Encoding Visual Attributes in Capsules for Explainable Medical Diagnoses

Rodney LaLonde[1] Drew Torigian[2] Ulas Bagci[1]

[1]Center for Research in Computer Vision
University of Central Florida

[2]Penn Medicine
University of Pennsylvania

Abstract

In high-risk domains, understanding the reasons behind machine-generated predictions is vital in assessing trust. In this study, we introduce a novel design of multi-task capsule network to provide explainable medical image-based diagnosis. Our proposed explainable capsule architecture, called X-Caps, encodes high-level visual attributes within the vectors of its capsules, then forms predictions based on these interpretable features. Since these attributes are independent, we modify the dynamic routing algorithm to independently route information from child capsules to parents. To increase the explainability of our method further, we propose to train our network on a distribution of expert labels directly, rather than the average of these labels as done in previous studies. This provides a meaningful metric of model confidence, punishing over/under confidence, directly supervised by human-experts' agreement. In our example high-risk application of lung cancer diagnosis, we conduct experiments on a large and diverse dataset of over 1000 CT scans, where our proposed X-Caps, a relatively small 2D capsule network, significantly outperforms the previous state-of-the-art deep dual-path dense 3D CNN in predicting visual attribute scores while also improving diagnostic accuracy. To the best of our knowledge, this is the first study to investigate capsule networks for making predictions based on human-level interpretable visual attributes in general and its applications to explainable medical image diagnosis in particular.

1. Introduction

Deep neural networks are often called black-boxes due to their difficult-to-interpret decisions. This is characteristic of a deeper trend in machine learning, where predictive performance typically comes at the cost of *interpretability*. Although deep learning (DL) has played a major role in a wide array of fields, there exist several which have yet to be comparably impacted: military, security, transportation, finance, legal, and healthcare among others [3, 20, 26]. At its core, DL owes its success to the joining of two essential tasks, *feature extraction* and *feature classification*, learned in a joint manner, usually through a form of backpropagation. Although this direction has dramatically improved the predictive performance on a diverse range of tasks, it has also come at a great cost, the sacrifice of human-level *explainability*. As features becomes less *interpretable*, and the functions learned more complex, model predictions become more difficult to explain. Several works have began to press towards this goal of explainable DL, as explored in Section 2, but the problem remains largely unsolved.

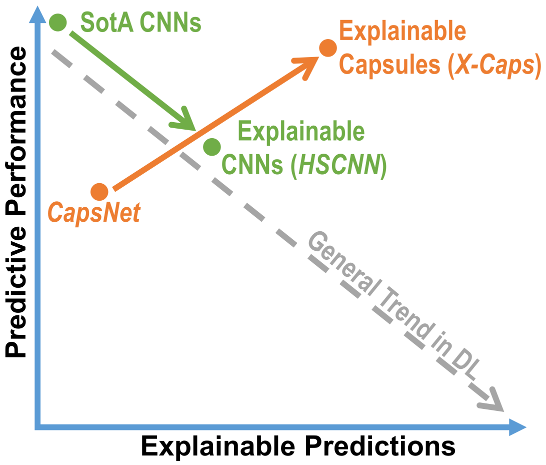

Figure 1: A symbolic plot showing the general trade-off between explainability and predictive performance in deep learning (DL). Our proposed explainable capsule network *X-Caps* rebuts the trend of decreasing performance from state-of-the-art (SotA) as explainability increases and shows it is possible to create more explainable models *and* increase predictive performance with capsule networks.

There has been a recent push in the community to move away from the *post-hoc interpretations* of deep models and instead create explainable models from the outset [27]. Since the terms *interpretable* and *explainable* are often used interchangeably, for the purposes of this study we want to be explicit about our definitions going forward. An *explainable* model is one which provides explanations for its predictions *at the human level* for a *specific task*. An *interpretable* model is one for which some conclusions can be drawn about the predictions of the model; however, they are not explicitly provided by the model and are typically at a lower level than human-level explanations.

For example, in image classification, when a deep model predicts an image to be of a cat, we can use a number of saliency-based, gradient-based, or other methods to attempt to *interpret* the prediction. However, when tasking a human with the same classification/explanation, the human will say it is a cat because it has four legs, paws, whiskers, fur, etc. Humans classify objects based on a taxonomy of characteristics/attributes. If our goal is to create *explainable* models, we should set out to design models which can explain their decisions using a similar set of "attributes" which humans can easily interpret. In this work, we create a model which explains its predictions using the same high-level visual attribute descriptions as human experts. Although we believe the proposed methods described are generic and can be applied to any classification problem in computer vision, we choose to focus on a high-risk application area where explainability is a critical lynch-pin holding back the adoption of DL in routine use: lung cancer diagnosis.

1.1. Explainable lung cancer diagnosis

Unlike detection tasks, diagnosis (classification) requires radiologists to explain their predictions through the language of high-level visual attributes, shown in Fig. 2. For DL-powered computer-aided diagnosis systems to be adopted by the healthcare industry and other high-risk domains, methods must be developed which can provide this same level of explainability currently provided by human experts (e.g. radiologists). Towards this goal, we propose a novel multi-task capsule architecture for learning visually-interpretable feature representations within the vectors of the capsules. We show even a relatively simple 2D capsule network, called *X-Caps*, can better capture high-level visual attribute information than a deep dual-path dense state-of-the-art 3D convolutional neural network (CNN) while improving diagnostic accuracy. To further increase the explainability of our framework, we propose to directly learn the distribution of radiologists' labels, rather than their average value as done in previous works, to provide a meaningful confidence metric on our predictions.

Lung cancer as a high-risk application: Lung cancer is the far-leading cause of cancer-related death in both men and women [7]. The National Lung Screening Trial showed that screening patients with low-dose computed tomography (CT) has reduced lung cancer mortality by 20% [35, 36]. However, only 16% of lung cancer cases are diagnosed at an early stage [11]. DL approaches such as 2D and 3D CNNs have been proposed to alleviate these challenges. Noticeably, some have achieved highly successful diagnosis results, comparable to or even better than expert level diagnosis [12, 38]. Nevertheless, the black-box nature of these previous studies has contributed to these methods not making their way into clinical routines. The purpose of this study is to fill this important research gap by creating explainable medical diagnoses through learning visually-interpretable features from medical images with new DL models, specifically a novel capsule network architecture.

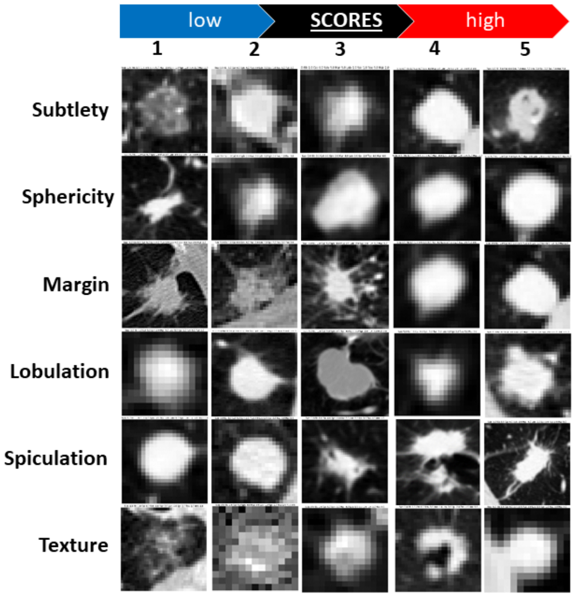

Figure 2: Lung nodules with high-level visual attribute scores as determined by expert radiologists. Scores were given from 1 – 5 for six different visual attributes related to diagnosing lung cancer.

1.2. Capsule neural networks and their visually-interpretable features

Capsule networks differ from traditional CNNs by replacing the scalar feature maps with vectorized representations, where these vectors are responsible for encoding orientation information, and thus provide *equivariance* to affine transformations on the input (as opposed to CNNs which are only equivarient to translation). These capsule vectors are then used in a dynamic routing algorithm which seeks to maximize the agreement between low-level and high-level features, not only in presence, but also in part-whole relationship agreement. In their introductory work [28], a capsule network (*CapsNet*) was shown to produce promising results on both the MNIST and CIFAR10 data sets; but more importantly, *CapsNet* was shown to encode high-level visually-interpretable feature representations of digits in MNIST (*e.g.* stroke thickness, skew, localized-parts) within the dimensions of its capsule vectors. While capsule networks are still a young area of research with many improvements to be made in terms of performance and accuracy, their ability to capture visually-interpretable features can be paramount in critical application domains that demand explainability of predictions. In this work, we show for the first time in the literature that even a simple 2D capsule network (*X-Caps*) can significantly outperform a deep dual-path 3D dense state-of-the-art CNN at capturing visually-interpretable high-level attributes.

1.3. Creating explainable capsules for diagnosis

In this study, we introduce a novel multi-task capsule network architecture for providing explainable diagnoses. These explanations come in the same form as human experts' explanations, namely radiologists provide six high-level visual attribute scores from lung nodules, as seen in Fig. 2, which they use when describing the malignancy of a nodule. In addition to providing these high-level explanations, our network both segments nodules and determines their malignancy score in a multi-task learning (MTL) approach. Our proposed architecture, called *X-Caps*, is an intuitive extension of the original *CapsNet*, where each dimension of the output capsule layer is supervised by a visual attribute label from multiple radiologists.

By *forcing* each dimension of the capsule vector output to embed a specific visually-interpretable feature, a significant benefit (explainable decision) is obtained by unraveling knowledge hierarchy inside deep networks. The multiple visual attributes are learned simultaneously with their associated weights being updated by both the radiologists visual interpretation scores as well as their contribution to the final malignancy score, and the segmentation reconstruction error. *X-Caps* malignancy predictive performance (without any pre-training) is on par with previously state-of-the-art deep pre-trained 3D CNN (*e.g.* [31, 32]), while also outputting visual attribute scores to explain network predictions, where nearly no previous works do so. We compare directly with *HSCNN* by Shen *et al.* [30], the only other work in the literature which attempts to provide explanations for lung cancer diagnosis, where *X-Caps* provides significantly higher attribute prediction accuracy.

Since radiologists' scores vary significantly between one another (inter-observer) for both malignancy and visual characteristics of a given nodule, it is not possible to train the proposed networks directly against these scores. Previous works train against the mean of the radiologists' scores and convert the mean to a binary label (malignant or benign); however, this throws away significant information. In our proposed method, we fit a Gaussian distribution of mean and variance equal to that of the radiologists' scores for a given nodule as the target of our networks' predictions. In this way, overconfidence by the network on more ambiguous nodules is punished in proportion to radiologists' disagreement, and likewise for under-confidence and strong radiologist agreement. This allows our method to produce classification scores across all five possible score values, rather than simply binary labels as in previous studies, while the variance of this predicted distribution is directly related to the confidence/reliability of the prediction, and thus provides an important additional metric to increase the explainability of our approach in high-risk applications.

1.4. Summary of our contributions

The contributions of this study are summarized as:

1. The first study, to the best of our knowledge, for directly learning an interpretable feature space by encoding high-level visual attributes within the vectors of a capsule network to perform explainable image-based diagnosis. We design a novel multi-task capsule network to encode visually-interpretable attributes within capsule vectors, then predict malignancy using these vectors to provide explanations *at the human-level*, in the same language used by radiologists.

2. Demonstrate even a simple 2D capsule network (*X-Caps*) trained from scratch can significantly outperform a state-of-the-art deep pre-trained dual-path 3D dense CNN at capturing visually-interpretable high-level attributes, while providing comparable malignancy accuracy to deep 3D CNNs.

3. Provide a meaningful confidence metric with our predictions by learning directly from expert label distributions to punish over/under confidence when human experts are in agreement/disagreement.

4. Modify the dynamic routing algorithm to independently route information from child capsules to parents when parent capsules are not mutually-exclusive.

The rest of the paper is organized as follows. In Section 2, we summarize related works in the literature pertaining to explainable DL, lung cancer diagnosis, and capsule networks in medical imaging. In Section 3, we introduce our proposed paradigm of learning visually-interpretable features via our newly designed multi-task capsule networks. In Section 4, we explain our experiments and results. We conclude our work with discussions in Section 5.

2. Related work

The majority of work in explainable deep learning has focused around *post hoc* deconstruction of already trained models. Two main approaches are primarily investigated, interpretation of the features learned by the networks and explaining deep networks' final predictions, at both the local (*i.e*. individual neurons) and global (*i.e*. entire layers/networks) levels. These approaches typically rely on human-experts to examine their results and attempt to discover meaningful patterns. While there are numerous studies on interpretable and explainable DL, we will attempt to faithfully cover the more prominent approaches. Following this, we will cover relevant lung cancer diagnosis and capsule-based works.

Visualization of features: Several works have attempted to examine network interpretability at the individual neuron level. Some of the earliest methods focused on visualizing individual filters and activation maps. While this can provide some insight into aspects of a network, such as dead neurons, the visualization of individual filters or feature maps are typically not interpretable at the human-level. Zeiler and Fergus [37] attached a deconvolutional network to network layers to map activations back to pixel space for visualization. Later, Springenberg *et al*. [34] used an all convolutional network and a guided-backpropagation algorithm to create much sharper visualizations which did not require the keys of the pooling operations. Mahendran and Vedaldi [22] focused more on layers of neurons and examine the representations learned by shallow and deep CNNs by inverting images using gradient descent. While these methods provide some insight into what CNNs learn, they are ultimately limited, as deep networks typically have hundreds of thousands of neurons and it is intractable to visually examine all or even large subsets of neurons in a network. Additionally, there is evidence to suggest these visualizations are unrelated to network predictions [25].

Receptive fields, input contributions: Beyond visualizing the features of CNNs, several methods have attempted to examine the effect of individual neurons or image regions on network outputs. In this first category, Girshick *et al*. [9] examined the receptive field of individual neurons and found the images which maximally activated each. Kindermans *et al*. [15] showed that [34, 37] (discussed above) did not create theoretically correct explanations for linear models, and created *PatternNet* and *PatternAttribution* to better visualize neuron activations. In the latter category, Kumar *et al*. [17] examined which input region correspond most strongly with each output class. An occlusion-based approach was used by Zeiler and Fergus [37] for masking out image regions to examine their contribution to the final output. One of the most popular methods of visualizing input contributions is *Grad-CAM* [29] which highlights the relative positive activation map of convolutional layers with respect to network outputs. Arguably, saliency detection can also fall into this category of determining input region importance. While these methods give important information related to designing networks and training data, they tell us very little about the internal representations being learned.

Feature spaces and GANs: Rather than looking at the individual neurons or image regions, several approaches instead focus on examining the feature spaces learned by deep networks. Generative adversarial networks (GAN) by Goodfellow *et al*. [10], show vulnerable regions of a learned feature space for a given network. In [5], Chen *et al*. creates a GAN-based method called *InfoGAN* to separate noise from the "latent code" in images. Using this method, they maximize the mutual information between the latent representations and the image inputs, encoding concepts such as rotation, width, and digit type for MNIST. In a similar way, capsule networks by Sabour *et al*. [28] (*CapsNet*) encode visually-interpretable concepts such as stroke thickness, skew, rotation, and others. These two methods are the most similar to the proposed approach. Lakkaraju *et al*. [18] attempt to discover a CNN's "blind spots" by sampling points in feature space in a weakly-supervised manner. While the other methods mentioned can provide some important clues about the feature space being learned, *InfoGAN* and *CapsNet* show the most promise for encoding and extracting visually-interpretable features.

Lung nodule classification: The majority of recent lung cancer diagnosis (nodule classification) studies have focused on deep 2D, multi-view, and 3D CNNs, with most works trained/tested on the publicly available LIDC-IDRI data set from Lung Image Database Consortium [2]. Buty *et al*. [4] extracted features from a pre-trained 2D multi-view CNN while encoding shape information though spherical harmonics (SH) to improve diagnostic accuracy from 79% (CNN) to 82% (CNN+SH). Hussein *et al*. [13] achieved a similar result, extracting deep features from a multi-view CNN then applying a Gaussian process regression strategy to achieve 82% accuracy. Li *et al*. [21] used a 3D deep CNN MTL learning approach, where attributes were predicted along with malignancy, and achieved a diagnosis accuracy of $80 - 83\%$, depending on the visual attributes chosen, although again no results were reported on the accuracy of predicting attributes.

Explainable lung cancer diagnosis: More recently, some deeper multi-crop [32], multi-scale [31], and denser dual-path multi-output [6] 3D CNNs, using methods such as curriculum learning [24] or gradient boosting machines [38] and complicated post-processing techniques [12], have been applied to push diagnosis accuracy to 87% – 92%. However, adding such techniques is beyond the scope of this work and would lead to an unwieldy enumeration of ablation studies necessary to understand the contributions between our proposed capsule architecture and such techniques. For a fair comparison in this study, we compare our method directly against *CapsNet* and explainable CNN approaches. Shen *et al*. [30] is one of the only works in the literature to attempt to create an interpretable framework by simultaneously predicting visual attribute scores along with malignancy. The authors used a deep dual-path dense 3D CNN to achieve an accuracy of 84%, however their results on individual attribute predictions were as low as 55%.

Capsule network-based medical diagnosis: It is worth noting, a number of recent studies have proposed using *CapsNet* for a variety of medical imaging classification tasks [1, 14, 33]. Nonetheless, since these methods nearly all follow the exact *CapsNet* architecture, or propose minor modifications which present nearly identical predictive performance [23]; hence, it is sufficient to compare only with *CapsNet* in reference to these works.

3. Capsules for encoding visual attributes

The goal of our proposed method is to model visual attributes using capsule neural networks for the important application domain of high-risk predictions. We apply our algorithm to CT lung data in order to provide the same explanations as radiologists for predicting malignancy, while simultaneously performing malignancy prediction and nodule segmentation/reconstruction. The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [2], described in more detail in Section 4, contains a collection of lung nodules with scores ranging from 1 – 5 across a set of visual attributes, indicating their relative appearance, and malignancy, as scored by up to four radiologists. These characteristics and scores are shown in Figure 2.

Our approach, referred to as *explainable capsules*, or *X-Caps*, was designed to remain as similar as possible to *CapsNet*, while allowing us to have more control over the visually-interpretable features learned by the capsule vectors. *CapsNet* already showed great promise when trained on the MNIST data set for its ability to model high-level visually-interpretable features. With this study, we examine the ability of capsules to model *specific* visual attributes within their vectors, rather that simply hoping some are learned successfully in the more challenging lung nodule data. As shown in Figure 3, *X-Caps* shares a similar overall structure as *CapsNet*, with the major differences being the addition of the supervised labels being provided for each dimension of the *X-Caps* vectors, the fully-connected layers for malignancy prediction, the reconstruction regularization also performing segmentation, and the modifications to the dynamic routing algorithm.

Building *X-Caps*: The first layer of our proposed explainable capsule network is a simple 2D convolutional layer which extracts the lowest-level features. Following this we form our primary capsules of 32 capsule types with $8D$ vector capsules. The primary capsules can be seen as either a convolution capsule layer with a single routing iteration or as grouping the feature maps of a convolutional layer and performing the non-linear squashing function from [28]. Following this, we form our X-Caps layer using a fully-connected capsule layer whose output is N capsule types, one for each of the visual-attributes we want to predict, by $16D$ length capsule vectors.

Unlike *CapsNet* where each of the parent capsules were dependant on one another (i.e. predicting digits or classes, if the digit is a five it cannot be a three), our parent capsules are (mostly) independent of each other (i.e. a nodule can score high or low in each of the attribute categories). For this reason, we needed to modify the dynamic routing algorithm presented in *CapsNet* to accommodate this significant difference. We will attempt to provide the high-level motivation before describing in algorithm in detail. The key change is the "routing softmax" employed by *CapsNet* forces the contributions of each child to send their information to parents in a manner which sums to one, which in practice effectively makes them "choose" a parent to send their information to. When the parents are co-dependant, this makes sense because we want to force a child to send all of their information to the three capsule or the five capsule, whichever one is present in the input. However, when computing prediction vectors for independent parents, we want a child to be able to contribute to all parent capsules for attributes which are present in the given input. With that motivation, the specific algorithm, which we unimaginatively call "routing sigmoid", is computed as follows,

$$r_{i,j} = \frac{\exp(b_{i,j})}{\exp(b_{i,j}) + 1}, \tag{1}$$

where $r_{i,j}$ are the routing coefficients determined by the dynamic routing algorithm for child capsule i to parent capsule j and the initial logits, $b_{i,j}$ are the prior probabilities that the prediction vector for capsule i should be routed to parent capsule j. Note the prior probabilities are initially set to 1 rather than 0 as in *CapsNet*, otherwise no routing could take place. The rest of the dynamic routing procedure follows the same as in [28].

Predicting malignancy from only visually-interpretable capsule vectors: In order to predict malignancy scores, we attach a fully-connected layer to our X-Caps attribute prediction vectors with output size equal to the number of scores (classes) with a softmax activation function. We wish to emphasize here, our final malignancy prediction is coming solely from the vectors whose magnitudes represent *visually-interpretable* feature scores. Every malignancy prediction score has a set of weights connected to the high-level attribute capsule vectors, and the activation from each tells us the exact contribution of the given visual attribute to the final malignancy prediction. Combining this information with the predicted scores for each attribute and our system can say, for any given nodule input, for example: *X-Caps* predicts with a 93% confidence value a malignancy score of 5 and the attributes 'sphericity' scoring highly and 'margin' scoring lowly contributed

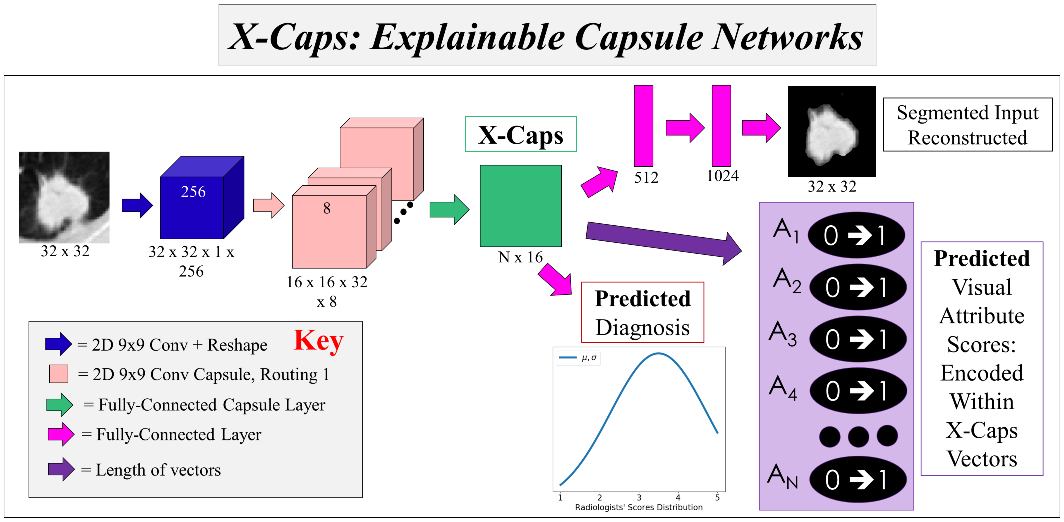
X-Caps: Explainable Capsule Networks
32 x 32
256
32 x 32 x 1 x 256
8
16 x 16 x 32 x 8
X-Caps
N x 16
512
1024
32 x 32
Segmented Input Reconstructed
Predicted Diagnosis
μ, σ
1
2
3
4
5
Radiologists' Scores Distribution
A1 0 → 1
A2 0 → 1
A3 0 → 1
A4 0 → 1
AN 0 → 1
Predicted Visual Attribute Scores: Encoded Within X-Caps Vectors
Key
= 2D 9x9 Conv + Reshape
= 2D 9x9 Conv Capsule, Routing 1
= Fully-Connected Capsule Layer
= Fully-Connected Layer
= Length of vectors

Figure 3: *X-Caps*: Explainable Capsule Networks. For a detected nodule, the proposed network (1) predicts high-level visual attributes of the nodule (purple box, bottom right), (2) segments the nodule and reconstruct the input image (pink, top right), and (3) diagnoses the nodule on a scale of 1 (benign) to 5 (malignant) based on the visually-interpretable high-level features encoded in the X-Caps (green square) capsule vectors (bottom middle). Note the malignancy diagnosis is not attempting to regress an average score, but rather is attempting to model the distribution of radiologists' scores in both mean and variance.

most strongly to this decision. During training, we directly compute the loss between the predicted malignancy score distribution and the distribution of radiologists' scores.

Loss and regularization: As in *CapsNet*, we also perform reconstruction of the input as a form of regularization. However, we extend the idea of regularization to perform a pseudo-segmentation, similar in nature to the reconstruction used by LaLonde and Bagci in [19]. Whereas in true segmentation, the goal is to output a binary mask of pixels which belong to the nodule region, in our formulation we attempt to reconstruct only the pixels which belong to the nodule region, while the rest are mapped to zero. More specifically, we formulate this problem as

$$R^{x,y} = I^{x,y} \times S^{x,y} \mid S^{x,y} \in \{0,1\}, \text{ and} \qquad (2)$$

$$\mathcal{L}_R = \frac{\gamma}{X \times Y} \sum_{x}^{X} \sum_{y}^{Y} \| R^{x,y} - O_r^{x,y} \|, \qquad (3)$$

where $\mathcal{L}_R$ is the supervised loss for the reconstruction regularization, γ is a weighting coefficient for the reconstruction loss, $R^{x,y}$ is the reconstruction target pixel, $S^{x,y}$ is the ground-truth segmentation mask value, and $O_r^{x,y}$ is the output of the reconstruction network, at pixel location (x, y), respectively, and X and Y are the width and height, respectively, of the input image. This adds another task to our MTL approach and an additional supervisory signal which can help our network distinguish visual characteristics from background noise. The malignancy prediction score, as well as each of the visual attribute scores also provide a supervisory signal in the form of

$$\mathcal{L}_a = \sum_{n}^{N} \alpha^n \| A^n - O_a^n \| \text{ and } \mathcal{L}_m = \beta \| M - O_m \|, \qquad (4)$$

where $\mathcal{L}_a$ is the combined loss for the visual attributes, A^n is the average of the attribute scores given by at minimum three radiologists for attribute n, N is the total number of attributes, α^n is the weighting coefficient placed on the n^{th} attribute, O_a^n is the network prediction for the score of the n^{th} attribute, $\mathcal{L}_m$ is the loss for the malignancy score, M is a Gaussian distribution over malignancy scores with mean and variance computed from scores given by at minimum three radiologists, O_m is the network prediction for the malignancy score distribution, and β is the weighting coefficient for the malignancy score. In this way, the overall loss for *X-Caps* is simply $\mathcal{L} = \mathcal{L}_m + \mathcal{L}_a + \mathcal{L}_R$. For simplicity, the values of each α^n and β are set to 1, and γ is set to $0.005 \times 32 \times 32 = 0.512$[1].

Uncertainty modeling of the visual scoring: All previous works in lung nodule classification follow the same strategy of averaging radiologists' scores for visual attributes and malignancy. To better model the uncertainty inherently present in the labels due to inter-observer variation, we propose a different approach: rather than simply

[1] Further tuning of these parameters could potentially lead to superior results but we did not have the computational resources to perform such an analysis for this study.

trying to regress the average of the values submitted by radiologists, or performing binary classification of these values rounded as above or below the score of 3, we attempt to predict the *distribution* of radiologists' scores. Specifically, for a given nodule where we have at minimum three radiologists' score values for each attribute and for malignancy prediction, we compute the mean and variance of those values and fit a Gaussian function to them, which is in turn used as the ground-truth for our classification vector. Nodules with strong inter-observer agreement produce a sharp peak, in which case wrong or unsure (*i.e.* low confidence score) predictions are severely punished. Likewise, for low inter-observer agreement nodules, we expect our network to output a more spread distribution and it will be punished for strongly predicting a single class label. This proposed approach allows us to model the uncertainty present in radiologists' labels in a way that no previous study has.

4. Experiments and results

Data, parameters, optimization: For our experiments, we used publicly available LIDC-IDRI data set [2]. The LIDC-IDRI includes 1018 volumetric CT scans, where each CT scan was interpreted by at most four radiologists by the LIDC-IDRI project team. Lung nodules were given scores by participating radiologists for each of six visual attributes and malignancy ranging from 1 to 5. For simplicity, and including malignancy indecision among radiologists, we excluded lung nodules from the consideration when their mean visual score was exactly 3. This left 1149 lung nodules to be evaluated (646 benign and 503 malignant). Table 1 shows the summary of visual score distribution of lung nodules evaluated by at least three radiologists.

Five-fold stratified cross-validation was performed to split the nodules into training and testing sets, with 10% of each training set set aside for validation and early stopping. All models were trained with a batch size of 16 using Adam [16] with an initial learning rate of 0.02 reduced by a factor of 0.1 after validation loss plateau. All code is implemented in TensorFlow and will be made publicly available. Consistent with the literature, predictions were considered correct if within ±1 of the radiologists' classification [12, 13].

Major findings: The experimental results summarized in Table 2 illustrate the prediction of visual attributes with the proposed *X-Caps* in comparison with a adapted version *CapsNet* and a deep dual-path dense 3D CNN (*HSCNN* [30]). To the best of our knowledge, *HSCNN* is the only other work in the literature which presents attribute-level predictions pursuant to creating explainable models through the modeling of high-level visual attributes for lung cancer diagnosis. Our results show that a simple 2D capsule network has the ability to model visual attributes far better than a state-of-the-art deep 3D CNN and even achieves better malignancy accuracy prediction. Further, we wish to emphasize the significance of *X-Caps* providing increased predictive performance *and* explainability over its *CapsNet* counterpart. This goes against the assumed trend in DL, illustrated with a symbolic plot in Figure 1, that explainability comes at the cost of predictive performance. We can observe this trend in the example of [38] achieving 90% accuracy in 2018 with a 3D Dual-Path network and [30] achieving 84% accuracy in 2019 with a 3D Dual-Path network but which also provides explainable predictions in the form of high-level visual attributes. While *X-Caps* is still slightly under-performing the best non-explainable models, it is reasonable to suspect that future research into deeper and more powerful capsule networks with tricks like residual or dense connections, normalization, and other novelties which were introduced to CNNs over the last several years, would allows explainable capsules to surpass these methods; we hope this study will promote such future works.

Table 1: Numbers within the table represent individual radiologists' scores. At the nodule level, there were 1149 nodules after removing those with less than three radiologists and those with mean score 3: 646 benign (< 3.0) and 503 malignant (> 3.0) nodule were used for training and testing in cross-validation.

	Visual Attribute Scores				
Attributes	**1**	**2**	**3**	**4**	**5**
subtlety	124	274	827	1160	1817
sphericity	10	322	1294	1411	1165
margin	174	303	512	1362	1851
lobulation	2394	924	475	281	128
spiculation	2714	789	336	174	189
texture	207	76	188	485	3246
malignancy	676	872	1397	658	599

Ablation studies: To analyze the impact of each component of our proposed approach, we performed ablation studies for: (1) learning the distribution of radiologists' scores rather than attempting to regress the mean value of these scores, (2) removing the reconstruction regularization from the network, and (3) performing our proposed "routing sigmoid" over the original "routing softmax" proposed in [28]. The results of each of these ablations is shown in Table 3 and we can see removing each component had a significant negative impact on our malignancy prediction results. This shows retaining the agreement/disagreement information among radiologists proved significantly useful, the reconstruction played a role in improving the network performance, and our proposed modifications to the dynamic routing algorithm were necessary for passing information from children to parents when the parent capsule types are independent. Lastly, although we defined the loss in Section 3 as mean squared error, we also experimented with cross-entropy, margin, and Kullback-Leibler diver-

Table 2: Prediction accuracy of visual attribute learning with capsule networks. Dashes (-) represent values which the given method could not produce. *X-Caps* significantly outperforms the state-of-the-art explainable method (*HSCNN*) at attribute modeling (the main goal of both studies), while also producing higher malignancy prediction scores, approaching state-of-the-art non-explainable methods performance.

	Attribute Prediction Accuracy %						**Malignancy**
	subtlety	sphericity	margin	lobulation	spiculation	texture	**Accuracy %**
Non-Explainable Methods							
3D Multi-Scale + RF [31]	-	-	-	-	-	-	86.84
3D Multi-Crop [32]	-	-	-	-	-	-	87.14
3D Multi-Out-DenseNet [6]	-	-	-	-	-	-	90.40
3D Dual-Path GBM [38]	-	-	-	-	-	-	90.44
CapsNet [28]	-	-	-	-	-	-	77.04
Explainable Methods							
3D Dual-Path-Dense *HSCNN* [30]	71.9	55.2	72.5	-	-	83.4	84.20
Proposed *X-Caps*	**90.39**	**85.44**	**84.14**	**70.69**	**75.23**	**93.10**	**86.39**

Table 3: Ablation studies for malignancy prediction accuracy: (1) regressing the mean score instead of predicting the distribution, (2) no reconstruction regularization, (3) using *CapsNet*'s "routing softmax" instead of the proposed "routing sigmoid", and (4) the proposed approach.

Mean Score	No Recon.	Routing Softmax	Proposed Method
83.09%	80.30%	80.69%	**86.39%**

gence loss functions. From our empirical analysis, these loss functions all performed comparably with each other, although a more systematic investigation would be needed to draw any firm conclusions. In our experience, capsule networks can be somewhat fragile and often some random initializations failed to converge to good performance. However, this might be more due to the small/shallow network size and its relation to the Lottery Ticket Hypothesis [8] than anything specific to capsules.

5. Discussions and concluding remarks

Deep leaning-generated predictions are mostly black-box in nature and not explainable; hence, not trusted by healthcare specialists. Available studies for explaining DL models, typically focus on *post hoc* interpretations of trained networks, rather than attempting to build-in explainability. This is the first study, to the best of our knowledge, for directly learning an interpretable feature space by encoding high-level visual attributes within the vectors of a capsule network to perform explainable image-based diagnosis. We approximate visually-interpretable attributes through individual capsule types, then predict malignancy scores directly based only on these high-level attribute capsule vectors, in order to provide malignancy predictions with explanations *at the human-level*, in the same language used by radiologists (i.e. visual attribute scores). Our proposed multi-task explainable capsule network, *X-Caps*, successfully approximated visual attribute scores significantly better than the previous state-of-the-art explainable diagnosis system, a deep dual-path dense 3D CNN (*HSCNN*), while also achieving higher diagnostic accuracy. While *X-Caps* still lags slightly in malignancy prediction performance compared with the best non-explainable CNNs, we are confident that as the field of capsule networks progresses, and similar advancements as those made with CNNs (*e.g.* residual/dense connections, normalization techniques) are made, more powerful capsule networks will be created to boost performance even further.

We found the proposed modifications to the reconstruction and dynamic routing algorithms both contributed significantly to the performance of our system, as well as the direct learning of radiologists' score distributions. While the direct modeling of the disagreement amongst experts aided our performance, its more significant contribution is in providing a means of estimating the confidence of network predictions via the variance in the output distribution, supervised to be large when expert disagreement is high and small when disagreement is low. We hope our work can provide radiologists with malignancy predictions which are explained via the same high-level visual attributes they currently use to explain malignancy, while also providing a meaningful confidence metric to advise when the results can be more trusted, thus allowing radiologists to quickly interpret and verify our predictions. Lastly, although we selected lung cancer diagnosis as a high-risk application domain for testing our method, we believe our approach should be generally applicable to any image-based classification task where high-level attribute information is available to provide explanations about the final prediction.

References

[1] Parnian Afshar, Arash Mohammadi, and Konstantinos N Plataniotis. Brain tumor type classification via capsule networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3129–3133. IEEE, 2018. 5

[2] S.G Armato III, G. McLennan, L. Bidaut, M. F McNitt-Gray, C. R Meyer, A. P Reeves, B. Zhao, D. R Aberle, C. I Henschke, E. A Hoffman, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2):915–931, 2011. 4, 5, 7

[3] Jason Bloomberg. Don't Trust Artificial Intelligence? Time To Open The AI 'Black Box'. http://www.forbes.com/sites/jasonbloomberg/2018/09/16/dont-trust-artificial-intelligence-time-to-open-the-ai-black-box/#6ceaf3793b4a, 11.16.2018. Forbes Magazine. 1

[4] Mario Buty, Ziyue Xu, Mingchen Gao, Ulas Bagci, Aaron Wu, and Daniel J Mollura. Characterization of lung nodule malignancy using hybrid shape and appearance features. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 662–670. Springer, 2016. 4

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016. 4

[6] Raunak Dey, Zhongjie Lu, and Yi Hong. Diagnostic classification of lung nodules using 3d neural networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 774–778. IEEE, 2018. 4, 8

[7] National Center for Health Statistics (US et al. Health, united states, 2016: with chartbook on long-term trends in health, 2017. 2

[8] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018. 8

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 4

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 4

[11] N Howlader, AM Noone, M Krapcho, D Miller, K Bishop, SF Altekruse, CL Kosary, M Yu, J Ruhl, Z Tatalovich, A Mariotto, DR Lewis, HS Chen, EJ Feuer, and KA Cronin. SEER Cancer Statistics Review, 1975-2013, National Cancer Institute. https://seer.cancer.gov/archive/csr/1975_2013/, 04.2018. 2

[12] Sarfaraz Hussein, Kunlin Cao, Qi Song, and Ulas Bagci. Risk stratification of lung nodules using 3d cnn-based multi-task learning. In *International Conference on Information Processing in Medical Imaging*, pages 249–260. Springer, 2017. 2, 4, 7

[13] Sarfaraz Hussein, Robert Gillies, Kunlin Cao, Qi Song, and Ulas Bagci. Tumornet: Lung nodule characterization using multi-view convolutional neural network with gaussian process. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 1007–1010. IEEE, 2017. 4, 7

[14] Tomas Iesmantas and Robertas Alzbutas. Convolutional capsule network for classification of breast cancer histology images. In *International Conference Image Analysis and Recognition*, pages 853–860. Springer, 2018. 5

[15] Pieter-Jan Kindermans, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: Patternnet and patternattribution. In *International Conference on Learning Representations (ICLR)*, 2018. 4

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[17] Devinder Kumar, Alexander Wong, and Graham W Taylor. Explaining the unexplained: A class-enhanced attentive response (clear) approach to understanding deep neural networks. In *IEEE Computer Vision and Pattern Recognition (CVPR) Workshop*, 2017. 4

[18] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *AAAI*, pages 2124–2132, 2017. 4

[19] Rodney LaLonde and Ulas Bagci. Capsules for object segmentation. *arXiv preprint arXiv:1804.04241*, 2018. 6

[20] Marianne Lehnis. Can We Trust AI If We Don't Know How It Works? http://www.bbc.com/news/business-44466213, 15.06.2018. BBC News. 1

[21] Xiuli Li, Yueying Kao, Wei Shen, Xiang Li, and Guotong Xie. Lung nodule malignancy prediction using multi-task convolutional neural network. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 1013424. International Society for Optics and Photonics, 2017. 4

[22] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015. 4

[23] Aryan Mobiny and Hien Van Nguyen. Fast capsnet for lung cancer screening. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 741–749. Springer, 2018. 5

[24] Aiden Nibali, Zhen He, and Dennis Wollersheim. Pulmonary nodule classification with deep residual networks. *International journal of computer assisted radiology and surgery*, 12(10):1799–1808, 2017. 4

[25] Weili Nie, Yang Zhang, and Ankit Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *International Conference on Machine Learning*, pages 3806–3815, 2018. 4

[26] Vyacheslav Polonski. People Don't Trust AI–Here's How We Can Change That.

http://www.scientificamerican.com/article/people-dont-trust-ai-heres-how-we-can-change-that/, 10.01.2018. Scientific American. 1

[27] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 1

[28] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866, 2017. 2, 4, 5, 7, 8

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 4

[30] Shiwen Shen, Simon X Han, Denise R Aberle, Alex A Bui, and William Hsu. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Systems with Applications*, 2019. 3, 4, 7, 8

[31] Wei Shen, Mu Zhou, Feng Yang, Caiyun Yang, and Jie Tian. Multi-scale convolutional neural networks for lung nodule classification. In *International Conference on Information Processing in Medical Imaging*, pages 588–599. Springer, 2015. 3, 4, 8

[32] Wei Shen, Mu Zhou, Feng Yang, Dongdong Yu, Di Dong, Caiyun Yang, Yali Zang, and Jie Tian. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, 61:663–673, 2017. 3, 4, 8

[33] Yan Shen and Mingchen Gao. Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization. In *International Workshop on Machine Learning in Medical Imaging*, pages 389–397. Springer, 2018. 5

[34] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 4

[35] National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011. 2

[36] David F Yankelevitz and James P Smith. Understanding the core result of the national lung screening trial. *New England Journal of Medicine*, 368(15):1460–1461, 2013. 2

[37] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 4

[38] Wentao Zhu, Chaochun Liu, Wei Fan, and Xiaohui Xie. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 673–681. IEEE, 2018. 2, 4, 7, 8