2021

# Hierarchical Visual Concept Interpretation for Medical Image Classification

Mohammed Khaleel
*Iowa State University*, mkhaleel@iastate.edu

Wallapak Tavanapong
*Iowa State University*, tavanapo@iastate.edu

Johnny Wong
*Iowa State University*, wong@iastate.edu

Junghwan Oh
*University of North Texas*

Piet de Groen
*University of Minnesota - Twin Cities*

## Recommended Citation

Khaleel, Mohammed; Tavanapong, Wallapak; Wong, Johnny; Oh, Junghwan; and de Groen, Piet, "Hierarchical Visual Concept Interpretation for Medical Image Classification" (2021). *Computer Science Conference Presentations, Posters and Proceedings*. 52.
https://lib.dr.iastate.edu/cs_conf/52

# Hierarchical Visual Concept Interpretation for Medical Image Classification

## Abstract

Most state-of-the-art local interpretation methods explain the behavior of deep learning classification models by assigning importance scores to image pixels based on how influential each pixel was towards the final decision. These interpretations are unable to provide further details to aid understanding of a complex concept in a domain such as medicine. We propose a novel Hierarchical Visual Concept (HVC) interpretation framework for CNN-based image classification models. As an explanation of the classification decision of a given image, HVC presents a concept hierarchy of most relevant visual concepts at multiple semantic levels. These concepts are automatically learned during training such that the lower-level concepts in the hierarchy support the corresponding higher-level concepts. Our quantitative and qualitative evaluation of the interpretation of VGG16 and ResNet50 classifiers on public and private colonoscopy image datasets shows very promising results.

## Keywords

Interpretation of deep neural networks, Hierarchical visual concept interpretation, Convolutional Neural Networks

## Disciplines

Computer Sciences | Systems Architecture

## Comments

This is a manuscript of a proceeding published as M. Khaleel, W. Tavanapong, J. Wong, J. Oh and P. de Groen, "Hierarchical Visual Concept Interpretation for Medical Image Classification," *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 2021, pp. 25-30, Virtual Event. doi:10.1109/CBMS52027.2021.00012. Posted with permission.

# Hierarchical Visual Concept Interpretation for Medical Image Classification

Mohammed Khaleel
*Department of Computer Science*
*Iowa State University*
Ames, IA, USA
mkhaleel@iastate.edu

Wallapak Tavanapong
*Department of Computer Science*
*Iowa State University*
Ames, IA, USA
tavanapo@iastate.edu

Johnny Wong
*Department of Computer Science*
*Iowa State University*
Ames, IA, USA
wong@iastate.edu

Junghwan Oh
*Department of Computer Science and Engineering*
*University of North Texas*
Denton, TX, USA
junghwan.oh@unt.edu

Piet de Groen
*Department of Medicine*
*University of Minnesota*
Minneapolis, MN, USA
degroen@umn.edu

***Abstract*—Most state-of-the-art local interpretation methods explain the behavior of deep learning classification models by assigning importance scores to image pixels based on how influential each pixel was towards the final decision. These interpretations are unable to provide further details to aid understanding of a complex concept in a domain such as medicine. We propose a novel Hierarchical Visual Concept (HVC) interpretation framework for CNN-based image classification models. As an explanation of the classification decision of a given image, HVC presents a concept hierarchy of most relevant visual concepts at multiple semantic levels. These concepts are automatically learned during training such that the lower-level concepts in the hierarchy support the corresponding higher-level concepts. Our quantitative and qualitative evaluation of the interpretation of VGG16 and ResNet50 classifiers on public and private colonoscopy image datasets shows very promising results.**

***Keywords*—*Interpretation of deep neural networks, Hierarchical visual concept interpretation, Convolutional Neural Networks.***

## I. Introduction

In recent years, Convolutional Neural Network (CNN) has been intensively explored for image recognition in several application domains including medicine [1]. Due to the success and the black-box nature of CNN-based models, interpretability or explanability of these models have received an influx of research attention. We use interpretability and explanability, interchangeably, as reviewed in [2]. Interpretation techniques provide insight into the internal working of the overall model (global interpretation) and/or how the model arrives to the class prediction for a specific input image (local interpretation). Most existing local interpretation methods summarized in [2] provide interpretation that highlights regions that support the class prediction of an input image. Recently, interpretation methods that provide multiple levels of concepts have begun to emerge [3, 4]. However, they require manually labeled concepts at various semantic levels. In medicine, manual labeling by domain experts of medical images at multiple semantic levels to create a sufficiently large training dataset for classification and interpretation is not feasible cost-wise.

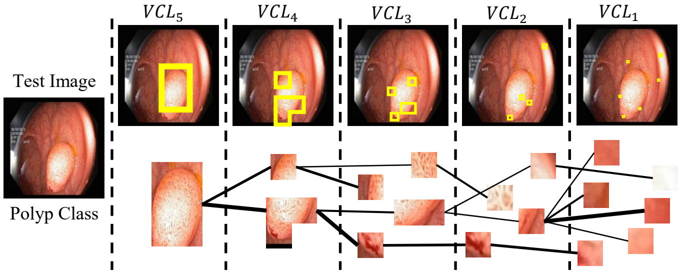


Fig. 1. A visual concept hierarchy interpretation for a polyp image detected by ResNet50 from the Kvasir V2 public dataset [5]. The hierarchy shows image patches at multiple semantic granularities detected by our visual concept layers: $VCL_5$ to $VCL_1$. Important texture and color are concepts detected by the visual concept layers $VCL_2$ and $VCL_1$, respectively. Thicker connecting lines indicate stronger influence of lower-level to higher-level visual concepts.

We propose a "Hierarchical Visual Concepts" (HVC) framework for interpretation of CNN-based image classification models. Fig. 1 shows an HVC explanation for the prediction that the input test image belongs to a polyp class by the ResNet50 classifier. Our interpretation framework provides bounding boxes over pixels that strongly influence the prediction of the class label by comparing them to the important patterns learned in each of the five semantic levels from the training dataset. At the top of the concept hierarchy $VCL_5$, the yellow bounding box indicates the region that causes ResNet50 to predict the entire image as a polyp image. Indeed, the interpretation shows that ResNet50 correctly uses the polyp region in the image for predicting the polyp class. At $VCL_4$, the two yellow boxes capture the curved edges of the polyp region, which contributes the most to the polyp region in $VCL_5$. Curved edges of polyps were found important for deriving hand-crafted features for polyp image detection [6]. At $VCL_3$, the yellow boxes focus on few edge segments of the polyp contour. At $VCL_2$, different polyp textures are highlighted. Finally, at $VCL_1$, six nodes represent a different shade of red and bright white. The interpretation at all the levels highlights different important characteristics of a polyp appearance.

Our contributions are summarized as follows.

- A new framework HVC that utilizes a Visual Concept Layer (VCL) to learn multi-level visual concepts (generalized feature representations) that capture the most important characteristics (for classification) of the class. HVC utilizes a new objective function to connect the learned visual concepts across different semantic levels: low-level concepts (color and texture), mid-level concepts (shape or parts of objects), and high-level concepts (objects of interest). HVC can interpret any CNN-based image classification model without reducing the classification accuracy. Most importantly, it does not require manual labeling of visual concepts for training, making HVC suitable for the domain where manual labeling is prohibitively expensive.
- A new local interpretation method that constructs visual concept hierarchies (trees) as a novel medium for interpretation. The tree branches reveal how well the lower-level concepts support the corresponding higher-level concepts, providing a more in-depth, structured explanation of a model's decision on a given input image. To the best of our knowledge, this is the first attempt toward automated construction of learned visual concepts at multiple granularities and visual concept hierarchies as explanation for CNN-based image classification.
- A qualitative and quantitative evaluation of HVC on interpreting VGG16 [7] and ResNet50 [8] models for two image classification tasks using both private and public colonoscopy datasets. The source code and experimental results of the public dataset are available on github.com/cml-cs-iastate/hvc.

## II. Related Work

Interpretable prototype-based classification models: Our visual concepts may be seen as closely related to "prototypes". Prototype-based classification methods [9, 10] use learned image prototypes for classification and reasoning. In [11], a self-interpretable classification model is trained to extract prototypes that are insensitive to small image perturbation. ProtoPNet [12] finds a prototype, a generalized representation of convolutional features of the last convolutional layer and uses the prototypes to make a final classification. ProtoPNet requires retraining of the entire baseline model. The authors stated that ProtoPNet marginally reduced the classification accuracy in comparison with the baseline models. ProtoPNet generates object-level prototypes only.

Local interpretation: Given an input image, the majority of the methods in this category highlight image region(s) supporting the predicted class [2]. A few recent methods [13-14] provide a contrastive explanation (e.g., image(s) most similar to the input image but of a different class). In [14], perturbation of the training images is done to find two sets of image regions from the images of the same class: one that must be present and the other that must be absent. Wang et al. [13] used the manually labeled Broden dataset [4] of color, texture, objects, and scenes for generic objects to build a hierarchy of concepts for local interpretation. Their method requires that each concept has a set of binary segmentation mask images and the concept label as ground truth. The method cannot detect other concepts beyond the ones already in the manually labeled dataset. To the best of our knowledge, there are no existing methods that automatically construct prototypes at different granularities, forming a hierarchical interpretation.

Global interpretation: The methods in this category attempt to reveal what image properties the neural network neurons or layers detect [2]. Zeiler and Fergus used deconvolution and un-pooling operations [15] to find patterns detected at intermediate layers. This method does not reveal relationships among the patterns across layers beyond spatial locations. In [16], high-level concepts are constructed by clustering image regions generated by an image segmentation method. The clustering uses the extracted features of these regions from a specific CNN layer. The Broden dataset [4] was used to identify which neurons in a given CNN detect which concepts using the intersection-over-union score between the predicted mask and the ground truth mask created manually. Unlike these works, HVC does not use image segmentation that requires appropriate parameter tuning and does not require manually labeled concept segmentation masks, avoiding the high cost of manual labeling.

## III. Hierarchical Visual Concept (HVC) Interpretation

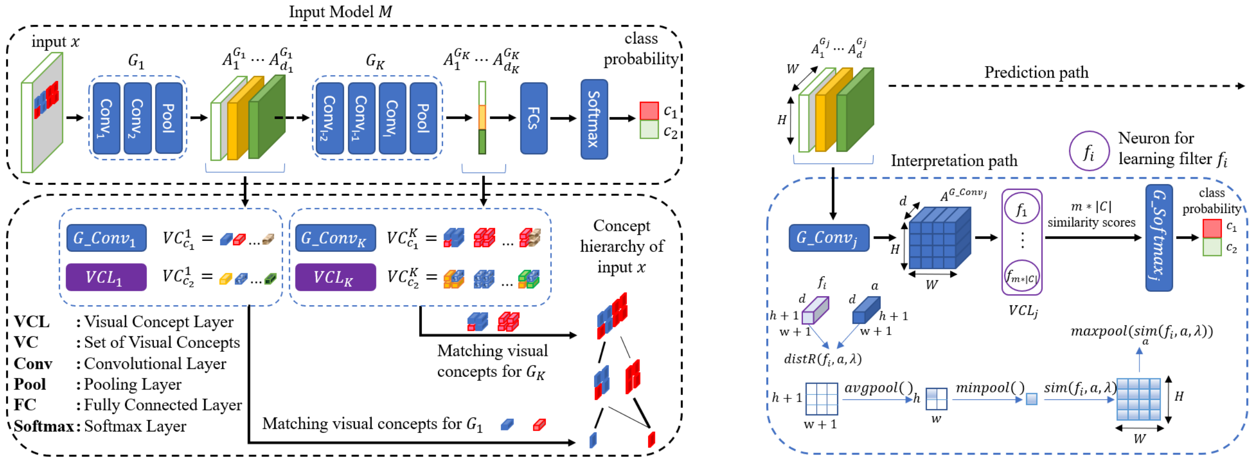
HVC learns distinct visual concepts automatically and generates a visual concept hierarchy as a new type of explanation of a CNN class prediction for a given image. The hierarchy shows inter-related multi-granularity concepts important for the prediction.

### *A. Overview of the interpretation framework*

HVC takes a pre-trained convolutional neural network $M$ with $l$ convolutional layers to classify an image into a class $c \in C$ where $C$ is the set of class labels. We group the consecutive convolutional layers of $M$ into $K$ groups, where $1 \leq K \leq l$. The number of groups ($K$) is user-defined and determines the number of levels in the concept hierarchy. For instance, when $K = 1$, there is only one group that includes all the layers before the fully connected layer of the CNN model. Thus, the concept hierarchy only includes high-level visual concepts like objects or their parts. Without domain knowledge defining levels of features (e.g., color or shape) that are more important for the target classification task, we recommend dividing the total number of convolutional layers into groups of roughly equal size. This will prevent any learning of visual concepts from combining features from too many or too few layers.

Fig. 2(a) shows an HVC architecture and the interpretation process for a binary classification task by a VGG16 classifier. For ease of exposition, we chose VGG16 as an example. The VGG16 first two convolutional layers, $Conv_1$ and $Conv_2$ with the pooling layer were grouped into the group $G_1$. The last group $G_K$ includes the last three convolutional layers, $Conv_{l-2}$, $Conv_{l-1}$, and $Conv_l$, and the pooling layer after them. The feature maps output of $G_j$ is input to our convolutional layer $G\_Conv_j$ followed by a visual concept layer $VCL_j$. $G\_Conv_j$ performs convolutional operations with a filter of size $1 \times 1$ followed by a sigmoid activation to output $d$ feature maps.

During training, we only train $G\_Conv_j$ and $VCL_j$ to learn $VC_c^j$, a set of visual concepts for each class $c \in C$ for each group $G_j$ of the pre-trained model. We do not retrain the pre-trained



(a)

(b)

Fig. 2. Example HVC architecture for VGG16 for a binary classification task; HVC supports $K$ semantic levels using one convolutional layer $G\_Conv_j$ and one visual concept layer $VCL_j$ for each semantic level $G_j$ for interpretation. (a) Interpretation process using the proposed layers (detailed in Section III.D). (b) Inner working of $VCL_j$. See details in Sections III.A - III.C.

classification model $M$ to avoid the drop in classification accuracy. For interpretation, HVC passes the feature maps output of each group $G_j$ of the pre-trained model $M$ on a test image to the corresponding $G\_Conv_j$ and $VCL_j$ to identify the most relevant visual concepts $M$ learned at that group. Afterwards, HVC constructs the concept hierarchy explanation of the classifier's prediction by connecting these visual concepts with differently assigned weights. Therefore, HVC is able to provide a more in-depth explanation of the model decisions across multiple semantic levels.

## B. *Visual Concept Layer (VCL)*

Fig. 2(b) depicts the inner working of the visual concept layer. Let $A^{G\_Conv_j}$ be the $d$ feature maps (of size $W \times H$) generated by $G\_Conv_j$. Each visual concept is learned using a filter $f \in \mathbb{R}^{(h+1)\times(w+1)\times d}$, where $h$ and $w$ are the predefined height and width of the visual concept, respectively; $1 < h < H$ and $1 < w < W$. To learn $m$ visual concepts per class, we configure each $VCL_j$ with $m * |C|$ neurons, one neuron for learning one visual concept for a particular class, where $|C|$ denotes the number of classes. The neurons output $m$ similarity scores for each class and are fully connected to the softmax layer of the group denoted as $G\_Softmax_j$ as shown in Fig. 2(b). Each filter $f$ is randomly initialized, but as the training progresses, its values eventually represent a distinct visual concept for its class. The learning is guided by our new loss function (detailed in Section III.C) that utilizes the similarity score that reflects the closeness between the visual concept this neuron learns and the most similar image patch in the image.

To obtain the similarity score, we slide each filter $f$ across $A^{G\_Conv_j}$ with a stride of one. We first calculate the similarity between $f$ and each feature map partition $a \in \mathbb{R}^{(h+1)\times(w+1)\times d}$ in $A^{G\_Conv_j}$ using (1).

$$sim(f, a, \lambda) = \frac{1}{1+\theta} \quad (1)$$

where $\theta = minpool\big(avgpool(distR(f, a, \lambda))\big)$.

Our $sim(f, a, \lambda)$ uses the distance function $distR(f, a, \lambda)$ to calculate the squared $L_2$ distance for each of the $d$ dimension vectors of $f$ and $a$ (Fig. 2(b)). To keep the most important features of the target concept in the central part of the filter, we multiply the distance value of every border element in the matrix with an importance constant $\lambda \in [0..1)$. Next, we apply average pooling, $avgpool()$, with the window size $h \times w$ and a stride of one to the output of $distR(f, a, \lambda)$ to calculate the average distance for each $h \times w$ region. Then, we apply $minpool()$ that returns the minimum average distance within the window size $h \times w$ as shown in Fig 2(b). Finally, we convert this distance into a similarity score between 0 and 1 using (1). Since we slide the filter $f$ across all partitions in $A^{G\_Conv_j}$, we have a matrix of similarity scores between $f$ and each of these partitions. Last, we apply $maxpool(\,)$ to select the highest similarity score from this matrix and denote it as $s^{vc}$ as shown in (2).

$$s^{vc} = maxpool\Big(\underset{a}{sim}(f, a, \lambda)\Big) \quad (2)$$

## C. *Training Algorithm and Loss Function*

Our training algorithm takes the pretrained model $M$ attached with the layers described in Section III.B. We set the weights for the connections between the neurons in $VCL_j$ and those in $G\_Softmax_j$ for each group $G_j$. We assign a value of 1 to the weight connection between these neurons for the same class and $-0.5$ otherwise, as done in [12]. During training, we fix these weights at their initial values, but train the weights of $G\_Conv_j$ and visual concepts of $VCL_j$ for each $G_j$ until convergence in terms of the loss $\mathcal{L}$ (3), starting from the latest $G_K$ and working backward to the earliest $G_1$. At the end of the training of $G\_Conv_j$ and $VCL_j$, we freeze the learned parameters of $G\_Conv_j$ and $VCL_j$ before training $G\_Conv_{j-1}$ and $VCL_{j-1}$. This to ensure that the visual concepts learned in the current iteration support the higher-level visual concepts learned in the previous iteration. We use a gradient descent optimization function to minimize the proposed loss function $\mathcal{L}$ of four terms: cross-entropy ($\mathcal{L}_{entropy}$), membership cost ($\mathcal{L}_{mem}$), diversity cost ($\mathcal{L}_{diversity}$), and clustering cost ($\mathcal{L}_{clst}$) [12] as defined in (3) to (6). $\mathcal{L}_{entropy}$ measures the classification error, which helps to extract convolutional features for each class. We introduce two new cost functions: membership cost and diversity cost. The notations used in these equations are defined in Table I.

$$\mathcal{L} = \mathcal{L}_{entropy} + \Gamma\, \mathcal{L}_{mem} + \alpha\, \mathcal{L}_{diversity} + \beta\, \mathcal{L}_{clst} \quad (3)$$

The membership cost as defined in (4) measures the similarity between the visual concepts of the same class across two consecutive groups $G_j$ and $G_{j+1}$ with respect to their corresponding most similar image patches. This cost function encourages the learning of visual concepts that are similar to the already learned visual concepts in the higher level. For each training image $i$, we use the true class $c$ of that image $i$ denoted as $y_{c,i}$ and compute $supp_{c,i}$, representing the average similarity between the visual concepts of the class $c$ across $G_j$ and $G_{j+1}$.

TABLE I. NOTATIONS FOR EQUATIONS 3 TO 8

| Notation | Explanation |
|---|---|
| $D$ and $\|D\|$ | Training image dataset and total number of images in $D$ |
| $C$ and $\|C\|$ | Set of class labels and total number of classes |
| $\alpha$, $\beta$, and $\Gamma$ | Term multipliers of value [0, 1] |
| $vc \in VC_c^j$ | Visual concept belonging to $VC_c^j$ |
| $p_i^{vc}$ | Patch of image $i$ most similar to the visual concept $vc$ |
| $s^{vc}$ | Similarity score of the visual concept $vc$ for a given image |
| $\|\|.,.\|\|_2^2$ | Squared $L_2$ distance between two vectors or flattened matrices |
| $d$ | Number of feature maps generated by $G\_Conv_j$ and the filter depth of visual concepts. |

$$\mathcal{L}_{mem} = \frac{1}{|D|m} \sum_{i \in D} \sum_{vc_1 \in VC_c^j} supp_{c,i}(vc_1) \quad where\ c = y_{c,i} \quad (4)$$

$$supp_{c,i}(vc_1) = \min_{vc_2 \in VC_c^{j+1}} \left( \left( \min_{p \in p_i^{vc_2}} \left\| p_i^{vc_1}, p \right\|_2^2 \right) \times \left(1 - s^{vc_2}\right) \right)$$

The term $\left\| p_i^{vc_1}, p \right\|_2^2$ is the squared $L_2$ distance between the patch of image $i$ most similar to the visual concept $vc_1$ and every part of the patch of image $i$ most similar to the visual concept $vc_2$ in $G_{j+1}$. This is required because the image patches of the visual concepts at $G_{j+1}$ have a larger width and height than those at $G_j$. We add term $(1 - s^{vc_2})$ to penalize the visual concept of the group $G_{j+1}$ for image $i$ that has a low similarity score. Minimizing these two terms encourages association of visual concepts across the concept hierarchy. Note that the membership cost takes only the content similarity of image patches into account, which allows for translation invariance.

We propose the diversity cost defined in (5) to guide the learning of the visual concepts of a particular class to capture different visual patterns in the training images of that class. The diversity cost is inversely proportional to the average minimum squared $L_2$ distance between two different visual concepts $vc_1$ and $vc_2$ of the same class. The diversity cost is very important for preventing the visual concepts designated for the same class from overfitting to the same image pattern, and thus results in more diverse learned concepts.

$$\mathcal{L}_{diversity} = -\frac{1}{|C|} \sum_{c \in C} IntraDist_c \quad (5)$$

$$IntraDist_c = \sum_{vc_1 \in VC_c^j} \min_{\substack{vc_2 \in VC_c^j \\ vc_2 \neq vc_1}} \|vc_1, vc_2\|_2^2$$

We adapt the clustering cost [12] shown in (6) to have parts of each training image of a class be represented by the visual concepts of the class. Given the class of each training image $i$, we compute the similarity score $s^{vc}$ of image $i$ using (2).

$$\mathcal{L}_{clst} = \frac{1}{|D|} \sum_{i \in D} \min_{vc \in VC_c^j} (1 - s^{vc}) \quad where\ c = y_{c,i} \quad (6)$$

Note that the loss function $\mathcal{L}$ impacts trainable parameters of the convolutional layer and visual concept layer per group $G_j$.

### *D. Local Interpretation with Concept Hierarchies*

We describe the steps to create a concept hierarchy (tree) of the relevant learned visual concepts as an interpretation of the class prediction of an input image $x$. First, we forward image $x$ through the convolutional layers of $M$ and the associated visual concept layers until reaching the last visual concept layer $VCL_K$. Suppose that the CNN classifier $M$ predicts that $x$ is of the class $c$. We initialize the output concept tree with a root node representing the entire image. Starting backward from $VCL_K$ to $VCL_1$, at each $VCL_j$, the method is as follows.

1) For each learned visual concept of the class $c$ of $VC_c^j$, find a single image patch that is most similar to it based on the similarity score in (2). Therefore, there are as many matching image patches as the number of visual concepts of $VC_c^j$. Create a corresponding tree node for each of the patches.

2) Create an edge between each newly created node with the node already in the concept tree. If the tree only has the root node, link the newly created nodes to the root node. Otherwise, create an edge between each newly created node and the leaf node with the highest membership score between them. The membership score between the nodes is calculated using the $supp_{c,i}()$ function defined in (4) given the image patches corresponding to these nodes. The membership score is used as the weight for the edge between these nodes and is reflected in the concept hierarchy using the thickness of the edge between nodes. Thicker edges mean higher membership scores.

3) At each level in the hierarchy, if the bounding boxes of the image patches associated with any two tree nodes at this level intersect or share borders in the image, we combine these patches into a bigger image patch and combine the corresponding tree nodes into a single node. This node inherits all the edges of their constituent nodes. If the root node only has one child node, we make that child node the new root node.

## IV. EXPERIMENTAL RESULTS

We evaluated the HVC framework qualitatively and quantitatively. We chose colonoscopy image datasets since we have the expertise to analyze the interpretation results. Due to space limitation, we present only the results on the EMIS-I dataset using the settings and values of hyperparameters chosen empirically as shown in Table II. The hyperparameter values and experimental results on the Kvasir dataset [5]. Both datasets do not have any patient identifiable information.

TABLE II. THE DESCRIPTION OF EXPERIMENTS PRESENTED IN THIS PAPER

| | |
|---|---|
| Private EMIS-I dataset | 5 classes with 2800 colonoscopy images per class; images resized to 224x224 (width × height) |
| Training, validation, testing split | 60%, 20%, 20% |
| Hyper-parameters | batch size:32, learning rate:0.001, momentum:0.9 |
| HVC hyper-parameters | Visual concept filter $h \times w = 2{\times}2$, $d = 128$, $m = 10$, $\lambda = 0.8$, $\alpha = 0.2$, $\beta = 0.8$, $\Gamma = 0.05$, $K = 2$ |
| HVC architecture | VGG16 pretrained on ImageNet and fine-tuned: $G\_Conv_1$ to $G\_Conv_2$ placed after block1_pool and block5_pool, respectively |
| Programing language | Python 3 using TensorFlow version 2.0 |

### A. *Qualitative Evaluation Results*

Fig. 3(a) shows a concept hierarchy interpretation of the snare class prediction by VGG16. The snare instrument used for polypectomy during colonoscopy has two parts, the thin wire and the sheath. This example shows interpretation when $K = 2$. At $VCL_1$ seven patches, three white bright patches of the snare instrument and four patches of the thin metal wire of the snare on the colon mucosa, were found at different locations in the image. These patches were found most similar to ten of the learned visual concepts at this visual concept layer. Observe that two of these patches are not part of the thin wire but are visually similar to the thin wire color. These patches are not errors since our membership cost does not rely on spatial locations. Seven leaf nodes in the concept hierarchy represent these seven patches. These leaf nodes are children of the node representing the combined image patch at the higher granularity level, $VCL_2$. The edge thickness in the concept hierarchy indicates that the dark color of the thin wire and the bright color and different texture of the sheath of the snare are equally important for classification of the snare instrument. Fig. 3(b) is an interpretation of the prediction of the forceps head. Similarly, two patches in $VCL_1$ are not on the forceps head region We see this as contrast rich explanation at this level. The forceps head not only has the silver color but needs to have a certain shape as captured in $VCL_2$.

### B. *Quantitative Performance Metrics and Results*

#### 1) *Diversity of the learned visual concepts*

We propose the intra-class diversity metric for class $c$ ($intraD_c$) defined in (7) and the inter-class diversity metric ($interD$) defined in (8). Recall the definitions of the variables in Table I. We normalize the metric values in the range [0, 1]. We trained HVC for the VGG16 classification model three times and averaged the performance over these runs.

$$intraD_c = \frac{1}{m^2 - m} \sum_{vc_1 \in VC_c^j} \sum_{\substack{vc_2 \in VC_c^j \\ vc_2 \neq vc_1}} \|vc_1, vc_2\|_2^2 \quad (7)$$

$$interD = \frac{1}{m^2(|C|^2 - |C|)} \sum_{c \in C} \sum_{vc_1 \in VC_c^j} \sum_{\substack{\tilde{c} \in C \\ c \neq \tilde{c}}} \sum_{vc_2 \in VC_{\tilde{c}}^j} \|vc_1, vc_2\|_2^2 \quad (8)$$

The $intraD_c$ metric is to evaluate the effectiveness of the diversity cost in preventing the visual concept filters of the same class from overfitting to the same concept. The larger the value for the class $c$, the more diverse the visual concepts are for that class. The $interD$ is the average distance between all pairs of $m$ visual concepts of different classes. The higher the value, the more different the learned visual concepts of one class from another. Ideally, $interD$ values should be far from zero and higher than the $intraD_c$ values since visual concepts of different classes should be more different.

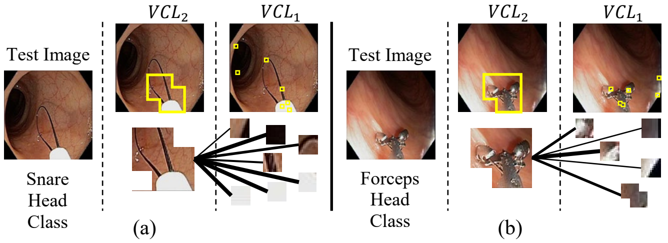


Fig. 3. Two-level hierarchical interpretation by HVC to explain the prediction of VGG16 for the snare head (a) and the forceps head (b) classes.

Table III shows these diversity values of the learned visual concepts for EMIS-I. The intra-class diversity values are higher for the group $G_1$ than those for the group $G_2$ for the first four classes. This is expected as the visual concept layer at $G_1$ learned low-level concepts like color and texture. These concepts make up the high-level concepts like the instrument objects in $G_2$. As the instruments share some common features for the cable part of the instrument (e.g., the shape, color, and texture of the cable), the high-level visual concepts ($G_2$) are less diverse. The non-instrument class has images of non-rigid colon walls. The color and texture learned in $G_1$ are less diverse for this class. The inter-class diversity values are higher than the intra-class diversity values as expected. Similar trends were found on Kvasir.

TABLE III. DIVERSITY METRICS ON THE EMIS-I TRAINING DATASET

| | **Intra-class diversity** | | | | | **Inter-class diversity** |
|---|---|---|---|---|---|---|
| | **Left Cable (C1)** | **Right Cable (C2)** | **Forceps Head (C3)** | **Snare Head (C4)** | **Non-instrument (C5)** | |
| $G_2$ | 0.14 | 0.29 | 0.15 | 0.19 | 0.27 | 0.79 |
| $G_1$ | 0.45 | 0.57 | 0.34 | 0.22 | 0.13 | 0.75 |

#### 2) *Representativeness of the learned visual concepts*

We propose a representative matrix. Each element in the representative matrix for each granularity level denotes the proportion of the training images in a class represented by the learned visual concepts of a class at that granularity. A training image is represented by a visual concept of a class when one of the visual concepts of that class is most similar to an image patch in the training image. Ideally, we expect a value of one along the diagonal of the representative matrix for each granularity level. In other words, all of the training images in a given class are represented by the learned visual concepts of that class only.

TABLE IV. REPRESENTATIVE MATRIX OF THE MOST SIMILAR VISUAL CONCEPT TO A TRAINING IMAGE PER CLASS

| | $G_2$ | | | | | $G_1$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **C1** | **C2** | **C3** | **C4** | **C5** | **C1** | **C2** | **C3** | **C4** | **C5** |
| **C1** | **1** | 0 | 0 | 0 | 0 | **0.95** | 0 | 0.02 | 0 | 0.03 |
| **C2** | 0 | **1** | 0 | 0 | 0 | 0.06 | **0.88** | 0.01 | 0 | 0.05 |
| **C3** | 0 | 0 | **1** | 0 | 0 | 0.1 | 0.01 | **0.67** | 0.01 | 0.21 |
| **C4** | 0 | 0 | 0 | **1** | 0 | 0.03 | 0.01 | 0.06 | **0.81** | 0.08 |
| **C5** | 0 | 0 | 0 | 0 | **1** | 0.13 | 0 | 0.14 | 0 | **0.73** |

Table IV shows the representative matrix of the visual concepts for EMIS-I. The values of 1 along the matrix diagonal elements show the perfect representation of the high-level concepts in $G_2$. However, at the low-level concepts in $G_1$, we see some confusion. Some training images in C2 (right cable class) are most similar to the visual concepts of C1 (left cable class) and C3 (forceps head class). This is because these three classes share common texture and color of the instrument cable. C3 class has images with a small part of the cable as well. About 5% of the training images in C2 are represented by the visual concepts of C5 (non-instrument class). This is due to colon folds with a strong light reflection similar to the strong light reflection coming from cable sheaths.

### 3) *Faithfulness of HVC interpretation*

We investigated whether the generated interpretation by HVC from the trained VGG16 model is indeed related to the classification decision made by the model. We used $K$=1 for HVC interpretation, which gives interpretation at the last convolutional layer. If a patch highlighted by HVC is indeed important for VGG16 class prediction of a given image, zeroing out all the pixels corresponding to this patch and giving the perturbed image to the VGG16 classifier should cause a drop in classification confidence. We measure the percentage of the number of training images where the VGG16 classification confidence drops after some perturbation. With the above perturbation, VGG16 dropped classification confidence in 98% of the EMIS-I training images. However, when zeroing out pixels randomly for the same number of pixels as that by HVC for each of the EMIS-I training images, VGG16 dropped classification confidence in only 87% of these images. The percentage was high even for the random method since we did not distinguish between a small drop or a large drop. Our HVC interpretation is more relevant to the classification decision made by the VGG16 classifier than random chance.

### 4) *Comparison with other interpretation methods*

To the best of our knowledge, there is no existing work that generates hierarchical interpretation for CNN-based image classification models without the need for manually created labels at multiple semantic levels. Ideally, we compare the learned visual concepts in one level, the object level. We chose ProtoPNet [12] since the learned representations (termed prototypes) are also in fixed-size patches. We defined a "relevance" score as the percentage of the area of the image patch, closest to the learned representation that intersects with the ground truth bounding box of the image object of the target class. We used the Oxford-IIIT Pet dataset [17] because it has a ground truth bounding box on the main object for each image. We reported the average of the relevance scores for all the images in the test set of the Oxford-IIIT Pet dataset. For a fair comparison, we set $K$ to one for HVC and set the number of learned representations to 10 for both HVC and ProtoPNet. For ProtoPNet, we used the same hyperparameters presented in [12]. HVC achieved an 84% relevance score, in comparison with 89% by ProtoPNet. The results show that both generated visual concepts are good. Although a 5% higher relevance score, ProtoPNet has a drop in classification accuracy from 83% to 79%. HVC classification accuracy remained 83%.

## V. Conclusion and Future Work

In this paper, we present a Hierarchical Visual Concept (HVC) interpretation framework for CNN-based image classification. HVC automatically learns and extracts relevant visual concepts at multiple semantic levels to construct a concept hierarchy as a novel medium for CNN classification interpretation. HVC does not require manual labeling beyond image labels and does not impact the classification accuracy, making it suitable for domains like medicine. The evaluation results show that HVC is promising. HVC is extensible to other domains. As future work, we will have domain experts evaluate the interpretation by HVC and impact of the number of visual concepts per class.

## Conflict of Interest and Aknowledgments

This work is partially supported by NIH Grant No. 1R01DK106130-01A1. Tavanapong, Wong, and Oh have equity interest and management roles in EndoMetric Corp. Dr. de Groen serves on the Scientific Advisory Board of EndoMetric Corp. Findings, opinions, and conclusions expressed in this paper do not necessarily reflect the view of the funding agency.

## Reference

[1] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical Image Analysis using Convolutional Neural Networks: A Review," Journal of Medical Systems, vol. 42, p. 226, 2018.

[2] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," in Proc. of IEEE Int'l Conf. on Data Science and Advanced Analytics (DSAA), Italy, 2018, pp. 80–89.

[3] D. Wang, X. Cui, and Z. J. Wang, "CHAIN: Concept-harmonized Hierarchical Inference Interpretation of Deep Convolutional Neural Networks". arXiv preprint: 2002.01660, 2020.

[4] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network Dissection: Quantifying Interpretability of Deep Visual Representations," in Proc. of IEEE CVPR, 2017, pp. 3319–3327.

[5] K. Pogorelov, K. Randel, C. Griwodz, S. Eskeland, T. de Lange, D. Johansen, and C. Spampinato, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in Proc. of ACM Multimedia Systems, USA, 2017, pp. 164–169.

[6] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. De Groen, "Part-Based Multiderivative Edge Cross-Sectional Profiles for Polyp Detection in Colonoscopy," IEEE JBHI, vol. 18, no. 4, pp. 1379–1389, 2014.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. of Int'l Conf. on Learning Representations, USA, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. of IEEE CVPR, USA, 2016, pp. 770-778.

[9] J. Bien and R. Tibshirani, "Prototype selection for interpretable classification," Annals of Applied Statistics, vol. 5, 2012.

[10] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in Proc. of AAAI, 2018.

[11] D. Alvarez-Melis and T. Jaakkola, "Towards Robust Interpretability with Self-Explaining Neural Networks," in Proc. of NeurIPS, USA, 2018.

[12] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, "This looks like that: deep learning for interpretable image recognition," in Proc. of NeurlIPS, Canada, 2019, pp. 8928–8939.

[13] A. Kanehira and T. Harada, "Learning to Explain With Complemental Examples," in Proc. of IEEE CVPR, USA, 2019, pp. 8595-8603.

[14] A. Dhurandhar, P. Chen, R. Luss, C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives," in Proc. of NIPS, USA, 2018.

[15] M. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Neural Networks," in Proc. of ECCV 2014, LNCS, vol. 8689, 2014.

[16] A. Ghorbani, J. Wexler, J. Zou, and B. Kim, "Towards Automatic Concept-based Explanations," in Proc. of NeurlIPS, Canada, 2019.

[17] O. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in Proc. of IEEE CVPR, USA, 2012, pp. 3498–3505.