Graph-based Analysis of Hierarchical Embedding Generated by Deep Neural Network

Korlan Rysbayeva, Romain Giot, Nicholas Journet

Graph-based Analysis of Hierarchical Embedding Generated by Deep Neural Network*

Korlan Rysbayeva[0000−0001−9798−3389], Romain Giot[0000−0002−0638−7504], and Nicholas Journet[0000−0002−6773−4071]

Univ. Bordeaux, Bordeaux INP, CNRS, LaBRI, UMR5800, F-33400 Talence, France
{korlan.rysbayeva,romain.giot,nicholas.journet}@u-bordeaux.fr

Abstract. In a previous work, we have developed a framework for the multimodal and hierarchical classification of images from soil remediation reports. We extended this work using Deep Metric Learning (DML) as an additional training step to improve embeddings quality and obtained 84.24% of weighted F1 score for the level 5th hierarchical level. However, the standard classifier performance metrics are insufficient to explain the decision process reasoning. So far of our knowledge, there are no methods to analyze hierarchical classification algorithms. In this work, we propose a method of graph analysis to describe the embeddings that represent the extended classifier, which we believe properly interprets the obtained results than classification metrics. We illustrate the method of analyzing hierarchical classification algorithms on private dataset, but the method remains generic enough to be used in other contexts.

Keywords: Graph analysis · Hierarchical embeddings · eXplainable Artificial Intelligence.

1 Introduction

Machine Learning (ML) is used in various applications [5] with an emphasis on Deep Learning (DL) [6] since a decade ago. Their models are error-prone due to various factors such as willingness to generalize, lack of expressiveness of the model, inappropriate training dataset. For these reasons, their architects, or users, dispose of various evaluation metrics [20] to assess the quality of the models. However, such metrics are only able to express its quality (*i.e.*, if it tends to do few or lots of errors); they are unable to explain errors and successes.

This is where eXplainable Artificial Intelligence (XAI) is relevant [1]. These techniques go beyond the standard evaluation by trying to explain these errors and successes. Several methods have been proposed by the ML [5] and the Visual Analytics (VA) [13] communities in parallel. We can classify them with methods doing a *single sample analysis* (mainly from the ML community) or a *database analysis* (mainly from the VA community). *Single sample analysis* regroups methods that compute: *features attribution* of the input sample for white

* This work is supported by Abai-Verne scholarship and Innovasol Consortium.

boxes [25,19,2] or black boxes [23], *factual* and *counterfactual examples* [16,15,14]. *Database analysis* regroup methods that compute *learned features* by a DL model [21] or methods that try to extract some behavior of the model [11,18]. Despite the studies in other areas, the majority of XAI works target image or text classification.

We have developed a multi-modal and hierarchical classifier [24], which is able to classify images in documents from soil depollution reports. Each image is (a) described by its raw pixels, optional text caption in the document and the OCR extracted text, and (b) classified along 5 hierarchy levels defined by soil remediation experts. We extend this work using Deep Metric Learning (DML) [12] as additional training step to improve the quality of embeddings and the overall recognition performance. The classifier performance is acceptable in regard to our application, but we lack the understanding of its behavior. In this paper, we are interested in analyzing this model within its usage context to understand if it behaves properly, or if its predictions are not consistent with the dataset. To do so, we will follow a *database analysis* to build and analyze a graph of embeddings, as well as a *single sample analysis* to collect counterfactuals from it.

So far of our knowledge, this is the first paper to describe the embedding relations of a multilevel classifier. Several papers in the literature stick to the use of umap or tsne projections of their embedding [22], whereas we use a graph-based approach.

2 Context

The classifier [24] we have developed aims to get the embeddings from data specified by multiple modalities and hierarchical structure. As mentioned earlier, it was recently extended with DML to improve the quality of its embeddings. Thereby, our model is completed in training in two successive steps, (i) with a multi-modal hierarchical classification system, where the last vector of embedding layers is extracted and presented to (ii) a deep metric learning system. Figure 1 presents the framework of the training process, where the *classification network* contains one branch for each modality (image, caption and embedded text) working in parallel. All three branches compose of feature extraction Ⓕ, embedding Ⓔ and classification Ⓒ layers. The feature extraction Ⓕ is specific for each modality. In embedding Ⓔ layer, the information from one hierarchy level is transferred to another in top-down manner by concatenating the activations of last embedding layers of one hierarchy level with feature representation of next level. Moreover, the embeddings of different modalities (image, caption, embedded text) are concatenated creating multi-modal embeddings for each hierarchy level and present to *DML network* for further training. For each hierarchy level, the final prediction of classification network is calculated by fusion of softmax tensors coming from three modalities by weighted averaging technique.

The described framework was tested on a real world private dataset. We processed 35 reports and automatically extracted 700 valid images with corresponding caption. Additionally, we manually extracted 500 images without a

caption. All images have been processed with Tesseract OCR engine [26] to obtain their embedded text. Images dimension range from 100×100 to 2000×2000 pixels and are resized to 256×256 pixels. The average caption length is 44 words, embedded text length is 100-300 words. Any sample in the dataset is assigned with one class of each level along its hierarchical path. The hierarchical classification with five levels is depicted in Figure 2. Level 1 labels correspond to *Cross section*, *Maps*, *Graphs and tables* and *Photos*. Node size is proportional to the number of samples per hierarchy level. The number of samples goes from 505 for class 0 of level 1 to 6 for class 8 of level 4.

Due to the low amount of data, prediction experiments have been executed using a stratified k-fold mechanism [7]. It means the dataset has been split to six subsets sharing the same ratio of samples per class than in the complete dataset. Five subset are used to train a model that serve for the inference with the 6th one. The classification results are computed globally by fusing 6 folds results. We report 84.24% of weighted F1 score for the 5^{th} hierarchical level. F1 score with weighted averaging is the output average accounted for the contribution of each class as weighted by the number of examples of that given class. However, the results we received were hard to interpret and understand. Consequently, we have analysed the model by describing the embeddings relation of different hierarchical levels by building a K-graph, and generating the similar and counterfactual examples. For that we have used part of the embeddings that was extracted at point ② shown in the Figure 1, precisely the embeddings coming from each separate levels.

3 Proposed Analysis

This section presents the data structure and main visual encoding we have chosen to use to represent the result of our model and the questions we want to answer.

3.1 Data structure

The database S contains around $1,2K$ samples described by three modalities, labeled on five levels. We are interested in the multi-modal embedding generated at each level: each sample $s_i \in S$ is described by five embeddings $\{emb_i^l, 1 <= l <= 5\}$ of size 512 for each level l and annotated by its groundtruth gt_i^l and prediction $pred_i^l$ that are level dependent.

For each level l, we build a proximity graph (also called k-graph) $G_l = (S, E_l)$ that encodes the nodes proximity related to the multi-modal embeddings of level l. Each node i represents a sample $s_i \in S$. There is an edge $(s_s, s_t) \in E_l$ between s_s and s_t if emb_t^l is among the k closest samples of emb_s^l in terms of Euclidean distance. Such mathematical object represents well the proximity between the objects.

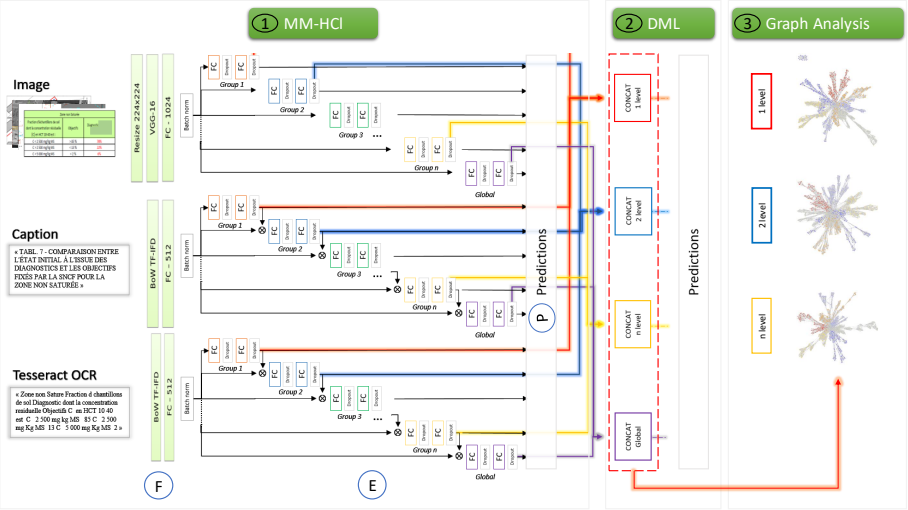


Fig. 1: Architecture of the analysed model. Training process is accomplishes in steps ① that corresponds to multi-modal hierarchical classification (MM-HCl) system following DML approach (step ②).For each hierarchical level the multi-modal hierarchical embeddings is extracted at step ② which is analyzed using K-graphs in step ③.

3.2 Visual Encoding

We want to depict graphs on screen to manually extract patterns and information. For this reason, we use the FM^3 [10] algorithm to compute the layout of each node on screen, followed by Fast Overlap Removal [9] to ensure there is no node-node overlap. The edges are colored in gray with alpha-transparency to reduce the visual clutter and the nodes can be colored according to the expected information (the groundtruth, the prediction, the fact it is an error, or any other metric). Our experiments have shown that such way of visualizing samples is of higher quality than the standard 2d projection with PCA, UMAP or T-SNE [17] depicted in a scatter plot.

The image representation of samples is not depicted on screen because it would take too much space. However, the graphs are visualized using the interactive tool Tulip [3] that allows to interactively obtain it (*i.e.* by hovering a node).

3.3 Questions of interest

Several questions arise and are treated in independent evaluations.

Does the embedding is consistent over the training folds ?Since prediction experiments have been executed using a stratified k-fold (6 folds) mechanism, we would like to verify if the embedding space is consistent among

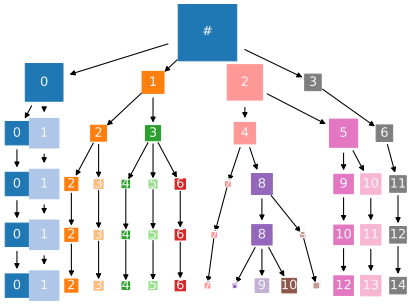


Fig. 2: Hierarchical classification of the data. 5 levels of classification are expected. Each level refines the concepts of the previous ones for a subset of the classes. Node size illustrates the balanced issue of the dataset.

the folds. If so, we can assume the embedding space is stable and the full dataset can be analyzed as a whole for the rest of the paper. Otherwise, it implies that the embedding is not stable other than on the trainings because it is sensitive to the training samples and each fold should be analyzed separately.

To verify this stability, the proximity graph is built using all samples, regardless of the fold they belong to. We then compare two representations of the drawn graph by coloring the nodes with their fold number and their groundtruth, and manually analyze visual patterns.

Does the embedding align with knowledge from the data ? The visualization of the k-graphs for each fold on 5 levels allows to quickly grasp the proximity of samples. We expect samples of the same class to be densely connected and depicted close together.

How the embeddings of one level perform on the other levels ? As we are addressing a hierarchical classification problem, we could expect the embeddings to follow a hierarchical pattern. To analyze them, we train some Random Forest classifiers to predict the classes of a target hierarchical level from the embeddings of a source level. To target the imbalanced data, for RF we have weighted classes such that rarely observed groups/classifications are more likely to be selected in bootstrap samples. We evaluate the classification using the weighted F1 score. We expect to obtain the best performance when the embedding and label levels are the same.

Does the embedding of successive levels makes sense ? We should observe patterns of interest. Another experiment consists of analysis of neighborhoods in the K-graph of the nodes over the layers. We expect to have similar neighbours of the same node at two successive layers. This can be verified by computing the Intersection Over the Union of the neighbors. It is a number from 0 to 1 that specifies the amount of overlap between the neighbors of one node at the current level and the next level. The more the network learns to represent the data the more it is consistent, and it relies on knowledge acquired on previous

levels, thereby we expect large distribution at higher IOU values than on deeper hierarchy levels.

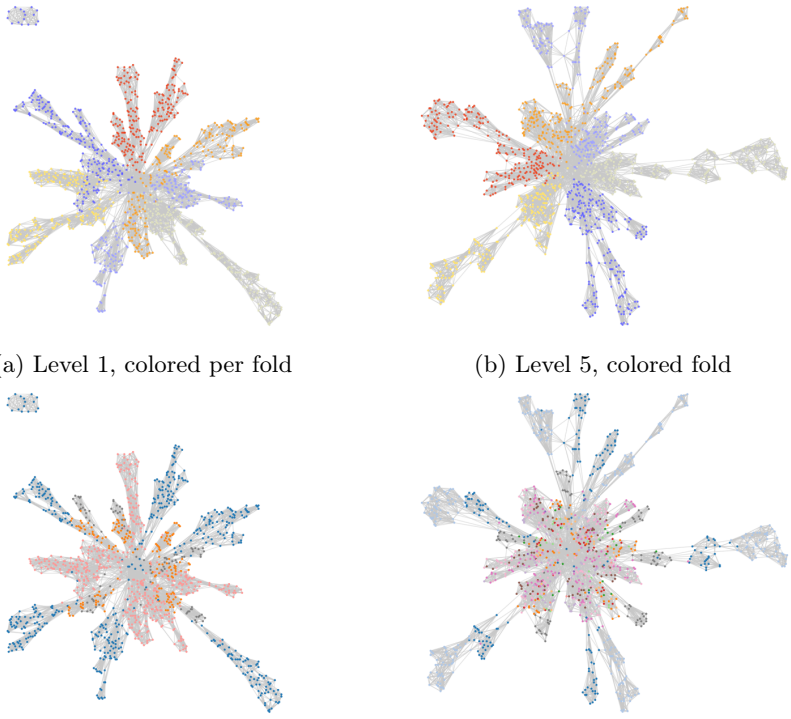
Does the embeddings influenced by modalities over different levels ? To be able to detect the similarity between samples, we propose to use Breadth First Search [8]. By selecting a node in the graph, the user can obtain the closest node by distance that is labelled with the same class for exemplars and with different class for counterfactuals. This is possible because we can assume that while three samples s_a, s_b, s_c follow graph distances constraints $distance_{G_l}(s_a, s_b) < distance_{G_l}(s_a, s_c)$, the embedding's Euclidean distances follow the constraints $norm\left(emb^l_{s_a} - emb^l_{s_b}\right) < norm\left(emb^l_{s_a} - emb^l_{s_b}\right)$. Thus, from the requested node, the exemplars and counterfactuals can be found by browsing the graph using a Breadth First Search. Similar examples (resp. counterfactuals) are collected by keeping only nodes belonging to the same (resp. to a different) prediction than the input; the search stop after collecting the appropriate number of samples. Eventually, the 1-top neighbors can be visited by ascending distance to input order. Since level 1 labels are visually distinguished, we expect to have similar and counterfactual examples, which are visually close. We also expect the embeddings contain more information from caption and embedded text for deeper hierarchy levels.

4 Results

The section presents the analysis results based on the evaluations presented in Section 3.3.

Does the embedding is consistent over the training folds ? Figure 3 presents the k-graph colored per fold and groundtruth for the first and last levels of the hierarchy. The color range in Figures 3a and 3b corresponds to the number of folds. For each fold from Figures 3a and 3b we can see the corresponding nodes in Figures 3c and 3d accordingly, where colors are defined by the groundtruth. For example, the nodes colored in yellow from Figure 3a contains 204 classified images (nodes), which corresponds to the same nodes at the same location in Figure 3c. Since we have four classes in level 1, we can see from Figure 3c that these 204 images are correctly projected among an equal number of classes. The same pattern applies for other folds (colors) of Figure 3a.

The representation of other levels follows the same trends. The patterns from the fold-colored graphs prove that the embedding is fold-sensitive: indeed, we clearly identify clusters of folds; it means that samples of the same fold are closer to each other than the samples from different folds. The comparison between patterns from gt-colored graphs and fold-colored graphs shows that the embedding is also able to recognize samples of the same class as soon as they come from the same fold. This is not an issue for an operational system, since only a single model (and thus family of embeddings) would be used, but the remaining analysis must then done on a fold basis rather than using the full dataset as a whole.



(a) Level 1, colored per fold

(b) Level 5, colored fold

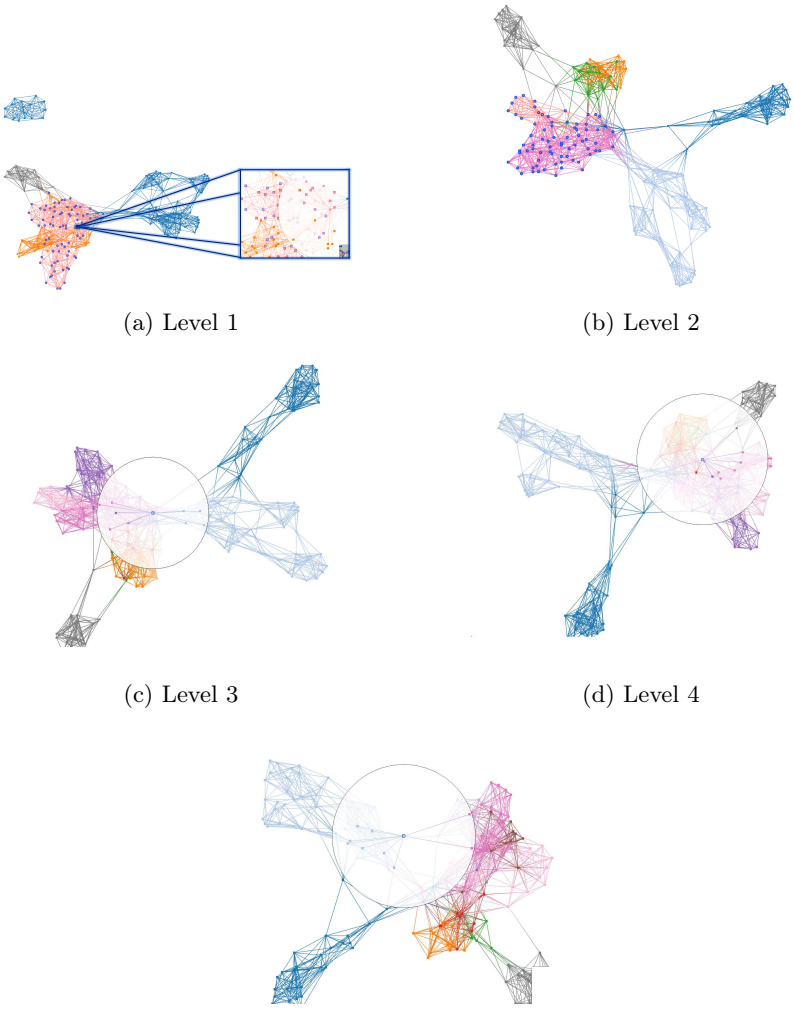
(c) Level 1, colored per groundtruth

(d) Level 5, colored per groundtruth

Fig. 3: Comparison of fold (a,c) and groundtruth (b,d) assignment on the full dataset for the first and last level of the hierarchical classification on the K-graph representation.

Does the embedding align with knowledge from the data ? Figure 4 illustrates the k-graph of the first fold with nodes colored by their groundtruth. For *Level* 1, according to the placement of samples, we can clearly see that embeddings of nodes are generated such that they have high intra-class similarity. However, the *Maps* (orange) and *Graph and tables* (pink) class embeddings are very close to each other and has common close neighbors (Figure 4a). In Figure 4b, we identify that selected samples of mentioned classes are clearly distance away from each other by embeddings on *Level 2*. Moreover, Figure 4c shows the example of a sample that has been wrongly classified on *Level 3*, but the embeddings of this sample progressively gets better in deeper hierarchy levels and still being wrongly predicted recognize by embedding towards correct class. However, it is true the other way around.

How the embeddings of one level perform on the other levels ? The random forest has been individually tested on each fold of the dataset with respect to results on the embedding consistency. Figure 5 presents the global (*i.e.*



(a) Level 1

(b) Level 2

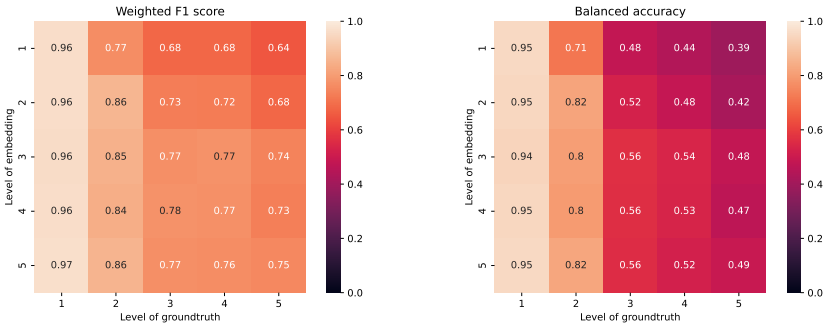
(c) Level 3

(d) Level 4

(e) Level 5

Fig. 4: K-graph of the embeddings from five levels of fold 1. The other folds have similar cluster behavior. The samples are colored according to their predicted labels on each level. In Figures (a) and (b), the selected samples (nodes with blue borders) are true (groundtruth) labels from class *Graphs and tables*. Figures (c-e) show the chosen example of a sample embeddings on one level that close to embeddings of other class samples in *Level 3*, but progressively gets better in deeper hierarchy levels.

computed on the whole result rather than on per fold aggregated result) weighted f1 score and the balanced accuracy for all the combinations of embedding/label



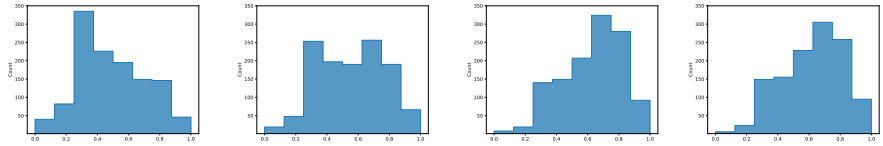
(a) Weighted F1 score (b) Balanced Accuracy

Fig. 5: Weighted F1-score of a Random Forest trained to recognize the label of level x when using the embedding of level y.

combinations. By looking at the weighted f1 score, we can make the following observations. Some semantics (even if it is still an open problem to properly identify them) have been properly extracted by the network. To predict the labels of a given level, it is better to use the embeddings of this specific level or the embedding of a deeper level (read the matrix per column). Indeed, some levels (1 and 3) have slightly better performance with the deepest embeddings; the network has not specialized enough properly generate these embeddings. By looking at the balanced accuracy, we can make the following observations. The problem gets more severe on deepest levels (look at the diagonal) until quickly reaching a point where several classes are not properly classified (level 3). There is a strong effect of the unbalanced dataset.

Does the embedding of successive levels makes sense ? According to Figure 5 the embeddings of a level systematically perform worst (or equal) on the next level than their true level (read the lines for left to right starting by the groundtruth of the same level): the network has learned the appropriate level of details to make the classification for this specific level. The embeddings of a level systematically perform better on the previous levels than their true level (read the lines for right to left starting by the groundtruth of the same level): the information provided by the next levels are consistent with the previous ones.

Figure 6 illustrates this aspect at a sample level by depicting the IOU of neighbors for the whole dataset. The graphs are shown for the first four levels, since IOU values are calculated among two successive levels. The distribution of IOU values calculated from 0 and 1 for all samples.By looking at the graphs, we see that the distribution of IOU increases to the right. For example, in Figure 6a, we see that around 350 samples have around $IOU = 0.4$, which means Level 1 and Level 2 samples have less common neighbors compared to Figure 6c where the largest distribution at $IOU = 0.8$. Since we have more classes at the Level 5 and the largest distribution at higher IOU among the same number of samples, we can they that the embeddings of Level 5 is consistent according to Figure 6d.



(a) Level 1 (b) Level 2 (c) Level 3 (d) Level 4

Fig. 6: IOU distribution of the first 4 levels. Taking into account the number of labels in the deepest hierarchical level, figure (d) shows that level 5 embeddings is consistent and acquire knowledge from embeddings of level 4 according to IOU value.

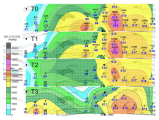
Does the embeddings influenced by modalities over different levels ? Tables 1,and 2 provide the results of breadth-first-search to extract examples of Similar and Counterfactual for different levels for the same sample on each table. We have selected samples from *Maps*, and *Graphs and tables* classes as targeted samples, since they are mostly mis-classified. Each row corresponds to examples for the same node but different hierarchy levels (1 and 5 for maps, 1, 3, 5 for graphs and tables).

Table 1 targets the examples from *Maps* class. For Level 1 examples, since we have only four classes on this level, the examples should be easily distinguished visually. However, we see clear difference in counterfactual examples, meaning that this photo has close embeddings with targeted map. Moreover, if we consider the caption and OCR information, the targeted example shows the geo-location of the field, whereas Similar example for Level 1 illustrates the concentration levels of pollutant on the specific area, and for level 5 the illustration of the geo-location.Thereby, we can conclude that the embeddings of level 5 is more defined that the embeddings of Level 1.

Table 2 targets the sample taken from *Graphs and tables* class that show the evolution of total injected volume in wells. Taking into account caption and embedded text of the provided samples, the Similar example for Level 1 shows the evolution of groundwater level (height of water) which is again easily distinguished visually, that it belongs to *Graph and tables* class. On the other hand, the Counterfactual example shows the map of geo-location of treated zone. For level 3, the Similar example illustrates the evolution of pollution, which is close to the evolution of injected water by the meaning observed from caption and embedded text, whereas the map of water level was selected as Counterfactual example. For level 5, the evolution of water level but given in table for Similar example check the meaning of thie sentence, whereas Counterfactual illustrates the contamination level, which is very close to the targeted sample label, but it is illustrated as a map.

The overall conclusion from the table is that for *graphs and tables* the tendency is the same, in lower hierarchy levels the embeddings contain more visual information, but in the deeper hierarchy level the embeddings have more information

Table 1: Exemplars of similar and counterfactual search for one node in level 1 and level 5 of *maps* class. Each sample is depicted by a thumbnail of its visual representation, an extract of its caption, and an extract of the text embedded in the image. *Some parts of the data are hidden under black box for confidentiality issues.

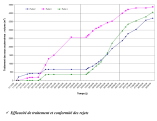
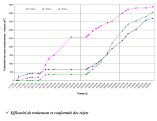
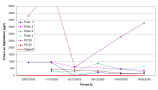
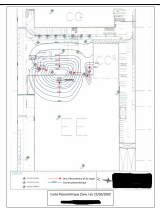
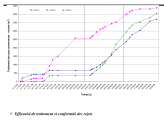
Example	**Similar**	**Counterfactual**
		
Figure 1 : Localisation du site (extrait de la carte [redacted], source : [redacted])	FIGURE 9- REPARTITION SPATIALE DE LA POLLUTION SUR LA PROFONDEUR 0-2M ENTRE T0, T+1 AN, T+2 ANS ET T+3 ANS	Figure 20 - Installation au droit du puits P4 avec une Spill buster
A RTS SU Se Zone d tude 48		
		
Figure 1 : Localisation du site (extrait de la carte [redacted], source : [redacted])		
A RTS SU Se Zone d tude 48		Legende Aire Potentiellement Contaminee Transformateurs jusqu en 1995 Transformateur apres 1995 Atelier thermique em APC9 necessation en 2002 Centrale de filtration des huiles us es Reseau des caniveaux de collecte enterres [redacted]

taken from captions and embedded text. The Counterfactual examples could be easily distinguished visually for lower levels, but the deeper the hierarchy level is the more similarity is observed regarding captions and embedded text.

5 Discussion

Studying the results with the k-graphs which are built using embeddings obtained by our classifier makes it possible to draw the following conclusions. It would be interesting to have a model with an improved stability other the training folds. However it is not straightforward to define a strategy to achieve it by keeping the same training data distribution. One solution could rely on the definition of dedicated losses.

Table 2: Exemplars of similar and counterfactual search for one node in level 1, 3, 5 of *graphs and table* class. Each sample is depicted by a thumbnail of its visual representation, an extract of its caption and an extract of the text embedded in the image. *Some parts of the data are hidden under black box for confidentiality issues.

Example	**Similar**	**Counterfactual**
		 
	Figure 7 - Evolution du niveau de la [redacted] entre janvier et juin 2017	FIGURE 1 : PLAN DU SITE ET LOCALISATION DES ZONES A TRAITER
Traitement des eaux souterraines volume m 4500 Puits 2 Puits 3 4 Puits 4 4000 3500 3000 2500 2000 1500 1000 500 SE ON ND KP EEE EE PE EP PEU PP EEE EE EE EE EE EE EE EE EE EE EE EE EE EE EE EE EE N D P	1 Ni le un FT M2 Sdual na3nex 4702 90 62 4702 90 67 4T02 90 T 4702 90 40 4702 90 70 4702 60 97 4702 60 07 LTOZ GO DT 4702 60 20 2T02 G0 T0 4TOZ p0 Gz LTOZ P0 6T LTOZ PO ET 4TOZ E0 40 TOZ EO TO TOZ ZO EZ TOZ ZO ET ATOZ 20 TT LTOZ 20 b0 LTOZ TO 62	ARTE nes e CMICUI LR LEE m en FE mi 22 Ne T
		
Traitement des eaux souterraines volume m 4500 Puits 2 Puits 3 4 Puits 4 4000 3500 3000 2500 2000 1500 1000 500 SE ON ND KP EEE EE PE EP PEU PP EEE EE EE EE EE EE EE EE EE EE EE EE EE EE EE EE EE N D P	Teneur en Naphtalene jg l 9000 4500 4000 3500 3000 2500 2000 1500 1000 500 Puits 1 s Puits 2 4 Puits 3 x Puits 4 x PZ 29 e PZ 35 Objectif 29 07 2005 11 10 2005 02 12 2005 29 12 2005 Temps ji 14 04 2006 16 05 2006	IT eue utmmares est solgo Carte Pi zom trique Zone du 22 06 2009 Site [redacted]
		
	l dans les eaux	
Traitement des eaux souterraines volume m 4500 Puits 2 Puits 3 4 Puits 4 4000 3500 3000 2500 2000 1500 1000 500 SE ON ND KP EEE EE PE EP PEU PP EEE EE EE EE EE EE EE EE EE EE EE EE EE EE EE EE EE N D P	1 000 0 7 2 000 0 5 0 5 0 5 1 0 02 0 02 JL S N 200 0 PE 200 7 0 5 0 02 0 5 0 5 0 02 0 5 0 5 0 5 0 05 0 5 0 5 1 0 0 03 0 5 0 5 0 58 0 02 0 02 0 02 0 02 0 02 0 02 0 04 0 02 0 02 0 12 0 12 0 05 0 12 0 12 0 EE S Ha O LQ Limite d	[redacted] Site de [redacted] 44 R habilitation in situ des sols Bilan de la phase pilote D de Legende EE TT 4 P rom re pres 3 nus 4 Sgncatve dun impec ao ga ump PA HO m reat 9 ns RE ou fi

The network slightly failed to provided level-specific embedding for level 1 and level 2 as using deepest embeddings allow to obtain better results. To overcome this issue, we should add a component to the loss that take into account this specialization. The unbalanced dataset effect should be overcame in the future by using dedicating methods [4].

A mixture of Sankey diagram and parallel coordinates would help to better understand the hierarchal treatment of the input samples by the network. The axes would be the depth of the hierarchy, the flows would be the samples clustered by prediction at each level.

The distribution of IOU showed the consistency of embeddings in deepest level taking into account the number of classes in the last hierarchy level, however, by the examplars and counterfactuals sometimes we observed that level 4 embeddings more truthworthy than the last hierarchy level for the prediction as well as defining semantic similarities. Thereby, adding one more modality from the text around the images could improve the network performance on the deepest hierarchical level. The other solution could be to re-define the level 4 and level 5 labels.

6 Conclusion

The training of multi-modal and hierarchical classifier [24] for images from soil remediation reports were extended by DML to improve the embeddings quality. The obtained results using classification network was acceptable, but we faced the problem of interpreting them by the classification metrics. Thereby, in this work we are interested in analyzing this model within its usage context to understand if it behaves properly or if its predictions are not consistent with our knowledge of the database. To do so, we describe the embedding relations of a multilevel classifier by database analysis to build and analyze a graph of embeddings as well as single sample analysis to collect counterfactuals from it.

First, we observed that the embeddings are fold-sensitive and not consistent among six stratified folds. The generated embeddings followed a hierarchical pattern, thereby it is better to use the embeddings of this specific level or the embedding of a deeper level. Moreover, we showed that the more the network learns to represent the data the more it is consistent, and it relies on the knowledge acquired on previous levels. Finally, we expect to see the impact of visual information on initial levels and more on semantic information by the influence of caption and embedded text. By generating exemplars and counterfactuals, we saw the pattern for Similar examples in which at lower hierarchy levels the embeddings contain more visual information, but the deeper in the hierarchy level the embeddings have more information taken from captions and embedded text.

For the future work, it is worth to study the consistency between the levels using the global embeddings, which we have not used in this work. These embeddings consider all classes in hierarchical tree at once and processes the hierarchy dependency in loss function. This global embeddings correspond to the last (violet) line of each modality in Figure 1.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access **6**, 52138–52160 (2018)

2. Ahmed Asif Fuad, K., Martin, P.E., Giot, R., Bourqui, R., Benois-Pineau, J., Zemmari, A.: Features understanding in 3d cnns for actions recognition in video. In: 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6 (2020)
3. Auber, D., Archambault, D., Bourqui, R., Delest, M., Dubois, J., Lambert, A., Mary, P., Mathiaut, M., Melançon, G., Pinaud, B., Renoust, B., Vallet, J.: Tulip 5, pp. 3185–3212. Springer New York, New York, NY (2018)
4. Batista, G.E., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter **6**(1), 20–29 (2004)
5. Dargan, S., Kumar, M., Ayyagari, M.R., Kumar, G.: A survey of deep learning and its applications: a new paradigm to machine learning. Archives of Computational Methods in Engineering **27**(4), 1071–1092 (2020)
6. Deng, L., Yu, D.: Deep learning: methods and applications. Foundations and trends in signal processing **7**(3–4), 197–387 (2014)
7. Diamantidis, N., Karlis, D., Giakoumakis, E.A.: Unsupervised stratification of cross-validation for accuracy estimation. Artificial Intelligence **116**(1-2), 1–16 (2000)
8. Dietterich, T.G., Michalski, R.S.: 3 - a comparative review of selected methods for learning from examples. In: Michalski, R.S., Carbonell, J.G., Mitchell, T.M. (eds.) Machine Learning, pp. 41–81. Morgan Kaufmann, San Francisco (CA) (1983)
9. Dwyer, T., Marriott, K., Stuckey, P.J.: Fast node overlap removal. In: International Symposium on Graph Drawing. pp. 153–164. Springer (2005)
10. Hachul, S., Jünger, M.: Drawing large graphs with a potential-field-based multilevel algorithm. In: International Symposium on Graph Drawing. pp. 285–295. Springer (2004)
11. Halnaut, A., Giot, R., Bourqui, R., Auber, D.: Deep dive into deep neural networks with flows. In: Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020): IVAPP. vol. 3, pp. 231–239 (2020)
12. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International workshop on similarity-based pattern recognition. pp. 84–92. Springer (2015)
13. Hohman, F., Kahng, M., Pienta, R., Chau, D.H.: Visual analytics in deep learning: An interrogative survey for the next frontiers. IEEE Transactions on Visualization and Computer Graphics (TVCG) (2018)
14. Karimi, A.H., Barthe, G., Schölkopf, B., Valera, I.: A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. arXiv preprint arXiv:2010.04050 (2020)
15. Keane, M.T., Kenny, E.M., Delaney, E., Smyth, B.: If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. arXiv preprint arXiv:2103.01035 (2021)
16. Kenny, E.M., Keane, M.T.: Explaining deep learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in xai. Knowledge-Based Systems **233**, 107530 (2021)
17. Kobak, D., Berens, P.: The art of using t-sne for single-cell transcriptomics. Nature communications **10**(1), 1–14 (2019)
18. Liu, M., Liu, S., Su, H., Cao, K., Zhu, J.: Analyzing the noise robustness of deep neural networks. In: 2018 IEEE Conference on Visual Analytics Science and Technology (VAST). pp. 60–71. IEEE (2018)
19. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning pp. 193–209 (2019)

20. Novaković, J.D., Veljović, A., Ilić, S.S., Papić, Ž., Milica, T.: Evaluation of classification models in machine learning. Theory and Applications of Mathematics & Computer Science **7**(1), 39–46 (2017)
21. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill (2017)
22. Rauber, P.E., Fadel, S.G., Falcao, A.X., Telea, A.C.: Visualizing the hidden activity of artificial neural networks. IEEE transactions on visualization and computer graphics **23**(1), 101–110 (2016)
23. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
24. Rysbayeva, K., Giot, R., Journet, N.: Hierarchical and multimodal classification of images from soil remediation reports. In: International Conference on Document Analysis and Recognition. pp. 160–175. Springer (2021)
25. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
26. Smith, R.: An overview of the tesseract ocr engine. In: ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition. pp. 629–633. IEEE Computer Society, Washington, DC, USA (2007)