples, the illustrations remain the same across editions; it is the captions that are added or changed, and this modification, in turn, changes the meanings of the image. A computational identification of captions gives an opportunity to scale up these analyses to see patterns that are difficult, if not impossible, to discern in the material form of the book, where comparisons between illustrations and captions are necessarily limited. The next section describes the AI methods we used to identify captions in this large dataset of historical illustrations. These methods can effectively be re-purposed across other datasets to add value to the searchability of illustrations. The section following suggests how this identification leads to new understandings of the significance and semantics of the captions of historical book illustrations, which are of relevance not only to scholars working in digital humanities but also to those whose interests lie in illustration studies and its attendant fields. 10

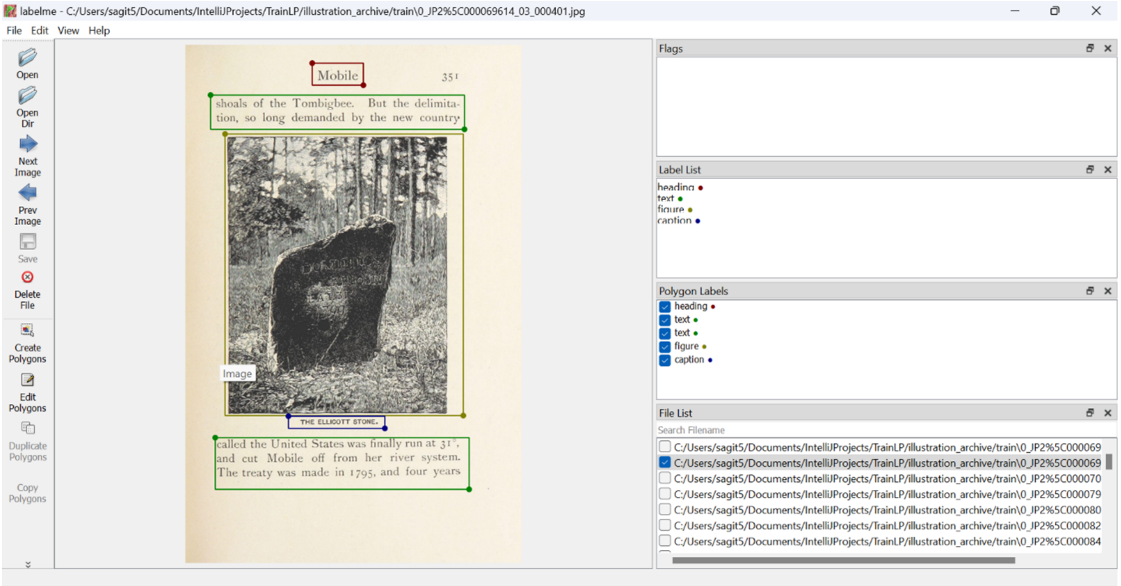
Using AI to Identify Captions

The first stage in our identification process was to isolate the captions from the main body of the text and the image, a task that has proven challenging for both machine learning and humanists attempting to formalise what a *caption* is. Zongyi Liu and Hanning Zhou provide a heuristics-based approach to caption finding, suggesting a series of rules to identify a caption based on size of the font, position relative to the image, and size of the caption itself [Liu and Zhou 2011]. Employing this method led to a high number of false positives and false negatives; our dataset presents wide variation in style of layouts, making it impossible to list a reliable set of heuristics. In order to isolate the caption, we needed to identify the other elements present on the page; that is, we needed to “parse” the layout of the page. For simplicity, we identified four building blocks: 11

1. Text, referring to the main body of text, such as paragraphs, introductions, and any text which does not fall into our two other textual categories;
2. Captions;
3. Headings, referring to titles, chapter titles, and any form of text appearing at the top of the page that might be shared across subsequent pages; and

4. Figures.

We then labelled a random subsample of 1,000 images from our dataset using labelme (https://github.com/wkentaro/labelme), a graphical annotation tool that allows users to draw bounding boxes on an image and assign them a label. The annotations of each page are then saved as a file that contains the coordinates of each bounding box and its respective label. Figure 1 shows a snippet of the labelling tool with one of our images. In the left panel, boxes can be drawn around the different sections of the page, highlighting its layout. On the right, each bounding box is labelled with one of the categories we have identified above. These annotations allow us to re-train a computer vision model to automatically identify the layout of a page and isolate a caption. 12


labelme - C:/Users/sagit5/Documents/IntelliJProjects/TrainLP/illustration_archive/train\0_JP2%5C000069614_03_000401.jpg

Mobile 351

shoals of the Tombigbee. But the delimitation, so long demanded by the new country

THE ELLICOTT STONE.

called the United States was finally run at 31°, and cut Mobile off from her river system. The treaty was made in 1795, and four years

Label List
heading ●
text ●
figure ●
caption ●

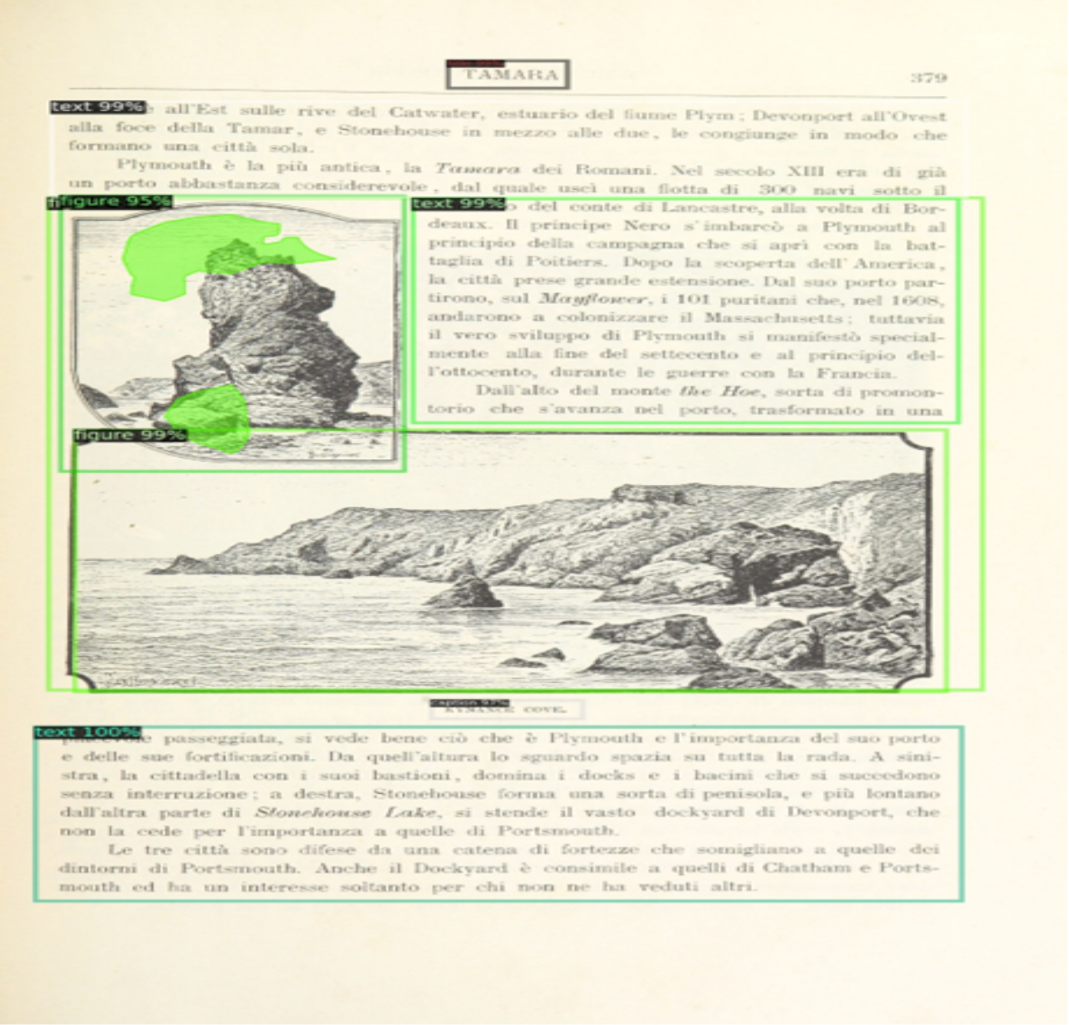
Polygon Labels
☑ heading ●
☑ text ●
☑ text ●
☑ figure ●
☑ caption ●

File List
Search Filename
☐ C:/Users/sagit5/Documents/IntelliJProjects/TrainLP/illustration_archive/train\0_JP2%5C000069
☑ C:/Users/sagit5/Documents/IntelliJProjects/TrainLP/illustration_archive/train\0_JP2%5C000069
☐ C:/Users/sagit5/Documents/IntelliJProjects/TrainLP/illustration_archive/train\0_JP2%5C000070
☐ C:/Users/sagit5/Documents/IntelliJProjects/TrainLP/illustration_archive/train\0_JP2%5C000079
☐ C:/Users/sagit5/Documents/IntelliJProjects/TrainLP/illustration_archive/train\0_JP2%5C000080
☐ C:/Users/sagit5/Documents/IntelliJProjects/TrainLP/illustration_archive/train\0_JP2%5C000082
☐ C:/Users/sagit5/Documents/IntelliJProjects/TrainLP/illustration_archive/train\0_JP2%5C000084

Figure 1. A snippet of the annotation process using the labelme tool.

With this annotated dataset, we turn to fine-tuning. In deep learning, fine-tuning is a transfer learning method where a pre-trained model is trained on new data to fine-tune it on a more specific task. In our case, that task is layout parsing on historical illustrated books. Layout parsing is the process of automatically detecting the layout structure of a page using computer vision. A highly effective recent model is Layout Parser (https://layout-parser.github.io/), which provides a range of models pretrained on diverse databases. We used the model trained on the PrImA dataset, which is is a Mask R-CNN model trained on newspapers that identifies text, figures, titles, tables, maths, and separators. Newspapers present a similarly busy and interlocked layout as historical illustrated books, meaning that this model most closely resembled our data. We ran training for 10,000 iterations with a learning rate of 0.001, achieving accuracy of 98% on the validation set. The majority of errors occured on pictorial capital letters: at the annotation stage, we decided for simplicity to include pictorial letters within the main body's bounding box labelled as text, but the model often labels them separately as figures. 13

We deployed our fine-tuned model on The Illustration Archive to predict the layout of each page and isolate the captions. The model's output consists of coordinates for the layout element of each page which can be visualised as bounding boxes. Figures 2–4 demonstrate an example output of the model: on each page, the model has predicted the position and category of a layout element, which is displayed with a confidence score as a percentage. We can see in Figure 2 that three bounding boxes with the label “figure” have been predicted, as the model is unsure whether the two illustrations should be considered separately. An analogous situation occurs in Figure 4 with the captions. 14



TAMARA 379

all'Est sulle rive del Catwater, estuario del fiume Plym; Devonport all'Ovest alla foce della Tamar, e Stonehouse in mezzo alle due, le congiunge in modo che formano una città sola.

Plymouth è la più antica, la *Tamara* dei Romani. Nel secolo XIII era di già un porto abbastanza considerevole, dal quale uscì una flotta di 300 navi sotto il del conte di Lancastre, alla volta di Bordeaux. Il principe Nero s'imbarcò a Plymouth al principio della campagna che si aprì con la battaglia di Poitiers. Dopo la scoperta dell'America, la città prese grande estensione. Dal suo porto partirono, sul *Mayflower*, i 101 puritani che, nel 1608, andarono a colonizzare il Massachusetts; tuttavia il vero sviluppo di Plymouth si manifestò specialmente alla fine del settecento e al principio dell'ottocento, durante le guerre con la Francia.

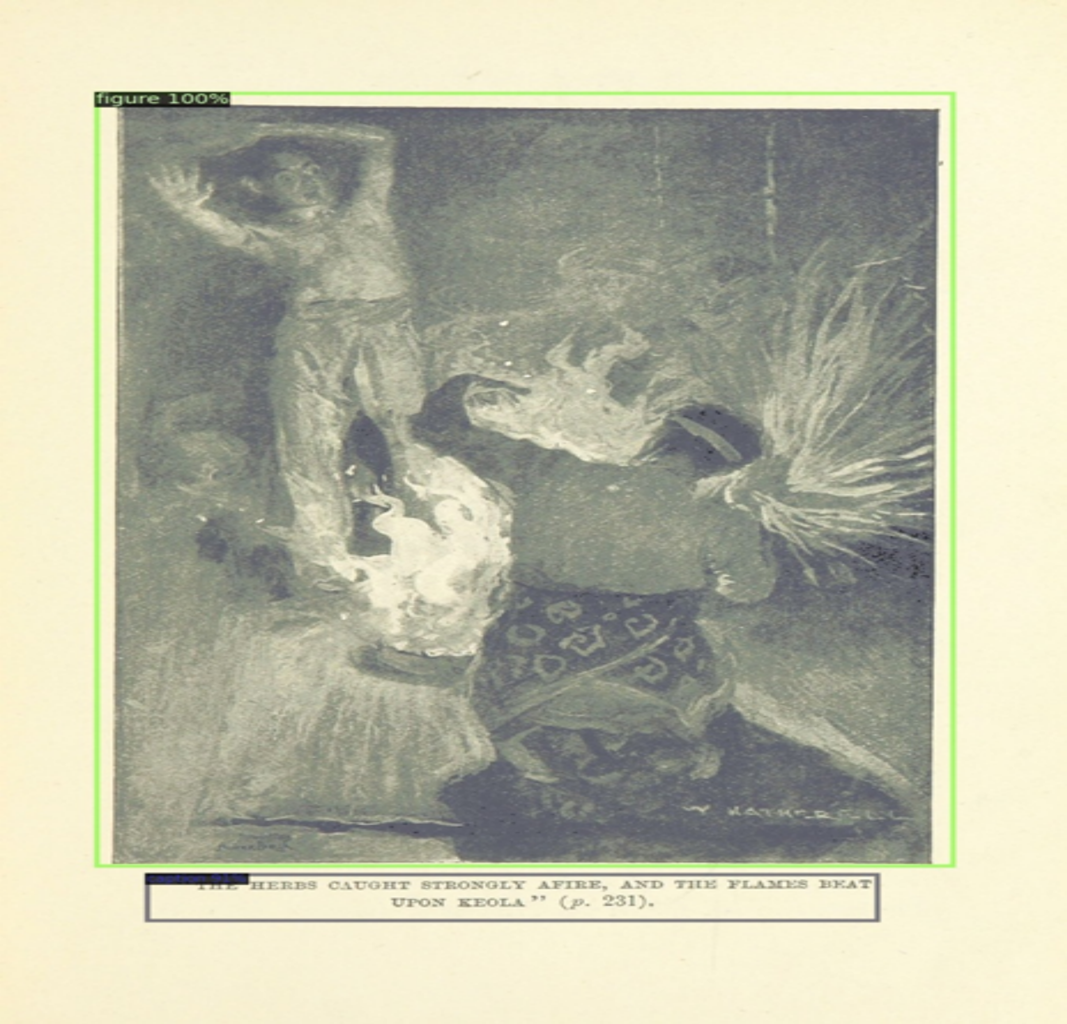
Dall'alto del monte *the Hoe*, sorta di promontorio che s'avanza nel porto, trasformato in una

KYNANCE COVE.

passeggiata, si vede bene ciò che è Plymouth e l'importanza del suo porto e delle sue fortificazioni. Da quell'altura lo sguardo spazia su tutta la rada. A sinistra, la cittadella con i suoi bastioni, domina i docks e i bacini che si succedono senza interruzione; a destra, Stonehouse forma una sorta di penisola, e più lontano dall'altra parte di *Stonehouse Lake*, si stende il vasto dockyard di Devonport, che non la cede per l'importanza a quelle di Portsmouth.

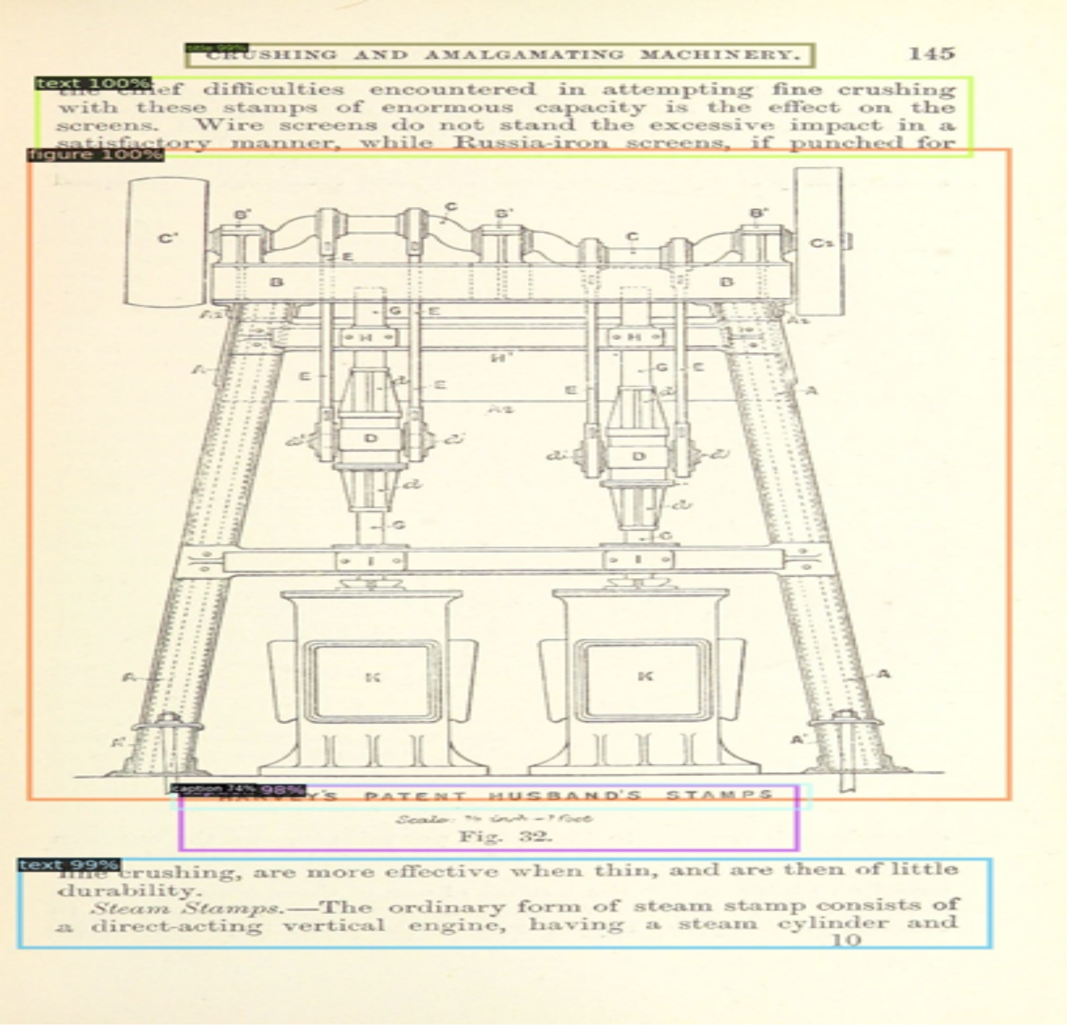
Le tre città sono difese da una catena di fortezze che somigliano a quelle dei dintorni di Portsmouth. Anche il Dockyard è consimile a quelli di Chatham e Portsmouth ed ha un interesse soltanto per chi non ne ha veduti altri.

Figure 2. An example of the output of the parsing tool.



“THE HERBS CAUGHT STRONGLY AFIRE, AND THE FLAMES BEAT UPON KEOLA” (*p.* 231).

Figure 3. An example of the output of the parsing tool.



CRUSHING AND AMALGAMATING MACHINERY. 145

the chief difficulties encountered in attempting fine crushing with these stamps of enormous capacity is the effect on the screens. Wire screens do not stand the excessive impact in a satisfactory manner, while Russia-iron screens, if punched for

HARVEY'S PATENT HUSBAND'S STAMPS

Scale: ½ inch = 1 foot

Fig. 32.

fine crushing, are more effective when thin, and are then of little durability.

Steam Stamps.—The ordinary form of steam stamp consists of a direct-acting vertical engine, having a steam cylinder and

10

Figure 4. An example of the output of the parsing tool.

After obtaining labelled layout structures, it becomes easier to perform optical character recognition (OCR) exclusively on the captions. We used pytesseract (https://pypi.org/project/pytesseract/), a Python library that is a wrapper for Google's Tesseract OCR engine. The main disadvantage of this tool is its speed when attempting to use multiple languages. For this reason, we allowed the tool to rely only on English at this stage, meaning that captions in foreign languages were not processed correctly; we are planning to fix this issue in a future iteration. 15

Through these computer vision techniques, we have obtained a dataset of captions that can be used as a search index for the illustrations in The Illustration Archive. Integrated as metadata on The Illustration Archive website, the captions allow users to explore the images, search by captions, and analyse the relation between the words in the captions and the content of the illustrations. The identification of the captions has also opened up possibilities to use image-to-text and text-to-image models such as CLIP (Contrastive Language-Image Pre-training: https://openai.com/research/clip) to cluster images with similar captions, identify objects in illustrations using CLIP text encodings as well as our fine-tuned caption encodings, and discover connections between language representation and visual representation. CLIP is a deep learning model that draws simultaneously from text and image. It is trained on vast amounts of data consisting of pairs of images and descriptive text, allowing it to train a text encoder and an image encoder to produce text and visual embeddings that are close to each other. In this way, a textual prompt can be used to search for images with embeddings resembling the textual prompt. We have used OpenCLIP ViT-L/14 model (Vision Transformer) to build a searchable index of the illustrations in The Illustration Archive, following the work of the Visual Geometry Group at the University of Oxford (see their WISE Image Search Engine, which uses AI tools to make image databases searchable by content: https://www.robots.ox.ac.uk/~vgg/software/wise/). 16

Analysing Captions

Developments in deep learning models mean that we are no longer wholly reliant on textual metadata (tags, bibliographic information, or, indeed, captions) to search the content of images in datasets of historical images. Back in 2004 when we started developing the Database of Mid-Victorian Illustration (https://www.dmvi.org.uk/), the only way of making the 900 or so images searchable to users was manually to tag their content. Captions, however, can still add value to content-driven visual or iconographic searches. Online archives of historical newspapers often use captions for searchability, although the captions of these illustrations do not necessarily function in the same way as the captions in books. In newspapers, the caption is generally a “heading” that appears at the top of the feature and is the title both for the illustration (if there is one) and the textual article. In the database Welsh Newspapers (https://newspapers.library.wales/), for instance, the title in the form of a heading has been identified as a distinct characteristic of the page. Any alternative captions underneath an image have been OCRed along with the full text and can be searched accordingly. A recent use of OCRed headlines, captions, and other text alongside deep learning models is the Newspaper Navigator, which searches the visual content of historical newspapers in the Library of Congress [Lee n.d.]. 17

We conducted a case study to evaluate the effectiveness of using our identified captions to search for specific illustrations in the dataset. In partnership with Lambeth Palace Library, we searched for book illustrations of cathedrals that would enrich the library's collections. We found 2,087 illustrations captioned with the word *cathedral* and 103 with the word *cathedrals*. Data-mining the captions in this way involves the same discrimination as might be adopted on a Google search, in the sense that we needed to anticipate the words that might be used in these captions (e.g., additional terms not covered by the word *cathedral*, such as York Minster, Westminster Abbey, and *chapterhouse*, which might signal a cathedral building). 18

In this study, searching the captions proved a highly effective way of identifying relevant illustrations from the dataset because the words of the captions describe in a broad sense what the illustrations depict: *cathedrals*. In some instances, a caption search can be *more* effective than using only a visual search. An image classification (as opposed to a caption) search would retrieve images of cathedrals, but it would also return churches since there are no specific architectural signifiers that mark out a cathedral from a church. Cathedrals are generally larger and more elaborate than churches, but their specific distinction — which would not usually be signalled in an illustration or in visual searches — is that cathedrals are the seat of a bishop. 19

Likewise, a caption search is able to retrieve illustrations that show features of cathedrals that are not iconographically specific to cathedrals and would not, therefore, be found in image-based searches that tend to cluster key features of cathedrals (external “grand” perspectives, spires, and high-domed interiors). A search via the captions revealed illustrations of floor plans of cathedrals and different perspectives where the cathedral is in the far distance, so not obviously identified as a cathedral, as well as many images of interiors and interior features (e.g., effigies, gargoyles, windows, staircases; see Figure 5). The captions here retrieve images of cathedrals that would not otherwise be found. 20

This example of the effectiveness of searching captions in large datasets raises an important point about the nature of captions, especially in illustrated non-fiction books. As contemporary labels attached to the images, captions can function as “expert tags”: authoritative and reliable indicators of what the illustration contains. They can, for instance, identify an otherwise unrecognisable interior setting as depicting “Bristol Cathedral” (Figure 5). 21

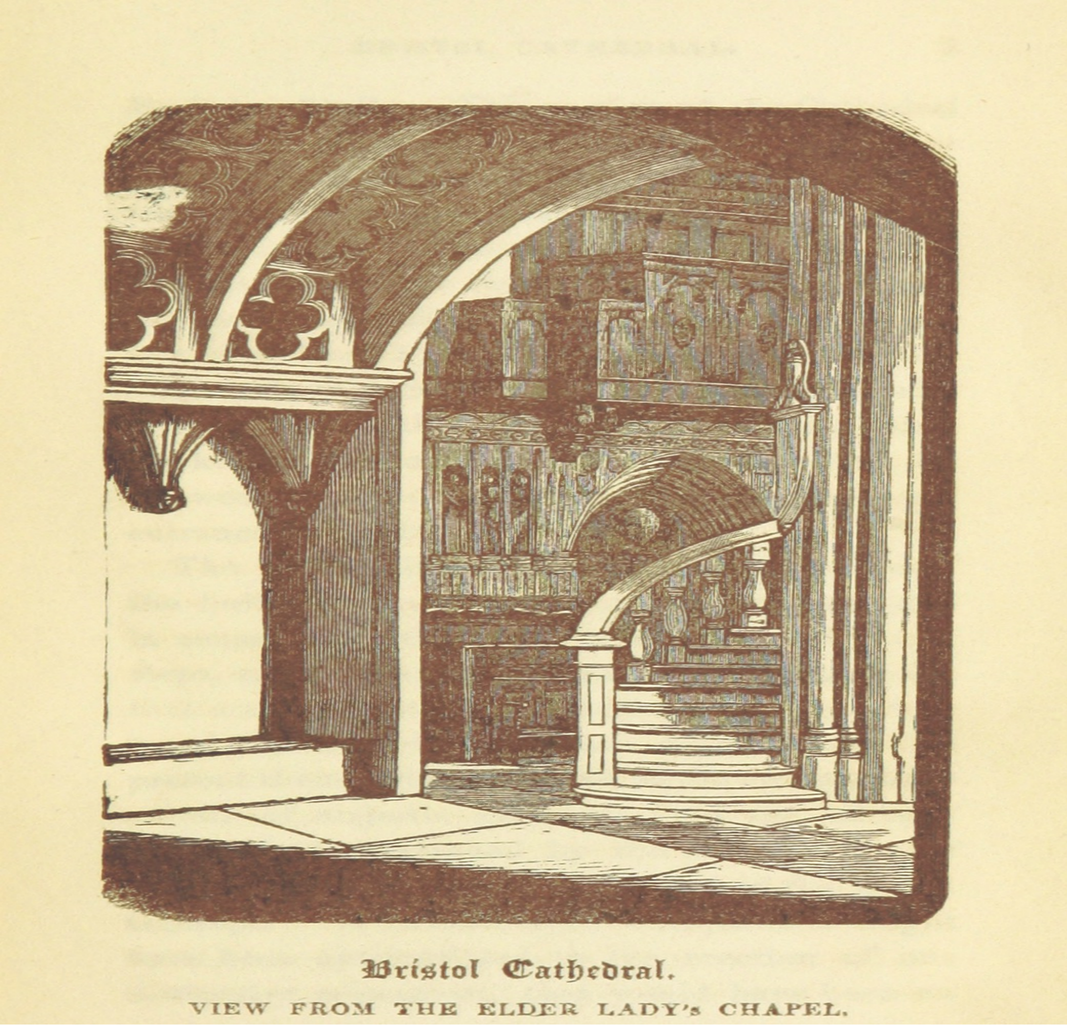


Figure 5. Example of an image retrieved by searching the captions for *cathedral.* From [Samuel Griffiths Tovey], *Cursory Observations on the Churches of Bristol. By an Occasional Visitor*, second edition (Bristol, England: Mirror Office, 1843), facing p. 6.

This aspect of the caption as a reliable descriptor of the image can add to the searchability of large illustration datasets. However, captions are far from neutral or ahistorical descriptors of the image, and it is this problematic relationship that comes to the fore in our research. By identifying hundreds of thousands of captions, we have effectively isolated a corpus of textual metadata that is culturally significant. Even a term as apparently straightforward as *cathedral* has a historical resonance: its use in a caption speaks to the establishment and embeddedness of (in this case) the Church of England, the rise of tourism in cathedral towns and cities in the UK, and the changing demographics brought about by industrialisation, as no new cathedrals were created for nearly 300 years until the Victorian period when the increased population in major industrial cities necessitated new cathedrals. There is a civic pride in the captioning of *cathedrals*, especially as illustrations of cathedrals most commonly appear in the context of tourist guidebooks and local histories, like the one in Figure 5. *Cathedral* is just one example of the cultural significance of the words used in captions. Others are more overtly political and unsettling, such as racist terms that are sometimes used. 22

We can interrogate this corpus of captions in multiple ways, including searching for the frequency distribution of words. Figure 6 shows a random subset of words sampled from different frequency bands. 23

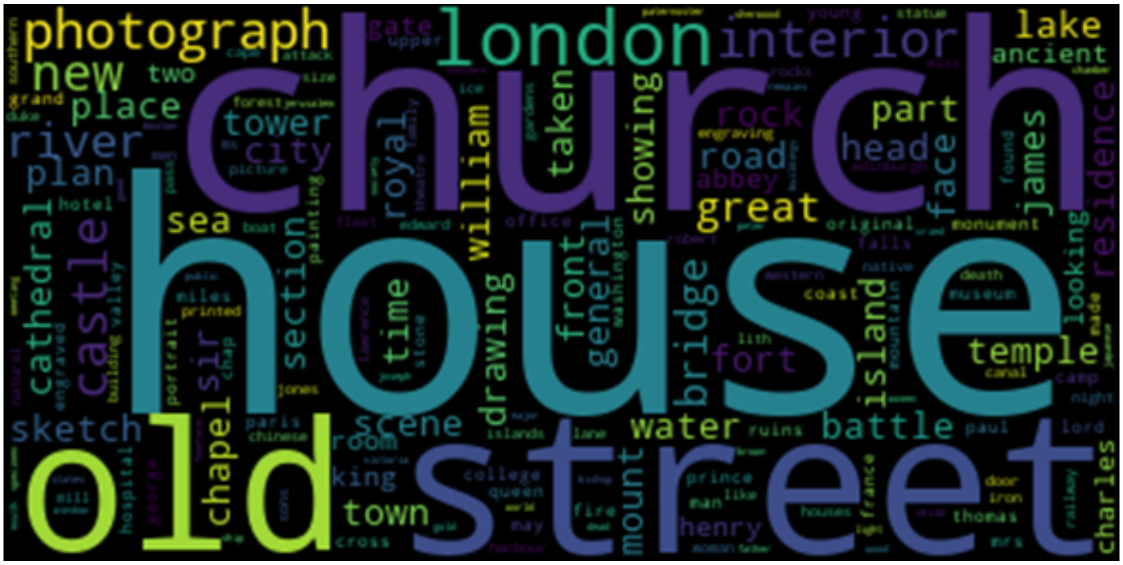


Figure 6. A word cloud obtained using Python from a random sample of words occurring in the corpus of captions.

24 The arbitrary nature of these words reflects the mixed genres that make up our illustration dataset, but, seen in the form of a word cloud, the words in the captions clearly foreground dominant nineteenth-century values. The most frequent keywords are *house* (in 3,043 captions) and *church* (in 2,688 captions); there is also an emphasis on the *old* (2,561 captions), further signalled by terms like *castle* (1,778 captions), *ancient* (592 captions) and *abbey* (556 captions); and the *new* (in 1,760 captions), signalled by *photograph* (2,009 captions). Aspects of the countryside and landscape are emphasised, alongside towns and cities (*London* is one of the top terms, used in 2,219 captions). Gender categories are particularly revealing, with 519 captions containing the word *man* and 307 containing *woman* (the plural suggests a similar margin: 274 *men* and 187 *women*); *Mr* appears 1,198 times, *Mrs* 497 times, and *Miss* 250 times.

25 Although it is highly likely that the illustrations in the dataset contain more illustrations of men than women, this is not necessarily the full picture. A comparison search for image features can prove illuminating here. An image feature search on a subset of 15,000 illustrations using CLIP found 524 men and 14 women. While, surprisingly, there were no groups of men, 2,487 groups of women and 1,562 groups of people, more generally, were recognised. More work needs to be done comparing different search mechanisms. Our analysis indicates that whilst there is an undoubted bias in the illustrations, there is also a bias in the captions, which might not necessarily be the same as the bias in the images. Identifying and isolating captions in this way exposes how captions emphasise and marginalise features of illustrations.

26 The caption's privileging and concomitant marginalising of pictorial details is, in fact, one of its major characteristics. The caption of Figure 5, “Bristol Cathedral. VIEW FROM THE ELDER LADY'S CHAPEL”, contains no mention of stairs, flagstones, archways, monuments, or any other architectural features. However descriptive and objective it appears, the caption is never fully reflective of the image because it is couched in another form: words, which cannot fully describe the visual features of an image. What captions emphasise and marginalise, then, is highly significant in that they provide a unique insight into how illustrations were viewed at the time, what were regarded as their most salient features, and what features were overlooked.

27 Using AI to identify the captions at scale allows us to look not only at the words of the caption, but also to trace patterns in how the captions signify in relation to the illustrations and the rest of the text. Our findings point to the fact that the caption makes its meanings in two main ways: it signifies in its conjunction with the image, and it acts as a point of connection, or bridge, between the illustration and the rest of the text. Either, or both, of these relationships can be at play in any captioned illustration, but they seem primarily to be determined by the genre of the book.

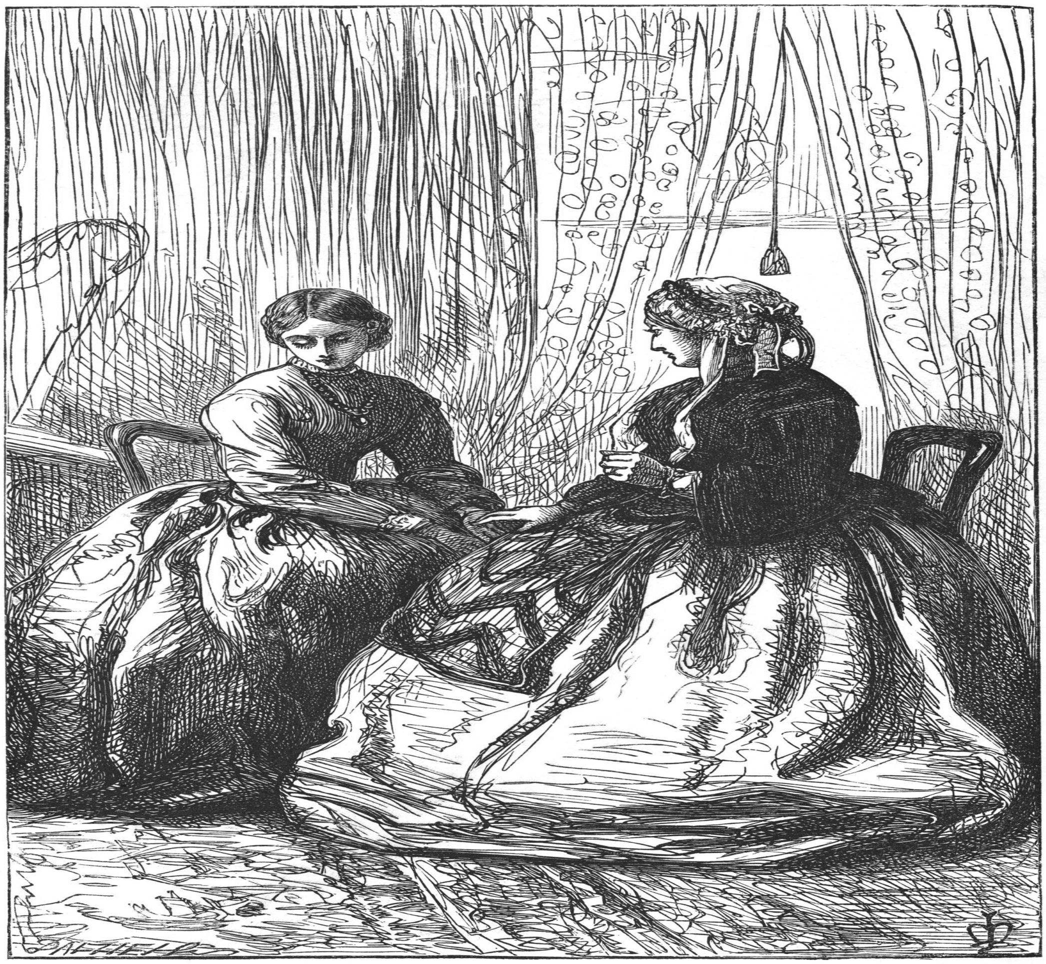
28 In texts conventionally classified as non-fiction, including works of history, geography, and science (in the broadest sense), the role of the caption and its relation to the image is characterised by the first of these models, with the

illustration and caption forming a complete signifying unit. In theory, this unit could be isolated from the rest of the text and still make sense. Although the caption “Bristol Cathedral” fails to mention certain aspects of the image, it nevertheless signifies in relation to the illustration, with the picture and the words of the caption constituting a signifying partnership.

This is the defining model across the range of non-fiction books in the dataset. In simple terms, the captions in these books describe the facts of what is represented in the illustrations; following the Latin etymology of *caption* they attempt to “capture” or “seize” the meanings of the picture. In our dataset, there are tens of thousands of botanical or zoological illustrations captioned with the names of the species, illustrative portraits of people captioned with their names, and illustrations of locations or buildings with captions that identify the locations and buildings they represent. In these examples, the signifying conjunction of the illustration and caption creates and enforces a notion of equivalence: the idea that the caption replicates in words what the illustration shows and vice versa (however illusory this notion is). 29

In illustrated fiction and literary texts (short stories, novellas, novels, plays, poetry, etc.), captions are usually citational, taking the form of quotes or deriving from the words of the text. These captions frequently state the characters' names and/or what they are doing: “Mr. Perry passed by on horseback” is a typical caption from a Victorian illustrated edition of Jane Austen's *Emma* (1896). (The designations *Mr.*, *Mrs.*, and *Miss* are more prevalent in the captions of fictional texts where they are used to name a character than they are in non-fiction captions.) In literary texts, the contexts and meanings of the illustrations and the captions depend on the text; they do not necessarily make sense in isolation like the non-fiction captioned illustrations described above. The role of these literary captions is to point to the specific episode being illustrated, a role that serves a practical function because book illustrations often appeared on different pages than the text that they depicted (thus many captions also include the relevant page numbers; see Figure 3). Instead of turning in towards the illustration, then, these captions point outside, acting as a bridge between the illustration and the text proper. 30

An example of this bridging technique is the caption “Never is a very long word”, which accompanies an illustration for Anthony Trollope's novel *Orley Farm* (Figure 7). 31



Never is a very long word.

Figure 7. John Everett Millais, “Never is a very long word”, engr. Dalziel Brothers. Illustration for Anthony Trollope, *Orley Farm* (London: Chapman and Hall, 1862), facing p. 77.

32 “Never is a very long word” is a direct quotation from the novel, and this caption is placed securely within quotation marks when it appears as a title on the contents page in the book edition. By citing the text of the novel, the caption works to direct the reader to the episode that is being illustrated, isolating the moment depicted in the illustration from several possibilities (the pose of the two seated ladies could fit any number of episodes). In this way, the caption works to connect the image to the rest of the text. As if to prove that an image cannot translate into words, the illustration does not depict the words “Never is a very long word”. How could these words even be illustrated in a picture? This caption, therefore, is a good example of where the use of captions to search for picture content would fail (the content of this image is manually tagged in the Database of Mid-Victorian Illustration). In terms of searchability, the use of captions is far more effective when searching across the non-fiction books in the dataset.

33 Our analysis, therefore, indicates that the caption does not signify in the same way across all books. Rather, there are specific conventions at play that are determined primarily by the broad generic classification of the book. An acknowledgement of these conventions allows us to recognise those instances where they are being adapted and manipulated, with implications for the meanings of the text and how it is read. An example of this is Figure 8, the frontispiece illustration for Trollope's *Orley Farm*. Unlike the caption in Figure 7, which follows the conventions of literary illustrations and quotes directly from the novel, Figure 8 draws instead on the conventions of the “factual” caption.

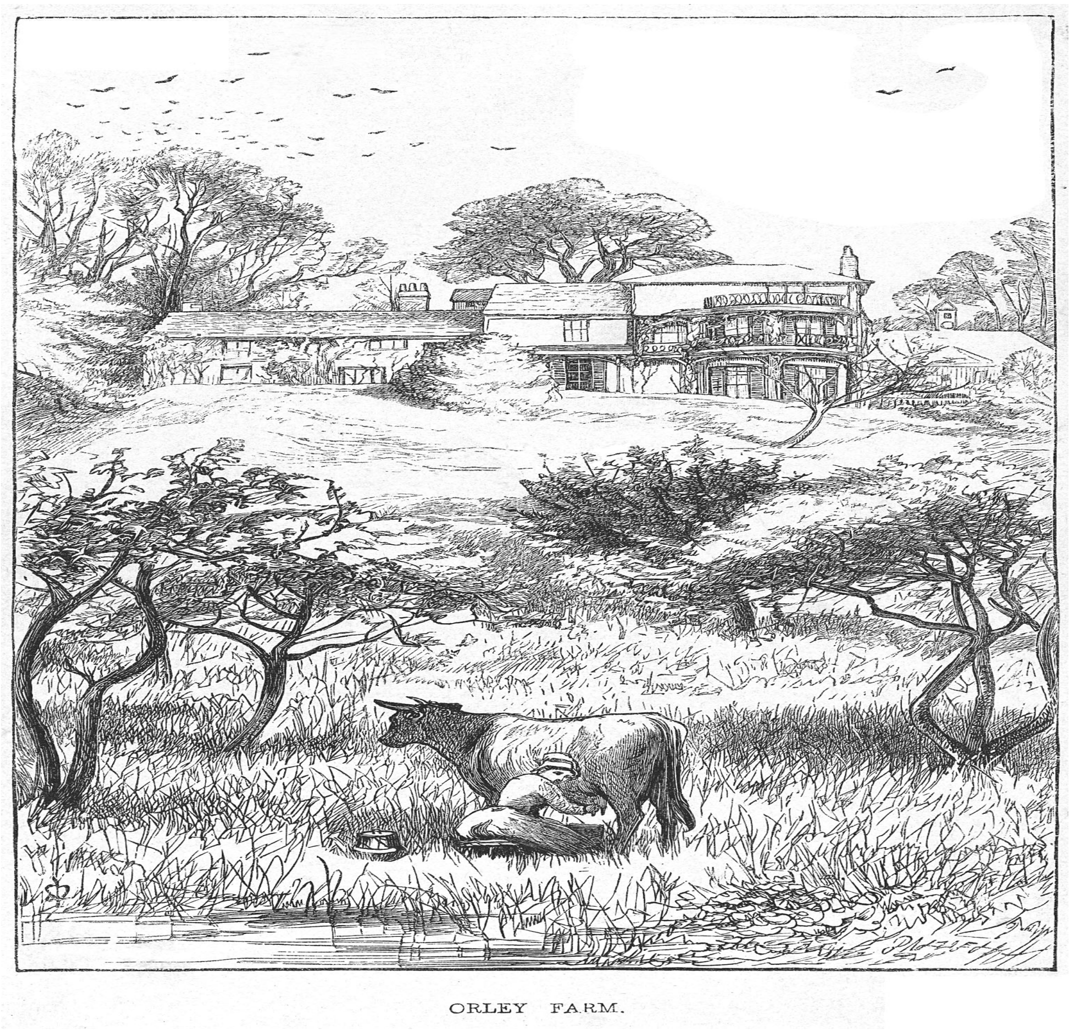


Figure 8. John Everett Millais, “ORLEY FARM”, engr. Dalziel Brothers. Illustration for Anthony Trollope, *Orley Farm* (London: Chapman and Hall, 1862), frontispiece.

34 The illustration in Figure 8 is based on a real location, the farmhouse where Trollope lived as a boy. The style of the image itself crosses generic categories: pictures of rural landscapes frequently appeared in both fictional and non-fictional books, including gift books, poetry anthologies, and topographical works. However, it is the caption that makes the link to a real geographic location. Capitalised to lend it more authority, “ORLEY FARM” mimics the style of captions in illustrated travel books that label the place illustrated (see, for example, Figure 9).

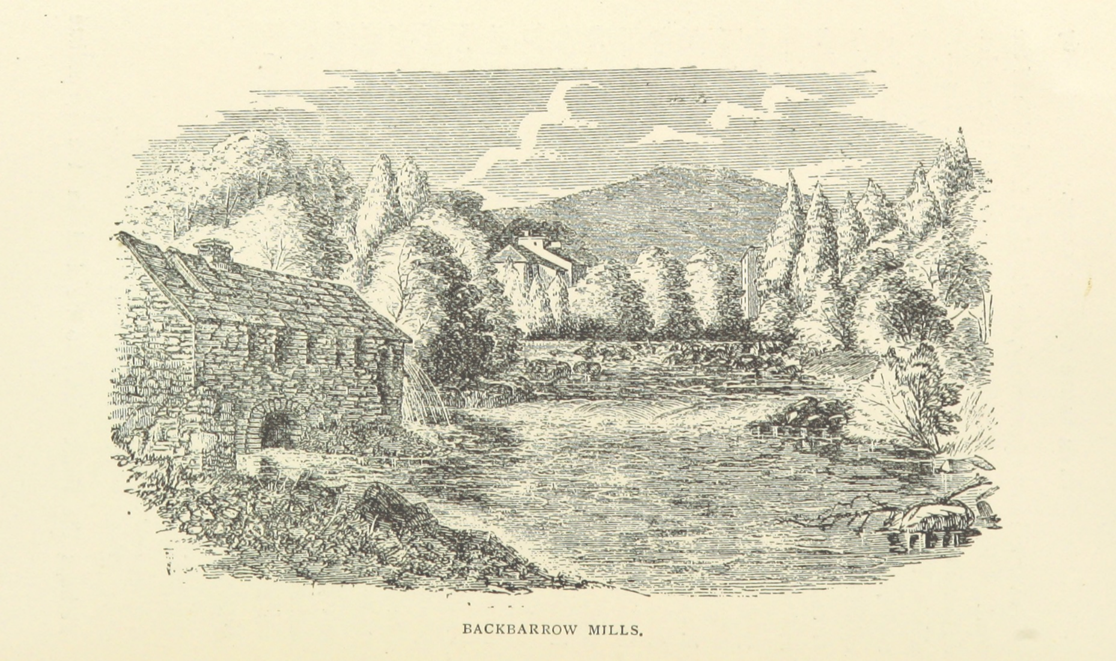


Figure 9. “BACKBARROW MILLS”, illustration for Edwin Waugh, *Rambles in the Lake Country and Other Travel Sketches*, ed. George Milner (Manchester and London: John Heywood, 1893), p. 68.

The association of the caption “ORLEY FARM” with the designatory captions of non-fiction books gives a sense of veracity to the novel at its outset (it is placed as the frontispiece illustration), and this has implications for the investment and immersion of readers in the imaginative world of the novel. The caption here functions alongside the realist conventions of the text and adds to the verisimilitude of the fictional world. Significantly, this caption is in marked contrast to the other captions in this novel, which either quote directly from the text (as in Figure 7) and/or identify the characters. The only caption that comes close is “Monkton Grange”, which appears underneath a picture of an old manor house where a group of people gather for a hunt. In this example, the image and the caption again function to situate the illustrated episode in a location that would have looked realistic to contemporary readers 35

Whilst Trollope's novel is, of course, a work of fiction, there are many eighteenth- and nineteenth-century illustrated books that cannot easily be classified as either fiction or non-fiction (e.g., memoirs, certain modes of travel writing, etc.). In these books, the captions often switch between different modes of address, dictating how the images are read. Charles Knight's multi-volume *Pictorial Edition of the Works of Shakspere* (1838-1843) is an interesting example of a book in which the literary text competes with historical images. Without their captions, many of these pictures could be viewed as fictional, but, with them, they are fixed in a time and place, such as “Room in Cleopatra's Palace” or “Part of Windsor Castle, built in the time of Elizabeth”. Likewise, some of the illustrations that might otherwise look historically, geographically, or botanically accurate are linked to the imaginative world of Shakespeare’s texts with captions that are direct quotations from the plays. Knight's edition veers dramatically between the factual and fictional, and this effect is produced as much by the captions as by the images. 36

Conclusion

Research into the significance and meanings of the captions of historical illustrations has been hampered by the materiality of the book and the difficulty of viewing multiple illustrations and their captions side by side and comparatively. AI tools rectify this by allowing captions to be identified and interrogated alongside each other and across tens of thousands of illustrated books. We suggest that it is the indeterminacy of the caption, its position on the threshold between the image and the text, which opens it up to computational identification and analysis. 37

Using AI to identify captions offers a mechanism for searching the content of illustrations in large datasets, and this can be more effective than using image-only classification searches. However, AI moves beyond its efficacy as a vehicle for 38 searching the content of pictures to reveal the significance of captions as words that describe — *and do not describe* — illustrations. Isolated as a dataset, captions are historically significant in their own right and can be interrogated in terms of their linguistic meanings and variations. AI also enables analysis of the caption in its complex dialogue with the illustration and the rest of the text, allowing recognition of signifying patterns and conventions across books and genres. For the first time, we can begin to trace at scale the ways in which the caption generates meanings and impacts on the reading and viewing process from its liminal position on the threshold.

Acknowledgements

39 We would like to thank Ian Harvey and Unai Lopez-Novoa for their foundational work on The Illustration Archive and captions. The Illustration Archive was funded by the AHRC. The current project is funded by the AHRC/NEH as a partnership between Cardiff University and the University of Wyoming.

Works Cited

Abbatelli 2018 Abbatelli, V. (2018) “Looking at captions to get the full picture: Framing illustrations in Italian editions of *Uncle Tom's cabin*”, *Image and Narrative*, 19(1), pp. 46-61.

Barthes 1977 Barthes, R. (1977) *Rhetoric of the image: Image, music, text*, ed. and trans. Stephen Heath. London: Fontana.

Bradski 2000 Bradski, G. (2000) “The OpenCV library”, *Dr. Dobb's*, 1 November. Available at: https://www.drdobbs.com/open-source/the-opencv-library/184404319 (Accessed: 11 May 2023).

DMVI n.d. *Database of mid-Victorian illustration* (n.d.). Available at: https://www.dmvi.org.uk/ (Accessed 11 May 2023).

Davies 2019 Davies, S. (2019) “‘A most venerable ruin’: Word, image and ideology in Guest's *Geraint the son of Erbin*”, *Studia Celtica*, 53(1), pp. 53-72.

Genette 1997 Genette, G. (1997) *Paratexts: Thresholds of interpretation*, trans. Jane E. Lewin. Cambridge: Cambridge University Press.

Hoffstaetter et al. 2022 Hoffstaetter, S. et al. (2022) *pytesseract 03.10*. Available at: https://pypi.org/project/pytesseract/ (Accessed: 11 May 2023).

Kim 2021 Kim, H. (2021) “Victorian400: Colorizing Victorian illustrations”, *International Journal of Humanities and Arts Computing*, 15(1-2), pp. 186-202.

Lee n.d. Lee, B.C.G. (n.d.) *Newspaper navigator*. Available at: https://news-navigator.labs.loc.gov/search (Accessed 11 May 2023).

Leetaru n.d. Leetaru, K. (n.d.) *500 years of book images*. Available at: https://blog.gdeltproject.org/500-years-of-the-images-of-the-worlds-books-now-on-flickr/ (Accessed 11 May 2023).

Liu and Zhou 2011 Liu, Z. and Zhou, H. (2011) “A simple and effective figure caption detection system for old-style documents”, *Proceedings of SPIE: The international society for optical engineering*. San Jose, CA, 24-29 January 2011. Bellingham, WA: SPIE-IS&T Electrionic Imaging, article #7874-28. Available at: https://www.researchgate.net/publication/221253773_A_Simple_and_Effective_Figure_Caption_Detection_System_For_Old-style_Documents (Accessed: 11 May 2023).

Radford et al. 2021 Radford, A. et al. (2021) “Learning transferable visual models from natural language supervision”, *arXiv*. https://doi.org/10.48550/arXiv.2103.00020 (Accessed: 11 May 2023).

Radford et al. n.d. Radford, A. et al. (n.d.) *CLIP*. Available at: https://github.com/openai/CLIP/?tab=readme-ov-file#readme (Accessed 11 May 2023).

Ruokkeinen, S. et al. 2023 Ruokkeinen, S. et al. “Developing a classification model for graphic devices in early printed books”, *Studia Neophilologica*. https://doi.org/10.1080/00393274.2023.2265985 (Accessed: 13 December 2023).

Shen et al. 2021 Shen, Z. et al. (2021) “LayoutParser: A unified toolkit for deep learning based document image analysis”, *arXiv*. https://doi.org/10.48550/arXiv.2103.15348 (Accessed: 11 May 2023).

Smits 2020 Smits, T. (2020) *The European illustrated press and the emergence of a transnational visual culture of the news, 1842-1870*. London: Routledge.

Sridhar, P. et al. n.d. Sridhar, P. et al. (n.d.) *WISE image search engine (WISE)*. Available at: https://www.robots.ox.ac.uk/~vgg/software/wise/downloads/wikiworkshop2023/sridhar2023wise.pdf (Accessed 20 December 2023).

The Illustration Archive n.d. *The illustration archive* (n.d.). Available at: https://illustrationarchive.cf.ac.uk/ (Accessed 11 May 2023).

Thomas 2017 Thomas, J. (2017) *Nineteenth-century illustration and the digital: Studies in word and image*. New York: Palgrave.

Wada n.d. Wada, K. (n.d.) *labelme: Image polygonal annotation with Python*. Available at: https://github.com/labelmeai/labelme (Accessed 11 May 2023).

Welsh Newspapers n.d. *Welsh newspapers* (n.d.). Available at: https://newspapers.library.wales/home (Accessed 11 May 2023).