Visualising Natural Disaster Image Embeddings

October 2025

By

Rashid Ahmed Pandor

Student number 202302628

Word count: 2499

Contents

1. Project background and purpose

1.1. Introduction

Major and catastrophic events like natural disasters have lots of images and videos uploaded to social media by a lot of people such as civilians, tourists and media companies. For humanitarian and charity organisations including emergency responders, these images and videos can be very useful to spread awareness globally to people. However, when there is so much information and data in terms of velocity and volume, evaluation becomes unreliable and unsupportable which can make important decisions also unreliable. This project aims to use the techniques of Artificial Intelligence to possibly overcome the current techniques, which are very highly unreliable, to make visual data understandable and more reliable to effectively spread awareness.

The title of this project is called “Visualising natural disaster image embeddings” and it will use AI techniques to process large image data from an incident. This work will use a pre-trained Neural Network to use high-dimensional features that are called embeddings.

The results will be an interactive visualisation software with images visualised in a 2D map. This will help users do quick exploratory analysis, identifying thematic patterns and clusters that might be impossible to discover through human inspection and evaluation. The Project Definition Document provides the background, rationale, motivation including the objectives and aims for a successful result.

1.2. Objectives

The aim of this project is finding an efficient and better solution than the one used right now. My solution will contain a visualisation using AI techniques to utilise large sets of high-quality images and plot them in a 2D map. This is also an exploration of how Artificial Intelligence could be used to help analyse and evaluate a crisis. Success will be decided on clear goals that have deadlines.

This first objective is to research and gather appropriate quality images through exploring the internet or finding open-source data to use as datasets. This will probably require a minimum of 5000 disaster-related images by the end of week 4 but the quantity might change depending on the strength of the current hardware.

The second objective is to create an effective data processing pipeline to use for the datasets. The next goal is to develop a facility to pre-process all the images in that dataset to produce a set of embeddings using a pre-trained neural network deep learning model by the end of week 7.

The third objective consists of creating a functional, interactive prototype of a visualisation built using the Streamlit framework. The prototype will contain the core of project’s delivery that presents the 2D UMAP view of the image embeddings and showing useful insights for the user by the end of week 13. We also need to implement a series of secondary goals by exploring clustering algorithms like HDBSCAN to search and assign thematic groups to the data (GeeksForGeeks, 2024).

1.3. Scope

The project is centred around the of software by converting a set of disaster-related images into an interactive, functional visualisation. The scope of the project is acquiring a bunch of quality images that can be used to be implemented in a data pipeline to get embeddings and to apply dimension reduction to achieve a 2D interactive projection. Not just that but also develop an operational prototype that is suitable for data exploration. The final aspect of this project includes an overall analysis and conclusion of the effectiveness of the prototype as well as the comprehensive report.

The project doesn't use real time data from the internet but actually uses existing open sourced datasets of images only. Furthermore, the visualisation does retain the abstract 2D similarity of the images. It also does not contain the locations in the world. At the end, the output is a prototype to be tested and proved and should not be assumed that it is industry-ready global scalable software package.

1.4. Deliverables

The deliverables of this project will also be evidenced by the presentation of the portfolio and its major results. They mark the quality of this research project not just the development phase but also the evaluation and understanding of the activity that were carried out in this project. The significant deliverables are an interactive visualisation program written in a Python based web prototype with streamlit and plotly libraries. The system will showcase a 2D projection of the image data set which allows the user to interact and explore with it on a web software. GitHub is used to store and maintain Source code and documentation including the entire history of the project.

The final report is another significant deliverable that contains a detailed and professional report that has the rationale, planning and delivery, methodology, analysis and evaluation as well as a reflective critique and at the end the entire project will be presented.

1.5. Constraints

This project gives lots of freedom of technical use but there are some significant externally imposed constraints. The strongest constraint here is the final project deadline which is in Trimester 2 which defines the project's overall timescale and progression. The project will need to strictly stay within the academic rules of the honours Stage project module and guidelines as well as the Computer Science AI programme competency that necessitates the application of innovative AI techniques to address a real problem. All of the development will also need to be in the bounds of the computing resources that will be available and will need to observe the data protection (GDPR) and copyright law of any data sets that will be used here (UK Government, 2025). Technology selection is not considered to be an external constraint but a personal choice.

1.6. Assumptions

My project has a series of real assumptions. I suppose that I can identify and gather a good quality of open source set of a minimum thousand images related to disaster that can be licensed with permission to be used for research purposes. We can defend this assumption via preliminary research of literature that appears to hold out the promise of their existence (De Abreu, 2023). We suppose that some images in the data set might be considered unreliable that could affect the outcome. Our project supposes that our requisite open source software libraries will be available over the period of this project and our current hardware will be able to meet the project requirements.

2. Project rationale and operation

2.1. Project benefits

This project will bring considerable value to lots of constituencies stakeholders. The major benefactors will be emergency services, government administration and humanitarian aid communities, the visualisation software in this project will represent a valuable software for improving awareness in a crisis environment. By putting raw imagery into an interactive map, the system will facilitate an automatic analysis which would be easier for humans. This will help people make better and earlier decisions. The second benefit would be the researcher and academic institutions including Computer Science communities, this project will bring a useful case study on using current AI technologies around crisis and emergency informatics and a useful comparison of suitability of various visualisation protocols.

2.2. Project operation

The project workflow will be in agile and iterative approach. Using the waterfall workflow for AI methods with a new and huge dataset would be inappropriate. This project will be organised as a series of weekly working cycles so that flexibility and the ongoing improvement of the data pipeline, pre-processing and visualisation can take place because of current results. This means that problems can be resolved whilst we are working with other parts and the direction of the project can be shifted due to the results from previous analysis and testing.

Managing the project every day will be retained through a few important methods. The master schedule which is laid out in the Gantt chart will be closely monitored against progressing work, weekly supervisor meetings which is used for feedback. GitHub will be used to maintain and document everything. The success will be approached by its flexibility for handling unexpected technical problems and making content advanced towards our goal.

2.3. Options

The roadmap of this project is supported through of continuous assessments of some core technologies to make sure that each part is appropriate to the task. The initial decision as to use a deep learning model for the embedding of images with two competent tools. The first is the multi-modal CLIP model (Radford et al., 2021) and the second is ResNet-50 design (He et al., 2015). The choice will be mad based on quality aesthetic coherency of the final visual clusters supported by quantitative measurements. After, we will compare three dimensionality reduction methods: UMAP, t-SNE and PCA. The assessment will be on the final visual functionality, as well as the performance using a sizeable dataset. Lastly, the visual functionality will be implemented either Streamlit, through its capability for rapid prototyping or using Dash for deeper customisation. This will be dependent on the individual features needed to achieve the project goals. This process ensures that the final decisions is not just workable but consciously and defensibly made through a selection from multiple possibilities.

2.4. Risk analysis and mitigation

An important aspect of project planning is future planning because it prepares for possible issues. Aive risk management is crucial in making sure that the project is progressing and meeting the goals aimed at. I have identified the vital risks which are data, technical and project management which are scored on the likelihood for possible impact. Every risk must have a clear and useful mitigation plan in preparation for possible setbacks. I have made a table below.

Risk ID	**Risk description**	**Likelihood**	**Impact**	**Mitigation plan**
R01	Data – Failure to get appropriate dataset	Medium	High	Immediately search for new datasets.
R02	Technical – AI model produces poor results	Medium	High	Compare two models (CLIP, ResNet) and select based on result.
R03	Hardware – PC performance isn't sufficient for training	Low	Medium	Use Google Colab's GPU/TPU or use PCs in DAIM.
R04	Personal – Coursework's impact project time	Medium	High	Follow the Gantt chart schedule strictly. 1-2 days per week on project.
R05	Users – Disturbed, upsetting Images	High	Medium	Give out warnings to users beforehand.

2.5. Resources required

This project will be developed as self-sufficient using already available resources. I intend to develop this on my computer which has appropriate specs. It is an open source software stack with the majority being written in Python along with Data science libraries like PyTorch, Scikit-learn that doesn't require licensing. Cloud Computing might be needed if current hardware is not sufficient.

2.6. Ethical and legal considerations

Ethical and legal considerations are very important and for our project. A full assessment is to be finished through the individual university checklist, but here we will recognise the main considerations. The main concern is data protection to meet the standards of GDPR (UK Government, 2025). The project will be using open source datasets that doesn't have identifiable data being processed. On the Ethics side, the dataset used needs to have a license that allows academic and research to comply with data possession and copyright. There might be algorithmic bias in pre-trained AI models in this project which I have acknowledged and that itself will be discussed in the final project report when we have the results.

2.7. Commercial considerations

This is a piece of academic work and real world applications have been considered, development cost other than my own time is negligible, this project will be made using open source software including a small fund set aside in case I need access to cloud computing environments. Potential markets for this project would be the government, emergency services and humanitarian organisations to try and make their decisions and response more quickly and effectively. This project could be used as a Software as a Service application for monitoring social media. The unique selling point of this product will be its in-depth semantic analysis of photographic and video data and not just keyword tracking.

3. Project methodology and outcomes

3.1. Initial project plan

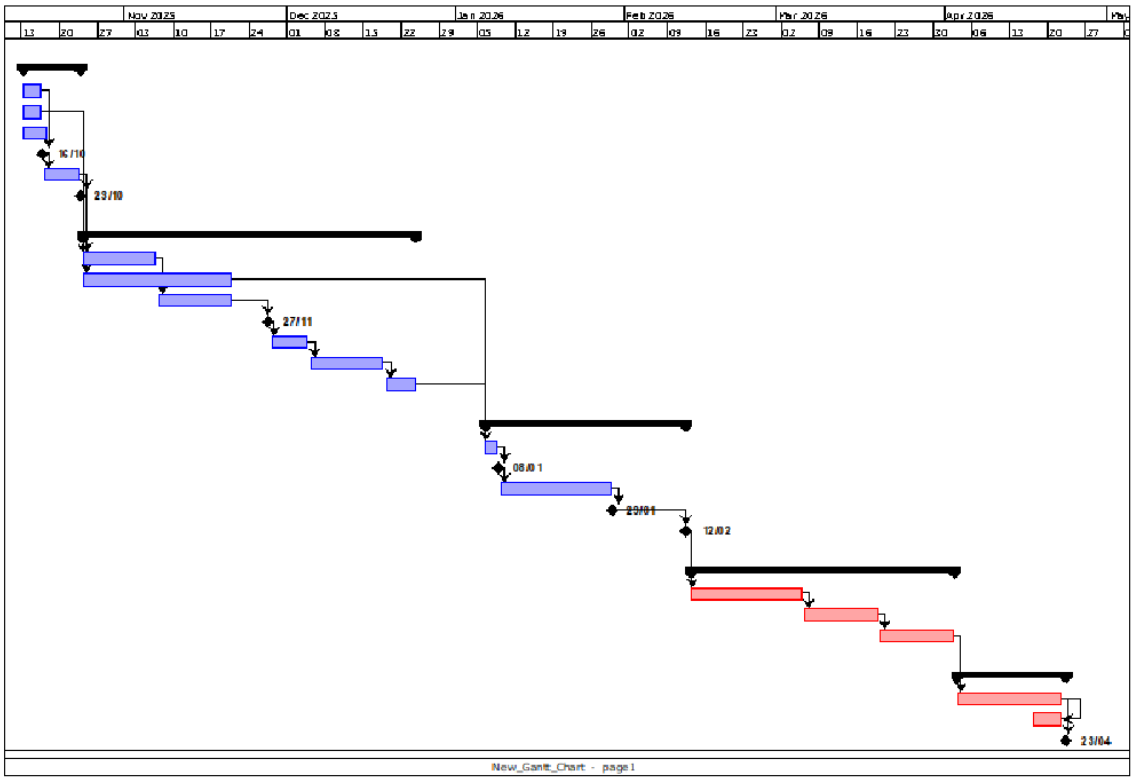
3.1.1. Tasks and milestones

Phase	ID	Task Description	Milestones
Initiation & Planning (Trimester 1, Weeks 1-4)	1.1	First Supervisor meeting & project planning - **Done**	
	1.2	Set up Git Repository and development Environment - **Done**	Find Suitable Datasets
	1.3	Research and get Suitable Image Dataset - **Done**	
	1.4	Complete Project Definition Document - **Done**	PDD Submitted
Research & Development (Trimester 1, Weeks 5-13)	2.1	Start In-depth Literature Review	
	2.2	Clean data and Implement Pre-Processing Pipeline	
	2.3	Evaluate & Choose Embedding Model	Models Evaluated
	2.4	Generate Embeddings for Full Dataset	
	2.5	Compare Reduction Techniques	
	2.6	Develop First Interactive Visualisation Prototype	First version of software
Interim Reporting (January – February)	3.1	Start writing Interim Report	Interim Report Submitted
	3.2	Utilise Supervisor Feedback	
Final Development & Evaluation (Trimester 2, Weeks 3-9)	4.1	Refine Visualisation Application	
	4.2	Implement Secondary Objectives	
	4.3	Conduct Full Evaluation	
Finalisation (Trimester 2, Weeks 10-12)	5.1	Write Final Dissertation Report	
	5.2	Prepare and Conduct Demo and Presentation	Final Submission

3.1.2. Schedule Gantt chart

		Name	Duration	Start	Finish	Predecessors
2		**Phase 1: Project Initiatio...**	**9 days**	**13/10/2025, 08:00**	**23/10/2025, 17:00**	
3		Complete Initial Ethics Check...	4 days	13/10/2025, 08:00	16/10/2025, 17:00	
4		Write and Complete PDD	4 days	13/10/2025, 08:00	16/10/2025, 17:00	
5		Research and Choose Dataset	5 days	13/10/2025, 08:00	17/10/2025, 17:00	
6		Submit PDD and Initial Ethic...	0 days	16/10/2025, 17:00	16/10/2025, 17:00	3;4
7		Set up GitHub Repository an...	5 days	17/10/2025, 08:00	23/10/2025, 17:00	6
8		Student Progress 1 - Milestone	0 days	23/10/2025, 17:00	23/10/2025, 17:00	7
10		**Phase 2: Research and De...**	**44 days**	**24/10/2025, 08:00**	**24/12/2025, 17:00**	
11		Data Cleaning and Pre-Proce...	10 days	24/10/2025, 08:00	06/11/2025, 17:00	4;9
12		Literature Review	20 days	24/10/2025, 08:00	20/11/2025, 17:00	7
13		Evaluate and Choose Embed...	10 days	07/11/2025, 08:00	20/11/2025, 17:00	11
14		Student Progress 2 - Milestone	0 days	27/11/2025, 08:00	27/11/2025, 08:00	13
15		Generate Embedding Models ...	5 days	28/11/2025, 08:00	04/12/2025, 17:00	14
16		Compare Dimensionality Re...	10 days	05/12/2025, 08:00	18/12/2025, 17:00	15
17		Develop Visualisation Prototy...	4 days	19/12/2025, 08:00	24/12/2025, 17:00	16
19		**Phase 3: Interim Reportin...**	**27 days**	**06/01/2026, 08:00**	**12/02/2026, 08:00**	
20		Write Interim Report	3 days	06/01/2026, 08:00	08/01/2026, 17:00	12;17
21		Interim Report Submitted	0 days	08/01/2026, 17:00	08/01/2026, 17:00	20
22		Prepare for Midway Review ...	15 days	09/01/2026, 08:00	29/01/2026, 17:00	21
23		Midway Review Meeting - M...	0 days	29/01/2026, 17:00	29/01/2026, 17:00	22
24		Student Progress 3 - Milestone	0 days	12/02/2026, 08:00	12/02/2026, 08:00	23
26		**Phase 4: Final Developme...**	**35 days**	**13/02/2026, 08:00**	**02/04/2026, 17:00**	
27		Finalise Visualisation Software	15 days	13/02/2026, 08:00	05/03/2026, 17:00	24
28		Add Secondary Objectives	10 days	06/03/2026, 08:00	19/03/2026, 17:00	27
29		Conduct Quantitative and Q...	10 days	20/03/2026, 08:00	02/04/2026, 17:00	28
31		**Phase 5: Finalisation and ...**	**14 days**	**03/04/2026, 08:00**	**23/04/2026, 08:00**	
32		Write Final Dissertation Report	14 days	03/04/2026, 08:00	22/04/2026, 17:00	29
33		Create Video and Finalise So...	4 days	17/04/2026, 08:00	22/04/2026, 17:00	32FF
34		Final Submission - Milestone	0 days	23/04/2026, 08:00	23/04/2026, 08:00	32;33

New_Gantt_Chart - page1



3.2. Project control

The project will be controlled and directed via a series of efficient practices. Progress tracking will primarily be achieved via the Gantt chart, which will be formally checked beginning of every week to make sure that the project stays on track. Supervisor meetings each week will be important for communicating progress, overcoming stalemates. To maintain the progress of my work, a computer logbook will be kept to record important decisions and results of experiments. Source code will be stored and maintained via GitHub and will be committed to the repository often which will show the full history of the project's evolution.

3.3. Project evaluation

To ensure that I'll have a means of proving the project successful, I have created a two part assessment approach. It will be an actual documentation that the project was successful and the end product works. I will make a quantitative assessment to establish hard metrics. I will be testing the major technical parts of the system. For example, I'll know how long it takes to create embeddings for the images and how fast each of the dimension-reduction algorithm runs. This will provide me with a clear idea of efficiency and it's a primary challenge identified in the brief. If I succeed in the secondary project tasks of clustering, I will use a measure such as Silhouette score to arrive at a quantification of how good clusters are.

The second and more important aspect will be a qualitative evaluation in which ill be utilising my own judgement to interpret my findings. My primary question answers is: "Do the clusters work? Do they cluster images together to form real world patterns such as 'flooded roads'"? This also includes comparing visual outputs from UMAP, PCA and t-SNE. This comparison will allow me to evaluate reasonable determination to which method was best suited to this task and dataset.

We will also have users testing this project at the end for feedback and improvements which will be documented.

All this combined will provide a detailed and robust measure of the success of this project.

4. References

1. GeeksforGeeks (2024). *UMAP: Uniform Manifold Approximation and Projection*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/umap-uniform-manifold-approximation-and-projection/.

2. GeeksforGeeks (2024). *Hierarchical DensityBased Spatial Clustering of Applications with Noise (HDBSCAN)*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/hdbscan/. (Accessed: 7 October 2025)

3. UK Government (2025). *Data Protection Act*. [online] Gov.uk. Available at: https://www.gov.uk/data-protection. (Accessed : 12 October 2025)

4. de Abreu, G. (2023). *Preliminary Literature Review: A Guide for Effective Research*. [online] Mind the Graph Blog. Available at: https://mindthegraph.com/blog/preliminary-literature-review/. (Accessed: 12 Oct. 2025).

5. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020 [cs]*. [online] Available at: https://arxiv.org/abs/2103.00020. (Accessed: 15 October 2025)

6. He, K., Zhang, X., Ren, S. and Sun, J. (2015). *Deep Residual Learning for Image Recognition*. [online] *arxiv.org*. Available at: https://arxiv.org/pdf/1512.03385 [Accessed 15 Oct. 2025].

5. Appendix a