DOI: 10.5862/JCSTCS/9
УДК 004.923

А.А. Хуршудов

ПРЕДСТАВЛЕНИЕ ТРЕХМЕРНЫХ ОБЪЕКТОВ С ПОМОЩЬЮ АНСАМБЛЯ ТРАНСФОРМИРУЮЩИХ АВТОАССОЦИАТОРОВ

A.A. Khurshudov

USING AN ENSEMBLE OF TRANSFORMING AUTOENCODERS TO REPRESENT 3D OBJECTS

Одна из ключевых задач машинного обучения в области компьютерного зрения – получение качественных представлений визуальных данных, остающихся устойчивыми к изменениям угла обзора, позиции в сцене, эффектов освещения или текстуры изображенного объекта. Существующие современные модели сверточных сетей, такие как GoogLeNet или AlexNet успешно решают эту задачу в некоторых условиях, формируя инвариантные представления, достаточные для эффективной классификации множества объектов. Некоторые исследователи (Хинтон, Крижевский и др.), однако предполагают, что используемый этими моделями подход, несмотря на впечатляющие результаты в задачах классификации, является фундаментально ошибочным по отношению к тому, что должна представлять собой эффективная зрительная система: инвариантные представления не способны реагировать на изменения положения объекта в пространстве. Упомянутые авторы предполагают, что целью любой качественной модели зрительной системы должна быть не инвариантность, а эквивариантность – способность изменять представление объекта предсказуемым образом в ответ на наблюдаемые пространственные преобразования.

В данной статье использована предложенная Хинтоном архитектура подобной эквивариантной модели трансформирующего автоассоциатора, модифицированная таким образом, чтобы обнаруживать низкоуровневые композиционные признаки в изображениях трехмерных объектов. С применением SVM-классификатора и использованием свойств трансформирующего автоассоциатора продемонстрирована возможность представления сложных трехмерных форм в виде ансамбля ограниченного количества автоассоциаторов, каждый из которых соответствует локальному признаку объекта. Благодаря способности трансформирующего автоассоциатора определять не только присутствие выученного признака, но и его пространственные параметры, становится также возможным соотносить вместе изображения одних и тех же объектов в условиях, существенно различных на уровне пикселей.

ТРАНСФОРМИРУЮЩИЙ АВТОАССОЦИАТОР; ОДНОРАЗОВОЕ ОБУЧЕНИЕ; ЭКВИВАРИАНТНОЕ ПРЕДСТАВЛЕНИЕ; КАПСУЛЫ.

One of the key goals of computer vision-related machine learning is to obtain high-quality representations of visual data resistant to changes in viewpoint, area, lighting, object pose or texture. Current state-of-the-art convolutional networks, such as GoogLeNet or AlexNet, can successfully produce invariant representations sufficient to perform complex multiclass classification. Some researchers, however, (Hinton, Khizhevsky, et al.) suggest that this approach, while being quite suitable for classification tasks, is misguided in terms of what an efficient visual system should be capable of doing: namely, being able to reflect spatial transformations of learned objects in a predictable way. The key concept of their research is equivariance rather than invariance, or the model's ability to change representation parameters in response to different poses and transformations of a model-specific visual entity.

This paper employs Hinton's architecture of transforming autoencoder neural networks to identify low-level spatial feature descriptors. Applying a supervised SVM classifier to these detectors, one can then represent a sufficiently complex object, such as a geometric shape or a human face, as a composition of spatially related features. Using the equivariance property, one can also draw distinctions between different object poses, e.g., a frontal face image or a profile image, and then, be able to learn about another, higher-leveled transforming autoencoder via the same architecture. To obtain initial data for first-level feature learning, we use sequences

of frames, or movies, and apply computer vision algorithms to detect regions of maximum interest and track their image patches across the movie. We argue that this way of learning features represents a more realistic approach to vision than general naive feature learning from a supervised dataset. The initial idea came from the concept of one-shot learning (by Fei-Fei et al.), that suggests a possibility of obtaining meaningful features from just one image (or, as in this study, a rather limited set of images supervised by time and order).

TRANSFORMING AUTOENCODER; ONE-SHOT LEARNING; EQUIVARIANT REPRESENTATION; CAPSULES.

Задача распознавания объектов окружающего мира по их визуальным изображениям представляет собой частный случай задачи классификации объектов по категориям (как заданным экспериментатором, так и определяемым алгоритмом). Эффективность решения соответствующих задач и классификации проблемной выборки считаются значительным критерием успешности моделей зрительных систем в целом. Типичная задача классификации представляется следующими условиями:

1. Выборка состоит из набора фотографий, равномерно распределенного по классам.

2. Каждая фотография промаркирована соответствующим классом, и модели могут использовать эти данные в качестве учителя.

3. Выборка делится случайным образом на две подгруппы: обучающую выборку и тестовую выборку (опционально возможно выделение отдельной подгруппы для проведения перекрестной проверки).

4. Обучающая выборка используется в качестве входных данных модели, параметры которой в ходе обучения подстраиваются таким образом, чтобы выходные данные модели соответствовали представленным учителем классам. Для этой цели, как правило, используется алгоритм обратного распространения ошибки.

Конкретным примером такой задачи является соревнование ImageNet [1], в котором используется выборка естественных изображений размера порядка миллиона снимков, сгруппированных по тысяче классов.

В случаях, когда выборка представляет собой достаточно репрезентативное множество фотографий, включающее в себя изображения, различающиеся по форме, освещению и положению в кадре, модель обучается так называемым инвариантным признакам, характеризующим принадлежность изображения к классу вне зависимости от перечисленных факторов. Обучение инвариантным признакам является желаемым результатом, позволяющим успешно распознавать изображения вне зависимости от преобразований, изменяющих внешний вид объекта, но не искажающих его классовой принадлежности. Некоторые исследователи, однако, замечают [2, 3], что, несмотря на эффективные результаты и высокий процент правильных предсказаний, соответствующие модели не отвечают фундаментальным свойствам зрительной системы.

Во-первых, модель, строящая свои предсказания на признаках, полученных из мультиклассовой выборки, становится уязвимой к аномальным ситуациям, выходящим за рамки выборки. Например, изображение на рис. 1 классифицируется моделью CaffeNet [15] как ягуар/леопард. Одновременно с этим модель способна усваивать тонкие различия, несущественные для наблюдателя-человека. Так, модели, показывающие высокие результаты на выборке ImageNet (такие как GoogLeNet, CaffeNet и AlexNet [14, 15]), способны провести различие между фотографиями гепарда и ягуара (отличительными признаками является структура пятен в расцветке шерсти), которых легко может спутать между собой человек без специальных познаний.

При этом такое поведение модели полностью соответствует цели обучения на выборке: в рамках заданных классов характерная пятнистая текстура оказывается достаточным признаком для определения класса. С точки зрения обученной модели пример на рис. 1 оказывается аномалией,

Рис. 1. Предсказания модели CaffeNet, обученной на выборке ILSVRC12.
В скобках указаны числовые значения классов

отклонением, включение которого в выборку позволило бы скорректировать ошибку. Однако подобный подход требует значительных затрат на ручное составление выборки и в основе своей отличается от того, как обучается головной мозг животных, не имеющих в естественной среде доступа к выборкам таких размеров.

Во-вторых, как отмечают Хинтон, Крижевский и Ванг [2], преимущество инвариантности моделей сверточных сетей оказывается фундаментальным недостатком, когда задача распознавания перестает ограничиваться классификацией фотографий по классам. Для большинства приложений зрительной системы, таких как ориентация в пространстве, анализ сцен, принятие решений в соответствии с увиденным, требуется уметь определить не только класс объекта в поле зрения, но и особенности его пространственного расположения, позу и прочие параметры. Использование последовательных слоев свертки и пулинга в сверточных сетях позволяет добиться высокой устойчивости к пространственным вариациям, но одновременно отбрасывает эту необходимую информацию без возможности восстановления. В своей работе авторы представляют архитектуру нейронной сети – трансформирующего автоассоциатора, способного помимо присутствия объекта на изображении в ходе обучения фиксировать эти пространственные параметры и предсказывать не только вероятность наличия объекта на изображении, но и его ориентацию.

Таким образом, можно определить несколько подзадач в решении задачи зрительного распознавания:

1. Признаки, используемые моделью для определения объектов, должны быть достаточными не только для классификации объектов в условиях ограниченной выборки.

2. Основываясь на данных об обучении естественных зрительных систем, можно сделать вывод о существовании признаков, которым модель может обучаться на нескольких (<10) примерах, и эффективно использовать их для распознавания. Для ускорения обучения и фиксации на наиболее характерных элементах изображения, модель должна уметь отыскивать и использовать такие признаки.

3. Требуется сформулировать алгоритм обучения таким локальным характерным признакам. Почти не подлежит сомнению, однако, то, что свойство эквивариантности должно быть в их описании определяющим. Многообещающим кандидатом выглядят локальные эквивариантные капсулы [2] – компактные нейронные сети архитектуры трансформирующего автоассоциатора, способные играть роль как детектора присутствия признака, так и регрессора, оценивающего его пространственные параметры.

Далее мы рассмотрим один из возможных способов обучения таким признакам и эффективность его использования при распознавании.

Трансформирующий автоассоциатор

Полное описание архитектуры в ее оригинальном варианте представлено в не-

скольких работах [2, 4, 5]. В данном разделе предлагается рассмотреть упрощенный обзор архитектуры и модификации.

Трансформирующий автоассоциатор как представитель класса автоассоциаторов представляет собой нейронную сеть, обучающуюся компактному представлению («коду») данных с помощью обратного распространения ошибки, используя при этом в качестве эталонного значения на выходе сети тот же набор данных, что и на входе. В общем случае автоассоциаторы используются для снижения размерности наряду с другими техниками, наподобие анализа главных компонент. При этом чем больше слоев у автоассоциатора, тем более сложное и нелинейное представление он способен выучить.

Основная проблема обучения автоассоциаторов состоит в том, что цель обучения – наиболее точным образом воссоздать входные данные, пропущенные через «бутылочное горлышко» скрытого слоя – не всегда позволяет получить осмысленный обобщенный код, подходящий для использования за пределами выборки. В противном случае автоассоциатор играет роль компрессора, бесполезного для распознавания. Существует несколько техник для решения этой проблемы, одна из которых заключается в том, чтобы обучать сеть с помощью пар трансформированных изображений и дополнительно добавлять численное значение трансформации (*dx* и *dy* для случаев трансляции, эйлеровы углы для трехмерного вращения или матрицу аффинного преобразования) к среднему слою автоассоциатора. Схема сети показана на рис. 2.

Здесь V и O – входной и выходной слои; R и G – слои снижения размерности, их составляющие элементы обучаются, соответственно, признакам для представления объекта (recognition units) и признакам для генерирования трансформации (generation units); I – «бутылочное горлышко» автоассоциатора, представленное одним нейроном P, кодирующим вероятность присутствия объекта на изображении, и нейронами I_α, I_β и I_γ, кодирующими эйлеровы углы поворота. Матрицы W_{VR}, W_{RI}, W_{IG} и W_{GO} представляют собой веса нейронов между соответствующими парами слоев.

Трансформирующий автоассоциатор обучается с подачей на вход и выход пары трансформированных изображений и добавлением к нейронам слоя I известных значений трансформации (в данном случае – изменения угла обзора).

Для обучения элемента P, однако, сети необходимо выработать дискриминирую-

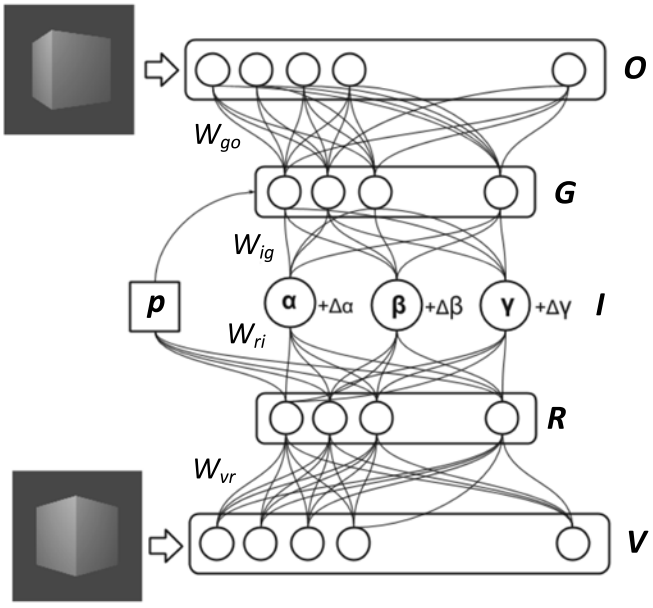

Рис. 2. Схема трансформирующего автоассоциатора

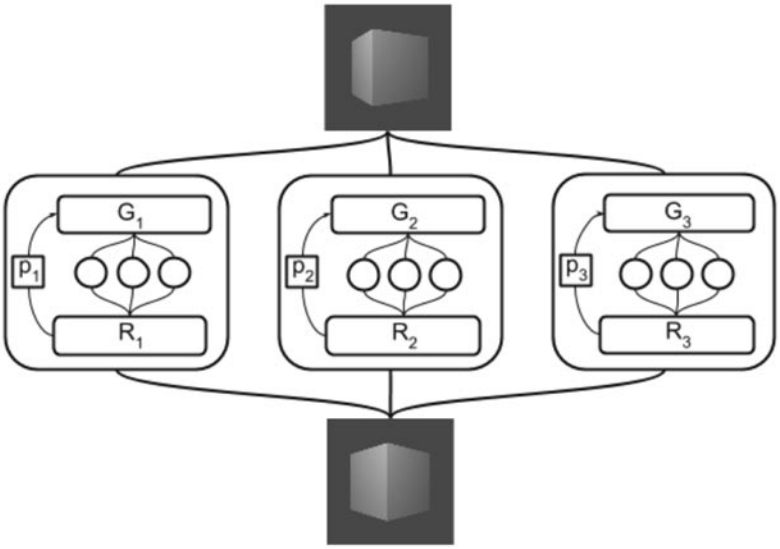

Рис. 3. Схема модели, состоящей из нескольких капсул

щие правила, позволяющие ей различать объекты разных категорий. Для этого в оригинальной работе предлагается тренировать трансформирующие автоассоциаторы совместно, формируя совокупную гипотезу в виде линейной суммы компонентов модели. Отдельные автоассоциаторы такой модели носят название «капсул» и активируются при присутствии на изображении своего класса объектов. В дальнейшем мы будем использовать термины «капсула» и «трансформирующий автоассоциатор» в качестве синонимов. Схема капсульной модели приведена на рис. 3.

В ходе обучения модели вычисляются градиенты следующих функций (активаций сети; приведены в порядке от входного слоя к выходному):

$$a_R = \sigma(xW_{VR}^c + b_{VR}^c); \quad (1)$$

$$a_I = \sigma(a_R W_{RI}^c + b_{RI}^c); \quad (2)$$

$$a_{I'} = a_I + T; \quad (3)$$

$$h = \sigma(a_{I'} W_{GO}^c + b_{GO}^c), \quad (4)$$

где $\sigma(z) = \dfrac{1}{1 - e^{-z}}$; c – индекс c-й капсулы; T – вектор трансформации (α, β, γ).

В качестве функции цены используются среднеквадратичное отклонение, функция кросс-энтропии или любые стандартные варианты, применяющиеся в нейронных сетях с обратным распространением.

Трансформирующий автоассоциатор как детектор признаков

Капсульная модель, представленная выше, предназначена для мультиклассовой классификации и обучается на выборке с фиксированным количеством категорий объектов. При этом в некоторых случаях капсулам не удается обучиться дискриминированным представлениям [5]: вместо этого они кооперируют между собой так, что каждая из них по отдельности обучается зашумленному, неявному и частичному представлению данных. Имея в качестве конечной цели решение проблем, перечисленных в начале статьи, предлагаем следующую модификацию для использования автоассоциатора:

1. Отдельная капсула обучается не на сложных, составных объектах, таких как лица, предметы или пейзажи, а на их низкоуровневых элементах: деталях лиц, углах, характерных неоднородностях.

2. Каждая капсула обучается на множестве трансформаций определенного низкоуровневого участка изображения с целью выработать характерные для определения трехмерной ориентации признаки. Конкретная капсула ничего не знает о других капсулах и других низкоуровневых объ-

ектах и выполняет задачу регрессора: для определенного участка изображения она способна определить его ориентацию в пространстве.

3. Затем, с использованием полученных капсулами признаков, обучим дискриминирующие детекторы, способные отличать регионы капсул друг от друга и от фоновых участков изображения. Каждый детектор, таким образом, выполняет в отношении своей капсулы одноклассовую классификацию (решая задачу «один против всех»).

4. Имея в распоряжении набор локальных регрессоров (капсул) и соответствующих детекторов, мы можем представить изображенный объект в виде ансамбля трансформирующих автоассоциаторов, связанных пространственными отношениями. Поскольку трансформирующие автоассоциаторы способны реагировать на изменение ориентации объектов, становится возможным представить одним и тем же ансамблем изображения, существенно отличающиеся на уровне пикселей (например, лицо в профиль и в фас), сохраняя при этом эквивариантность полученного представления.

Модифицируем архитектуру автоассоциатора следующим образом: в модели будет принимать участие только одна капсула, и за неимением необходимости удалим элемент сети P, отвечающий за вероятность обнаружения объекта. Таким образом мы избавляем сеть от необходимости самой по себе принимать классификационные решения и должны следить за тем, чтобы данные, поступающие на вход капсулы, относились строго к изображениям одного низкоуровневого объекта. Эта задача требует отдельного рассмотрения.

Сбор данных для обучения капсул

Во множестве работ в области психофизиологии отмечалась характерная особенность зрительных систем концентрировать внимание на неоднородных, выделяющихся участках поля зрения [6]. В качестве первого этапа извлечения данных определим соответствующие области на исходном изображении при помощи определителя Гессе [7] и найдем все точки (x, y, t), удовлетворяющие условию:

$$detHL(x, y, t) = t^2(L_{xx}L_{yy} - L_{xy}^2), \qquad (5)$$

где $L(x, y, t)$ – изображение как функция от переменных x и y при фиксированном масштабе t; H – матрица Гессе для этой функции; (x, y, t) – координаты, соответствующую центру и радиусу интересующих регионов.

Альтернативный способ решения этой задачи предлагает детектор заметности (saliency) Кадира–Брэди [8], используемый в т. ч. в работах по машинному обучению зрительных систем [9] и определяющий неоднородные участки изображения с помощью вычисления информационно-теоретической энтропии На рис. 4 приведены примеры участков неоднородности, обнаруженных детекторами.

В некоторых работах (Фей-Фей и др.

Рис. 4. Участки, обнаруженные двумя детекторами.
Слева – детектор Кадира–Брэди, справа – определитель Гессе. Оба детектора в числе прочих регионов находят черты лица

[9, 10]), рассматривающих обучение без выборки (one-shot learning), соответствующие участки используются для построения констелляционной модели на нескольких (<10) изображениях с выравненными позами, отыскивая комбинацию неоднородных участков изображения, которая наилучшим образом описывает тестовую группу. Мы воспользуемся другим подходом под обобщенным названием «время как супервизор» [11]. Рассмотрим последовательность кадров, составляющих видеофрагмент. Для первого кадра производится поиск участков неоднородности указанным выше способом. Для каждого последующего кадра проводим трекинг обнаруженных участков с помощью таких методов компьютерного зрения, как оптический поток Лукаса-Канаде или детектор ORB [14]. Результатом будут последовательности участков изображения, где каждый из элементов представляет собой результат некоторой трехмерной трансформации предыдущего (рис. 5).

Если численное выражение трансформации известно (его можно определить с помощью использующегося метода трекинга – например, оптического потока, либо контролируя движение камеры), то полученные последовательности могут использоваться для обучения трансформирующих автоассоциаторов, каждый из которых обучается реагированию на определенную функциональную деталь изображения.

Стоит обратить внимание также на то, что хотя множество участков неоднородности в первом кадре представляют собой случайные эффекты освещения и текстуры (см. рис. 4), в процессе трекинга такие участки естественным образом отбраковываются, исчезая при повороте объекта. Результирующие последовательности представляют собой естественные составные части объекта, что можно наблюдать на примере человеческого лица.

Обучение капсул

Каждый автоассоциатор обучается на одной последовательности из полученных на предыдущем этапе путем составления всевозможных пар элементов последовательностей с определением трансформации между парами. Для общего случая трехмерного объекта трансформация задается тремя значениями α, β и γ – углами Эйлера, соответствующими вращениям камеры (в рассматриваемом примере вращение ограничено одной осью). Каждый автоассоциатор минимизирует функцию цены $J(x)$:

$$J(x) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}(h(x^{(i)}) - y^{(i)})^2\right) + \\ + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(W_{ij}^{(l)})^2, \quad (6)$$

где m – количество изображений на входе; $h(x^{(i)})$ – выходное значение сети для i-го изображения; λ – параметр регуляризации; $W_{ij}^{(l)}$ – вес связи между i-м нейроном слоя l и j-м нейроном слоя $l-1$.

В процессе обучения отмечено, что добавление к $J(x)$ критерия разреженности увеличивает точность реконструкции автоассоциатора, в некоторых случаях до 40 %. Для выражения критерия разреженности используем дивергенцию Кульбака–Лейбьера:

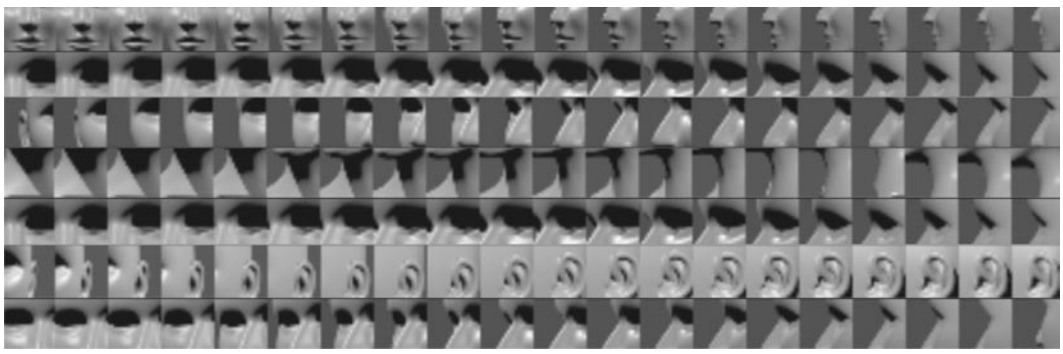

Рис. 5. Последовательности участков неоднородности, обнаруженные при вращении трехмерной модели лица. Вращение ограничено одной осью

$$KL(\rho|\widehat{\rho}_j) = \rho log \frac{\rho}{\widehat{\rho}_j} + (1-\rho) log \frac{1-\rho}{1-\widehat{\rho}_j}, \quad (7)$$

где $\widehat{\rho}_j = \frac{1}{m}\sum_{i=1}^{m}(a_j^{(l)} x^{(i)})$ – среднее значение активации j-го нейрона в слое l; ρ – константа, параметр разреженности.

В обученной капсуле-автоассоциаторе веса нейронов слоя представления (слой R) соответствуют различным состояниям, которые принимает ограниченный полем капсулы участок изображения под воздействием заданных трансформаций.

Признаки, полученные капсулой, сами по себе не являются дискриминативными. Так как назначение капсул двояко – они должны как находить на изображении «свои» участки, так и предсказывать их позу – обучим на обнаруженных признаках детектор, представляющий собой одноклассовый SVM (допустимо использование других алгоритмов, реализующих классификацию «один против всех»).

Построение ансамбля и результаты

Группа обученных капсул оказывается способной находить соответствующие участки даже в тех случаях, когда объект рассматривается с различных углов, а изображения значительно различаются на пиксельном уровне (рис. 6).

Такой ансамбль капсул затем может использоваться для обнаружения композиционного объекта (в рассматриваемом примере – лицо человека). Ключевым достоинством использования капсул является то, что модель не требует обучения отдельных признаков для лиц в анфас и лиц в профиль, в отличие от общераспространенных методов распознавания лиц [13] (на рис. 6 одним цветом отмечены активации одинаковых капсул). Более того, каждая капсула несет информацию не только о том, активирует ли ее выделенный участок изображения, но и способна предсказать параметры трансформации, которым подвернута видимая капсулой сущность.

Так, изображение лица в профиль может быть описано с помощью капсул как «глаз, вид слева», «линия губ, вид слева» и «ухо, вид слева». Предсказаний автоассоциатора оказывается более чем достаточно для того, чтобы сделать предсказание касательно ориентации составного объекта лица в целом – более того, наличие дублирующихся предсказаний капсул может оказаться полезным в тех случаях, когда часть признаков не видна (скрыта другими объектами).

Соответствующий ансамбль признаков может в дальнейшем использоваться для обнаружения объектов несколькими способами:

алгоритмически ищется пространство возможных сочетаний признаков, которые могут встретиться на изображении лица (данный подход рассмотрен Фей-Фей в [10], для решения предложено использовать классификатор Байеса и алгоритм максимизации ожиданий);

метод, предлагаемый нами: использовать тот факт, что капсулы на самом деле располагаются не на плоскости, а в трехмерном пространстве. Если нанести их на трехмер-

Рис. 6. Участки изображения, отмеченные капсулами. Особенности трансформирующего автоассоциатора дают возможность распознавать одни и те же сущности с разных углов обзора

ную модель лица (которая может быть получена как в ходе первого этапа обучения с использованием оптического потока [12], так и сторонними способами), то становится возможным представить любую фотографию лица как результат трехмерной трансформации между данным и неким эталонным изображением. Используя активации капсул первого уровня в качестве параметров модели, на этих данных можно обучить еще один, более высокоуровневый трансформирующий автоассоциатор. Преимущество метода состоит в его потенциальной способности к наращиванию: в дальнейшем автоассоциатор, реагирующий на изображение лица, можно совместить с другими автоассоциаторами, обученными на изображениях корпуса/рук/прочих отдельных частей тела, и увеличивать глубину модели, обучая ее эквивариантно реагировать на полноценные человеческие фигуры.

Постепенное развитие эквивариантных моделей, по мнению многих авторов, – будущее если не машинного обучения в целом, то его отрасли, связанной с компьютерным зрением. Несомненно, рассмотренная модель в текущем виде не способна составить конкуренции таким моделям на базе сверточных сетей, как GoogLeNet, AlexNet и CaffeNet, но эта ограниченность проистекает именно из того факта, что классы изображений для этих моделей (леопард/человек/здание и т. д.) представляют собой сложные, многокомпонентные объекты, требующие совместной работы множества капсул. Возможности полноценно обученной эквивариантной модели, однако, обещают значительно больше того, на что способны современные сверточные сети. Так, модель могла бы различать сложные позы (человек с поднятой рукой или сидящий человек), предсказывать действия объектов на изображении и принимать решения, выходящие далеко за пределы тех, на которые способны модели с инвариантной индикацией объектов.

Среди преимуществ рассмотренной модели обучения можно назвать следующие:

полное отсутствие необходимости в учителе – модель способна работать с любыми трехмерными объектами, для которых существуют участки неоднородности;

результаты, опробованные на человеческих лицах и моделях геометрических тел, показывают эффективность идентификации искомых объектов, сопоставимую с существующими моделями, такими как [13] (для сопоставления использовалась база LFW);

трансформирующие автоассоциаторы – новый, но сравнительно известный алгоритм, успевший продемонстрировать успешные результаты как в области компьютерного зрения, так и в других сферах машинного обучения.

Среди недостатков, в первую очередь, имеет смысл отметить слабую изученность теоретических оснований эквивариантного обучения (трансформирующий автоассоциатор на данный момент – единственная достаточно известная архитектура, решающая эту задачу). Некоторые авторы [5] отмечают неспособность трансформирующего автоассоциатора справляться с определенными категориями трансформаций.

СПИСОК ЛИТЕРАТУРЫ

1. **Deng J. et al.** Imagenet: A large-scale hierarchical image database //Computer Vision and Pattern Recognition. IEEE Conf. on. 2009. Pp. 248–255.

2. **Hinton G.E., Krizhevsky A., Wang S.D.** Transforming auto-encoders //Artificial Neural Networks and Machine Learning. Springer Berlin Heidelberg, 2011. Pp. 44–51.

3. **Kivinen J.J., Williams C.K.I.** Transformation equivariant Boltzmann machines //Artificial Neural Networks and Machine Learning. Springer Berlin Heidelberg, 2011. Pp. 1–9.

4. **Jaitly N., Hinton G.E.** A new way to learn acoustic events //Advances in Neural Information Processing Systems. 2011. Vol. 24.

5. **Wang S.** Learning to Extract Parameterized Features by Predicting Transformed Images. 2011.

6. **Underwood G., Foulsham T.** Visual saliency and semantic incongruency influence eye movements when inspecting pictures //The Quarterly journal of experimental psychology. 2006. Vol. 59. No. 11. Pp. 1931–1949.

7. **Liu J., White J.M., Summers R.M.** Automated detection of blob structures by Hessian analysis and object scale //Image Processing. 17th IEEE Internat. Conf. on. 2010. Pp. 841–844.

8. **Shao L., Kadir T., Brady M.** Geometric and photometric invariant distinctive regions detection //Information Sciences. 2007. Vol. 177. No. 4. Pp. 1088–1122.

9. **Fei-Fei L., Fergus R., Perona P.** One-shot learning of object categories // Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2006. Vol. 28. No. 4. Pp. 594–611.

10. **Fe-Fei L., Fergus R., Perona P.** A Bayesian approach to unsupervised one-shot learning of object categories //Computer Vision, 2003. Proc. 9th IEEE Internat. Conf. on. 2003. Pp. 1134–1141.

11. **George D., Jaros B.** The HTM learning algorithms [Электронный ресурс]/ URL: http://numenta. com/for–developers/education/ Numenta_HTM_Learning_Algos. pdf

12. **Mae Y. et al.** Object tracking in cluttered background based on optical flow and edges // Pattern Recognition, Proc. of the 13th Internat. Conf. on. IEEE. 1996. Vol. 1. Pp. 196–200.

13. **Lienhart R., Maydt J.** An extended set of haar-like features for rapid object detection //Image Processing. Proc. Internat. Conf. on. IEEE. 2002. Vol. 1. – Pp. I-900-I-903.

14. **Krizhevsky A., Sutskever I., Hinton G.E.** Imagenet classification with deep convolutional neural networks //Advances in neural information processing systems. 2012. Pp. 1097–1105.

15. **Jia Y. et al.** Caffe: Convolutional architecture for fast feature embedding //Proc. of the ACM Internat. Conf. of Multimedia. 2014. Pp. 675–678.

REFERENCES

1. **Deng J. et al.** Imagenet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition. IEEE Conference on*, 2009, Pp. 248–255.

2. **Hinton G.E., Krizhevsky A., Wang S.D.** Transforming auto-encoders. *Artificial Neural Networks and Machine Learning*, Springer Berlin Heidelberg, 2011, Pp. 44–51.

3. **Kivinen J.J., Williams C.K.I.** Transformation equivariant Boltzmann machines. *Artificial Neural Networks and Machine Learning*, Springer Berlin Heidelberg, 2011, Pp. 1–9.

4. **Jaitly N., Hinton G.E.** A new way to learn acoustic events. *Advances in Neural Information Processing Systems*, 2011, Vol. 24.

5. **Wang S.** *Learning to Extract Parameterized Features by Predicting Transformed Images,* 2011.

6. **Underwood G., Foulsham T.** Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *The Quarterly journal of experimental psychology*, 2006, Vol. 59, No. 11, Pp. 1931–1949.

7. **Liu J., White J.M., Summers R.M.** Automated detection of blob structures by Hessian analysis and object scale. *Image Processing 17th IEEE International Conference on,* 2010, Pp. 841–844.

8. **Shao L., Kadir T., Brady M.** Geometric and photometric invariant distinctive regions detection. *Information Sciences*, 2007, Vol. 177, No. 4, Pp. 1088–1122.

9. **Fei-Fei L., Fergus R., Perona P.** One-shot learning of object categories. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 2006, Vol. 28, No. 4, Pp. 594–611.

10. **Fe-Fei L., Fergus R., Perona P.** A Bayesian approach to unsupervised one-shot learning of object categories. *Computer Vision, Proceedings 9th IEEE International Conference on. IEEE*, 2003, Pp. 1134–1141.

11. **George D., Jaros B.** The HTM learning algorithms. Available: *http://numenta. com/for–developers/education/Numenta_HTM_Learning_Algos. pdf*

12. **Mae Y. et al.** Object tracking in cluttered background based on optical flow and edges. *Pattern Recognition, Proceedings of the 13th International Conference on, IEEE*, 1996, Vol. 1, Pp. 196–200.

13. **Lienhart R., Maydt J.** An extended set of haar-like features for rapid object detection. *Image Processing, Proceedings International Conference on, IEEE*, 2002, Vol. 1, Pp. I-900-I-903.

14. **Krizhevsky A., Sutskever I., Hinton G.E.** Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012, Pp. 1097–1105.

15. **Jia Y. et al.** Caffe: Convolutional architecture for fast feature embedding. *Proceedings of the ACM International Conference on Multimedia*, 2014, Pp. 675–678.

ХУРШУДОВ Артем Александрович – *аспирант кафедры информационных систем и программирования Кубанского государственного технологического университета.*
350072, Россия, Краснодарский край, г. Краснодар, ул. Московская, д. 2.
E-mail: art1783@gmail.com

KHURSHUDOV Artem A. *Kuban State Technological University.*
350072, Moskovskaya Str. 2, Krasnodar, Krasnodar krai, Russia.
E-mail: art1783@gmail.com