Hierarchical Multi-Task Learning for Healthy Drink Classification

Homin Park
BIGHEART
National University of Singapore
Singapore
bighp@nus.edu.sg

Homanga Bharadhwaj
Department of Computer Science
Indian Institute of Technology Kanpur
India
homangab@cse.iitk.ac.in

Brian Y. Lim
Department of Computer Science
National University of Singapore
Singapore
brianlim@comp.nus.edu.sg

***Abstract*—Recent advances in deep convolutional neural networks have enabled convenient diet tracking exploiting photos captured with smartphone cameras. However, most of the current diet tracking apps focus on recognizing solid foods while omitting drinks despite their negative impacts on our health when consumed without moderation. After an extensive analysis of drink images, we found that such an absence is due to the following challenges that conventional convolutional neural networks trained under the single-task learning framework cannot easily handle. First, drinks are amorphous. Second, visual cues of the drinks are often occluded and distorted by their container properties. Third, ingredients are inconspicuous because they often blend into the drink. In this work, we present a healthy drink classifier trained under a hierarchical multi-task learning framework composed of a shared residual network with hierarchically shared convolutional layers between similar tasks and task-specific fully-connected layers. The proposed structure includes two main tasks, namely sugar level classification and alcoholic drink recognition, and six auxiliary tasks, such as classification and recognition of drink name, drink type, branding logo, container transparency, container shape, and container material. We also curated a drink dataset, Drink101, composed of 101 different drinks including 11,445 images overall. Our experimental results demonstrate improved classification precision compared to single-task learning and baseline multi-task learning approaches.**

I. Introduction

Drinking sugar-sweetened and alcoholic drinks may be pleasing, but it is detrimental to our health when consumed without moderation. Many studies, including ones from the School of Public Health at Harvard University[1], have shown that frequent consumption of unhealthy drinks could lead to the development of obesity epidemic, metabolic syndrome, fatty liver diseases, and brain damages [1], [2]. In fact, dietitians and clinicians emphasize the importance of making healthier drink choices to curb their detrimental effects [3]. However, many people have misconceptions about what is healthy and what is not [4]. For example, fruit juices are often perceived as healthy when most of them are loaded with as much sugar as a sugary soft drink.

One of the promising ways to help make healthier drink choices is to use photo-based mobile diet logging and tracking applications that increase users' awareness of their behaviors, thus promoting healthy behavior changes [5], [6]. However, none of the state-of-the-art applications present a feasible drink classifier accurate enough to successfully categorize the healthiness of drinks. After an extensive analysis of drink images and related studies, we found that such an absence is due to the following properties of drinks.

Unlike solid foods, drinks conform to the shape of their containers and thus have higher visual variance. Also, visual cues are often distorted, occluded, and limited by various container properties. For example, bottles may be colored or covered with branding materials distorting and occluding visual cues. In case of a ceramic cup or a mug, visual cues are only marginally evident through the top of the container. Such properties contribute to higher intra-class variances (same drink looks very different) and lower inter-class variances (different drinks look very similar) compared to solid food items. Furthermore, the nutrient composition of a drink is almost impossible to correctly identify when automated systems are used without user intervention because ingredients often blend into the drink. Despite the recent advances in convolutional neural networks (CNN), such properties are extremely challenging to address when a classifier is trained under single-task learning (STL) framework [7], [8].

To address this challenging computer vision problem, we consider employing a multi-task learning (MTL) framework which has been demonstrated to be feasible across many applications of machine learning, e.g., computer vision [7], [9]–[13] and natural language processing [14]. Unlike STL, MTL shares representations between different, yet related, tasks through shared network architectures in order to help improve the performance of the primary tasks [15]. In fact, MTL architectures, employing CNNs, have successfully demonstrated their feasibility in addressing diet related problems, such as cooking recipe retrieval [7] and dietary assessment [13]. Most of these works employed a simple structure of shared convolutional layers and task-specific fully-connected layers. However, numerous studies have emphasized the importance of considering the hierarchical relationship between different tasks since all tasks are not equally related and do not share similar representation [16]–[18].

In this work, we propose a healthy drink classifier using hierarchical multi-task learning (HMTL) framework composed

[1]https://www.hsph.harvard.edu/nutritionsource/healthy-drinks/sugary-drinks/

Fig. 1: Healthy drink classification example showing the inferred attributes of the drink, which can be used to determine its healthiness.

of a shared residual network and shared hierarchical convolutional layers between similar tasks and task-specific fully-connected layers. Two primary tasks, namely sugar level classification and alcoholic drink recognition, are used to categorize the healthiness of a drink considering their significant impact on our health. In addition, we introduce six auxiliary tasks, including classification and recognition of drink name, drink type, branding logo, container transparency, container shape, and container material. Examples of inferred drink attributes are shown in Figure 1.

Note that we do not attempt to explicitly measure the amount of sugar or alcohol in drinks, which cannot be done with a system solely relying on drink images. Furthermore, the ultimate decision on whether a drink is healthy or not should be made by the end users and domain experts. The significance of drinking unhealthy beverages varies by individual, e.g., drinking a Coke can be dangerous for diabetic patients while others might not feel the same effect.

We also curated a drink dataset, Drink101, composed of 101 different drinks including a total of 11,445 images. Our experimental results demonstrate improved classification precision compared to single-task learning and baseline multi-task learning approaches. In summary, our main contributions are as follows:

- We propose multiple drink attributes and exploit their hierarchical relationships through shared convolutional layers, demonstrating accurate classification performances
- We propose a healthy drink classifier exploiting hierarchical multi-task learning (HMTL) framework
- We curate a drink dataset called Drink101 and make it available to the public

II. Related Work

The proposed healthy drink classifier encompasses several strands of research, such as diet recognition, multi-task learning, and hierarchical relationship of different tasks.

A. Diet Recognition and Dataset

Numerous diet recognition technologies exploiting conventional single-task CNN (SCNN) have been proposed in recent years. This includes several mobile applications, such as Nibble [6] and Im2Calories [19]. Various other works have also demonstrated accurate diet recognition models, such as DeepFood [20], FoodAi[2], NutriNet [8], and DietLens [7]. However, none of these technologies can accurately recognize drinks due to two major issues. First, drinks have high visual variations [7], [8]. Second, drinks are often omitted or treated as a sub-type of foods from various datasets. For example, Food101 [21], FoodLog [22] and UEC256 [23] do not include any drink items while VIREO [24], and NutriNet [8] cover very small number of drink categories. ChinFood1000 [25] claims to include 91 drink categories out of 1,000 classes, but this is a proprietary dataset. Therefore, motivated by the significance of making healthier drink choices and the absence of feasible technologies, we propose a healthy drink classifier and a drink image dataset, Drink101.

B. MTL and Hierarchical Relationship of Tasks

By simultaneously learning different yet related tasks using shared deep neural network layers, Multi-task learning (MTL) has been demonstrated to be effective across various machine learning applications, including computer vision [10]–[12] and natural language processing [14]. In fact, Chen et. al. demonstrated that food recipes can be better retrieved using MTL with related tasks, namely food categorization, ingredient recognition, and cooking method recognition [7]. Furthermore, a dietary assessment technique was introduced very recently exploiting MTL capabilities [13].

However, all tasks are not equally related and do not share exactly the same representations [15], [17]. Multiple studies have demonstrated the feasibility and necessity of embedding the hierarchical relationship of different tasks to address complex and challenging machine learning problems. For example, a hierarchical Bayesian model exploiting a latent task hierarchy was proposed recently [17]. A Joint Many-Task Model proposed the use of pre-defined hierarchical architecture composed of several natural language tasks as a model for MTL [16]. HD-CNN was proposed to demonstrate the feasibility of hierarchical deep convolutional neural networks for large scale visual recognition [26]. Furthermore, Lu et. al. introduced a fully-adaptive feature sharing technique for deep multi-task networks which automatically identifies feasible hierarchical relationships among different tasks [9].

III. Multi-task Deep Learning

Due to the limited visual cues in drink images, training a CNN on the task of healthy drink classification would entail significant parameter tuning, optimization and large number of iterations to achieve acceptable performances. However, such a fine-tuned model is prone to overfitting. In our work, we formulate healthy drink classification as a multi-task deep

[2]http://foodai.org/

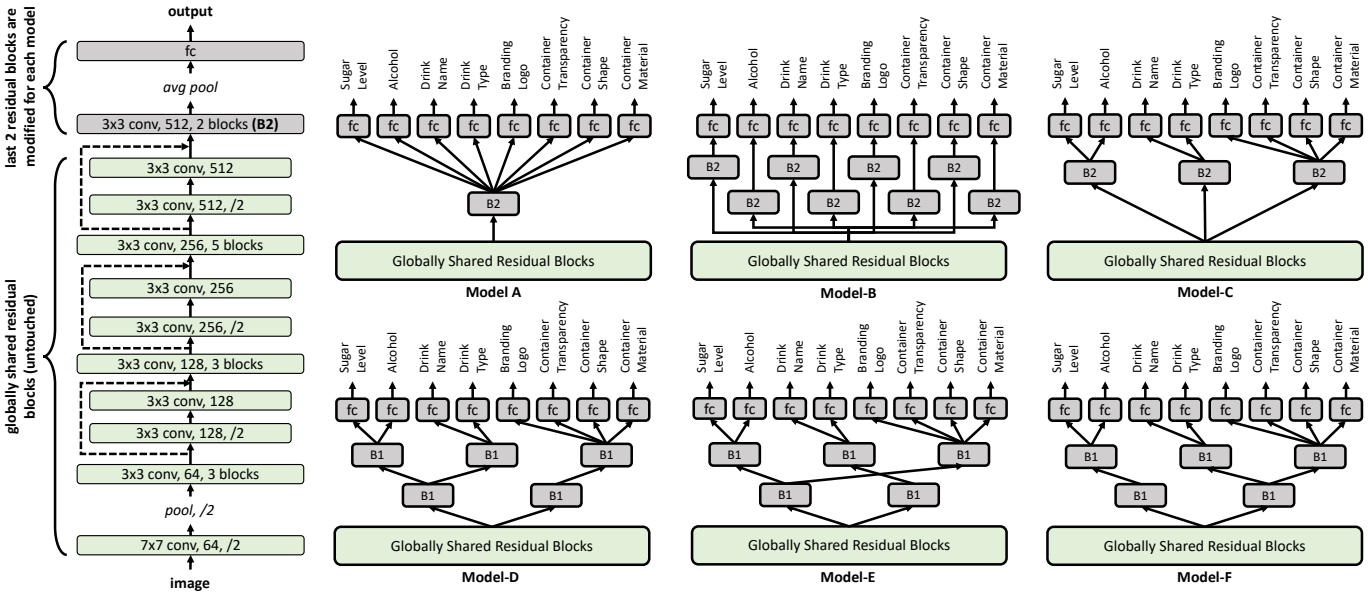


Fig. 2: The six architectures we considered for MTL. These are modifications of ResNet34 (shown on the left) which is described in [27]. 'fc' denotes a fully connected layer, B1 denotes one residual block consisting of two convolutional layers and B2 denotes two residual blocks.

learning problem where the two primary tasks are coupled with six auxiliary tasks for simultaneous learning. Details of each task and the proposing HMTL architecture are described in the following sub-sections.

A. The Primary and the Auxiliary Tasks

Our work employs two primary tasks, sugar level classification and alcoholic drink recognition, which can be used to determine whether a drink is healthy or not. To improve the classification and recognition performances, we identify and utilize six auxiliary tasks, namely, drink name classification, drink type classification, branding logo recognition, container transparency classification, container shape classification, and container material classification. These auxiliary tasks provide additional contexts to the visual cues and additional knowledge regarding the relationships of different drinks.

The drink name is the most fine-grained measure for classifying the healthiness of the drink. Training a STL model on drink names is the typical way of training food classification models [21]. However, due to a large number of drink name categories and limited visual cues, a STL framework trained on drink name alone will have poor classification accuracy. Instead, we leverage the representation learned from this task to improve the performance of the primary tasks.

The drink type is of coarser granularity than drink name. This generalizes the categorization of drinks into common drink types, such as coffee, dairy, juice, and soda among others. We determined the categorization from a thematic analysis after analyzing thousands of drink names. Therefore, this represents knowledge external to the drink images and facilitates the model to draw correlations by external supervisory information over and above the visual cues obtained from drink image alone. Just as generalizations of drink names into these types can help people quickly infer the drink healthiness, we hypothesize that providing these knowledge associations will help to improve the speed of training and model accuracy for drink healthiness.

The third auxiliary task is of branding logo recognition. This task is not to identify what is written in the logo but only to infer if there is a logo in the drink image. The intuition for this task is that the model will learn to draw correlations such as the fact that logos are usually present in branded drinks and such drinks typically tend to have high sugar level and/or are alcoholic.

The fourth, fifth and sixth auxiliary tasks pertain to the attributes of the container in which the drink is kept. Some drinks are typically served or consumed in specific containers. For example, wine is typically served in a wineglass, coffee in a mug, and cans typically contain soda. In fact, recognizing a wineglass should increase the likelihood that the drink is alcoholic. This association can help with identifying the drink name or type and consequently the drink healthiness. Unlike drink type, drink container is determined from the visual cues in the image. We manually generated labels of drink container attributes, namely transparency, shape, and material in our image dataset. In this work, we did not specify generic names of different containers but described their appearance through three tasks including transparency, shape, and material. For example, a glass beer mug would be labeled as transparent, cup with handle, and glass.

B. Architecture Design

From the above discussion, it is clear that all the eight tasks are different, yet have a high correlation to support other tasks.

This motivates the use of private layers for each task and some shared layers for all the tasks in a CNN. The influence of one task over another is modeled through the learning of parameters in the shared layers. In potential CNN architectures, the initial convolutional layers should be globally shared among all task because we need the fundamental visual features to be available to all the tasks of the network. This helps ensure that no task remains significantly under-learned (i.e., insufficiently trained) during training.

In our case, since the eight tasks widely differ in the number of class labels, so the rate of successful training on the eight tasks is different. Hence, we have at least one block of fully connected (FC) layers privately for each task in order that the training of one task does not become a bottleneck to the training of the other tasks.

After analyzing numerous studies on multi-task CNN (MCNN), we postulate that having private convolutional layers for each task (or shared by only a subset of all tasks) is essential because different shapes or visual features may be significant for specific tasks. Guided by this intuition, we design various MTL architectures, incorporating hierarchical design, as described in Figure 2. In the hierarchical structures (Models D, E and F), the rationale for shared convolutional layers lower in the hierarchy is to capture *global* features relevant for larger groups of tasks while the convolutional layers shared among tasks higher up in the hierarchy are expected to incorporate *local* features relevant for smaller groups of tasks. We group tasks by leveraging semantic relations between them as described below.

Model-A has only one private FC layer for each task and shares all convolutional layers. In addition to private FC layers, Model-B includes two private residual blocks for each task, because we hypothesize that different shapes or visual features may be significant for specific tasks. For example, the drink container classifier may learn a cup handle concept, but this is not relevant for drink types or drink names; a drink type classifier may learn an effervescence (bubbles) texture to recognize soda or beer, but this would be less relevant for the drink container classifier.

Model-C adopts a hybrid-split structure by sharing the final residual block between semantically related tasks. In this model, two primary tasks (sugar level classification and alcoholic drink recognition) share two residual blocks (B2) since both aim to detect ingredients in the drink for which visual cues are not conspicuous. Separate B2s are also shared between drink name and drink type classification tasks, and among branding logo recognition and the three tasks of container property identification. We extend this intuition in models D, E and F by having hierarchically shared convolution layers among different task groups. Indeed, we demonstrate empirically that the hierarchical architectures (D, E, and F) achieve the best performance.

C. Implementation

Residual networks (ResNet) [27] have been particularly successful in image recognition tasks, and we implement our system to be composed of shared residual blocks. The primary reason for their high performance is that by virtue of the short-circuit connections, a sufficiently deep architecture can also be trained well. The backbone architecture on which we build our model is ResNet34. In Figure 2, the architecture up to the globally shared residual blocks is exactly the same as ResNet34. Since all the eight tasks are multi-class classification problems, we compute the cross-entropy loss of individual tasks and combine them to form the global loss. Let N denote the total number of images during training and T denote the total number of tasks, which is 8 in our case. The global loss function is then defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_t L_t \tag{1}$$

where, $\gamma_1, ..., \gamma_8$ are weighing parameters for the individual losses such that $\sum_{t=1}^{8}\gamma_t = 1$. We train our network to back-propagate gradients from the global loss function into all the eight branches and the shared layers. Let p_c^i denote the probability of observation i (i.e. the i^{th} image) being classified in category c and let y_c^i be a binary label whose value is 1 if class c is indeed correct for image i. Now, L_t can be defined as:

$$L_t = -\sum_{c=1}^{C} y_c^i log(p_c^i) \tag{2}$$

The formulation of each $L_t \forall t \in [1, ..., 8]$ is similar. With the above loss functions, we perform experiments with the proposed architectures in Figure 2 as described in the Experiments section. In our implementation, in order to leverage the benefits of a residual network, instead of having the basic block as one convolutional layer, ReLU, and MaxPool combination, we use a residual block [27] as the smallest building unit in the architecture. As illustrated in Figure 2, B2 indicates two residual blocks while B1 refers to one residual block.

IV. Drink Image Dataset: Drink101

A well-curated image dataset is essential for evaluating a healthy drink classifier, to implement a sound training routine, and to extract representative information of suitable features. In this section, we describe how we curated and labeled our drink image dataset, Drink101[3].

A. Curating Drink101

To determine a list of drinks to collect images of, we searched for most frequently consumed beverages in the world and collected images from various search engines, including Google, Baidu, Flickr, and Instagram. Next, we analyzed collected images to identify the prevalent and representative container types for each drink. For example, we found that Coke was frequently served in a can and a transparent bottle with a branding logo. To ensure the quality of the dataset, we filtered out blurry images, consisted of multiple types of drinks and containers, and were composed of other objects

[3]Please contact the authors for the access

that are more prominent than the drink itself (e.g., person's face or solid food item). Note that our dataset also includes numerous professionally photographed images. After going through the curation processes, our dataset finally includes 101 unique drink names (130 unique combinations of drink name and container type) and 11,445 images overall.

B. Drink Healthiness

The healthiness of each drink category is described using two attributes, namely sugar level and alcoholic. According to the guidelines proposed by School of Public Health at Harvard University[4], drinks can be tagged with one of three labels, namely red, yellow, and green, representing drinks with high sugar level (>1g of sugar per ounce) to be drunk "sparingly or infrequently," drinks with low sugar level (<1g/oz) as a "better choice," and drinks with little or no sugar (∼0g/oz) as the "best choice," respectively. Additionally, we include another label named alcoholic due to its negative impact on our health. Other than having high calories, alcohol can cause minor issues, such as drowsiness, headaches, and anemia, or major problems, including coma, nerve damage, liver damage, and other heart-related diseases, when consumed for a long time [2]. We determine the sugar level and alcoholic nature of each drink from the nutrient dataset of MyFitnessPal[5] and Health Promotion Board of Singapore[6].

C. Statistics of Drink101

Each category is labeled with various attributes, namely sugar level (no, low, and high), alcohol (present and absent), drink type (coffee, dairy, juice, soda, tea, water, beer, spirit, and wine), branding logo (present and absent), container transparency (opaque, semi-transparent, and transparent), container shape (bottle, can, cup, cup with handle, cup with stem, and pack), and container material (aluminum, ceramic, glass, paper, plastic). The statistics for each attribute are as follows. Sugar level – no sugar 34.6%, low sugar 13.8%, and high sugar 51.5%. Alcohol – present 28.5% and absent 71.5%. Drink type – coffee 6%, dairy 12%, juice 15%, soda 21%, tea 11%, water 4%, beer 3%, spirit 20%, and wine 4%. Branding logo – present 38% and absent 62%. Container transparency – opaque 21%, semi-transparent 7%, and transparent 72%. Container shape – bottle 23%, can 8%, cup 45%, cup with handle 12%, cup with stem 10%, and pack 2%. Container material – aluminum 8%, ceramic 8%, glass 63%, paper 2%, and plastic 19%.

V. Experiments

In this section, we evaluate the accuracy of our hierarchical multi-task learning (HMTL) framework for healthy drink classification with respect to several baselines. We train our model on both single tasks and all combinations of multiple tasks and observe the results from each mode of training. Since our primary task is to classify the healthiness of drinks, we present our results mainly in terms of sugar level classification and alcoholic drink recognition and report the accuracy on auxiliary tasks (drink type, drink name, branding logo, and drink container attributes) where appropriate. We show that our HMTL framework helps in making better predictions on the two primary tasks as compared to learning them alone through an SCNN or non-hierarchical MTL architectures. We trained our proposed and baseline models on 80% (randomly selected images) of the full Drinks101 dataset and tested on the remaining 20%.

A. Comparison against STL Architectures

To verify whether MTL is suitable for healthy drink classification, we first train STL models for each task and compare them against MTL models shown in Figure 2. Our experiment results shown in Table I indicate that the precision@k (P@k) value for each STL model is relatively lower than that of MTL models. Specifically, P@1 values of the primary tasks (sugar level and alcohol) are higher in all MTL models compared to the respective STL counterparts. For example, there is a 3% increase in P@1 for the sugar level classification task from STL-Sugar to Model-A, and a 12% increase for Model-F. In case of alcoholic drink recognition there is a 5% increase in P@1 from STL-Alcohol to Model-A and a 16% increase for Model-F. This substantiates our claim for the use of MTL architectures. Another interesting observation from Table I is the fact that the precision values of all other (auxiliary) tasks also increase from STL to the MTL setting although by varying magnitudes. This suggests that the tasks of sugar-level detection and alcoholic drink recognition have not improved *at the cost of* the other tasks, but *along with* them.

B. Comparison among MTL Architectures

To identify the most feasible MTL architecture for healthy drink classification, we train and evaluate all models shown in Figure 2. In addition, we also implement and evaluate a MTL architecture proposed by Chen et. al. [28] to compare its feasibility against our models. Differing from our work, Chen et. al. used a VGG-16 network with task-specific fully-connected layers. We denote this architecture as MTL-VGG in the remaining part of this section. Furthermore, we also implemented an automatic tensor decomposition method (Tensor-Fact) proposed in [29] which is one of the state-of-the-art adaptive HMTL frameworks. A number of other HMTL solutions are not included in our evaluation since we cannot make apples-to-apples comparisons due to the different problem formulations. Note that all performance measurements shown in Table I are drawn from optimal $\gamma_t \forall t \in [1, ..., 8]$ arrived at after conducting a grid search in the range [0,1].

As evident in Table I, the P@1 values of Model-F for both sugar-level classification and alcoholic drink recognition are the highest. Overall, the hierarchical structures D, E and F perform much better than the other models. This performance enhancement is evident not just on the two primary tasks but also on the auxiliary tasks. We attribute this improved performance to better collaboration between tasks. The shared

[4]https://www.hsph.harvard.edu/nutritionsource/healthy-drinks/

[5]https://www.myfitnesspal.com/

[6]http://focos.hpb.gov.sg/eservices/ENCF/

TABLE I: Comparison of Precision@k for various tasks with different architectures. For STL, a separate network per task was used for training. MTL-VGG is a VGG based architecture [28] trained on all eight tasks. Tensor-Fact is an adaptive MTL method proposed in [29], which we apply to the ResNet34 architecture. Model-A has private FC layers alone for each task, model B has private residual block for each task, model C has private residual blocks for groups of tasks. Models D, E and F are hierarchical structures, with different hierarchies of residual blocks shared between tasks. All the architectures are described in Figure 2.

Model	Sugar Level	Alcohol	Drink Name		Drink Type		Logo	Trans-parency	Shape	Material
	P@1	P@1	P@1	P@2	P@1	P@2	P@1	P@1	P@1	P@1
STL-Sugar	0.77	x	x	x	x	x	x	x	x	x
STL- Alcohol	x	0.74	x	x	x	x	x	x	x	x
STL-DrinkName	x	x	0.33	0.36	x	x	x	x	x	x
STL-DrinkType	x	x	x	x	0.43	0.50	x	x	x	x
STL-Logo	x	x	x	x	x	x	0.62	x	x	x
STL-Transparency	x	x	x	x	x	x	x	0.33	x	x
STL-Shape	x	x	x	x	x	x	x	x	0.33	x
STL-Material	x	x	x	x	x	x	x	x	x	0.34
Tensor-Fact	0.83	0.84	0.39	0.44	0.41	0.51	0.72	0.38	0.36	0.38
MTL-VGG	0.79	0.76	0.36	0.38	0.44	0.53	0.68	0.34	0.35	0.36
Model-A	0.80	0.79	0.35	0.39	0.43	0.55	0.67	0.36	0.37	0.35
Model-B	0.81	0.81	0.37	0.44	0.45	0.51	0.67	0.35	0.46	0.37
Model-C	0.83	0.84	0.36	0.45	0.46	0.57	0.62	0.37	0.45	0.37
Model-D	0.85	0.86	0.36	0.48	0.46	0.58	0.68	0.38	0.47	0.38
Model-E	0.86	0.85	0.37	0.45	0.44	0.49	0.73	0.36	0.42	0.49
Model-F	**0.89**	**0.90**	0.41	0.43	0.47	0.54	0.69	0.42	0.44	0.43

residual blocks in the higher layers are responsible for extracting *global* features relevant for large groups of tasks while the penultimate residual blocks are responsible for extracting *local* features relevant for specific small groups of tasks. Figure 3 shows the convergence of each model during training. We see that Model-F convergences to the lowest validation loss during training, indicating that it is the fastest to train in addition to converging at a better optima.

From Table I, it is also evident that Models B and C perform better than Model-A re-affirming our belief that there are task-specific relevant features which are best extracted by having private residual blocks for groups of tasks (in Model-C) or for each task (in Model-B). In Model-A, there is no scope for incorporation of task-specific features since only FC layers are private. The improved performance of Model-C compared to Model-B is indicative of the fact that the semantic grouping of tasks for sharing residual blocks helps in better leveraging of features from drink images between tasks.

We trained MTL-VGG on the drink dataset with tuned hyper-parameters, but found it to have lower precision for the primary tasks than Models B, C, D, E and F. This might be because of superior learning ability of residual nets to better propagate the gradients in light of insufficient contextual cues. Another reason could be the sharing of only-fully connected layers, which might not be a problem with food images due to abundant visual cues but fails to perform well with drinks.

Another interesting result is that the Tensor-Fact model performs much worse compared to model F on all the tasks. This is primarily because the automatic tensor decomposition method proposed in [29] is not scalable to large networks

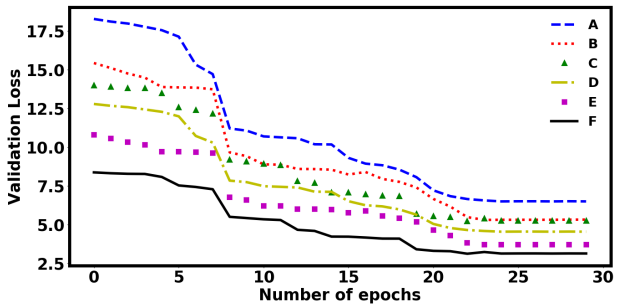


Fig. 3: Comparison of convergence of each of the MTL models, i.e. Model-A to Model-F, during training. The plot shows variation of test error with number of training epochs.

(for example a residual network as used in our work) due to combinatorial explosion of many possible connections. Moreover, the method does not distinguish between primary and auxiliary tasks and hence does not have tunable parameters for appropriately weighing the tasks.

C. Impact of Hyper-parameter γ

The feasibility of MTL models is further demonstrated by analyzing the impact of hyper-parameter γ. First, to understand the performance gap between STL and MTL, we trained all MTL models by weighing each task equally, assigning same γ for all the tasks. As shown in Table II, the sub-optimal performance of MTL models, derived from untuned hyper-parameters, still exceeds conventional STL models. Another interesting observation is that the performance of the hierarchi-

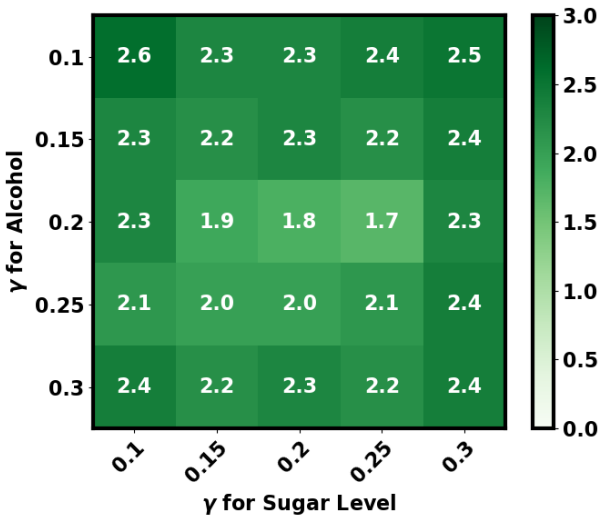
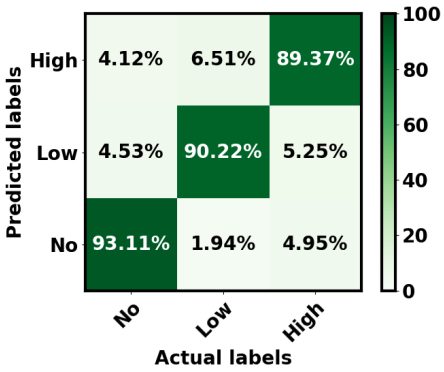


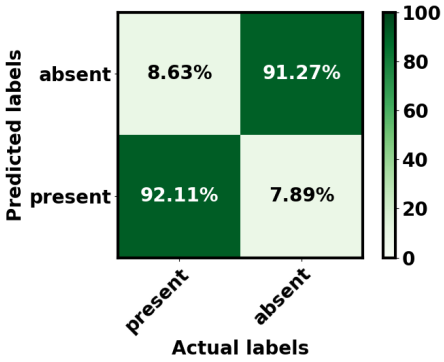
Fig. 4: Comparison of the overall test loss in Model-F after 30 epochs of training. This is the sum of individual cross-entropy loss over all tasks. The x-axis shows the value of γ corresponding to sugar level classification and the y-axis shows the value of γ corresponding to alcoholic drink recognition with the other γ's set equal such that $\sum_t^8 \gamma_8 = 1$.

cal structures D, E and F are better than the other three naive models (A, B and C). This indicates that the better modelling prowess of hierarchical MTL architectures is independent of parameter tuning.

To further improve the performance of the HMTL models, we searched for an optimal hyper-parameter setting. Figure 4 shows the overall cross-entropy error of Model-F when varying the γ with respect to the two primary tasks. Other γ values are kept equal such that the sum equals 1 for each training instance. So, for the coordinate (0.1,0.2) in Figure 4, the loss value shown is for the setting $\gamma_3 = 0.2$, $\gamma_4 = 0.1$ and $\gamma_1 = \gamma_2 = \gamma_5 = \gamma_6 = \gamma_7 = \gamma_8 = 0.12$. We observe from the heatmap that the overall loss (on the validation set upon convergence of the model) increases when the γ corresponding to a particular primary task is either increased to a very high value (max=1) or decreased to a very low value (min=0). As a γ increases and approaches 1 for a task, the framework approaches the STL setting for that task because the contributions of the other tasks are de-emphasized. As a γ decreases and approaches 0 for a task, that task is not trained effectively as the corresponding loss is too low for effective gradient updates. We found that, for effective training, the γ values of both the primary tasks should be in the range between 0.15 and 0.25. Since we have a large number of hyperparameters (total 8), we show variation with only those corresponding to the primary tasks in the heatmap in Figure 4.



(a) Sugar Level



(b) Alcohol

Fig. 5: Confusion matrices generated with Model-F for sugar-level prediction and alcoholic drink recognition with each region % normalized per prediction.

TABLE II: Comparison of Precision@1 for all MTL models with γ set equal for all tasks, i.e. $\gamma_t = 0.125$ for each task t

	Sugar Level	Alcohol	Drink Name	Drink Type	Logo	Trans-parency	Shape	Material
	P@1	P@1	P@1	P@1	P@1	P@1	P@1	P@1
A	0.77	0.76	0.34	0.43	0.66	0.37	0.37	0.36
B	0.79	0.78	0.38	0.44	0.65	0.37	0.46	0.38
C	0.81	0.81	0.37	0.46	0.64	0.38	0.46	0.39
D	0.84	0.84	0.36	0.47	0.68	0.39	0.47	0.45
E	0.86	0.86	0.37	0.45	0.73	0.35	0.45	0.52
F	0.88	0.87	0.42	0.48	0.69	0.44	0.45	0.45

D. Confusion Matrices for the Primary Tasks

To understand how our model successfully classifies drink healthiness and makes errors, we created the confusion matrices as shown in Figure 5b. These confusion matrices of our Model-F for sugar-level classification and alcoholic drink

recognition tasks for all drinks. While overall results are reasonable accurate, our model tends to overestimate the low sugar level as high by 6.51%. This has the consequence of accusing consumers of drinking highly sweetened beverages even though they are drinking low sugar drinks.

VI. Conclusion

Recognizing the importance of moderating unhealthy drink consumption, we have developed a healthy drink classifier exploiting hierarchical multi-task learning framework. To address the challenges of recognizing drinks, we trained our classifier with two primary tasks, namely sugar level classification and alcoholic drink recognition, and six auxiliary tasks. We also presented the Drink101 dataset which is a drink image dataset curated for healthy drink classification labeled with multiple drink attributes. We investigated different multi-task residual network architectures to exploit the knowledge-base of hierarchical relationships among different tasks. Our work can be used to help make healthier drink choices. However, when someone's belief in a drink's healthiness is challenged by our model, e.g., we say orange juice is high in sugar while the user believes it is healthy, there must be an effective strategy of explanations to persuade and nudge the user. In fact, as a future work, we are planning to expand our framework by addressing the nuances of explainability which can be made feasible by exploiting the multiple attributes used in our model.

Acknowledgment

We thank Wendi Ren for assistance in curating Drink101 dataset. This work was carried out in part at the BIGHEART research center and funded by the National Research Foundation, Singapore, Ministry of Education, Singapore. The Titan Xp used for this research was donated by the NVIDIA Corporation.

References

[1] G. Bray and B. Popkin, "Calorie-sweetened beverages and fructose: what have we learned 10 years later," *Pediatric obesity*, vol. 8, no. 4, pp. 242–248, 2013.

[2] O. A. Parsons, "Alcohol abuse and alcoholism." 1996.

[3] D. F. Tate, G. Turner-McGrievy, E. Lyons, J. Stevens, K. Erickson, K. Polzien, M. Diamond, X. Wang, and B. Popkin, "Replacing caloric beverages with water or diet beverages for weight loss in adults: main results of the c hoose h ealthy o pt i ons c onsciously e veryday (choice) randomized clinical trial–," *The American journal of clinical nutrition*, vol. 95, no. 3, pp. 555–563, 2012.

[4] A. Pan, V. S. Malik, T. Hao, W. C. Willett, D. Mozaffarian, and F. B. Hu, "Changes in water and beverage intake and long-term weight changes: results from three prospective cohort studies," *International journal of obesity*, vol. 37, no. 10, p. 1378, 2013.

[5] F. Cordeiro, E. Bales, E. Cherry, and J. Fogarty, "Rethinking the mobile food journal: Exploring opportunities for lightweight photo-based capture," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 3207–3216.

[6] B. Y. Lim, X. Chng, and S. Zhao, "Trade-off between automation and accuracy in mobile photo recognition food logging," in *Proceedings of the Fifth International Symposium of Chinese CHI*. ACM, 2017, pp. 53–59.

[7] J.-j. Chen, C.-W. Ngo, and T.-S. Chua, "Cross-modal recipe retrieval with rich food attributes," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 1771–1779.

[8] S. Mezgec and B. Koroušić Seljak, "Nutrinet: A deep learning food and drink image recognition system for dietary assessment," *Nutrients*, vol. 9, no. 7, p. 657, 2017.

[9] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. S. Feris, "Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification." in *CVPR*, vol. 1, no. 2, 2017, p. 6.

[10] E. M. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification." in *AAAI*, 2017, pp. 4068–4074.

[11] J. Wang, J. Tian, L. Qiu, S. Li, J. Lang, L. Si, and M. Lan, "A multi-task learning approach for improving product title compression with user search log data," 2018.

[12] S. Pandey, T. Agarwal, and N. C. Krishnan, "Multi-task deep learning for predicting poverty from satellite images," 2018. [Online]. Available: https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16441/16388

[13] Q. Liu, Y. Zhang, Z. Liu, Y. Yuan, L. Cheng, and R. Zimmermann, "Multi-modal multi-task learning for automatic dietary assessment," 2018. [Online]. Available: https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16235/15949

[14] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8599–8603.

[15] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[16] K. Hashimoto, C. Xiong, Y. Tsuruoka, and R. Socher, "A joint many-task model: Growing a neural network for multiple nlp tasks," *arXiv preprint arXiv:1611.01587*, 2016.

[17] H. Daumé III, "Bayesian multitask learning with latent hierarchies," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 135–142.

[18] E. M. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification." in *AAAI*, 2017, pp. 4068–4074.

[19] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy, "Im2calories: towards an automated mobile vision food diary," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1233–1241.

[20] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma, "Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment," in *International Conference on Smart Homes and Health Telematics*. Springer, 2016, pp. 37–48.

[21] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *European Conference on Computer Vision*. Springer, 2014, pp. 446–461.

[22] T. Miyazaki, G. C. de Silva, and K. Aizawa, "Image-based calorie content estimation for dietary assessment," in *Multimedia (ISM), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 363–368.

[23] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *European Conference on Computer Vision*. Springer, 2014, pp. 3–17.

[24] Z.-Y. Ming, J. Chen, Y. Cao, C. Forde, C.-W. Ngo, and T. S. Chua, "Food photo recognition for dietary tracking: System and experiment," in *International Conference on Multimedia Modeling*. Springer, 2018, pp. 129–141.

[25] Z. Fu, D. Chen, and H. Li, "Chinfood1000: A large benchmark dataset for chinese food recognition," in *International Conference on Intelligent Computing*. Springer, 2017, pp. 273–281.

[26] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu, "Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2740–2748.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 32–41.

[29] Y. Yang and T. Hospedales, "Deep multi-task representation learning: A tensor factorisation approach," *arXiv preprint arXiv:1605.06391*, 2016.