Artificial Intelligence in XPRIZE DeepQ Tricorder

Edward Y. Chang, Meng-Hsi Wu, Kai-Fu Tang, Hao-Cheng Kao, and Chun-Nan Chou
HTC AI Research & Healthcare
eyuchang@gmail.com

ABSTRACT

The DeepQ tricorder device developed by HTC from 2013 to 2016 was entered in the Qualcomm Tricorder XPRIZE competition and awarded the second prize in April 2017. This paper presents DeepQ's three modules powered by artificial intelligence: symptom checker, optical sense, and vital sense. We depict both their initial design and ongoing enhancements.

KEYWORDS

Artificial Intelligence, DeepQ, XPRIZE Tricorder, Medical IoT

ACM Reference format:
Edward Y. Chang, Meng-Hsi Wu, Kai-Fu Tang, Hao-Cheng Kao, and Chun-Nan Chou. 2017. Artificial Intelligence in XPRIZE DeepQ Tricorder. In *Proceedings of April, XPRIZE, 2013-17 (Tricorder Competition),* 8 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In Q1 2013, the XPRIZE foundation announced the Tricorder competition sponsored by Qualcomm [18]. The competition was to develop a portable device weighing less than five pounds that is able to accurately diagnose 12 common diseases[1] and capture 5 real-time vital signs[2]. The device was to be operated by a home user, independent of a health care worker or facility, and in a way that provides a compelling consumer experience [18]. The aim of this competition is to improve healthcare accessibility, especially in under privileged regions of the world.

The competition received 312 pre-registered entries from 38 countries in August 2013. Twenty-nine teams submitted their proposals by May 2014, and ten were selected in August 2014 to enter the qualifying round. The ten teams submitted their Tricorder devices for the first deadline in June 2015, and then for the final deadline in September 2016. In October 2016, six finalists were announced at the XPRIZE annual meeting. The two final winners were announced on April 12th, 2017 [19].

Dr. Chung-Kang Peng[3] founded the Dynamic Biomarker Group (DBG) in 2013 for the purpose of entering Qualcomm XPRIZE Tricorder competition. HTC Research & Healthcare was invited to be a collaborator and sponsor because of HTC's expertise in hardware integration, industry design, user experience research, and artificial intelligence [2]. In this article, we omit the first three elements that made DeepQ meeting both the weight constraint and user-experience requirements. We focus our presentation on the three modules[4] supported by artificial intelligence: symptom checker, optical sense, and vital sense.

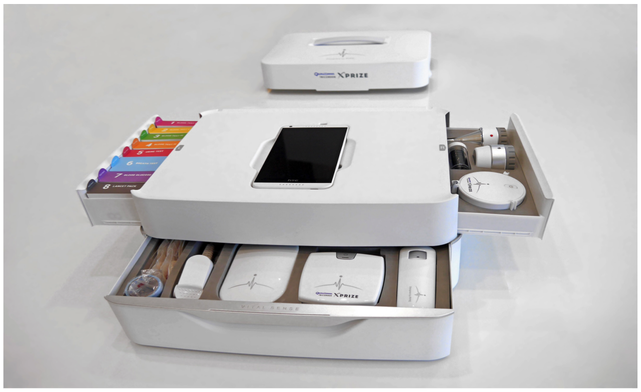

Figure 1: DeepQ Tricorder. DeepQ consists of four compartments. On the top is an HTC mobile phone, which acts as a data hub and also runs the symptom checker. The drawer on the right-hand-side contains optical sense. The drawer on the lower-front contains vital sense and breath sense. The left-hand-side drawer contains blood/urine sense.

(1) *Symptom checker.* Working with Dr. Peng's team at Harvard, HTC and Dr. Andrew Ahn developed a symptom checking module, which uses a Q&A session to recommend a user to perform the most relevant self-tests using DeepQ instruments. Symptom checker is designed to achieve good user experience and high accuracy. The DeepQ team has subsequently advanced symptom checker's accuracy using context-aware reinforcement learning, and expanded its disease coverage from 12 to all common diseases.
(2) *Optical sense.* Optical sense is composed of a wireless camera base and two lenses for diagnosing otitis media and

[1] The twelve required diseases are anemia, urinary tract infection,diabetes, atrial fibrillation, stroke, sleep apnea, tuberculosis, chronic obstructive pulmonary disease (COPD), pneumonia, otitis media, leukocytosis, and hepatitis A.

[2] Five vital signs include blood pressure, Electrocardiography (heart rate/variability), body temperature, respiratory rate, and oxygen saturation

[3] Dr. Peng, then was the founding Dean of College of Health Sciences and Technology at National Central University, on leave from Harvard Medical School. (Peng left NCU and returned to Harvard in August, 2014.)

[4] DeepQ consists of five modules, depicted in Figure 1. In additional to the three presented in this article, the other two are blood/urine sense, and breath sense.

Tricorder Competition, 2013-17, XPRIZE
© 2017 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

melanoma (two of the 12 diseases required by XPRIZE), respectively. The novel technology behind our diagnosis algorithm is a combined scheme of deep learning and transfer learning.

(3) *Vital sense.* XPRIZE requires monitoring five vital signs (heart rate/variability, respiration rate, blood pressure, oxygen saturation, and body temperature) continuously, and audits two diseases (atrial fibrillation and sleep apnea). We have developed our vital sense module to meet these requirements and derived heart rate variability, respiration rate, and blood pressure conditions around the basis of the electrocardiogram (ECG). In this article, we focus our presentation on the module that enables the atrial fibrillation detection, which is a result from the joint effort of HTC Research and Dr. Chung-Kang Peng [3]. To make further improvement in accuracy and to support wearable ECG patches, we recently completed a two year data annotation effort, which aims to provide the public a dataset with scale, diversity, and quality to facilitate artificial intelligence research on arrhythmia detection.

The remainder of this paper is organized as follows: Section 2 presents our reinforcement learning algorithm to predict potential diseases with limited inquiries and on-going enhancements. Section 3 depicts the transfer learning algorithm that supports otitis media and melanoma diagnosis. Section 4 describes our data collection endeavor for improving heart disease diagnosis. We offer our concluding remarks in Section 5.

2 CONTEXT-AWARE SYMPTOM CHECKER

To identify a patient's disease or absence of disease, the DeepQ kit uses a symptom checker to query a user and then recommend appropriate tests. Symptom checking first inquires a patient with a series of questions about their symptoms, and then attempts to predict some potential diseases. Two design goals of a symptom checker are high accuracy and good user experience. Good user experience consists of two requirements. First, the interactions between the symptom checker and patients must be intuitive. Second, the number of inquires should be minimal.

We have further enhanced the symptom checker module in the DeepQ kit and proposed CASC, which stands for Context-Aware Symptom Checker [24, 25]. CASC enhances the deployed symptom checker on our XPRIZE DeepQ kit in two aspects. First, CASC employs reinforcement learning (RL), a more effective algorithm than traditional Bayesian inference and decision trees [11] to model reward and penalty instead of using the theory of information gain [9, 10]. CASC in particular addresses the large-disease class challenge with RL when we expand the disease coverage from 12 (required by XPRIZE) to about 650 common diseases. Second, CASC considers a patient's contextual information including but not limited to *who*, *when* and *where* aspects.

2.1 Scalable Reinforcement Learning

To expand disease coverage from 10 to all common diseases, CASC faces two challenges. First, classification accuracy is expected to drop because of higher probability of similar symptoms resulting from different diseases. Second, the computation time to train a classifier increases. To tackle these challenges, we divide a body into P parts, and request a user to indicate which parts of her/his body exhibit what symptoms.

Table 1: The set $\mathcal{P}$ of anatomical parts.

head	neck	arm
chest	abdomen	back
pelvis	buttock	leg
skin	general symptoms	

Let $\mathcal{I}$, $\mathcal{D}$ and $\mathcal{P}$ denote the sets of symptoms, diseases and anatomical parts, respectively. Table 1 shows the set $\mathcal{P}$. Given a part $p \in \mathcal{P}$, we use $\mathcal{D}_p \subseteq \mathcal{D}$ to denote the set of diseases that is contained by p. For two parts p and q, the disease sets of these two parts may overlap, i.e., $\mathcal{D}_p \cap \mathcal{D}_q \neq \phi$. For example, the disease *food allergy* can happen in parts *neck*, *chest*, *abdomen*, and so on. We use $\mathcal{I}_p \subseteq \mathcal{I}$ to denote the set of symptoms that is involved in p. Similarly, $\mathcal{I}_p \cap \mathcal{I}_q \neq \phi$ for two parts p and q.

A diagnosis process is a sequential decision problem of an agent that interacts with a patient. At each time step, the agent inquires about a certain symptom $i \in \mathcal{I}$ of the patient. The patient then responds with true/false to the agent indicating whether the patient suffers from symptom i. In the meantime, the agent can integrate user responses over time steps to propose subsequent questions. At the end of the process, the agent receives a scalar reward if it can correctly predict the disease with a limited number of inquiries (every addition inquiry deduces a penalty from the reward). The goal of the agent is to maximize the reward. In other words, the goal is to correctly predict the patient disease $d \in \mathcal{D}$ by the end of the diagnosis process with limited number of inquiries. Reinforcement learning [23] is most suitable to address this problem, since RL can not only model the reward of accurate prediction, but also the penalty of asking each additional question.

Formally, in RL terms [23], our CASC agent receives a state s_t at time step t; then it chooses an action from a discrete action set $\mathcal{A}$ according to a policy π. In our formulation, $\mathcal{A} = \mathcal{I} \cup \mathcal{D}$. Based on the action $a_t \in \mathcal{A}$ chosen by the agent, it receives a reward r_t, where $r_t = 1$ if $a_t \in \mathcal{D}$ and a_t predicts the correct disease, or $r_t = -1$ if a_t is repeated; otherwise $r_t = 0$. The agent attempts to maximize the discounted return $R_t = \sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'}$, where $\gamma \in [0, 1]$ is a discount factor. The symptom checking process terminates when the action $a_t \in \mathcal{D}$.

The state-action Q-value function [23] is defined as $Q^{\pi}(s, a) = \mathbb{E}[R_t \mid s_t = s, a_t = a, \pi]$, referring to the expected return of performing an action a in a state s, along with a policy π. Since the Q-value can be divided into a current reward and a next-step Q-value using dynamic programming, it can be rewritten into the following recursive definition: $Q^{\pi}(s, a) = \mathbb{E}_{s'}[r + \gamma \mathbb{E}_{a' \sim \pi(s')}[Q^{\pi}(s', a')] \mid s, a, \pi]$. The optimal Q-value is defined as $Q^{*}(s, a) = \max_{\pi} Q^{\pi}(s, a)$. Also, it can be shown that the optimal Q-value obeys the Bellman equation: $Q^{*}(s, a) = \mathbb{E}_{s'}[r + \gamma \max_{a'} Q^{*}(s', a') \mid s, a]$. Lastly, the optimal deterministic policy can be defined by

$$\pi^{*}(s) = \arg\max_{a \in \mathcal{A}} Q^{*}(s, a).$$

2.2 Modeling Context

The contextual information includes but is not limited to three aspects about a patient: *who*, *when*, and *where*. The who aspect includes a person's demographic information (e.g., age and gender), hereditary (characterized by genetic data), and medical history. The when aspect can be characterized by a distribution of diseases in the time of year (e.g., season or month). The where aspect can be characterized by a distribution of diseases from coarse to fine location granularities (e.g., by country, city, and/or neighborhood). Any joint distributions of any combinations of the who, when and where aspects can be formulated and quantified into a context-aware model. This context-aware model is then utilized to perform three tasks in a symptom-checking framework to improve diagnosis accuracy:

1. Initialization.
   Supposing a patient does not provide any information, the symptom checker can initialize a symptom-checking dialog by inquiring about a symptom (or symptoms) that exhibits the highest probability based on the contextual information (joint context distribution).
2. Inquiries. Once one or some symptoms have been collected from the patient, the symptom-checking algorithm generates next symptom inquiries based on contextual information in additional to given answers to the previous inquiries.
3. Predictions. At the end of each symptom checking iteration, the symptom checker predicts potential diseases by jointly considering both contextual information and answered symptoms.

To model context, we modify the deep Q-network (DQN) [15], which is a function approximator of Q-functions. The DQN is essentially a neural network representing $Q(s, a; \theta)$ with parameters θ. To mimic real doctors who may have different specializations, we devise our model to be an ensemble model of different anatomical parts: $\mathcal{M} = \{m_p \mid p \in \mathcal{P}\}$. There are 11 anatomical parts in $\mathcal{P}$ as shown in Table 1. Each model m_p is a DQN specialized for symptom checking.

The model m_p accepts a state $s = [b^T, c^T]^T$, where b denotes the symptom statuses inquired by our model and c denotes the contextual information possessed by a patient. Formally, we describe the encoding scheme of b as follows: First, each symptom $i \in \mathcal{I}_p$ can be one of the following statuses: *true*, *false*, and *unknown*. We can use a three-element one-hot vector $b_i \in \mathbb{B}^3$ to encode the status of a symptom i. Second, the status of a symptom is determined based on the following rule. If a user responded yes to a symptom inquired by our model, that symptom is marked as true. On the other hand, if the user responded no, the symptom is marked as false. Symptoms not inquired by our model are marked as unknown. Finally, the vector b then concatenates all the symptom statuses into a Boolean vector, i.e., $b = [b_1^T, b_2^T, \dots, b_{|\mathcal{I}_p|}^T]^T$.

The other part of a state s is the contextual information c that currently comprises the *age*, *gender*, and *season* information of a patient. (Any other *who*, *when*, and *where* information can be easily incorporated.) Here, we denote $c = [c_{age}, c_{gender}, c_{season}]^T$. First, the age information $c_{age} \in \mathbb{N}$ is useful because some diseases have higher possibilities on babies whereas some have higher possibilities on adults. For example, meningitis typically occurs on children, and Alzheimer's disease on the elderly. Second, the gender information $c_{gender} \in \mathbb{B}$ is important because some diseases strongly correlate with gender. For example, females may have problems in uterus, and males may have prostate cancer. Third, the season information $c_{season} \in \mathbb{B}^4$ (a four-element one-hot vector) is also helpful because some diseases are associated with seasons.

Given a state $s = [b^T, c^T]^T$, our model m_p outputs the Q-value of each action $a \in \mathcal{A}_p$. In our definition, each action a has two types: an inquiry action ($a \in \mathcal{I}_p$) or a diagnosis action ($a \in \mathcal{D}_p$). If the maximum Q-value of the outputs corresponds to an inquiry action, then our model inquires the corresponding symptom to a user, obtains a feedback, and proceeds to the next time step. The feedback is incorporated into the next state $s_{t+1} = [b_{t+1}^T, c^T]^T$ according to our symptom status encoding scheme. Otherwise, the maximum Q-value corresponds to a diagnosis action. In the latter case, our model predicts the maximum-Q-value disease and then terminates.

Since each model m_p is independent, we can train eleven different models m_p simultaneously, with each is in charge of an anatomical part p. More specifically, we use the DQN training algorithm [15] proposed by Mnih et al. The loss function is defined as $L_j(\theta_j) = \mathbb{E}_{s,a,r,s'}[(y_j - Q(s, a; \theta_j)^2]$, where target $y_j = r + \gamma \max_{a'} Q(s', a'; \theta^-)$ is evaluated by a separate *target network* [15] $Q(s', a'; \theta^-)$ with parameters θ^-. The variable j is the index of training iteration. To improve training stability and convergence, the target network is fixed for a number of training iterations. The parameters θ can be updated by the standard backward propagation algorithm.

2.3 Preliminary Experimental Results

We used 650 out of 801 SymCat's symptom-disease database to conduct experiments. Our preliminary experiments did not yet consider contextual information. Whereas the details of data preparation and experiment setup are documented in [25], Table 2 shows the experimental results. We compare our P-part model, which we call the anatomical model, with the traditional monolithic model. The first column shows the number of diseases we selected for each anatomical part. The second column shows the total number of diseases among 11 anatomical parts. The column in *anatomical model* shows accuracies and average inquiry steps of our proposed ensemble model. The column in *monolithic model* shows the same statistics produced by a single model that supports the total number of diseases. From the table, we can see that our ensemble model achieves significantly higher accuracy than the traditional single model approach.

3 OPTICAL SENSE

Otitis media (OM) and melanoma are image-based diagnoses. With the success of deep learning, one may consider that the problem can easily be solved by employing deep learning with abundant training data. Unfortunately, in the domain of medicine, training data can be scarce, and approaches such as data augmentation are not applicable. In our case, the training data available to us are 1) 1,195 OM images collected by seven otolaryngologists at

Table 2: Experimental results on anatomical and monolithic models.

Task	$\|\mathcal{D}_p\|$	$\|\bigcup_p \mathcal{D}_p\|$	ω	Anatomical Model				Monolithic Model			
				Top 1	Top 3	Top 5	#Steps	Top 1	Top 3	Top 5	#Steps
Task1	25	73	1	48.12	59.01	63.23	7.17	39.42	43.13	44.97	1.64
Task2	50	136	2	34.59	41.58	45.08	7.06	27.49	29.16	30.28	1.48
Task3	75	196	3	25.46	29.63	31.82	5.98	2.08	2.81	4.19	3.42
Task4	100	255	4	21.24	24.56	26.15	6.94	0.73	1.46	2.19	3.37

Cathay General Hospital[5], Taiwan [20] and 2) 200 melanoma images from the PH^2 dataset [14]. Transfer representation learning is a plausible alternative, which can remedy the insufficient training data issue. The common practice of transfer representation learning is to pre-train a convolutional neural network (CNN) on a very large dataset (called source domain) and then to use the pre-trained CNN either as an initialization or a fixed feature extractor for the task of interest (called target domain). The source domain from which representations are transfered to our two target diseases is ImageNet [4].

What are symptoms or characteristics of OM and melanoma? OM is any inflammation or infection of the middle ear, and treatment consumes significant medical resources each year [17]. Several symptoms such as redness, bulging, and tympanic membrane perforation may suggest an OM condition. Color, geometric, and texture descriptors may help in recognizing these symptoms. However, specifying these kinds of features involves a hand-crafted process and therefore requires domain expertise. Often times, human heuristics obtained from domain experts may not be able to capture the most discriminative characteristics, and hence the extracted features cannot achieve high detection accuracy. Similarly, melanoma, a deadly skin cancer, is diagnosed based on the widely-used dermoscopic "ABCD" rule [22], where A means asymmetry, B means border, C color, and D different structures. The precise identification of such visual cues relies on experienced dermatologists to articulate. Unfortunately, there are many congruent patterns shared by melanoma and nevus, with skin, hair, and wrinkles often preventing noise-free feature extraction.

3.1 Transfer Representation Learning

We started with unsupervised codebook construction. On the large ImageNet dataset, we learned the representation of these images using a variant of deep CNN, AlexNet [12], which contains eight neural network layers. The first five layers are convolutional and the remaining three are fully connected. Different hidden layers represent different levels of abstraction concepts. We utilized AlexNet in Caffe [8] as our foundation to build our encoder to capture generic visual features.

For each image input, we obtained a feature vector using the codebook. The information of the image moves from the input layer to the output layer through the inner hidden layers. Each layer is a weighted combination of the previous layer and stands for a feature representation of the input image. Since the computation is hierarchical, higher layers intuitively represent higher concepts. For images, the neurons from lower levels describe rudimentary perceptual elements like edges and corners, whereas the neurons from higher layers represent aspects of objects such as their parts and categories. To capture high-level abstractions, we extracted transfer-learned features of OM and melanoma images from the fifth, sixth and seventh layers, denoted as pool5(P5), fc6 and fc7, respectively.

Once we had transfer-learned feature vectors of the 1, 195 collected OM images and 200 melanoma images, we performed supervised learning by training a support vector machine (SVM) classifier [1]. We chose SVMs to be our model since it is an effective classifier widely used by prior works. Using the same SVM algorithm lets us perform comparisons with the other schemes solely based on feature representation. As usual, we scaled features to the same range and found parameters through cross validation. For fair comparisons with previous OM works, we selected the radial basis function (RBF) kernel.

To further improve classification accuracy, we experimented with two feature fusion schemes, which combine OM features hand-crafted by human heuristics (or model-centric) in [20] and our melanoma heuristic features with features learned from our codebook [21]. In the first scheme, we combined transfer-learned and hand-crafted features to form fusion feature vectors. We then deployed the supervised learning on the fused feature vectors to train an SVM classifier. In the second scheme, we used the two-layer classifier fusion structure proposed in [21]. In brief, in the first layer we trained different classifiers based on different feature sets separately. We then combined the outputs from the first layer to train the classifier in the second layer.

3.2 Experimental Results

Two sets of experiments were conducted to validate our idea. In this subsection, we first report OM classification performance by using our proposed transfer representation learning approach, followed by our melanoma classification performance. Then, we elaborate the correlations between images of ImageNet classes and images of disease classes by using a visualization tool to explain why transfer representation learning works.

For fine-tuning experiments, we performed a 10-fold cross-validation for OM and a 5-fold cross-validation for melanoma to train and test our models, so the test data are separated from the training dataset. We applied data augmentation, including random flip, mirroring, and translation, to all the images.

For the setting of training hyperparameters and network architectures, we used mini-batch gradient descent with a batch size of 64 examples, learning rate of 0.001, momentum of 0.9 and weight decay of 0.0005. To fine-tune the AlexNet model, we replaced the

[5] The dataset was used under a strict IRB process. The dataset was deleted by April 2015 after our experiments had completed.

fc6, fc7 and fc8 layers with three new layers initialized by using a Gaussian distribution with a mean of 0 and a std of 0.01. During the training process, the learning rates of those new layers were ten times greater than that of the other layers.

Our 1, 195 OM image dataset encompasses almost all OM diagnostic categories: normal; AOM: hyperemic stage, suppurative stage, ear drum perforation, subacute/resolution stage, bullous myringitis, barotrauma; OME: with effusion, resolution stage (retracted); COM: simple perforation, active infection. Table 3 compares OM classification results for different feature representations. All experiments were conducted using 10-fold SVM classification. The measures of results reflect the discrimination capability of the features.

The first two rows in Table 3 show the results of human-heuristic methods (hand-crafted), followed by our proposed transfer-learned approach. The eardrum segmentation, denoted as 'seg', identifies the eardrum by removing OM-irrelevant information such as ear canal and earwax from the OM images [20]. The best accuracy achieved by using human-heuristic methods is around 80%. With segmentation (the first row), the accuracy improves 3% over that without segmentation (the second row).

Rows three to eight show results of applying transfer representation learning. All results outperform the results shown in rows one and two, suggesting that the features learned from transfer learning are superior to that of human-crafted ones.

Interestingly, segmentation does not help improve accuracy for learning representation via transfer learning. This indicates that the transfer-learned feature set is not only more discriminative but also more robust. Among three transfer-learning layer choices (layer five (pool5), layer six (fc6) and layer seven (fc7)), fc6 yields slightly better prediction accuracy for OM. We believe that fc6 provides features that are more general or fundamental to transfer to a novel domain than pool5 and fc7 do. (Section 3.3 presents qualitative evaluation and explains why for OM fc6 is ideal.)

We also directly used the 1, 195 OM images to train a new AlexNet model. The resulting accuracy was only 71.8%, much lower than applying transfer representation learning. This result confirms our hypothesis that even though CNN is a good model, with merely 1, 195 OM images (without the ImageNet images to facilitate feature learning), it cannot learn discriminative features.

Two fusion methods, combining both hand-crafted and transfer learning features, achieved a slightly higher OM-prediction F1-score (0.9 over 0.895) than using transfer-learned features only. This statistically insignificant improvement suggests that hand-crafted features do not provide much help.

Finally, we used OM data to fine-tune the AlexNet model, which achieves the highest accuracy. For fine-tuning, we replaced the original fc6, fc7 and fc8 layers with the new ones and used OM data to train the whole network without freezing any parameters. In this way, the leaned features can be refined and are thus more aligned to the targeted task. This result attests that the ability to adapt representations to data is a critical characteristic that makes deep learning superior to the other learning algorithms.

Table 3: OM classification experimental results

	Method	Accuracy(std)	F_1-Score
1	Heuristic w/ seg	80.11%(18.8)	0.822
2	Heuristic w/o seg	76.19%(17.8)	0.79
3	Transfer w/ seg (pool5)	87.86%(3.62)	0.89
4	Transfer w/o seg (pool5)	88.37%(3.41)	0.894
5	Transfer w/ seg (fc6)	87.58%(3.45)	0.887
6	Transfer w/o seg (fc6)	88.50%(3.45)	0.895
7	Transfer w/ seg (fc7)	85.60%(3.45)	0.869
8	Transfer w/o seg (fc7)	86.90%(3.45)	0.879
9	Feature fusion	89.22%(1.94)	0.90
10	Classifier fusion	89.87%(4.43)	0.898
11	Fine-tune	90.96%(0.65)	0.917

3.3 Qualitative Evaluation - Visualization

In order to investigate what kinds of features are transferred or borrowed from the ImageNet dataset, we utilized a visualization tool to perform qualitative evaluation. Specifically, we used an attribute selection method, SVMAttributeEval [7] with Ranker search, to identify the most important features for recognizing OM and melanoma. Second, we mapped these important features back to their respective codebook and used the visualization tool from Yosinski et al. [26] to find the top ImageNet classes causing the high value of these features. By observing the common visual appearances shared by the images of the disease classes and the retrieved top ImageNet classes, we were able to infer the transferred features.

Fig. 2 demonstrates the qualitative analyses of four different cases: the Normal eardrum, acute Otitis Media (AOM), Chronic Otitis Media (COM) and Otitis Media with Effusion (OME), which we will now proceed to explain in turn. First, the normal eardrum, nematode and ticks are all similarly almost gray with a certain degree of transparency, features that are hard to capture with only hand-crafted methods. Second, AOM, purple-red cloth and red wine have red colors as an obvious common attribute. Third, COM and seashells are both commonly identified by a calcified eardrum. Fourth, OME, oranges, and coffee all seem to share similar colors. Here, transfer learning works to detect OM in an analogous fashion to how *explicit similes* are used in language to clarify meaning. The purpose of a simile is to provide information about one object by comparing it to something with which one is more familiar. For instance, if a doctor says that OM displays redness and certain textures, a patient may not be able to comprehend the doctor's description exactly. However, if the doctor explains that OM presents with an appearance similar to that of a seashell, red wine, orange, or coffee colors, the patient is conceivably able to envision the appearance of OM at a much more precise level. At level fc6, transfer representation learning works like finding similes that can help explain OM using the representations learned in the source domain (ImageNet).

Our transfer representation learning experiments consist of the following five steps:

(1) Unsupervised codebook construction: We learned a codebook from ImageNet images, and this codebook construction is "unsupervised" with respect to OM and melanoma.

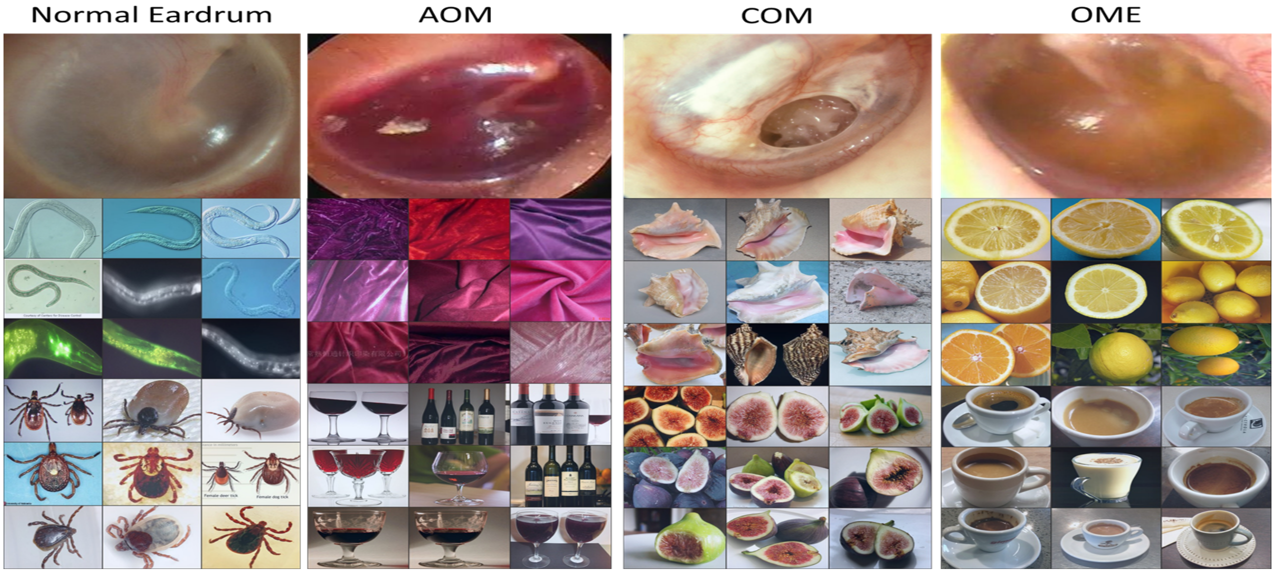

Figure 2: The visualization of helpful features from different classes corresponding to different OM symptoms(from left to right: Normal eardrum, AOM, COM, OME)

(2) Encode OM and melanoma images using the codebook: Each image was encoded into a weighted combination of the pivots in the codebook. The weighting vector is the feature vector of the input image.
(3) Supervised learning: Using the transfer-learned feature vectors, we then employed supervised learning to learn two classifiers from the 1,195 labeled OM instances or 200 labeled melanoma instances.
(4) Feature fusion: We also combined some heuristic features of OM (published in [21]) and ABCD features of melanoma with features learned via transfer learning.
(5) Fine tuning: We further fine-tuned the weights of the CNN using labeled data to improve classification accuracy.

As we will show in the remainder of this section, step four does not yield benefit, whereas the other steps are effective in improving diagnosis accuracy. In other words, these two disease examples demonstrate that features modeled by domain experts or physicians (the model-centric approach) are ineffective. The data-driven approach of big data representation learning combined with small data adaptation is convincingly promising.

3.4 Observations

Our experiments on transfer learning provided three important insights on representation learning.

(1) Low-level representations can be shared. Low-level perceptual features such as edges, corners, colors, and textures can be borrowed from some source domains where training data are abundant. After all, low-level representations are similar despite different high-level semantics.
(2) Middle-level representations could be correlated. Analogous to explicit similes used in language, an object in the target domain can be "represented" or "explained" by some source domain features. In our OM visualization, we observed that that a positive OM may display appearances similar to a seashell or amoeba with colors of red wine, oranges, or coffee, features learned and transferred from the ImageNet source domain.
(3) Representations can adapt to a target domain. Even though in the small data training situation the amount of data is insufficient to learn effective representations by itself, given representations learned from some big-data source domains, the small data of the target domain can be used to align (e.g., re-weight) the representations learned from the source domains to adapt to the target domain.

4 VITAL SENSE

One of the 12 required diseases to be diagnosed by the XPRIZE Tricorder competition is *atrial fibrillation (AF)*. AF is one common type of serious rhythmic *arrhythmia*, which results from a very fast and irregular contraction of the atria. An arrhythmia condition is often preceded by the events of ectopic heartbeats and has four main types: premature beats, supraventricular, ventricular and bradyarrhythmias.

Interpreting electrocardiograms (ECGs) is an inexpensive and noninvasive way for cardiologists to assess the cardiac conduction system and diagnose arrhythmia. A physician can determine the abnormal cardiac activities at each part of the heart by measuring the time intervals between fiducial points on the ECG. Measuring the amount of a wave that travels through the heart muscle can help infer which part of the heart being hypertrophic, and precisely which type of arrhythmia has occurred.

It is certainly desirable to be able to detect arrhythmia automatically with non-invasive wearables and a continuously running computer algorithm. However, two fundamental challenges prevent the existing solutions [13] to be highly accurate. First, the variability of a patient's underlying wave morphology, cardiac rhythms and artifacts conspire to make the analysis difficult. This difficulty is further elevated by the inter-patient diversity. Second, we again face the same *small data* problem that we come up against in detecting OM and melanoma (Section 3). Most research groups that developed and evaluated their automatic classification algorithms based on the MIT-BIH Arrhythmia Database (MITDB), which is the first generally freely available standard test material for arrhythmia detection analysis [16]. Though the MITDB database is regarded as the most representative and invaluable database for developing automated arrhythmia detectors and is listed as one of the evaluation standards by the Association for the Advancement of Medical Instrumentation (AAMI), the database consists of merely 48 ECG records from 47 subjects. Each record is slightly over 30 minutes in length and contains two ECG leads.

While the MITDB database has been an invaluable benchmark, the small number of unique individuals in this database characterizes the limited variability and insufficiency for exhaustive studies. Furthermore, the MITDB database has a significant limitation that 60% of its ECG recordings were obtained from inpatients. In other representative and public arrhythmia datasets such as MIT-BIH Supraventricular Arrhythmia Database (SVDB) [6] and St. Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database (INCART) [5], the number of unique patients is also in the two-digit scale size and disease annotations are not audited by a cardiologist. Importantly, all aforementioned databases did not consider outpatient ECG measures, which may contain countless motion artifacts and data loss. The deterioration of signal quality usually impose significant difficulty on reliable arrhythmia detection. Contemplating these weaknesses, we in 2015 began to construct a new dataset that is about 10 folds the size of current MITDB database to fit three main purposes:

(1) *Scale*: Large-scale data in terms of greater unique patient numbers
(2) *Diversity*: Inpatient and outpatient ECG measures with detailed beat-by-beat, rhythm and heartbeat fiducial points annotations. We intentionally consider three different activities and motion intensities, namely lying down, sitting, and walking. These modes can facilitate training a classifier for wearable ECG patches.
(3) *Quality*: Complement the MITDB database in the exhaustive development and evaluation of the arrhythmia detector

4.1 The DeepQ Arrhythmia Database

The DeepQ Arrhythmia Database (DeepQ) is being developed with Taipei Veteran General Hospital, Taiwan. We attempt to include a large variety of realistic arrhythmic ECG recordings that are observed in clinical practice and outpatients. The entire data collection process complies with the human participants guideline and regulation of the Institutional Review Board (IRB)[6]. During a clinical visit or cardiac examination, patients were asked if they were willing to participate in the data collection. Participating patients were explained through the guidance and the IRB informed consents were obtained before their ECG examinations. We used a water-resistant, non-invasive single-lead ECG device adopted from our Tricorder XPRIZE vital sense module for this data collection. DeepQ ECG patch was worn on the left chest in the Modified Lead-II configuration. We also placed the Philip 1810 series holter on the participant along with our DeepQ ECG recording device for comparison throughout the collection process. All patients were instructed to rest for at least five minutes during the initial setup and between each activity session. Throughout the ECG measurement, each patient was engaged in a sequence of three five-minute activities, namely lying down, sitting, and walking and contributed three recordings to the database. This protocol ensures a smooth transition between different activity intensities. In the walking session, participants were allowed to walk freely around the facility to mimic the recordings in outpatient situations.

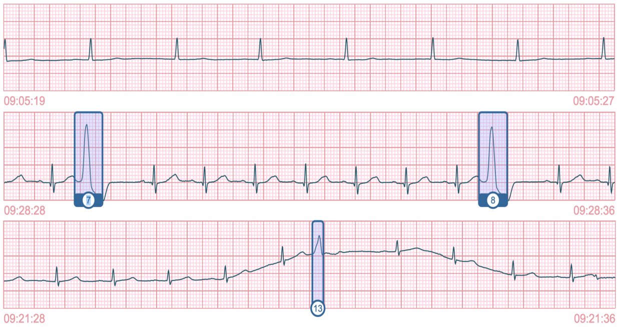

Figure 3: From top to bottom: normal sinus rhythm, two PVC beats, a PVC event during a walking session.

The same ECG module as we designed in the Qualcomm Tricorder XPRIZE competition was used for this data collection. The ECG data was sampled at a frequency of 250 Hz with with 24-bit resolution and wirelessly transferred via BLE to a remote smartphone receiver for temporary storage. The recorded ECG signals were then uploaded to our private server and later compared with Philips 1810-series holters for quality assurance by the certified cardiographic technicians before any annotations. Figure 3 shows three ECG excerpts from the DeepQ database. All records were annotated using a web interface tool designed for this task. Labeling works were initially carried out by a group of certified cardiographic technicians and then verified by a cardiologist. To ensure quality and consistency, annotation rules were devised; the cardiographic technicians were guided through the use of web tool and supervised by a senior technician. There are three label categories in this database: beat-by-beat, rhythm episodes and heartbeat fiducial points. The beat-by-beat annotation protocol is compatible with the AAMI recommendations. Along with the beat-by-beat class annotation, each heartbeat's P, QRS, and T fiducial points are also marked, if present. A strip is marked for rhythm-level labels in which the beginning and the end of the strip correspond to the onset and offset of an abnormality segment.

Currently this database contains 897 annotated records from 299 unique patients. Each record is about five minutes in duration, from one activity and includes a compact clinical summary with technical information about the recording. Comparing the DeepQ dataset with the other three public databases (MITDB, SVDB, and INCART), DeepQ's number of unique subjects and records outnumber them. DeepQ not only has at least twice the amount of unique patients but also has more heartbeats in the AAMI supraventricular ectopic beats

[6] IRB reference number: 2015-03-001A

Table 4: Database details

Database	#Records	#Subjects	#SVEB Records	#VEB Records	#AF Records	AAMI Heartbeat class N	SVEB	VEB	Sample Rate (Hz)	Duration (min)	#Channels
DeepQ	**897**	**299**	**71**	**125**	**48**	**303233**	**4158**	**8616**	**250**	**15**	**1**
MITDB	48	47	32	37	11	90125	2781	7009	360	30	2
SVDB	78	78	73	67	-	161902	12083	9897	128	30	2
INCART	75	32	18	69	3	153517	1958	19991	257	30	12

(SVEB) and ventricular ectopic beats (VEB) classes compared to the MITDB database. Table 4 summarizes this comparison. In addition to the beat-level classes, the rhythm-level arrhythmia collection includes 48 unique atrial fibrillation (AF) or atrial flutter (AFL) cases, four first degree atrioventricular block (1°AVB), three second degree atrioventricular block (2°AVB), one paroxysmal supraventricular tachycardia (PSVT), six supraventricular tachycardia (SVT), one sinoatrial block (SAB), two ventricular tachycardia (VT), and one junctional rhythm cases.

5 CONCLUDING REMARKS

DeepQ made the milestone set forth by the XPRIZE foundation by putting together an integrated device that can drastically improve healthcare accessibility. This article presented three modules: symptom checker, optical sense, and vital sense that are powered by the latest artificial intelligence techniques including deep learning, transfer learning, and reinforcement learning. In particular, the medical domain encounters the challenge of small data. Though we showed that our active reinforcement learning and transfer representation learning algorithm to be promising, much work remains to further improve disease diagnosis accuracy. The DeepQ arrhythmia database is a marked milestone achieved by our two year annotation effort. We plan to make this DeepQ dataset publicly available to advance medical research in developing outpatient, mobile arrhythmia detectors.

ACKNOWLEDGEMENT

We would like to thank the following physicians for their contributions to DeepQ modules relevant to symptom checker, optical sense, and vital sense: Dr. Andrew Ahn (Beth Israel Deaconess Hospital), Dr. Shih-Lin Chang (Veteran General Hospital), Dr. Chung-Kang Peng (Harvard University), and Dr. Pa-Chun Wang (Cathay General Hospital).

REFERENCES

[1] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2, 3 (2011), 27.

[2] Edward Y Chang. 2011. *Foundations of large-scale multimedia information management and retrieval: mathematics of perception.* Springer Science & Business Media.

[3] Madalena Costa, Ary L Goldberger, and C-K Peng. 2005. Multiscale entropy analysis of biological signals. *Physical review E* 71, 2 (2005), 021906.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 248–255.

[5] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet. *Circulation* 101, 23 (2000), e215–e220.

[6] Scott David Greenwald, Ramesh S Patil, and Roger G Mark. 1990. Improved detection and classification of arrhythmias in noise-corrupted electrocardiograms using contextual information. In *Computers in Cardiology 1990, Proceedings.* IEEE, 461–464.

[7] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning* 46, 1-3 (2002), 389–422.

[8] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia.* ACM, 675–678.

[9] Ron Kohavi. 1996. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA.* 202–207. http://www.aaai.org/Library/KDD/1996/kdd96-033.php

[10] Igor Kononenko. 1993. Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence* 7, 4 (1993), 317–337. https://doi.org/10.1080/08839519308949993

[11] Igor Kononenko. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine* 23, 1 (2001), 89–109. https://doi.org/10.1016/S0933-3657(01)00077-X

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems.* 1097–1105.

[13] Eduardo José da S Luz, William Robson Schwartz, Guillermo Cámara-Chávez, and David Menotti. 2016. ECG-based heartbeat classification for arrhythmia detection: A survey. *Computer methods and programs in biomedicine* 127 (2016), 144–164.

[14] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. 2013. PH 2-A dermoscopic image database for research and benchmarking. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE.* IEEE, 5437–5440.

[15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. 2013. Playing Atari with Deep Reinforcement Learning. *CoRR* abs/1312.5602 (2013). http://arxiv.org/abs/1312.5602

[16] George B Moody and Roger G Mark. 2001. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine* 20, 3 (2001), 45–50.

[17] American Academy of Pediatrics Subcommittee on Management of Acute Otitis Media et al. 2004. Diagnosis and management of acute otitis media. *Pediatrics* 113, 5 (2004), 1451.

[18] Qualcomm. 2013. XPRIZE Tricorder site. http://tricorder.xprize.org. (2013).

[19] Qualcomm. 2017. XPRIZE Tricorder Winning Teams. http://tricorder.xprize.org/teams. (2017).

[20] Chuen-Kai Shie, Hao-Ting Chang, Fu-Cheng Fan, Chung-Jung Chen, Te-Yung Fang, and Pa-Chun Wang. 2014. A hybrid feature-based segmentation and classification system for the computer aided self-diagnosis of otitis media. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE.* IEEE, 4655–4658.

[21] Chuen-Kai Shie, Chung-Hisang Chuang, Chun-Nan Chou, Meng-Hsi Wu, and Edward Y. Chang. 2015. Transfer representation learning for medical image analysis. *IEEE EMBC* (2015), 711–714.

[22] W Stolz, A Riemann, AB Cognetta, L Pillet, W Abmayr, D Holzel, P Bilek, F Nachbar, and M Landthaler. 1994. ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma. In *European Journal of Dermatology.* 521–527.

[23] R.S. Sutton and A.G. Barto. 1998. *Reinforcement learning: An introduction.* Vol. 116. Cambridge Univ Press.

[24] Kai-Fu Tang and Edward Y. Chang. 2017. A Context-Aware Symptom Checking Framework. In *Patent filed with HTC.*

[25] Kai-Fu Tang, Hao-Cheng Kao, Chun-Nan Chou, and Edward Y. Chang. 2016. Inquire and Diagnose: Neural Symptom Checking Ensemble using Deep Reinforcement Learning. In *Proceedings of NIPS Workshop on Deep Reinforcement Learning.*

[26] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).