PatFig Dataset Documentation

Dana Aubakirova
ENS Paris-Saclay
dana.aubakirova@ens-paris-saclay.fr

Kim Gerdes
Qatent
Université Paris-Saclay, LISN (CNRS)
kim@qatent.com

Abstract

The PatFig dataset is the curated version of the dataset presented in Aubakirova et al. [2023] containing over 18,000 figures from over 7,000 European invention patent applications, the corresponding short and long captions, reference numerals, their corresponding terms, and the minimal claim set that describes the interactions between the components of the image. It covers the period from January 1, 2020, to December 31, 2020.

1 Background

The terminology used throughout this dataset is consistently adopted from the language and conventions found in patent applications. A *patent* is a well-structured document typically including the textual sections *title, abstract, background, brief summary of the invention, detailed description*, and the *claims*, as well as *drawings* and the patent *classification code*. The patent images were extracted from the *drawings* section of these documents, with a single patent application potentially comprising multiple images. *Claims* define the scope of protection conferred by a patent. They are legally enforceable statements that describe the features or processes that distinguish the invention from prior art. Each claim consists of a single sentence; a patent encompasses a sequence of claims, typically arranged in order of decreasing breadth.

In this dataset, the term *minimal set of claims* refers to the specific set of claims that span all terms or elements depicted in an image, essentially summarizing the interactions among these elements within a single image. The Cooperative Patent Classification (CPC), an evolved version of the International Patent Classification (IPC) system, offers a global standard for categorizing patents across various technological fields. The IPC framework is structured into eight primary categories, each symbolizing a broad technology sector. These categories are hierarchical representations of the broadest possible areas of technology Zuo et al. [2023]. For the purposes of this dataset, the IPC system is utilized, with a detailed exposition of this classification framework provided in Table 1.

A *patent figure* is a central component of patent applications, often providing a more efficient medium for conveying complex scientific or technical information than text alone Carney and Levin [2002], Ganguly et al. [2011]. They comprise technical drawings, block diagrams, flow charts, plots, and grayscale photographs.

Challenges arise, as seen in Figure 1c, when multiple figures are encompassed within a single image. The spatial arrangement of these subfigures and their overlapping captions complicates classical rule-based and image-processing based separation approaches requiring advanced image segmentation methods. Such complexities also impede the accurate analysis of figure descriptions and the delineation of relevant sentences for each subfigure. The need for the figure rotation and padding is supported by samples such as in Figure 1b, which provides examples of an inverted caption. Therefore, based on the analysis of the figure caption's OCR result, we rotate the image to realign captions correctly. Then we square the image by resizing it from (3508, 2592) to (3508, 3508) pixels.

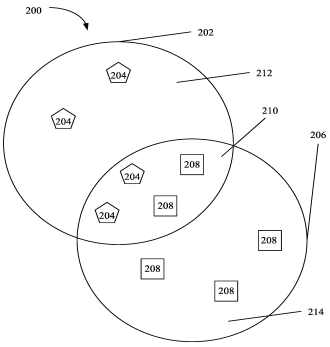

FIG. 2

(a) The illustration of the system for improving transmission in wireless networks.

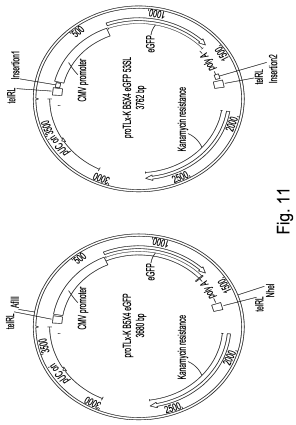

Fig. 11

(b) A plasmic map.

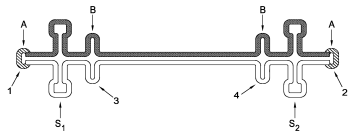

Fig. 12

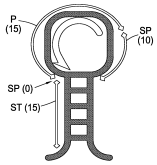

Fig. 13

(c) An exemplary structure of a closed linear DNA template and stem loop.

Figure 1: 1a is an example of an abstract figure conveying complex ideas; 1b the examples with the inverted caption, 1c is a compound figure.

A *short description* of the patent figure is a concise statement that identifies a specific illustration within a patent document and summarizes its content or function in relation to an embodiment of the invention. It usually follows a standard list format, separated by new lines, enabling a rule-based extraction method. These descriptions often appear in a section titled "Brief Description of Drawings." Typical short descriptions start with a figure number and a brief explanation, e.g., "Fig. 1 depicts a bottle pourer per an embodiment."

A *long description* of the patent figure is a broader narrative, limited to 500 tokens in this dataset, detailing the intricate interactions and relationships among the elements depicted in the illustration, derived from the analytical examination of the patent's detailed description section. Its extraction poses more challenges due to the varied structure of patent application descriptions. Addressing this, our method involved text normalization, searching for repeated figure number references, and extracting relevant sections until the start of another paragraph or a different figure mention.

Table 1: The eight IPC section categories.

Section	Title
A	HUMAN NECESSITIES
B	PERFORMING OPERATIONS; TRANSPORTING
C	CHEMISTRY; METALLURGY
D	TEXTILES; PAPER
E	FIXED CONSTRUCTIONS
F	MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING
G	PHYSICS
H	ELECTRICITY

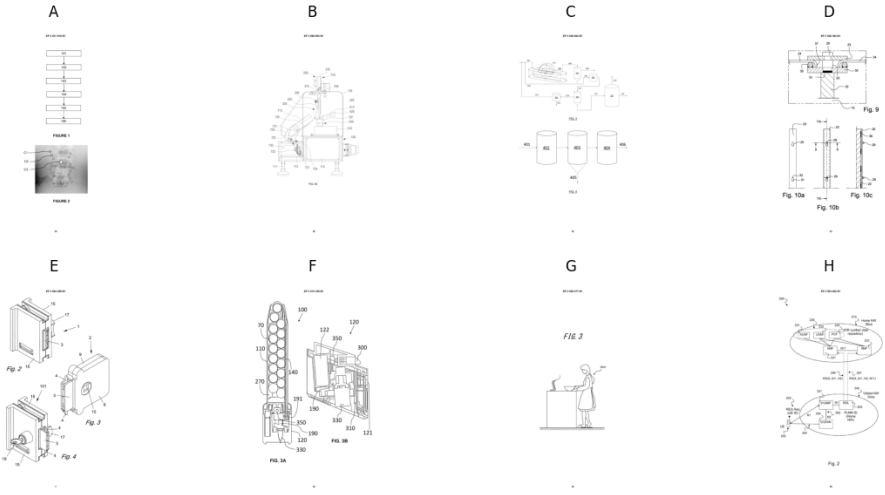

Figure 2: The samples of images from each IPC section

2 Description

The dataset is divided into training and test sets, with relevant statistics detailed in Table 2. Table 3 illustrates the distribution across IPC sections. To ensure the integrity of the test results, patents included in the training set do not overlap with those in the test set.

Each image in the dataset is accompanied by both short and long descriptions. Given the dataset's varied lengths of descriptions, we used binning optimization to find the best ranges for the distribution of token counts. We applied the Freedman-Diaconis rule to determine the optimal bin width, which is based on the interquartile range (IQR) of the token counts. The IQR helps measure the data's spread. We then calculated the number of bins by dividing the total range of token counts by the bin width and rounding up to the nearest whole number.

For short descriptions, we set limits between 10 and 40 tokens, excluding any outside this range as outliers. Long descriptions were considered those exceeding 50 tokens, with an upper threshold of 500 tokens. Descriptions falling outside these specified limits were similarly removed from the dataset.

In cases where images contain multiple subfigures, their descriptions are concatenated, separated by new lines. Consequently, the token count for a single figure's description may exceed the typical maximum of 40 tokens for short descriptions and 500 tokens for long descriptions, due to the aggregated nature of the text for compound figures.

Table 2: Statistics for train and test sets

Set	Total Images	Compound Figures	Unique Patents
Train	17,386	5,204	7,185
Test	998	473	600

Table 3: CPC Class Distribution in Training and Testing Datasets

CPC Class	Frequency in Train Set	Frequency in Test Set
H	4,711	254
G	4,474	252
B	3,409	219
A	2,559	160
F	1,208	69
C	717	27
E	217	11
D	78	2
Y	13	4

Below we provide a detailed explanation of each column of the metadata.

1. **image_name**: This column contains the file names of images extracted from patent documents. Each image is a visual representation related to the patent's content, such as a diagram or a scheme.
2. **pubNumber**: The publication number of the patent document from which the image is derived. A single patent may contain several images.
3. **title**: The title of the patent document provides a concise summary of the invention or the technological field to which the invention belongs.
4. **figs_norm**: A normalized figure identifier that refers to the figure number or label within the patent document. This helps in correlating the image with its specific reference in the document.
5. **descFig**: A short description of the figure, providing an initial overview or summary of what the figure depicts in the context of the patent.
6. **description**: A long description of the figure, elaborating on the components, functionalities, and interactions depicted. This description is crucial for understanding the technical nuances of the invention.
7. **descFig_token_count**: The number of tokens (i.e., words or meaningful units of text) in the short figure description.
8. **description_token_count**: The number of tokens in the long description of the figure.
9. **draft_class**: This column indicates the type of diagram or illustration, such as a flow diagram, schematic, etc. This classification can help in understanding the nature of the visual representation. These are rough retrieved versions of the figure classes retrieved from the text. Note that for around 3,000 figures, the types/viewpoints are not available.
10. **cpc**: the category of the patent figure (see 2), symbolizing a broad technology sector.

11. **relevant_terms**: A dictionary-like structure listing terms relevant to the figure, with keys being element identifiers in the figure (e.g., part numbers) and values being lists of terms associated with those elements. This is useful for parsing and understanding the key components and concepts depicted in the figure.
12. **min_claims**: Specific claims that span all terms or elements depicted in an image, essentially summarizing the interactions among these elements. When the OCR fails to detect the reference numerals within the image, or the image doesn't include any, it is impossible to retrieve the terms from the description, so we take the first claim by default.
13. **compound**: A boolean indicator (True/False) specifying whether the figure contains subfigures.
14. **references**: A list of element identifiers in the figure, providing a quick reference guide to important components or parts depicted in the image, that were extracted with OCR.
15. **figs_per_image**: The number of subfigures represented in each figure. Some images may contain multiple figures, each depicting different aspects or embodiments of the invention.

3 Collection Process

Qatent's internal Solr sol database contains complete textual patent data from the European Patent Office (EPO) including publication number, title, abstract, claim, IPC (patent classification), inventors, patent family, applicants, id, and complete description. Based on this database, we initiated the image acquisition process by retrieving the publication numbers within the time range from January 1, 2020, to December 31, 2020. Subsequently, using Espacenet esp, the EPO's patent search website, we collected a total of 62,513 patent images corresponding to 15,645 unique patents based on the patent publication numbers, enabling for linking of the images to their respective textual patent data.

4 Data Preprocessing

Text processing: We fetch the full descriptions of patents from a Solr database using a specified query, focusing on publications within a given date range. The descriptions are then analyzed to identify and extract sections with titles in uppercase letters, indicative of significant segments such as detailed or brief figure descriptions, claims, and detailed descriptions. This analysis facilitates the segregation of text into categorized content, aiding the extraction of claims, description, brief description section, and IPC section of the retrieved patent applications for further analysis. We extract the short descriptions from the BRIEF DESCRIPTION OF DRAWINGS section, and the long descriptions from different textual variants of the DETAILED DESCRIPTION OF DRAWINGS section. We also try to extract semantic information (object and viewpoint) from short figure descriptions. The details of the text processing are elaborated in Section 3.2 and 3.3 of Aubakirova et al. [2023]. The result of the text processing step is an organized relational database, which includes metadata at both the document and figure levels for each figure, as well as information on the object and viewpoint for each figure, provided that such details are present in the retrieved short descriptions.

Image processing: The image processing branch identifies compound figures and recognizes figure labels. The process can be summarized in the following steps: (1) Figure label detection involves the use of OCR technology to locate and extract the text and positions of labels; (2) In instances where label text is vertically aligned, the corresponding image is rotated and resized to a standard square format; (3) The identification of compound figures is achieved by assessing the number of figure identifiers extracted; (4) This is followed by linking labels to their respective figures; (5) The final step is aligning metadata with individual figures. For details, refer to Section 3.4 in the main paper.

5 Intended Use

We utilized the PatFig Dataset for the patent image captioning task, with detailed discussions on the experiments and selected results referenced in Aubakirova et al. [2023]. Future work will extend to the Visual Question Answering (VQA) task. This will involve the creation of a visual question-and-answer set, leveraging the metadata information. The structure for this task is exemplified in Table 4.

Table 4: Visual Question Answering Format Representation

Type	Question	Answer
VQA	What is the IPC section of the patent?	The IPC section is A.
Caption	What is the title of the patent?	Trans-Septal Implantable Medical Device.
VQA	Can you list the components in the figure?	The figure illustrates several components, including a 'power source' labeled 24, a 'coupling' mechanism labeled 72, and the 'IMD' (Implantable Medical Device) itself labeled 20, among others.
VQA	How many figures are present in the image?	There is 1 figure present.
Caption	Please, provide a short description for each figure.	FIG3 is depicted as a schematic block diagram of an implantable medical device (IMD), in accordance with the patent disclosure, showcasing its internal and external components.
Caption	Please, provide a long description for each figure.	FIG3 offers a comprehensive illustration of the IMD, intended for deployment within a heart chamber, adjacent to the septum. The detailed view includes various components such as the housing, power source, and adjustment mechanisms, elaborating on their functions and placement.

For the image captioning task, our approach mirrored the structure outlined for the VQA task, focusing specifically on questions related to the short and long descriptions of patent images. This methodology ensures a consistent framework for both tasks, facilitating comparative analysis and the application of findings from image captioning to the development of the VQA task.

PatFig was built automatically using machine-learning and deep-learning methods. Therefore, the data contains noise, which was characterized in the main paper. One can remove the compound figures, or use human evaluation to create a gold subset of the data if one wants to customize the dataset. Yet, our experiments have shown that it was possible to improve the system's performance using noisy data.

The dataset should not be used for generating fake patent figures and documents.

6 License

PatFig is licensed under a Creative Commons Attribution-NonCommercial 2.0 Generic International License described at https://creativecommons.org/licenses/by-nc/2.0/.

References

European Patent Office – Espacenet Patent Search. `https://worldwide.espacenet.com/patent/` Accessed on August 19, 2023.

Apache solr. `https://solr.apache.org`. Accessed on August 19, 2023.

Dana Aubakirova, Kim Gerdes, and Lufei Liu. Patfig: Generating short and long captions for patent figures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2843–2849, 2023.

Russell N Carney and Joel R Levin. Pictorial illustrations still improve students' learning from text. *Educational psychology review*, 14:5–26, 2002.

Debasis Ganguly, Johannes Leveling, and Gareth JF Jones. United we fall, divided we stand: A study of query segmentation and prf for patent prior art search. In *Proceedings of the 4th workshop on Patent information retrieval*, pages 13–18, 2011.

You Zuo, Benoît Sagot, Kim Gerdes, Houda Mouzoun, and Samir Ghamri Doudane. Exploring data-centric strategies for french patent classification: A baseline and comparisons. In *18e Conférence en Recherche d'Information et Applications\\16e Rencontres Jeunes Chercheurs en RI\\30e Conférence sur le Traitement Automatique des Langues Naturelles\\25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 349–365. ATALA, 2023.