Метод чередования обучаемых параметров

А. А. Хусаенов[1]

В работе предлагается метод повышения качества обучения сверточных искусственных нейронных сетей (ИНС) за счет разделения параметров по их возможности расширения рецептивного поля. При обучении ResNet50 достигается увеличение точности за счет чередуемой остановки обучения в 4-х слоях, расширяющих рецептивное поле.

Показано, что повышение обобщающей способности модели при использовании предложенного метода достигается за счет устранения избыточного вклада отдельных существенных (окклюзивных) элементов изображения при формировании карт признаков. В пользу указанных предположений приводятся результаты экспериментов в задаче transfer learning и рассуждения относительно существования указанной проблемы.

Предлагаемые подходы могут оказаться полезными, в частности, при обучении ИНС на малых данных или дистилляции обучающего множества, где проблемы переобучения на отдельных окклюзивных признаках имеют высокую значимость.

Ключевые слова: сверточная искусственная нейронная сеть, рецептивное поле нейрона, проблемы переобучения моделей, окклюзия признаков в сверточных искусственных нейронных сетях

1. Введение

В задачах распознавания образов в сверточных искусственных нейронных сетях (ИНС) существует проблема переобучения на отдельных окклюзивных (существенных) признаках [1, 2, 3, 4] (п.3.3) исходного изображения.

Например, одна из популярных проблем распознавания «леопардовый диван»: за счет явно-выраженного образа текстуры игнорируются общие признаки объекта (то есть объект «диван» распознается как объект «леопард»)

Подобное переобучение приводит к понижению обобщающей способности модели при отсутствии или недостаточно полной аугментации данных. Указанная проблема переобучения связана с существенными поте-

[1] *Хусаенов Артем Азатович* — аспирант, м.н.с. кафедры математической теории интеллектуальных систем мех.-мат. ф-та МГУ; e-mail: a.khusaenov@mail.ru

Khusaenov Artem Azatovich — postgraduate student, junior research fellow, Moscow State University, faculty of Mechanics and Mathematics, Mathematical Theory of Intelligent Systems department

рями в точности распознавания, если при аугментации существенных признаков обучающего множества будут отсутствовать случаи, которые могут встречаться во время эксплуатации моделей. Причем указанные ошибки не будут выявлены во время проверки модели в случае, если в проверочном множестве соответствующая аугментация существенных признаков также отсутствует.

В задачах понижения вычислительной сложности ИНС и существенного сокращения объема обучающего множества (малые данные) [5, 6, 7, 8] рассматриваемая проблема переобучения усугубляется, а ее выявление во время валидации по-прежнему остается невозможным при недостаточной аугментации данных.

Существующие методы устранения указанной проблемы в большинстве своем сводятся к регуляризации параметров ИНС и более полной аугментации данных, что во многом позволяет сократить переобучение при достаточном числе эпох. Однако, методов решения проблем окклюзии карт признаков, возникающей из-за отдельных существенных элементов изображения, на сегодняшний день существует не так много (например, attention-механизмы [9]).

В данной работе предлагается метод оптимизации указанного переобучения за счет сокращения вклада отдельных окклюзивных элементов изображения при формировании карт признаков сверточной ИНС. Предполагается, что вклад указанных признаков может быть сокращен в слоях, производящих значительное увеличение рецептивного поля исходного изображения. Демонстрируется увеличение обобщающей способности карт признаков при использовании предлагаемого подхода в задачах transfer learning с несколькими типами архитектур ИНС.

2. Рецептивное поле

2.1. Естественно-биологические инварианты зрительного восприятия

С точки зрения естественно-биологического описания восприимчивых областей естественных нейронов, рецептивное поле зрительных рецепторов определяется как область в поле зрения, где зрительные нейроны реагируют на визуальные стимулы [10].

Базовое свойство зрительного восприятия [10] заключается в следующей особенности передачи сигнала рецептору: когда свет достигает визуального датчика, такого как сетчатка, информация, необходимая для определения свойств окружающего мира, содержится не в значениях интенсивности изображения в одной точке, а в соотношениях между значениями интенсивности в разных точках [10]. Вводя, теперь, более формальное описание, рецептивное поле зрительного нейрона может быть определено как область поля зрения (область визуальных датчиков) на визуальные стимулы которого он реагирует [11].

Основная мотивация вычислительной теории рецептивных полей [10, 12] заключается в учете свойств проекции 3-мерных объектов на 2-мерный датчик освещенности (сетчатку), где данные изображения могут подвергаться базовым преобразованиям следующего вида [10]:

локальные масштабные преобразования, вызванные объектами разных размеров и на разных расстояниях для наблюдателя (1)

локальные аффинные преобразования, вызванные изменениями направления обзора относительно объекта (2)

локальные преобразования Галилея, вызванные относительными движениями между объектом и наблюдателем (3)

локальные мультипликативные преобразования интенсивности, вызванные изменениями освещенности (4)

Тогда, поскольку зрительная система способна поддерживать стабильное восприятие окружающей среды в условиях указанной выше аугментации сигнала, одним из ключевых требований к математической формализации является устойчивость модели к таким преобразованиям.

2.2. Инварианты обработки зрительных образов в искусственных нейронных сетях

С точки зрения моделирования процессов зрительного восприятия в сверточных ИНС [13] от модели требуется устойчивость к указанным преобразованиям и достижение обобщающей способности при достаточно полном и аугментированном обучающем множестве. Однако, исходя из предпосылок к усилению отдельных сигналов в пользу минимизации функционала потерь, очевидно предполагать, что отдельные свойства, сохраняющиеся во всех объектах одного класса, будут вносить существенный вклад в карты признаков сверточной ИНС. Особенно, если эти свойства изображения являются устойчивыми (инвариантными) к преобразованиям (1)–(4) – то есть остаются неизменными во всем обучающем множестве. Для упрощения повествования подобные элементы изображения будем называть **существенными элементами** или **существенными сигналами**.

Подобные усиления сигналов можно наблюдать [14] в автоассоциативных ИНС (автоэнкодерах), где для осуществления обратного отображения (декодирования) с наименьшими потерями ИНС избыточно усиливает сигналы отдельных признаков, имеющих сильную корреляцию с целевым признаком.

При формировании карт признаков во всех слоях сверточной ИНС может сохраняться избыточный вклад подобных существенных элементов. Об этом свидетельствуют исследования, посвященные окклюзии изображений в сверточных ИНС [1, 2, 3, 4].

Предполагается, что вклад подобных существенных элементов усиливается при формировании карт признаков в слоях ИНС, производящих расширение рецептивного поля исходного изображения (п.2.4).

В данной работе предлагается метод чередования обучаемых параметров ИНС, который позволяет увеличить обобщающую способность модели за счет ослабления вклада отдельных подобных сигналов путем заморозки отдельных слоев, увеличивающих рецептивное поле. Предполагается, что рассматриваемый эффект повышения обобщающей способности может быть достигнут при заморозке параметров указанных слоев ИНС на последних этапах обучения.

2.3. Вычисление рецептивного поля в искусственной сверточной нейронной сети

Рецептивное поле нейрона [6-8] – это область входного изображения в сверточной ИНС, от которой зависит реакция этого нейрона. На рис. 1 представлен пример [6] рецептивного поля для 2-х слоев свертки с ядром 3x3.

Пусть сверточная ИНС имеет L слоев. Выходные сигналы l-го слоя ($l = \overline{1, n}$) будем обозначать как f_l (то есть f_0 – это входное изображение, а f_n – сигналы последнего слоя). Каждый сверточный слой l имеет 3 параметра свертки:

- k_l - размер ядра свертки
- s_l – шаг ядра свертки
- p_l – размер паддинга

Рассмотрим случай одномерного входного сигнала. Размер рецептивного поля нейронов из слоя l будем обозначать как r_l. В работе [16] предлагаются подходы к вычислению исходного рецептивного поля для рассматриваемого случая.

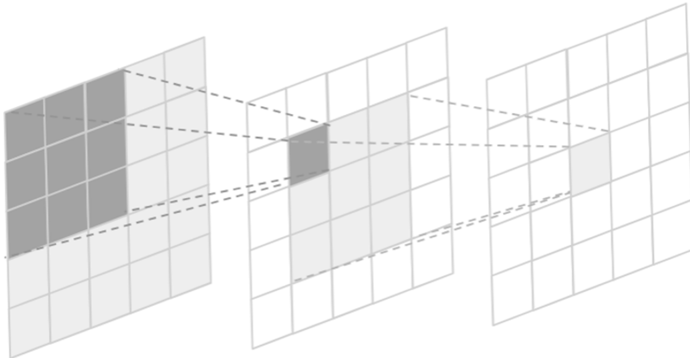


Рис 1. Пример рецептивного поля в сверточных слоях ИНС [16]
(1-й слой - входное изображение, 2-й слой - свертка с ядром 3x3, 3-й слой - свертка с ядром 3x3. Темно-серый - рецептивное поле одного нейрона из второго слоя, светло-серый - рецептивное поле для одного нейрона третьего слоя)

Рецептивное поле l-го сверточного слоя для 1-мерного входного вектора может быть вычислено [16] следующим образом

$$r_{l-1} = s_l \cdot r_l + (k_l - s_l) \tag{5}$$

Представленный подход предполагает рекурсивное вычисление [16] размера исходного рецептивного поля r_0 на основе параметров (s_i, k_i), где $i = \overline{0, L}$ (учет паддинга p_i будет представлен далее).

$$r_0 = \sum_{l=1}^{L} \left((k_l - 1) \prod_{i=1}^{l-1} s_i \right) + 1 \tag{6}$$

Причем, как будет рассмотрено позднее, размер рецептивного поля для слоев подвыборки в данном случае [16] аналогично представляется с помощью параметров s_i и k_i.

Рассмотрим, теперь, вывод [16] значений координат рецептивного поля. Пусть u_l и v_l – это крайний левый и крайний правый координаты рецептивного поля в слое l соответственно. Координаты рецептивного поля для $l-1$ слоя могут быть выражены [16] следующим образом

$$u_{l-1} = -p_l + u_l \cdot s_l \tag{7}$$

$$v_{l-1} = -p_l + v_l \cdot s_l + k_l - 1 \tag{8}$$

Координаты исходного рецептивного поля могут быть выражены рекурсивно [16].

$$u_0 = u_L \prod_{i=1}^{L} s_i - \sum_{l=1}^{L} p_l \prod_{i=1}^{l-1} s_i \tag{9}$$

$$v_0 = v_L \prod_{i=1}^{L} s_i - \sum_{l=1}^{L} (1 + p_l - k_l) \prod_{i=1}^{l-1} s_i \tag{10}$$

Для удобства вычисления исходного рецептивного поля введем совокупный сдвиг S_l (страйд) и совокупный паддинг P_l [16]

$$S_l = \prod_{i=l+1}^{L} s_i \tag{11}$$

$$P_l = \sum_{m=l+1}^{L} p_m \prod_{i=l+1}^{m-1} s_i \tag{12}$$

Тогда, теперь, координаты исходного рецептивного поля могут быть выражены следующим образом

$$u_0 = -P_0 + u_L \cdot S_0 \tag{13}$$

$$v_0 = u_0 + r_0 - 1 \tag{14}$$

Размер рецептивного поля после применения подвыборки (пулинга) с ядром размера k_{L+1} в последнем слое может быть выражен следующим образом

$$r_0 = 1 + \sum_{l=1}^{L} \left((k_l - 1) \prod_{i=1}^{l-1} s_i \right) + (k_{L+1} - 1) \prod_{i=1}^{L} s_i \tag{15}$$

Нетрудно заметить, что **слой подвыборки значительно расширяет рецептивное поле**. Этим замечанием мы воспользуемся далее в п.2.3.

В таблице 1 представлен пример вычисления размера рецептивного поля для ИНС AlexNet [15]

Таблица 1. Пример вычисления рецептивного поля для ИНС AlexNet [15]

l	**Тип слоя**	r_l	k_l	s_l
8	max pooling	1	3	2
7	convolution	3	3	1
6	convolution	5	3	1
5	convolution	7	3	1
4	max pooling	9	3	2
3	convolution	19	5	1
2	max pooling	23	3	2
1	convolution	47	11	4
0	**input**	**195**	-	-

В таблице 2 [16] представлены размеры рецептивного поля для различных архитектур сверточных ИНС.

Таблица 2. Размер рецептивного поля для некоторых ИНС [16] (где $|l_0|$ - длина входного вектора)

Модель	r_0	$\lvert l_0 \rvert$	S_0	P_0
alexnet v2	195	224	32	64
vgg 16	212	224	32	90
mobilenet v1	315	224	32	126
mobilenet v1 075	315	224	32	126
resnet v1 50	483	224	32	239
resnet v1 101	1027	224	32	511
resnet v1 152	1507	224	32	751
resnet v1 200	1763	224	32	879
inception v2	699	224	32	318
inception v3	1311	224	32	618
inception v4	2071	224	32	998
inception resnet v2	3039	224	32	1482

Необходимо заметить [16], что по мере развития моделей рецептивное поле исходного изображения увеличивается [16] (рис.2).

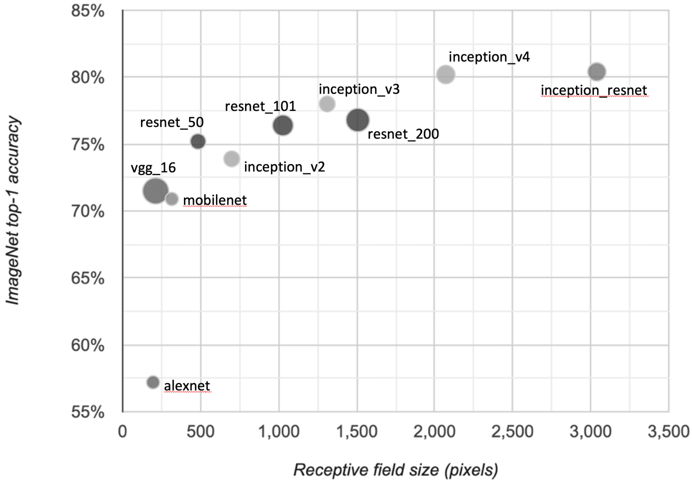


Рис 2. график качества классификации в зависимости от размера рецептивного поля [16]

2.4. Слои увеличения рецептивных полей

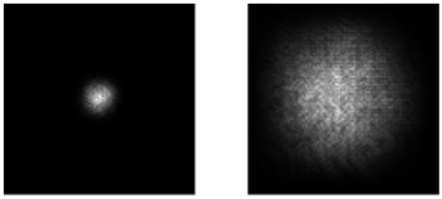
В работе [16] рецептивное поле определяется как область исходного изображения, сигналы которой участвуют в формировании карт признаков. Как было замечено ранее, рецептивное поле значительно увеличивается за счет слоев подвыборки.

При этом существуют слои, производящие значительное увеличение рецептивного поля не в смысле размера области исходного изображения, а относительно вклада его элементов в качество классификации.

Часть рецептивного поля, вносящая существенный вклад в качество классификации, будем называть *действительным (effective receptive filed)* [17]. Рецептивное поле, оцениваемое в смысле координат исходной области изображения, в дальнейшем будем называть *вычисляемым.*

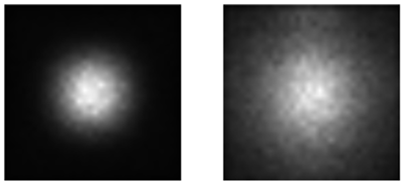
В работе [17] демонстрируется увеличение действительного рецептивного поля как за счет отдельных слоев некоторой сверточной ИНС (рис.3), так и в процессе обучения указанной ИНС (рис.4). В качестве меры влияния отедельных входных сигналов $x^n_{(i,j)}$ некоторого слоя n на выходные сигналы $y^n_{(i,j)}$, оценивается частная производная $\partial y^n_{(i,j)}/\partial x^n_{(i,j)}$ [17].

Предполагается, что вклад существенных элементов изображения усиливается в слоях ИНС, производящих значительное расширение вычисляемого рецептивного поля (п.2.3). При формировании карт признаков отдельные сигналы, полученные из вычисляемого рецептивного поля, могут ослабляться в пользу сигналов, полученных на основе существенных элементов изображения, сохраняющихся во всех объектах обу-



(а) (б)

Рис 3. Рецептивное поле в зависимости от слоя ИНС [17]
(а) - рецептивное поле сверточного слоя; (б) - рецептивное поле слоя подвыборки (пуллинга)



(а) (б)

Рис 4. Рецептивное поле до и после обучения ИНС [17]
(а) - рецептивное поле до обучения; (б) - рецептивное поле после обучения (пуллинга)

чающего множества (п.2.2). Особенно, если эти элементы устойчивы (инварианты) к преобразованиям (1)-(4).

Таким образом, небольшие размеры **действительного рецептивного поля** по отношению к размерам входного изображения или **вычислимого рецептивного поля** можно рассматривать признаком недостаточного обучения ИНС. Понятно, что слои увеличения рецептивных полей играют особую роль в расширении действительного рецептивного поля.

3. Преодоление переобучения

3.1. Определение переобучения

В качестве базового представления проблемы переобучения модели рассмотрим *закон смещения-дисперсии* [18, 19] для задачи классификации.

Пусть $\{x_1, x_2, \ldots, x_n\}$ – некоторое обучающее множество, где каждому объекту x_i соответствует некоторое вещественное число y_i (принадлежность к классу, учитель). При этом существует целевая зависимость y, которая определена как на указанном множестве $\{x_1, x_2, \ldots, x_n\}$, так и за его пределами. Причем целевая зависимость может быть представлена следующим образом.

$$y(x_i) = f(x_i) + \epsilon \tag{16}$$

где $f(x) : R^p \to R^k$ есть некоторая функция, а ϵ - есть случайная величина (шум). Будем считать, что ϵ имеет нулевое среднее $M\epsilon = 0$ и дисперсию σ.

Данную формализацию можно рассматривать следующим образом: для некоторых объектов обучающего множества $\{x_1, x_2, ..., x_n\}$ доступны ответы $\hat{f}(x)$, на основе которых необходимо аппроксимировать целевую зависимость y за пределами указанного множества $\{x_1, x_2, ..., x_n\}$.

В таком случае задача машинного обучения заключается в нахождении приближающей функции (модели) $a(x)$, которая с допустимой точностью аппроксимирует целевую зависимость y как на всем обучающем множестве $\{x_1, x_2, \ldots, x_n\}$, так и за его пределами. В качестве ошибки аппроксимации будем рассматривать среднеквдатратическое отклонение $(M(y) - a(x))^2$.

$$\begin{aligned}
M(y(x) - a(x))^2 &= M\big(y^2(x) - 2y(x)a(x) + a^2(x)\big) = \\
&= M\big(a^2(x)\big) - 2M\big(y(x)a(x)\big) + M\big(y^2(x)\big) = \\
&= M\big(a^2(x)\big) + M\big(y^2(x)\big) - 2M\big((f+\epsilon)a(x)\big) = \\
&= M\big(a^2(x)\big) + M\big(y^2(x)\big) - 2M\big(fa(x)\big) - 2M\big(\epsilon a(x)\big) = \\
&= M\big(a^2(x)\big) - \big(M(a(x))\big)^2 + \big(M(a(x))\big)^2 + \\
&\quad + M\big(y^2(x)\big) - \big(M(y(x))\big)^2 \\
&\quad + \big(M(y(x))\big)^2 - 2M\big(fa(x)\big) = \\
&= D\big(y(x)\big) + D\big(a(x)\big) + \big(M(y(x))\big)^2 + \\
&\quad + \big(M(a(x))\big)^2 - 2M\big(fa(x)\big) = \\
&= D\big(y(x)\big) + D\big(a(x)\big) + \big(M(f)\big)^2 - \\
&\quad - 2M\big(fa(x)\big) + \big(M(a(x))\big)^2 = \\
&= D\big(a(x)\big) + \big(M(f - a(x))\big)^2 + D\big(y(x)\big) = \\
&= variance\big(a(x)\big) + bias\big(f, a(x)\big) + \sigma^2
\end{aligned}$$

где

- $bias\big(\hat{f}, a(x)\big) = \big(M(\hat{f} - a(x))\big)^2$ - *смещение* модели, то есть ошибка относительно заданного множества точек $\{x_1, x_2, ..., x_n\}$

- $variance\big(a(x)\big) = D\big(a(x)\big) = M\big(a^2(x)\big) - \big(Ma(x)\big)^2$ - *дисперсия* модели, то есть разброс значений относительно среднего на заданном множестве

- $\sigma^2 = D(y)$ - дисперсия целевой зависимости, рассматриваемая как *неустранимая ошибка*

Закон смещения - дисперсии предполагает [19]: чем больше точек $(x_1, x_2, ..., x_n)$ захватывает модель $a(x)$, тем ниже смещение (bias), однако выше ее дисперсия (variance). Если предполагать, что при росте сложности модели увеличивается число точек, которые она способна захватить с допустимой точностью, то данный закон может быть проиллюстрирован следующим образом (рис.5) [19].

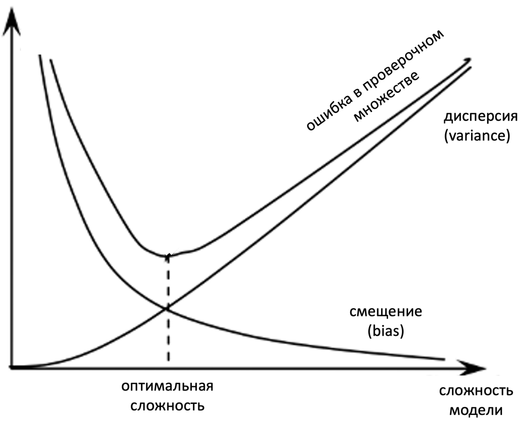


Рис 5. Иллюстрация закона смещения-дисперсии [19]

Тогда проблема переобучения может быть выражена следующим образом:

Определение: *переобучением ИНС будем называть рост ошибки на проверочном множестве при увеличении сложности модели.*

Для моделей ИНС данный закон может быть выражен в терминах сложности ИНС, где сложность модели определяется числом разделяющих гиперповерхностей (нейронов).

3.2. Избыточность модели или недостаточность данных

В данном параграфе приводятся рассуждения, подкрепляющие введенное ранее определение переобучения (п.3.1.).

Избыточная сложность модели в случае ИНС может описываться как избыточное число разделяющих гиперповерхностей при высокой гранулированности групп объектов в пространстве признаков (рис.6).

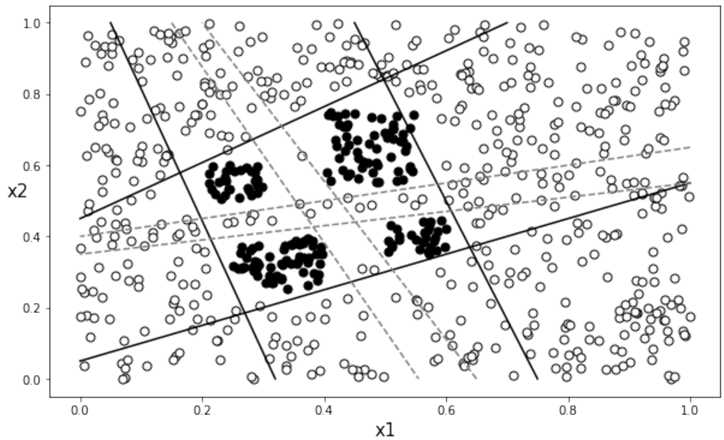


Рис 6. Пример избыточной сложности модели при гранулированных классах

При высокой гранулированности обучающего множества и при существовании отдельных признаков, по которым объекты разных классов становятся легко отделимыми, появляется набор существенных переменных для функции классификации (рис.7).

В таком случае переобучение ИНС возникает не только при длительном обучении, но и при излишнем дроблении признакового пространства из-за избыточной сложности модели.

3.3. Переобучение на окклюзивных признаках

В задаче распознавания образов отдельные элементы изображения, сохраняющиеся во всем множестве объектов некоторого класса, являются существенными признаками. В особенности, если эти признаки являются устойчивыми (инвариантными) к преобразованиям (1)-(4), так как их образ сохраняется неизменным во всех объектах обучающего множества.

О существовании проблем переобучения на существенных признаках (п.3.2) в задачах распознавания образов свидетельствуют эксперименты по оценке окклюзии изображений [1, 2, 3, 4] в сверточных ИНС. О подобных проблемах так же свидетельствует множество экспериментов с

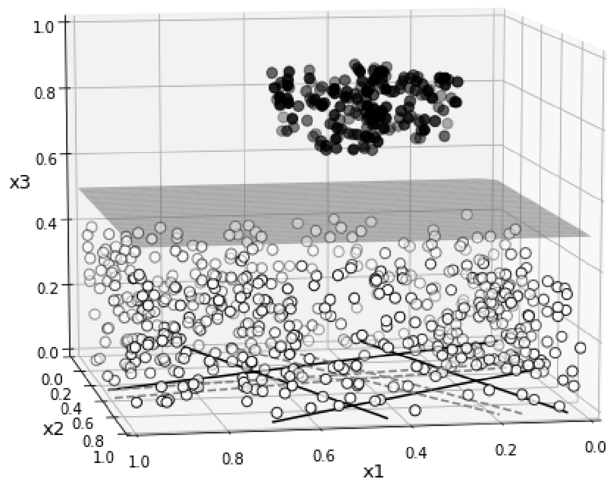


Рис 7. Пример существенной переменной (x_3) в пространстве признаков

зашумлением изображений (атаки на ИНС), приводящим к значительному понижению точности [20].

Базовые подходы по оценке окклюзии [1] позволяют определить влияние отдельных существенных признаков на точность классификации. При зашумлении отдельных элементов изображения оценивается изменение точности классификации в обученной ИНС. Производя указанную операцию итерационно по всем областям изображения, возможно оценить степень влияния отдельных элементов изображения на финальное предсказание [4].

Эксперименты по окклюзии изображений [1, 2, 3, 4] демонстрируют проблему ухудшения обобщающей способности модели за счет избыточного внимания на отдельных признаках изображения. То есть, проблема переобучения в данном случае рассматривается как проблема избыточного вклада отдельных окклюзивных элементов изображения в карты признаков сверточной ИНС.

4. Метод чередования обучаемых параметров

4.1. Описание метода

В данной работе предлагается метод чередования обучаемых параметров ИНС, позволяющий увеличить обобщающую способность модели за счет ослабления вклада отдельных существенных сигналов (п.2.2).

Предполагается, что рассматриваемое переобучение можно регуляризировать за счет попеременной остановки обучения слоев, увеличивающих вычисляемое рецептивное поле. В таком случае возможно усилить участие других менее существенных элементов изображения при формировании карт признаков, то есть увеличить действительное рецептивное поле, что приведет к повышению обобщающей способности модели. Данную технику чередования обучаемых параметров будем далее называть **обучением с фиксированным рецептивным полем (RFF – receptive field freeze)**.

4.2. Эксперименты

Для демонстрации метода рассматривается классическая задача transfer-learning. Данный подход выбран, поскольку позволяет оценить повышение обобщающей способности модели на уровне формирования карт признаков, а не классификатора в последнем слое.

ИНС обучается на некотором наборе изображений (указано далее), принадлежащих множеству базовых классов. Далее производится обучение последнего (выходного) классификатор-слоя ИНС на группе классов, не входящих в базовое обучающее множество. Обучение ИН на исходном наборе классов (до transfer learning) для упрощения изложения будем называть базовым обучением.

Эксперимент имеет следующий вид (рис.8):

Шаг 1. Базовое обучение модели

Базовое обучение модели в течении n эпох

Шаг 2. Receptive Field Freeze - обучение (RFF)

Дообучение модели в двух вариациях:

- **модель 1:** базовое обучение ИНС в эпохе $(n + 1)$
- **модель 2:** RFF обучение ИНС в эпохе $(n + 1)$

Шаг 3. Transfer learning

На основе каждой из 2-х моделей производится transfer-learning (с одинаковыми гиперпараметрами)

Для чистоты эксперимента на шаге 1 на протяжении первых n эпох обучается 1 общая модель (шаг 1). Затем на эпохе $(n + 1)$ производится разветвление на 2 модели (шаг 2). Далее для каждой модели производится transfer learning с одинаковым гиперпараметрами, одинаковой

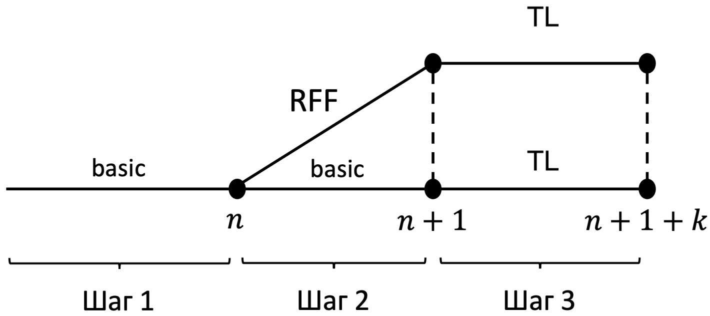


Рис 8. Схема эксперимента

стартовой инициализацией весов в последнем классификатор-слое и общим генератором случайных чисел.

На основе указанной схемы далее производятся 3 эксперимента: базовый эксперимент, эксперимент с преодолением переобучения ИНС, эксперимент с глубокой ИНС ResNet50.

4.2.1. Базовый эксперимент

В базовом эксперименте используется датасет CIFAR-100. Рассматривается архитектура ИНС, состоящая из 2-х типов блоков: блоки сверток без подвыборки и блоки с подвыборкой. Архитектура ИНС представлена в таблице 3.

В рамках обучения с фиксированными рецептивными полями (receptive field freeze, RFF) в указанной ИНС фиксируются параметры слоев, производящих свертки сразу после операции подвыборки (max pooling), то етсь слои 7 и 13. Шаги эксперимента указаны в таблице 4.

Для оценки устойчивости метода, датасет CIFAR-100 был разбит на 10 подмножеств классов. Каждое из подмножеств было разбито на 2 группы классов: классы для базового обучения и классы для transfer-learning. Итоговые группы подмножеств представлены в таблице 5.

Таблица 3. Архитектура базовой ИНС

N	Слой	Размер карты	Кол-во ядер	Размер ядра
1	Convolutional	(32,32)	32	(3,3)
2	BatchNorm	(32,32)	-	-
3	ReLU	(32,32)	-	-
4	Convolutional	(32,32)	32	(3,3)
5	ReLU	(32,32)	-	-
6	MaxPooling	(16,16)	32	(2,2)
7	Convolutional	(16,16)	64	(3,3)
8	BatchNorm	(16,16)	-	-
9	ReLU	(16,16)	-	-
10	Convolutional	(16,16)	64	(3,3)
11	ReLU	(16,16)	-	-
12	MaxPooling	(8,8)	64	(2,2)
13	Convolutional	(8,8)	64	(3,3)
14	BatchNorm	(8,8)	-	-
15	ReLU	(8,8)	-	-
16	Convolutional	(8,8)	64	(3,3)
17	ReLU	(8,8)	-	-
18	Flatten	4096	-	-
19	Danse	64	-	-
20	ReLU	64	-	-
21	Danse	8	-	-

Таблица 4. Схема экспериментов 1.1-1.10

Номер эпохи	**Модель 1**	**Модель 2**
0	инициализация параметров	
[1,4]	базовое обучение	
5	базовое обучение	RFF-обучение
6	transfer-learning	transfer-learning

Таблица 5. Разбиение классов в экспериментах 1.1-1.10

N	**Классы для базового обучения**	**Классы для transfer-learning**
1.1	яблоко, хомяк, ребенок, медведь, часы, пчела, велосипед, бутылка	жук, бобр
1.2	мост, автобус, верблюд, девушка, банка, замок, гусеница, крупный рогатый скот	чаша, мальчик
1.3	шимпанзе, облако, кровать, таракан, диван, гора, чашка, динозавр	стул, крокодил
1.4	дельфин, камбала, лес, лиса, бабочка, дом, кенгуру, клавиатура	слон, аквариумная рыбка
1.5	лампа, газонокосилка, леопард, лев, ящерица, омар, черепаха, кленовое дерево	краб, мотоцикл
1.6	мышь, гриб, дубовое дерево, апельсин, орхидея, выдра, пикап, сосновое дерево	пальмовое дерево, груша
1.7	равнина, тарелка, мак, опоссум, дикобраз, енот, луч, дорога	кролик, ракета
1.8	роза, море, тюлень, землеройка, небоскреб, улитка, змея, паук	скунс, акула
1.9	белка, поезд, сладкий перец, стол, танк, телевизор, тигр, трактор	телефон, трамвай
1.10	подсолнух, форель, шкаф, кит, ивовое дерево, волк, женщина, червь	мужчина, тюльпан

Таким образом, для оценки устойчивости метода, эксперимент повторялся 10 раз на разных классах изображений. На каждом из подмножеств классов производился эксперимент следующего вида. Для валидации отводилось 20% объектов из каждой группы классов для transfer learning. В качестве метрики оценивается accuracy валидации (val_accuracy).

Таблица 6. Результаты экспериментов 1.1-1.10

Номер эксп-та	**Модель 1**	**Модель 2**	**delta**
1.1	0,640	0,645	+**0,005**
1.2	0,590	0,595	+**0,005**
1.3	0,675	0,680	+**0,005**
1.4	0,765	0,760	$-0,005$
1.5	0,700	0,745	+**0,045**
1.6	0,890	0,905	+**0,015**
1.7	0,795	0,800	+**0,005**
1.8	0,880	0,875	$-0,005$
1.9	0,640	0,655	+**0,015**
1.10	0,790	0,825	+**0,035**
avarage delta			+**0,012**

На графиках ниже (рис.9) представлены показатели точности обучения и валиации.

Нетрудно заметить, что в экспериментах 1.4 и 1.8 (где точность валидации модели 2 оказалась ниже) модель 1 достигает наибольшей обобщающей способности относительно прочих экспериментов (так как показатели точности валидации сопоставимы с точностью обучения). В многих остальных случаях (1.1, 1.3, 1.6, 1.7, 1.9) ИНС оказалась близка к фазе переобучения на первых 4 эпохах.

Исходя из этого наблюдения, естественным образом возникает предположение, что RFF обучение позволяет избежать ранней стадии переобучения, которая выражается высокой окклюзией изображения. Эксперимент с этим предположением рассматривается в следующем разделе.

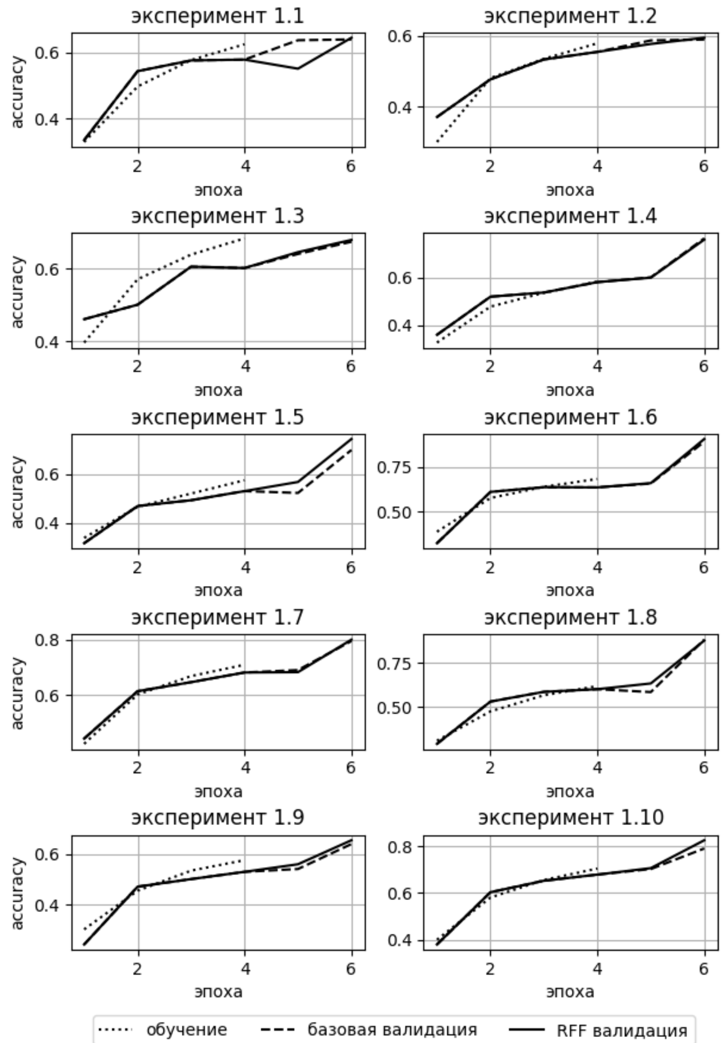


Рис 9. Результаты экспериментов 1.1-1.10

4.2.2. Эксперимент с преодолением переобучения

Для оценки предположения, рассмотренного в прошлом разделе, производится ряд экспериментов с базовым обучением в n эпох, где значение n определяется стадией переобучения ИНС. В каждом из 10 экспериментов базовое обучение производится до тех пор, пока ИНС не войдет в фазу переобучения на некоторой эпохе n. То есть, эксперимент имеет следующий вид:

Таблица 7. Схема экспериментов 2.1-2.10

Номер эпохи	**Модель 1**	**Модель 2**
0	инициализация параметров	
$[1,n-1]$	базовое обучение	
n	базовое обучение	RFF-обучение
$n+1$	transfer-learning	transfer-learning

Как и в экспериментах прошлого раздела, для валидации transfer learning отводилось 20% объектов из каждой группы классов. В качестве метрики оценивается accuracy при валидации (val_accuracy)

Таблица 8. Результаты экспериментов 2.1-2.10

Номер эксп-та	**Модель 1**	**Модель 2**	**delta**
2.1	0,640	0,635	$-0,005$
2.2	0,690	0,670	$-0,020$
2.3	0,830	0,850	+ **0,020**
2.4	0,900	0,905	+ **0,005**
2.5	0,825	0,830	+ **0,005**
2.6	0,870	0,900	+ **0,030**
2.7	0,755	0,795	+ **0,040**
2.8	0,855	0,905	+ **0,050**
2.9	0,830	0,850	+ **0,020**
2.10	0,740	0,725	$-0,015$
avarage delta			+**0,013**

На графиках ниже (рис.10) представлены показатели точности обучения и валиации. Анализ и выводы приведены в параграфе 4.3.1.

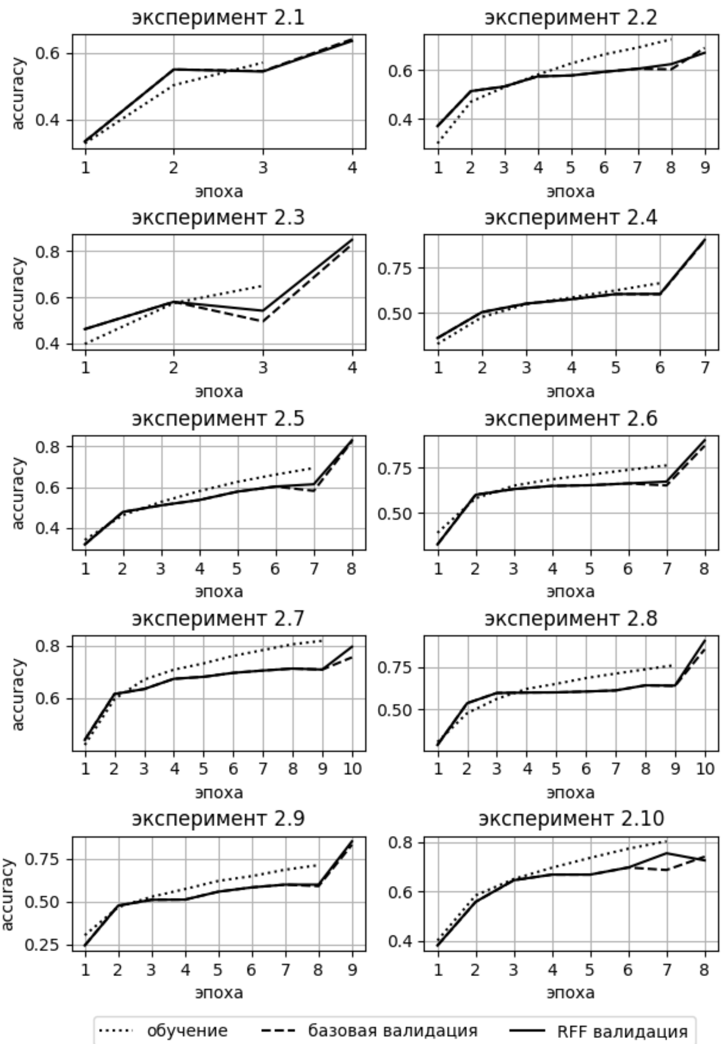


Рис 10. Результаты экспериментов 2.1-2.10

4.2.3. Эксперимент с глубокой нейронной сетью

В экспериментах №1.1-1.10 и №2.1-2.10 эмитировались слои глубоких ИНС, производящие сжатие карт признаков (то есть слои, расширяющие рецептивное поле). В данном разделе эксперимент повторяется для глубокой ИНС ResNet50.

Из датасета ImageNet были выбраны 10 случайных классов. Далее из указанного подмножества классов случайным образом были выбраны 2 класса для transfer learning. В ИНС ResNet50 в рамках RFF-обучения были зафиксированы параметры в слоях, понижающих размер карты признаков (таблица 9) : 8, 40, 82, 144. Шаги эксперимента указаны в таблице 9.

Таблица 9. Слои, понижающие размер карты признаков в ResNet50

Номер слоя в ResNet50	**Размер входной карты признаков**	**Размер выходной карты признаков**
8	(114,114)	(56,56)
40	(56,56)	(28,28)
82	(28,28)	(14,14)
144	(14,14)	(7,7)

Таблица 10. Схема эксперимента 3

Номер эпохи	**Модель 1**	**Модель 2**
0	инициализация параметров	
$[1,n-1]$	базовое обучение	
n	базовое обучение	RFF-обучение
$n+1$	transfer-learning	transfer-learning

Как и в экспериментах прошлых разделов, для валидации transfer learning отводилось 20% объектов из каждой группы классов. В качестве метрики оценивается accuracy при валидации (val_accuracy)

Таблица 11. Результаты эксперимента 3

Номер эксп-та	**Модель 1**	**Модель 2**	**delta**
3	0,6827	0,7346	+**0,052**

На графике ниже (рис.11) представлены показатели точности обучения и валидации. Анализ и выводы приведены в параграфе 4.3.1.

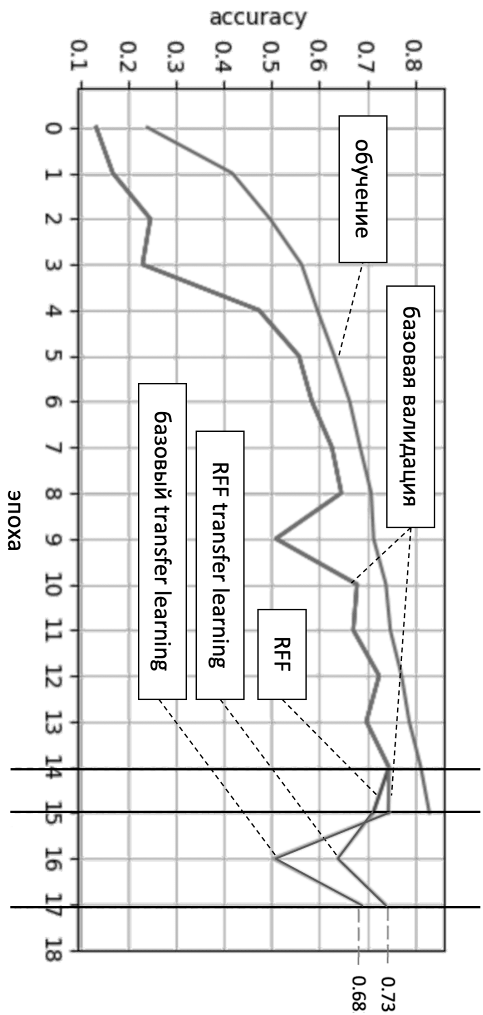


Рис 11. Результаты эксперимента 3

4.3. Результаты экспериментов

4.3.1. Увеличение точности

В таблице 12 представлены результаты всех экспериментов (п.4.2.). В рамках выбранных архитектур ИНС наблюдается увеличение обобщающей способности моделей.

Таблица 12. Результаты всех экспериментов

N	**тип модели**	**кол-во эпох**	**критерий остановки**	**M1**	**M2**	**delta**
1.1	базовая	6	число эпох	0,640	0,645	+ **0,005**
1.2	базовая	6	число эпох	0,590	0,595	+ **0,005**
1.3	базовая	6	число эпох	0,675	0,680	+ **0,005**
1.4	базовая	6	число эпох	0,765	0,760	$-0,005$
1.5	базовая	6	число эпох	0,700	0,745	+ **0,045**
1.6	базовая	6	число эпох	0,890	0,905	+ **0,015**
1.7	базовая	6	число эпох	0,795	0,800	+ **0,005**
1.8	базовая	6	число эпох	0,880	0,875	$-0,005$
1.9	базовая	6	число эпох	0,640	0,655	+ **0,015**
1.10	базовая	6	число эпох	0,790	0,825	+ **0,035**
2.1	базовая	4	переобучение	0,640	0,635	$-0,005$
2.2	базовая	9	переобучение	0,690	0,670	$-0,020$
2.3	базовая	4	переобучение	0,830	0,850	+ **0,020**
2.4	базовая	7	переобучение	0,900	0,905	+ **0,005**
2.5	базовая	8	переобучение	0,825	0,830	+ **0,005**
2.6	базовая	8	переобучение	0,870	0,900	+ **0,030**
2.7	базовая	10	переобучение	0,755	0,795	+ **0,040**
2.8	базовая	10	переобучение	0,855	0,905	+ **0,050**
2.9	базовая	9	переобучение	0,830	0,850	+ **0,020**
2.10	базовая	8	переобучение	0,740	0,725	$-0,015$
avarage delta						+**0,013**
3	ResNet50	17	число эпох	0.6827	0.7346	+**0,052**
avarage delta						+**0,052**

Увеличение обобщающей спосоности достигается на уровне слоев, формирующих карты признаков (поскольку эксперименты производятся в рамках задачи transfer learning).

Результаты, полученные в экспериментах с преодалением переобучения могут сигнализировать о повышении обобщающей способности модели за счет ослабления вклада отдельных окклюзивных признаков изображения.

В эксперименте с глубокой ИНС ResNet50 (п.4.2.3.) достигается более высокий прирост точности, чем в ИНС со значительно меньшей архитектурой (п.4.2.1, п.4.2.2.).

4.3.2. Оценка времени обучения

В экспериментах при переходе от модели 1 к моделе 2 не зафиксировано повышение времени обучения. В таблице 13 демонстрируется результаты профилирования времени.

Предположительно, незначительные изменения скорости обучения (на всем числе эпох) вызваны техническими издержками на фиксирование и пропуск отдельных связей при распространении сигнала в ИНС.

Таблица 13. Изменение времени обучения

N	**тип модели**	**M1 (sec)**	**M2 (sec)**	**delta (sec)**
1.1 – 1.10	базовая	67.0	67.5	+0.7%
2.1 – 2.10	базовая	108.0	108.5	+0.4%
3	ResNet50	30499	30828	+0.1%

5. Вывод

Реализована техника, при которой обучение сверточной ИНС происходит с чередуемой остановкой обучения в слоях, производящих расширение рецептивного поля. На примере задачи transfer leaning продемонстрировано увеличение обобщающей способности модели во множестве экспериментов с базовой архитектурой ИНС.

Для глубокой архитектуры ResNet50 получено увеличение точности на 5% за счет остановки обучения всего для 4 слоев и в рамках всего 1 эпохи.

Список литературы

[1] Zeiler M.D., Fergus R., Fleet D., Pajdla T., Schiele B., Tuytelaars T., “Visualizing and Understanding Convolutional Networks.”, *Computer Vision – ECCV 2014.*, 2014, № 8689.

[2] Kortylewski A., He J., Liu Q., Yuille A., “Compositional Convolutional Neural Networks: A Deep Architecture with Innate Robustness to Partial Occlusion”, *CVPR 2020*, 2020.

[3] Kortylewski A., Liu Q., Wang A., Sun Y., Yuille A., “Compositional Convolutional Neural Networks: A Robust and Interpretable Model for Object Recognition under Occlusion”, *International Journal of Computer Vision*, 2021, № 129, 736-760.

[4] Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A., “Learning Deep Features for Discriminative Localization”, *CVPR*, 2016.

[5] Hinton, G., Vinyals, O., Dean, J., “Distilling the Knowledge in a Neural Network”, *NIPS*, 2015.

[6] Wang T., Zhu J., Torralba A., Efros A., “Dataset Distillation”, 2018..

[7] Sucholutsky, I., Schonlau M., “Soft-Label Dataset Distillation and Text Dataset Distillation”, 2019.

[8] Medvedev D., D’yakonov A., “New Properties of the Data Distillation MethodWhen Working With Tabular Data”, 2020.

[9] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R. Martin, Ming-Ming Cheng, “Attention Mechanisms in Computer Vision: A Survey”, *Computational Visual Media*, 2022, № 8, 331–368.

[10] Lindeberg T., “A computational theory of visual receptive fields”, *Biological Cybernetics*, 2013, № 107, 589–635.

[11] Hubel DH, Wiesel TN, “Receptive fields of single neurones in the cat’s striate cortex”, *Physiol*, 1959, № 147, 226-238.

[12] Lindeberg T., “Normative theory of visual receptive fields”, *Heliyon*, **1**:7 (2021).

[13] LeCun Y., Haffner P., Bottou L. Bengio Y., “Object Recognition with Gradient-Based Learning”, *Lecture Notes in Computer Science*, 1999, № 1681.

[14] Хусаенов А.А.,, “Автоассоциативные нейронные сети в задаче классификации с усеченным множеством”, *Интеллектуальные Системы. Теория и приложения*, **2**:26 (2022).

[15] Krizhevsky.A, Sutskever.I, Hinton G.E.,, “ImageNet Classification with Deep Convolutional Neural Networks”, *Advances in Neural Information Processing Systems*, 2012, № 25.

[16] Araujo A., Norris W., Sim J.,, “Computing Receptive Fields of Convolutional Neural Networks”, *Distill*, 2019.

[17] Luo W, Li Y., Urtasun R., Zemel R.,, "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks", *Neural Information Processing Systems*, 2016, № 29.

[18] Geman S., Bienenstock E., Doursat R., "Neural networks and the bias/variance dilemma", *Neural Computation*, 1992, № 4.

[19] Lean Yu, Kin Keung Lai, Shouyang Wang, Wei Huang, "A Bias-Variance-Complexity Trade-Off Framework for Complex System Modeling", *Computational Science and Its Applications*, 2006.

[20] Alparslan Y., Alparslan K., Keim-Shenk J., Khade S., Greenstadt R.,, "Adversarial Attacks on Convolutional Neural Networks in Facial Recognition Domain", *IEEE Access*, 2021.

Learning parameters rotation method
Khusaenov A.A.

A method is proposed for improving the quality of training of convolutional artificial neural networks (ANN) by dividing the parameters according to their ability to expand the receptive field. For example, ResNet50 accuracy increase is achieved by only 4 layers freeze.

It is shown that the model generalizing ability increase is achieved by eliminating the excessive contribution of individual significant (occlusive) image elements in the feature maps. In favor of these assumptions the results of experiments in the transfer learning task and reasoning about the existence of this problem are presented.

The proposed approaches can be useful in training ANNs on small data or distillation of the training set, where the problems of overfitting on individual occlusive features are of high importance.

Keywords: convolutional artificial neural network, neuron receptive field, model overfitting problems, feature occlusion in convolutional artificial neural networks

References

[1] Zeiler M.D., Fergus R., Fleet D., Pajdla T., Schiele B., Tuytelaars T., "Visualizing and Understanding Convolutional Networks.", *Computer Vision – ECCV 2014.*, 2014, № 8689.

[2] Kortylewski A., He J., Liu Q., Yuille A., "Compositional Convolutional Neural Networks: A Deep Architecture with Innate Robustness to Partial Occlusion", *CVPR 2020*, 2020.

[3] Kortylewski A., Liu Q., Wang A., Sun Y., Yuille A., “Compositional Convolutional Neural Networks: A Robust and Interpretable Model for Object Recognition under Occlusion”, *International Journal of Computer Vision*, 2021, № 129, 736-760.

[4] Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A., “Learning Deep Features for Discriminative Localization”, *CVPR*, 2016.

[5] Hinton, G., Vinyals, O., Dean, J., “Distilling the Knowledge in a Neural Network”, *NIPS*, 2015.

[6] Wang T., Zhu J., Torralba A., Efros A., “Dataset Distillation”, 2018..

[7] Sucholutsky, I., Schonlau M., “Soft-Label Dataset Distillation and Text Dataset Distillation”, 2019.

[8] Medvedev D., D’yakonov A., “New Properties of the Data Distillation MethodWhen Working With Tabular Data”, 2020.

[9] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R. Martin, Ming-Ming Cheng, “Attention Mechanisms in Computer Vision: A Survey”, *Computational Visual Media*, 2022, № 8, 331–368.

[10] Lindeberg T., “A computational theory of visual receptive fields”, *Biological Cybernetics*, 2013, № 107, 589–635.

[11] Hubel DH, Wiesel TN, “Receptive fields of single neurones in the cat’s striate cortex”, *Physiol*, 1959, № 147, 226-238.

[12] Lindeberg T., “Normative theory of visual receptive fields”, *Heliyon*, **1**:7 (2021).

[13] LeCun Y., Haffner P., Bottou L. Bengio Y., “Object Recognition with Gradient-Based Learning”, *Lecture Notes in Computer Science*, 1999, № 1681.

[14] Khusaenov A.A., “Autoassociative neural networks in a classification problem with truncated dataset”, *Intelligent systems. Theory and applications*, **2**:26 (2022) (In Russian).

[15] Krizhevsky.A, Sutskever.I, Hinton G.E.,, “ImageNet Classification with Deep Convolutional Neural Networks”, *Advances in Neural Information Processing Systems*, 2012, № 25.

[16] Araujo A., Norris W., Sim J.,, “Computing Receptive Fields of Convolutional Neural Networks”, *Distill*, 2019.

[17] Luo W, Li Y., Urtasun R., Zemel R.,, "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks", *Neural Information Processing Systems*, 2016, № 29.

[18] Geman S., Bienenstock E., Doursat R., "Neural networks and the bias/variance dilemma", *Neural Computation*, 1992, № 4.

[19] Lean Yu, Kin Keung Lai, Shouyang Wang, Wei Huang, "A Bias-Variance-Complexity Trade-Off Framework for Complex System Modeling", *Computational Science and Its Applications*, 2006.

[20] Alparslan Y., Alparslan K., Keim-Shenk J., Khade S., Greenstadt R.,, "Adversarial Attacks on Convolutional Neural Networks in Facial Recognition Domain", *IEEE Access*, 2021.